National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”
Faculty of Informatics and Computer Science
Department of Computer Engineering

Security, Fault Tolerance, Intelligence

International Conference ICSFTI2019
May 14−15, 2019

Kyiv
Ukraine
Igor Sikorsky Kyiv Polytechnic Institute
2019

UDC 004.056+004.052.3+004.8](062)
T44

Recommended by the Academic Council of FICS Igor Sikorsky Kyiv Polytechnic Institute (Protocol No 12 dated June 10, 2019)

Security, Fault Tolerance, Intelligence: proceedings of the International Conference ICSFTI2018, Kyiv, Ukraine, May 14–15, 2018. – Kyiv : Igor Sikorsky Kyiv Polytechnic Institute, publisning house "Polytechnica", 2019. – 214 p.

The collection presents the works of the scientific conference in the field of computer engineering, software engineering and technical education. The conference is devoted to the memory of the outstanding scientist, professor of the department of computer science prof. V. Shyrochyn.

For teachers, graduate students, students in the field of informatics and computer science.

ISBN 978-966-622-956-7

http://comsys.kpi.ua

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра обчислювальної техніки

Безпека. Відмовостійкість. Інтелект

Міжнародна науково-практична конференція ICSFTI2019

14−15 травня 2019 р.

Київ
Україна
КПІ ім. Ігоря Сікорського
2019

УДК 004.056+004.052.3+004.8](062)
Б40

Рекомендовано Вченою радою ФІОТ КПІ ім. Ігоря Сікорського (Протокол № 12 від 10.06.2019 р.)

Безпека. Відмовостійкість. Інтелект: зб. пр. міжнародної наук.-практ. конф. ICSFTI2018, Київ, Україна, 14–15 трав. 2019 р. – Київ : КПІ ім. Ігоря Сікорського, Вид-во «Політехніка», 2019. – 214 с.

Подано праці міжнародної наукової конференції з комп'ютерної інженерії, інженерії програмного забезпечення та технічної освіти, яка присвячена видатному вченому, викладачу кафедри обчислювальної техніки професору В. П. Широчину.

Для викладачів, аспірантів, студентів галузі інформатики та обчислювальної техніки.

ISBN 978-966-622-956-7

CONTENT

Plenary Section

Section 1. SEC (Security of computer systems and networks. Fault-tolerant distributed computing)

Section 2. RT (Internet of Things, Real-Time Systems)

Section 3. AI (Intelligent Systems. Machine learning, big data)

Section 4. GN (Global networks, grid and cloud systems)

Plenary Section

UDC 004.383

Anatoliy Sergiyenko, Yuriy Vinogradov,
Olexiy Molchanov, Chojadurdy Jepbarov

MALICIOUS HARDWARE IN FPGA

Сергієнко Анатолій, Виноградов Юрій,
Молчанов Олексій, Джепбаров Ходжадурди

ЗЛОНАМІРЕНО СТВОРЕНЕ АПАРАТНЕ ЗАБЕЗПЕЧЕННЯ У ПЛІС

A survey of malicious hardware in FPGA is considered. The methods for malicious hardware searching and preventing its loading into FPGA are highlighted as well. A conclusion is made that FPGA is the most-safe device against malicious hardware loading. A stack processor structure is proposed which can be used for monitoring the malicious hardware.

Keywords: FPGA, stack processor, side channel.
Fig.: 0. Tabl.: 1. Bibl.: 7.

У статті оглядаються можливості впровадження злонамірено створеного апаратного забезпечення (ЗСАЗ) у програмованих логічних інтегральних схемах (ПЛІС), способів їх виявлення та перешкоджання їх впровадженню. Робиться висновок про те, що ПЛІС є найбільш захищені від впровадження в них ЗСАЗ. Запропоновано блок моніторінгу ЗСАЗ на базі стекового процесора.

Ключові слова: ПЛІС, стековий процесор, побічний канал.
Рис.: 0. Табл.: 1. Бібл.: 7.

Introduction. Malicious hardware is the hardware, which is designed and implemented to violate the security and reliability of the computer systems. Many issues of malicious hardware in ASICs are described in [1]. Recently, more attention is paid to malicious hardware which is implemented in the field programmable gate arrays (FPGAs) [2,3]. The development and implementation of projects for FPGA has the same stages as the creation of the software: compilation, and debugging of the programs written in a high-level programming language, loading of the codes into the equipment. Therefore, the well-known malicious software and malicious hardware have much in common. But the second one needs more investigations.

Types of malicious hardware. The classification of malicious hardware comes from the classification of malicious software. A *backdoor* or a *trapdoor* allows the access to a computer system that is not subject to its specifications. It can be implemented during the development of the computer system or during its upgrade.

A *kill switch* is a malicious artificial object that allows the attackers to interfere with the correct functioning of the hardware. The kill switch causes a failure in the form of an inability to perform the necessary functions. It can be installed as part of the FPGA hardware or the FPGA configuration.

An *FPGA virus* is part of the FPGA configuration, which can cause the short-circuiting of the outputs of the internal gates and, as a result, the overheating and failure of the chip [4].

Trojan hardware is a wide variety of malicious hardware. Like the software Trojans, this malicious hardware is implemented by third parties and is intended to make a violation of the normal device operation, such as slowing down, or transmitting or receiving important information through a hidden channel [5].

By the size and number of involved inputs-outputs, the malicious hardware blocks are divided into *small* and *large* ones. A large block is much easier to activate and detect than a small one. But it has a functionality that provides the solutions that cause great losses.

Malicious hardware can be activated from the outside *through a hidden channel* or *from inside*. The latest malicious hardware is divided into *always-on* and *on-demand* activated hardware. The always-on malicious hardware is permanently activated and may interfere the functioning of the computer system at any time. The on-demand malicious hardware is inactive until a certain condition is met. This may be the fixation of a given internal logical state or state of input signals or overflow of an internal counter. Such malicious hardware cannot be detected until its activation.

The *information transmitting* malicious hardware is designed to send secret information from a protected area to the outside. For example, a certain change in the network parameters of the encryption block allows the spy to determine the encryption key through the analysis of power consumption of the computer system.

Ability to implement malicious hardware in FPGA. Theoretically, malicious hardware can be embedded in FPGA during crystal manufacturing. But due to the fact that at the production stage there is no thought about where and how the FPGA would be used, what kind of firmware would be implemented, the implementation of malicious hardware is too difficult [6].

The FPGA firmware can always be designed in a state of secrecy. Therefore, a project for FPGA is protected, if it is not acquainted with an intruder. Since the operation of FPGA is based on the information in the configuration file, the standard methods of the information protection are applied to it.

The attacks on the hardware, which are used for the conventional chips, are unacceptable for FPGAs because they lead to the configuration destruction. In order to inject the malicious hardware into FPGA, the malefactor must have the initial or the precisely reconstructed project.

Nevertheless, the additional protection measures are being taken in responsible applications. The simplest measure is the one-time programming, when the full access

to the configuration is not possible. During the last two decades, FPGA encrypts the configuration file, which allows it to be loaded into FPGA multiple times after shutting down and turning on the power without any unauthorized copying. This additionally makes it impossible to reconstruct the project.

In addition, the partial FPGA reconfiguration, which allows the intruder to attach the parts of the project, is prohibited in the encrypted mode. Also, the requirement to turn off the power before downloading a new configuration destroys the previous configuration. The configuration file key is stored in battery-operated RAM of FPGA, which is automatically erased when it is turned off. The known attack methods on this key disable this power and break the cipher since the key is stored at the depth of the crystal under several layers of dielectrics and metallization.

Thus, perhaps the only opportunity to insert the malicious hardware into the FPGA is its attachment during the design process. For example, the designer can insert an IP core from a third-party vendor with the malicious hardware hidden in it. Such an IP core can be embedded in an automated software program generator that invisibly invades this malicious hardware to the project [6].

Ways to detect malicious hardware. There are three basic approaches to detect malicious hardware in FPGA. They are an analysis of project files, testing with the automatic test generation and analysis of signals from the side channels [5]. But without knowledge of the logical network, its location, and activation method of the malicious hardware, it is virtually impossible to find a test sequence that detects the presence of this malicious hardware.

The analysis of the information leakage in the side channels allows the investigator to detect the presence of malicious hardware. For example, the feedback signals such as power consumption, infrared radiation, and radio frequency can be analyzed. The efficiency of this method is increased if the special modules which are introduced in the chip that contribute to this analysis. These may be the voltage and temperature sensors, delay measurement modules placed at different points of the chip, as well as the built-in self-testing circuit [2].

In contrast to the complex analysis of ASIC, the detection of malicious hardware in FPGA is trivial. Any correction of the configuration file changes its control code. In addition, the existing software tools make it possible to compare the project versions at different design stages to distinguish the unwanted logic networks [6].

Prevention of malicious hardware introduction. The harmful effect of the malicious hardware can be made impossible by the special measures of the structural design. So, a redundant method can counteract the kill switches, and the introduction of the infrastructure modules like the parametric sensors help to detect the malicious hardware.

The most of FPGA projects are built around the general purpose processor which runs some user programs under some RTOS. This processor and its memory remain the target for the attacks from both malicious hardware and software. The most

effective way is the spatial and logical isolation of the responsible blocks like this processor from the rest of the computer system. This is the essence of the moats and drawbridges method. This isolation greatly complicates the configuration of the hidden channel and simplifies its detection.

Block of monitoring the malicious hardware. To remove a side channel through a module with shared access, for example, the shared memory, the mechanism for access confirmation is usually used. Then, the request monitoring block denies access to the shared module from other modules that do not have this right, including the malicious hardware [2].

A small 8-bit configured microcontroller, such as SM8, described in [7] is proposed to serve as a block of monitoring the malicious hardware. Its features are very small hardware volume, short instructions (8-bit and16-bit instructions), which are implemented for one or two clock cycles. Table 1 shows the parameters of the SM8 core configured in FPGAs of different series.

The microcontroller core has a very small share in the total hardware volume and moderate speed, which is enough for its functioning. It could not seriously infer the computer system project configured in FPGA. From this point of view, this core can serve as the malicious hardware as well. Therefore, its investigations can help to investigate the problem of malicious hardware revealing.

Table 1

Parameters of the SM8 microcontroller core

FPGA series	***Hardware volume***	***Share of total volume***	***Maximum clock frequency, MHz***
Xilinx XC7K480T	181 LC	0.038%	250
Intel 10M50	1164 LE	2.3%	150
Intel 10CX220	210 ALM	0.095%	230

Conclusions. In recent times, a lot of attention is being paid to the malicious hardware problem due to the fact that it poses a serious threat to the security of the state and people. The detecting or removing the malicious hardware embedded in FPGA is a very complicated task. However, the implementation of malicious hardware in the FPGA-based computing systems is considerably complicated compared to ASICs. The cheapest and most reliable way to prevent the implementation of malicious hardware in FPGA is to perform the appropriate security measures when designing a configuration file for the project. The structural and logical measures that prevent the introduction of malicious hardware are redundant design, spatial and logical isolation of critical units, access control to shared resources, minimization of noise switching, implementation of infrastructure modules to counteract malicious hardware. The development, detection, counteraction, and prevention of malicious hardware is a range of tasks that require further multilateral scientific and technical research.

References

1. Сергієнко А. М. Злонамірено створене апаратне забезпечення / А. М. Сергієнко // Information Technology and Security. – 2012. – № 1. – С. 93 – 100.

2. Huffmire T., Irvine C., Nguyen T.D., Levin T., Kastner R., Sherwood T. Handbook of FPGA Design Security. – Springer. – 2010. – 177 p.

3. Security Trends for FPGAs. From Secured to Secure Reconfigurable Systems / Editors: B. Badrignans, G. Gogniat, J. L. Danger, L. Torres, V. Fischer. – Springer, 2011. – 196 p.

4. Hadzic I., Udani S., Smith J. FPGA viruses // Proceedings of the 9-th Int. Workshop on Field-Programmable Logic and Applications, FPL'99, Glasgow, UK. – 1999. – P. 291 – 300.

5. Wang X., Tehranipoor M., Plusquellic J. Detecting malicious inclusions in secure hardware: challenges and solutions // Proc. IEEE Workshop on Hardware Oriented Security and Trust, HOST, Anaheim, CA, June, 2008, –2008. –P.15-19.

6. Trimberger S. Trusted design in FPGAs // Proceedings of the 44-th Design Automation Conference, San Diego, CA, USA, DAC 2007, June 4 – 8, 2007, – P. 1.5 – 1.8.

7. Molchanov O. Microprocessor Architecture for Serial Port Communications / Sergiyenko A., Orlova M. // Advances in Computer Science for Engineering and Education II. – Springer. – 2019. – P. 238 – 246.

Authors

Anatoliy Sergiyenko – professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: aser@comsys.kpi.ua

Сергієнко Анатолій Михайлович – професор, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Juriy Vinogradov – assistant, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: vinograd@comsys.ntu-kpi.kiev.ua

Виноградов Юрій Миколайович – асистент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Oleksiy Molchanov – post-graduate student, Application Specific System Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: oleksii.molchanov@gmail.com

Молчанов Олексій Андрійович – аспірант, кафедра спеціалізованих комп'ютерних систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Chojadurdy Jepbarov – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

Джепбаров Ходжадурди – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНІ АНОТАЦІЇ

Сергієнко Анатолій, Виноградов Юрій,
Молчанов Олексій, Джепбаров Ходжадурди

ЗЛОНАМІРЕНО СТВОРЕНЕ АПАРАТНЕ ЗАБЕЗПЕЧЕННЯ У ПЛІС

Актуальність теми дослідження. На відміну від злонамірено створенного програмного забезпечення, злонамірено створеному апаратному забезпеченню (ЗСАЗ) присвячується недостатньо уваги. В той же час ЗСАЗ може нанести не менше шкоди, ніж його програмний аналог, а методи і заходи для протидії ЗСАЗ розвинуті недостатньо.

Аналіз останніх досліджень і публікацій. Протягом останніх років з'являється все більше статей присвячених класифікації та заходам виявлення та протидії ЗСАЗ особливо у замовлених НВІС. Але проблемам протидії ЗСАЗ в програмованих логічних інтегральних схемах (ПЛІС) присвячено мало наукових робіт.

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена огляду, вивченню та аналізу проблем, які пов'язані з впровадженням ЗСАЗ у ПЛІС.

Постановка завдання. Завданням є пошук перспективного напрямку дослідження ЗСАЗ у ПЛІС та створення блоку моніторингу наявності ЗСАЗ та протидії його впливу на ресурси, які розділяються.

Викладення основного матеріалу. Розглянуто класифікацію ЗСАЗ, які впроваджуються у ПЛІС, шляхи, ймовірність та ефективність їх впровадження та заходи для їх виявлення і знешкодження. Запропоновано блок моніторингу наявності ЗСАЗ та протидії його впливу на ресурси, які розділяються. Блок виконано у вигляді ядра мікроконтролера зі стековою архітектурою, який має як мінімізовані апаратні витрати, так і невеликий об'єм вбудованого програмного забезпечення.

Висновки. Встановлено, що ПЛІС є достатньо захищеною мікросхемою від впровадження ЗСАЗ. Найбільш ймовірним джерелом ЗСАЗ у ПЛІС є віртуальні модулі, які поставляються сторонніми компаніями. Для протидії таким ЗСАЗ запропоновано блок моніторингу наявності ЗСАЗ та протидії його впливу на ресурси, які розділяються.

Ключові слова: ПЛІС, стековий процесор, побічний канал.

UDC 378.1

Victor Poriev

SOME ASPECTS OF BUILDING AN INTELLIGENT SOFTWARE SYSTEM TO SUPPORT EDUCATIONAL PROCESS

Віктор Порєв

ДЕЯКІ АСПЕКТИ ПОБУДОВИ ІНТЕЛЕКТУАЛЬНОЇ ПРОГРАМНОЇ СИСТЕМИ ДЛЯ ПІДТРИМКИ НАВЧАЛЬНОГО ПРОЦЕСУ

Some aspects of building an intellectual information system to support the learning process are considered in this article. The concept of individual student portrait is introduced based on the graphical representation of timeliness of the tasks. It is proposed to use artificial neural networks to classify types of success by analyzing time series.

Key words: individual student portrait, intellectual information system, neural networks, support of educational process, time series classification.

Fig.: 7. Tabl.: 0. Bibl.: 6.

У статті розглядаються деякі аспекти побудови інтелектуальної інформаційної системи для підтримки навчального процесу. Вводиться поняття індивідуального портрету студента на основі графічного відображення вчасності виконання завдань. Пропонується використовувати штучні нейронні мережі для класифікації типів успішності шляхом аналізу часових рядів.

Ключові слова: аналіз, класифікація часових рядів, інтелектуальна інформаційна система, нейронна мережа, навчання нейронних мереж, підтримка навчального процесу.

Рис.: 7. Табл. 0. Бібл.: 6.

Relevance of the research topic. The development of artificial intelligence systems provides a wide range of opportunities for improving information systems supporting the learning process, giving them new functionalities. The main purpose of the article is to improve the planning and conduct of the educational process in the system of higher education through the introduction of methods and technologies of artificial intelligence.

Formulation of the problem. Limited human capabilities and the need to cover the vast amounts of information, knowledge and experience that accumulate over decades lead to the emergence of computer intelligence systems that can not only accumulate, but also analyze data and processes. This fully applies to information systems to support educational activities. Providing the possibility of such systems to

automatically save, in addition to traditional data, and time series, raises the problem of analysis and classification of large volumes of such information.

Actual scientific researches and issues analysis. As a result of the analysis of literary and advertising sources devoted to the systems of support for the automation of the educational process, we can conclude that these systems are based on the platforms of database management systems; and the main function that is implemented in such systems - the storage and accumulation of curricula [1, 2].

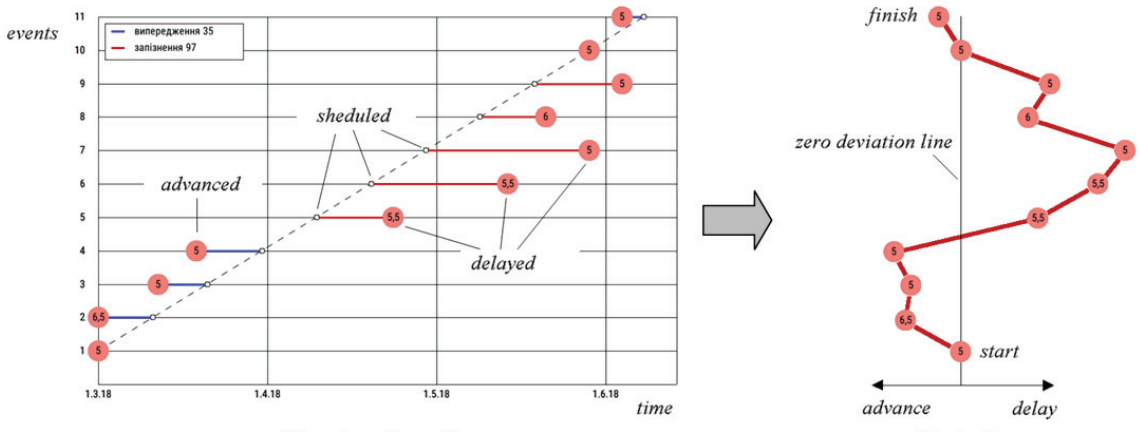
In recent years, research on the implementation of ideas, methods and means of artificial intelligence, in particular, the creation of intelligent systems that are themselves able to study [3] are actively conducted.

Closely related is the study of time series classification methods [4, 5].

Uninvestigated parts of general matters defining. New features provide information systems for storing information about student tasks. Analysis of such data allows to reveal regularities of development of educational processes in time. There are still unexplored questions regarding the methods and means of such analysis.

The research objective. The purpose of this article is to study the suitability of artificial neural networks to solve the tasks of the classification of time series, which describe the processes of educational activity. An important task is to find a format for organizing time data for the best perception of their artificial neural network. It is also necessary to solve the issues of choosing the configuration of the neural network and determine the parameters of the training network to provide the required level of reliability of the classification of time series in the educational process.

The statement of basic materials. The computer system for the maintenance of the educational process must memorize the time of each grade. Then the student's behavior can be described by a time series. It is proposed to line the timely execution vertically, and from it horizontally to the right or left to indicate the delay or advance. If you depict the time of each event in such a system of coordinates, it is convenient to describe the student's behavior in the form of a polyline — a certain symbol. We get a time "portrait" of the student (Fig. 1)



Events – time diagram Portrait

Fig. 1. Student behavior in time

You can classify the types of such portraits. Figure 2 shows some types of portraits.

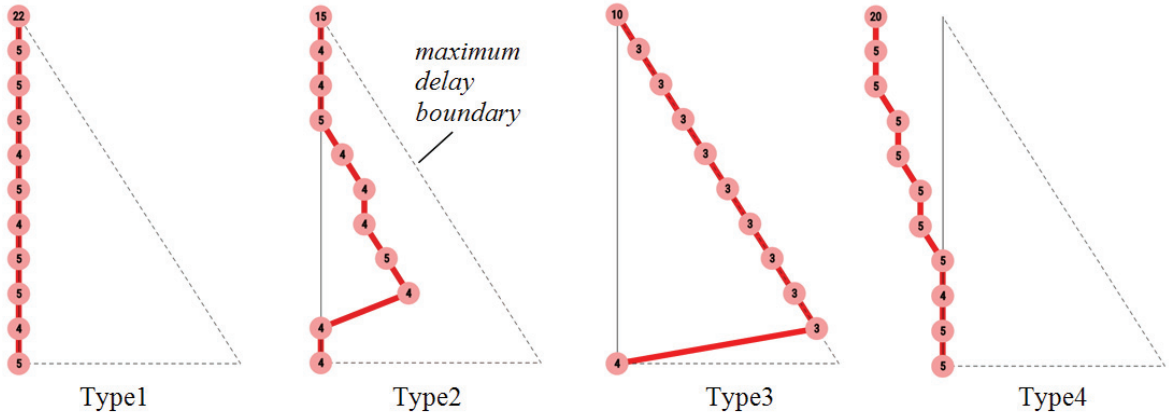


Fig. 2. Some types of the portraits

Type 1 means perfectly timely execution of tasks. The vertical line may be slightly curved within certain tolerances in time.

Type 2 indicates an episode of significant delay in the execution of tasks, for example, due to illness. Such an episode can happen at different time, as shown below.

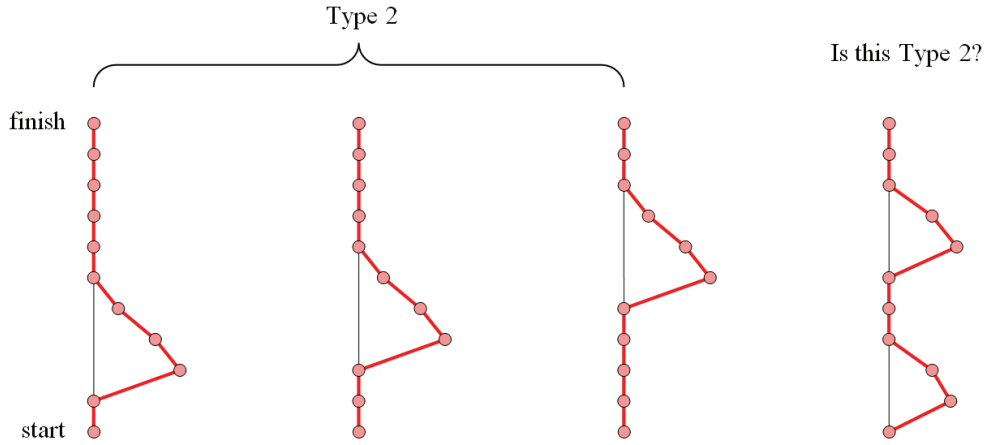


Fig. 3. Episodes of significant lateness – classified as Type2

The extreme case of type2 is type 3. Type 4 represents the early completion of tasks.

The above types are quite idealized. In reality, there can be an infinite set of portraits of behavior. It is necessary to somehow classify the types of behavior, given the fact that the boundaries for each type may be fuzzy. It is often quite difficult for a person to analyze and classify data when it needs to be processed quickly.

One should not think that a comprehensive description of student activity is an end in itself and is used only for fixing and documenting. It is obvious that the behavior of the student is determined not only by the characteristics of the student. At a minimum, there are three more important factors that determine student behavior — this is society, family, and the education system. It is important to be able to identify trends and use them to improve processes.

A neural networks may be useful for solving such problems.

General model structure. For the classification of the types of academic performance, a neural network of the multilayer perceptron type was chosen with separate outputs for each type of classification. From the point of view of memory costs and speed, one network with several outputs is better. From the point of view of recognition accuracy, it is more expedient to use a separate network with one output for each type of classification.

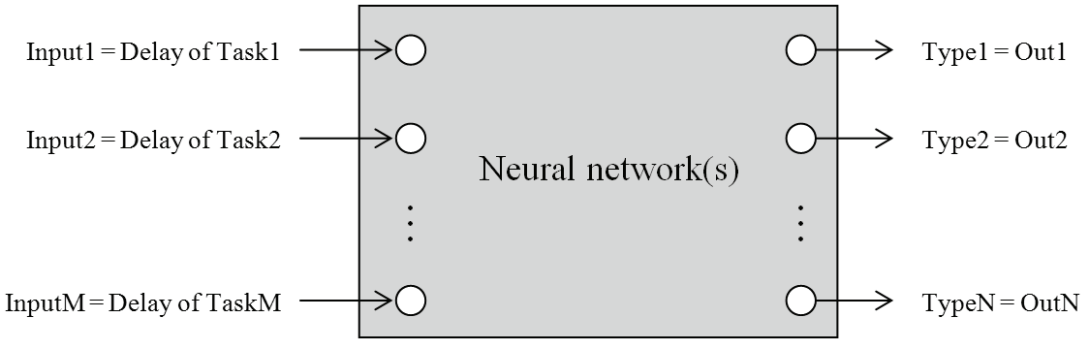


Fig. 4. General model structure

Thus, the main parameters of the classification model are the number of student behaviors that a neural network should be able to recognize. Another important parameter will be the length of the time series being analyzed, i.e., the number of tasks, or, generally speaking, the events of the occurrence of which need to be analyzed. It is also required to provide that a system configured and trained, for example, to analyze the implementation of 10 tasks, could analyze the implementation of a different number of tasks without significant restructuring. This can in certain cases be solved by interpolation along the time axis.

The neural network is implemented by the author of the article in the form of a C++ program module for Windows for research purposes, as well as a Java class for embedding in Android applications.

Experiments. For the training of the neural network, the results of assigning tasks to several groups of students were used. Data sets were formed, each of which is classified by one of 11 types of academic achievement.

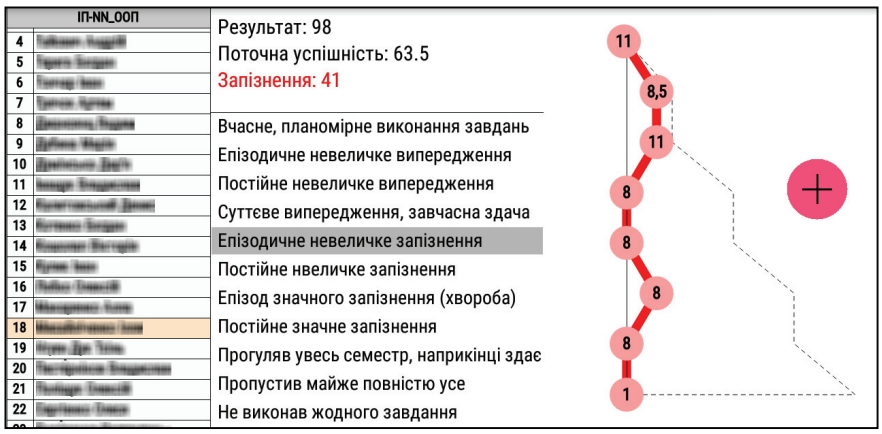


Fig. 5. Example of student grading

The training was performed by the method of *back-propagation errors* [6]. When calculating the increments of the weights of connections Δw, the following training parameters were used: ε – is the learning rate and α – is the coefficient taking into account the change in the weight increment of the given link at the previous iteration. The formula for incrementing the connection weight from neuron A to neuron B is as follows:

$$\Delta w_{A-B}^{i} = \varepsilon \cdot \delta_B \cdot OUT_A + \alpha \cdot \Delta w_{A-B}^{i-1},$$

where: OUT_A is the output value of neuron A; δ_B is the fraction of the error calculated for the neuron B taking into account the layer in which the neuron is located.

The learning process is designed as a sequence of epochs. In each epoch ($Epoch_i$) the network scans all input data sets (Set_j) selected for training. Thus, $Epoch_i = \{Set_1, Set_2, \ldots, Set_n\}$. During the network training cycle, for each epoch, the total square error of the dataset was calculated separately for each output. For one m-th network output, which indicates the corresponding classified m-th type, the error was calculated by the formula:

$$\Delta epoch_{i,m} = \frac{1}{n}\sum_{j=1}^{n}\left(OUTideal_{j,m} - OUTactual_{j,m}\right)^2,$$

where: *OUTideal* is the desired value, *OUTactual* is the output of the network.

Below are graphs of changes in the error for one of the classified types during 10000 epochs of learning for different values of the parameters ε and α.

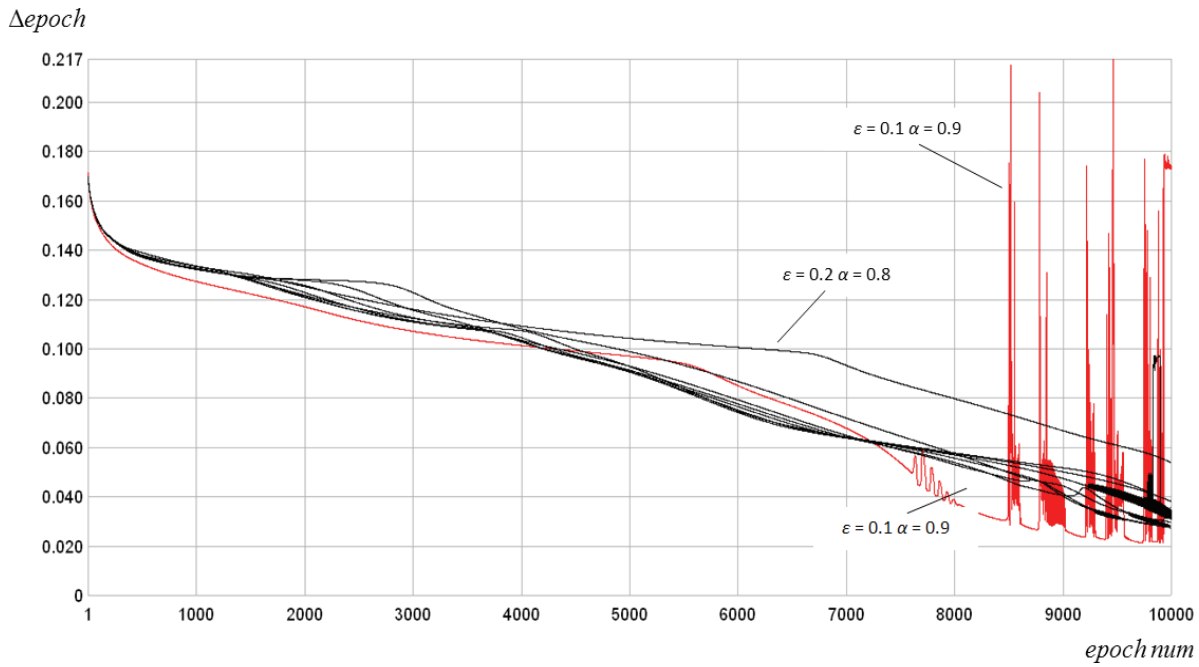


Fig. 6. Change errors during training

As you can see, the choice of the parameters ε and α significantly affects the learning outcome. So, for example, the worst situation occurs when $\varepsilon = 0.1$, $\alpha = 0.9$ -

the process is unstable, and after about 8000 epochs of learning, the error may even increase.

A significant influence on the training of the neural network may be the choice of the initial values of the links weights. Random values were selected from -0.5 to 0.5. Fig. 7 Figure illustrates the differences in training at ε = 0.7, α = 0.3 for 6 different sets of initial random weight values

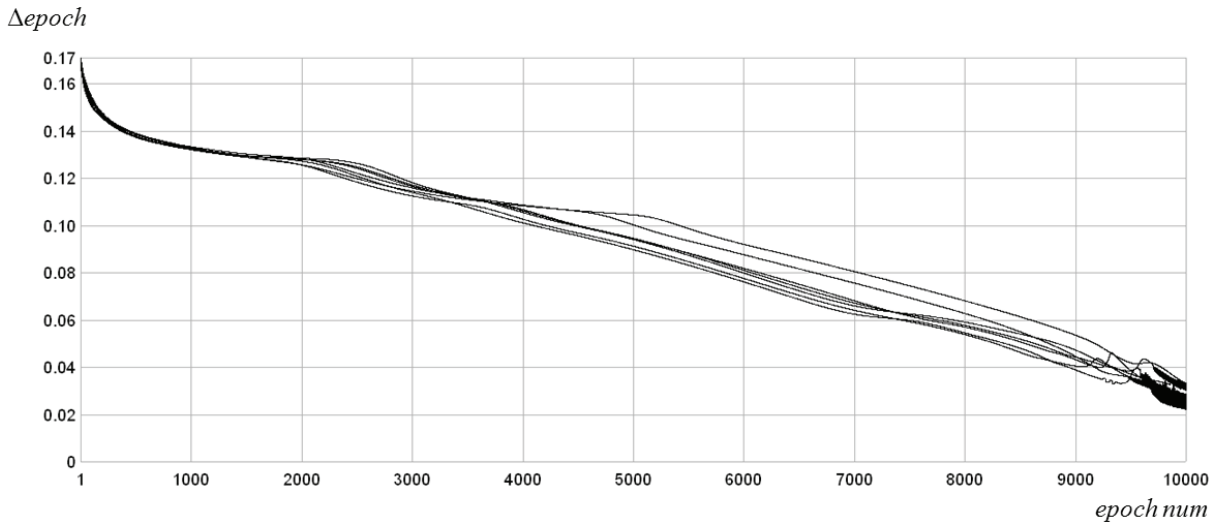


Fig. 7. Influence of the initial values of the bond weights on learning

It should be noted that the nature of training may vary for different types of classification. For some of them, when the task is reduced to the identification of fragments of typical curves of temporal dependencies, it may be advisable to use the technique of convolutional neural networks [5].

Conclusions. The article proposed to store and analyze the description of student achievements in the form of a student's trick in time. The possibilities of using neural networks for the analysis of the regularities of the educational process were explored. The influence of individual factors on the achievement of the possibilities of reliable classification of students' behavior patterns was revealed.

References

1. АІС "Навчальні плани". КБІС НТУУ "КПІ". URL: http://kbis.kpi.ua/kbis /index.php?option=com_content&task=view& id=84&Itemid=91

2. ПОЛІТЕК СОФТ. Програмне забезпечення для вищих навчальних закладів України. [Електронний ресурс]. – Режим доступу: http://www.politek-soft.kiev.ua/

3. И. И. Казмина, Е.В. Нужнов. Интеллектуальная поддержка образовательных процессов на уровне специальности (профиля) // Открытое образование. – 2013. – №6. – С.80-84.

4. Time Series Classification Web-Site. URL: http://timeseriesclassification.com.

5. Zhicheng Cui, Wenlin Chen e Yixin Chen. “Multi-Scale Convolutional Neural Networks for Time Series Classification”. A: CoRR abs/1603.06995 (2016). URL: http://arxiv.org/abs/1603.06995.

6. Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. "Learning representations by back-propagating errors". Nature. Vol 323. October 1986. 533–536pp. doi:10.1038/323533a0. ISSN 1476-4687.

Authors

Victor Poriev – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: v_porev@ukr.net

Порєв Віктор Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

В. М. Порєв

ДЕЯКІ АСПЕКТИ ПОБУДОВИ ІНТЕЛЕКТУАЛЬНОЇ ПРОГРАМНОЇ СИСТЕМИ ДЛЯ ПІДТРИМКИ НАВЧАЛЬНОГО ПРОЦЕСУ

Актуальність теми дослідження. Розвиток систем штучного інтелекту надає широкі можливості для вдосконалення інформаційних систем підтримки навчального процесу, наділення їх новими функціональними можливостями.

Постановка проблеми. Обмеженість людських можливостей і необхідність охоплення величезних обсягів інформації, знань та досвіду, які накопичуються упродовж десятків років призводить до необхідності появи комп'ютерних інтелектуальних систем, які здатні не тільки накопичувати, а й аналізувати дані та процеси. Це повною мірою стосується інформаційних систем для підтримки освітньої діяльності. Забезпечення можливості таких систем автоматично зберігати окрім традиційних даних також і часові ряди породжує проблему аналізу та класифікації великих обсягів такої інформації.

Аналіз останніх досліджень і публікацій. В результаті аналізу літературних джерел, присвячених системам підтримки автоматизації навчального процесу, можна зробити висновок, що ці системи будуються на платформах систем керування базами даних; і основна функція, яка реалізується у таких системах – зберігання та накопичення навчальних планів.

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена питанням стосовно методів і засобів аналізу інформаційними системами часових рядів, які описують освітню діяльність.

Постановка завдання. Завданням є знайти формат впорядкування часових даних, вибрати конфігурацію нейронної мережі та визначити методику та параметри навчання мережі для забезпечення потрібного рівня достовірності класифікації часових рядів в освітньому процесі.

Викладення основного матеріалу. Висвітлено необхідність зберігання та аналізу часових рядів для навчальної діяльності. Запропоновано формат індивідуального портрету діяльності студента у часі. Запропоновано використовувати штучні нейронні мережі для аналізу таких даних. Досліджено вплив окремих параметрів на навчання мережі, яка класифікує типи поведінки студентів.

Висновки. Розглянуто підхід до створення інтелектуальної інформаційної системи для підтримки навчального процесу. Запропоновано методологічну основу для побудови елементів комп'ютерної інтелектуальної інформаційної системи. Наведені приклади реалізації окремих функцій системи підтримки навчального процесу.

Ключові слова: аналіз, класифікація часових рядів, інтелектуальна інформаційна система, нейронна мережа, навчання нейронних мереж, підтримка навчального процесу.

UDC 004.032.26

Mykhailo Novotarskyi

ASYNCHRONOUS METHOD FOR ACTOR-CRITIC REINFORCEMENT LEARNING

Михайло Новотарський

АСИНХРОННИЙ МЕТОД НАВЧАННЯ З ПІДКРІПЛЕННЯМ ДЛЯ АКТОР-КРИТИК АРХІТЕКТУР

В роботі розглянуто метод глибокого навчання з підкріпленням, яке ґрунтується на застосуванні асинхронного підходу при використанні градієнтного спуску. На основі запропонованого методу побудовано актор-критик алгоритм, що характеризується більшим стабілізуючим ефектом при навчанні у порівнянні з існуючими паралельними методами. Крім того, запропонований підхід дозволяє розпаралелювати процес навчання шляхом застосування властивості багатоядерності сучасних комп'ютерів.

Ключові слова: навчання з підкріпленням, актор-критик алгоритм, паралельне навчання, асинхронний градієнтний спуск.

Рис.: 9. Бібл.: 6.

The method of deep reinforcement learning is considered in the work, which implements an asynchronous approach with the use of gradient descent. An actor-critic algorithm based on the proposed method is constructed. Parallel asynchronous algorithm is characterized by a greater stabilizing effect during learning process compared with existing parallel methods. In addition, the proposed approach allows parallelizing the learning process by applying the multi-core properties of modern computers.

Key words: reinforcement learning, actor-critic algorithm, parallel learning, asynchronous gradient descent.

Fig .: 9. Bibl.: 6.

Introduction. In connection with the growing popularity of algorithms concerning the field of artificial intelligence, the number of new approaches to the development of reinforcement learning algorithms has significantly increased.

At the same time, deep neural networks are a promising means of implementing artificial intelligence algorithms. However, application of reinforcement learning algorithms to the mentioned networks is problematic, which is caused by fundamentally unstable learning processes. In order to overcome this problem considerable efforts have been made in recent years [1,2]. The main result of the study

was that the main direction of increasing the sustainability of the learning process is to use the previous experience of the agent to correct the following actions.

However, this approach also has its own negative sides, because it requires additional memory and additional calculations per agent activity. Therefore, the main direction of further development of these algorithms lies in the field of parallelization. It is known that the fundamental difference between parallel algorithms from sequential is the presence of a phase of interaction in which parallel fragments must perform information exchange for the further successful advancement of local sub-processes that form a common parallel process. From the point of view of the total time of the algorithm, the interaction phase has a negative effect and therefore it is natural to combine this phase with the phase of computation, which leads to asynchronous parallel algorithms.

One of the well-known approaches that uses a parallel asynchronous algorithm for generalized reinforcing learning is Gorila (General Reinforcement Learning Architecture) [3]. In the Gorila algorithm, each agent is represented as a local sub-process that operates in a separate copy of the medium with its own memory for the accumulation of experience, and a learner who selects data from memory and calculates loss gradients. The gradients are asynchronously transmitted to the central parameter server, which updates the main copy of the model. Updated policy settings are sent to agent-learner at certain intervals. Such an approach makes sense for distributed systems because it is characterized by coarse-grained parallelism.

The algorithm "Sarsa" was proposed in [4]. This algorithm, as in the previous case, uses several parallel sub-processes such as "agent-learner" to accelerate the overall learning process. Each local sub-process shows a separate operation and periodically exchanges information about the learning outcomes using direct communications with other sub-processes. This approach allows for flexible computational organization, but leads to a significant increase in the time of interaction when there is an increase in the number of local sub-processes.

On the basis of the generalization of the above-mentioned approaches, a framework of asynchronous algorithms was proposed in [5] to parallelize the reinforcement learning process. In this framework, the parallel development algorithm "Sarsa" was further developed, algorithms of the actor-critic method and the n-stepping methods were considered.

The comparative characteristics of learning outcomes based on a wide range of research in various fields of practical application have shown that actor-critic methods have the best prospects. In this paper, the algorithm of parallel asynchronous learning based on the actor-critic method is considered. This algorithm allows for scaling and can be implemented without significant changes, so in multicore processors as in cloud environments.

Basics of the reinforcement learning theory. The reinforcements learning is used in algorithms to achieve a certain complex goal by maximizing the target

parameter. As a rule, the goal is achieved by performing a significant number of steps in a given direction. The choice of the right direction of movement is achieved by choosing the appropriate rewards and punishments.

The actor-critic structure consists of two main parts, as shown in fig.

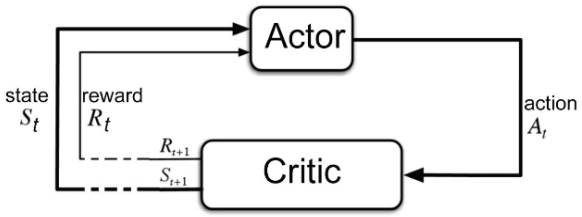


Fig. 1. The actor-critic structure

The terminology for such algorithms includes such concepts as actors, critics, states, rewards, etc. The actor performs certain activities that form a set of activities A. As a result of these actions, the actor changes his current state to a state that is an element of the set of possible states S. If these activities move the learning process into the right direction, then the actor receives an reward from the set of rewards R. The actor chooses the next activity based on an analysis of the current state by applying a policy π. The essence of π policy is to provide the actor with the greatest rewards from the next activity. In addition to the short-term benefits of current activity, the actor is also guided by the long-term profit that he can get in the long run after a certain amount of activity. Therefore, we will determine $V^{\pi}\left(s\right)$ as expected long-term profits from the current state s because of π policy implementation. There is one more long-term profit Q, which differs from V that depends not only on the current state, but also on the current activity $a \in A$. So $Q^{\pi}\left(s,a\right)$ means the long-term profits from the current state, if the activity is carried out as a result of policy π. Thus at any moment t, the policy π maps the state s_t to the activity a_t. After performing the activity a_t, the actor goes into the next state s_{t+1} and receives a reward, which is represented by a scalar value r_t. After k steps, starting with the step t, the actor can accumulate total profits $R_t = \sum_k \gamma^k r_{t+k}$, where $\gamma \in \left[0,1\right]$ is the discount factor. The main objective of the actor is to maximize profits from each state s_t. Consequently, long-term profit from the implementation of activity a_t in a state s_t is determined from the expression $Q\left(s_t,a_t\right) = \max_{\pi} Q^{\pi}\left(s_t,a_t\right)$ by choosing the maximum long-term profit achieved by any policy. With the use of artificial neural networks, as a universal approximator, we obtain the optimal function of long-term profit $Q\left(s,a\right) \approx Q\left(s,a;\theta\right)$. For this case, the parameter θ is optimized by the

iterative process of minimizing loss functions, where i -th loss function can be given by an expression

$$L_i\left(\theta_i\right)=\mathrm{E}\left(r+\gamma\max_{a'}Q\left(s_{t+1},a_{t+1};\theta_{i-1}\right)-Q\left(s_t,a_t;\theta_i\right)\right)^2,$$

where s_{t+1} is the state that arises after the state s_t and a_{t+1} is the activity that goes after the activity a_t.

This is one-step learning method, since the update of the value of the profit takes place after every step in value $r+\gamma\max_{a_{t+1}}Q\left(s_{t+1},a_{t+1};\theta\right)$.

This is the simplest approach since the reword r depends directly on the previous state and activity. Its main disadvantage is that the history of previous states s and activities a can only be taken into account indirectly through updates $Q\left(s,a\right)$. This leads to a slower learning due to reducing the sustainability of the learning process.

The natural way to increase the stability of the learning process is to apply multi-step methods. In the case of using an artificial neural network to approximate the function of long-term profit $Q\left(s,a\right)\approx Q\left(s,a;\theta\right)$, we receive an update after n steps that corresponds to the value

$$r_t+\gamma r_{t+1}+\ldots+\gamma^{n-1}r_{t+n-1}+\max_a\gamma^n Q\left(s_{t+n},a\right).$$

Therefore, the parameter θ can be optimized by minimizing loss functions $L_i\left(\theta_i\right)$ only after n steps, where the loss function can be given by expression

$$L_i\left(\theta_i\right)=\mathrm{E}\left(r_t+\sum_{k=1}^{k=n-1}\gamma^k r_{t+k}+\max_a\gamma^n Q\left(s_{t+n},a,\theta_{i-1}\right)-Q\left(s_t,a_t;\theta_i\right)\right)^2.$$

A significant disadvantage of this approach is the low degree of system response to changing environmental conditions. Because the result is formed on the basis of previous steps. Thus, the multi-step method can increase the stability of the learning method by increasing the number of calculations and increasing the response time to a situation that arises due to changes in the parameters of the environment.

Asynchronous learning. In this paper, it is proposed an approach to reinforcement training that combines the benefits of one-step and multi-step methods for the actor-critic structures. The main direction of further development reinforcement learning algorithms for deep neural networks lies in the field of parallelization. Therefore, in this approach, it is suggested to use parallel computing with shared RAM. This allows you to effectively use the multi-core architecture of modern

processors. At the same time, it is also important to have an effective computation organization in such a parallel structure. It is known that the fundamental difference between parallel algorithms from sequential is the presence of an interaction phase. The organization of parallel computing with synchronous and asynchronous interaction of processes is shown in fig. 2.

We can see that in the case of asynchronous computing we have the advantage of combining the computation and interaction phases.

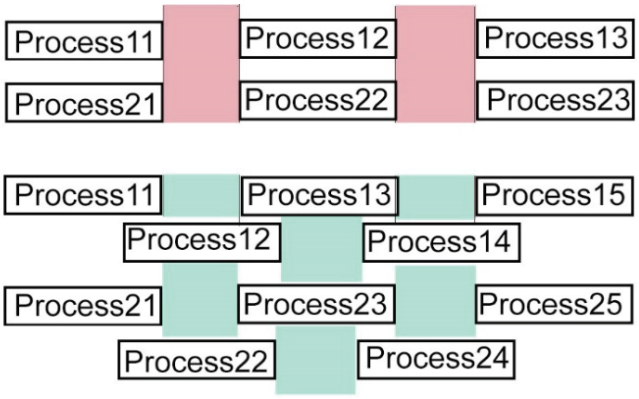


Fig. 2. Synchronous and asynchronous calculations

Our parallel asynchronous actor-critic of the structure is shown in fig. 3.

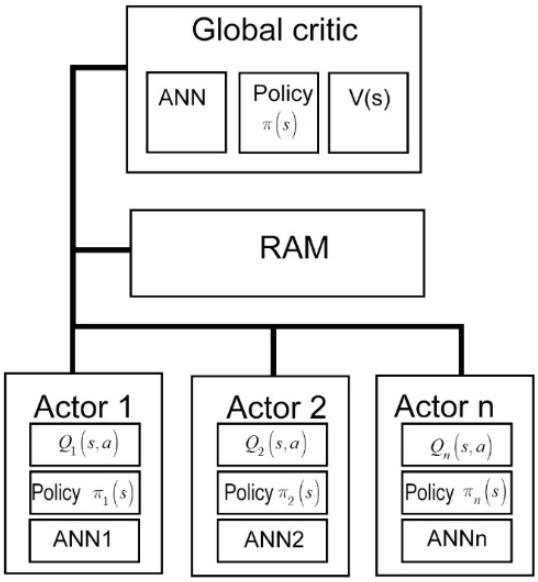


Fig.3. Parallel asynchronous actor-critic structure

The proposed parallel structure contains n actors that can simultaneously accumulate experience of the environment, guided by local politics $\pi_i(s_t)$. As shown in [6], such an approach can be considered as an actor-critic architecture for which a separate policy π is typical for actors and the critic is determined by some value

function $V\left(s_t\right)$. Then the value $R_t - V\left(s_t\right)$ can be considered as an estimate of some advantage for activity a_t in the state s_t.

If we denote by $A\left(s_t, a_t, \theta\right)$ the function of the advantage generated by artificial neural networks with parameters θ, then we obtain the general expression for parallel asynchronous method

$$A\left(s_t, a_t; \theta\right) = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^k V\left(s_{t+n}; \theta\right) - V\left(s_t; \theta\right).$$

The parallel asynchronous algorithm is proposed for reinforcement learning which is based on the described approach. The pseudocode of this algorithm is shown in fig. 7.

```
Set shared and local parameters θ_t and θ_{t+1}
Set the shared an local counter T = 0 and t = 1
While T < T_max:
    dθ = 0; θ_{t+1} = θ_t; t_start = t
    Get(s_t)
    While s_t != s_final or t - t_start < t_max:
      Calculate a_t according to π(a_t | s_t; θ_{t+1})
      T+=1; t+=1
    If s_t == s_final: R=0 else R = V(s_t, θ_{t+1})
    For i in {t-1,…,t_start}:
        R = γr_i + R; Calculate θ_{t+1} = dθ + ∇_{θ_t} A(s, a; θ)
    Asynchronous update of the θ parameters.
```

Fig. 7. Pseudo-code for parallel asynchronous algorithm

Practical implementation with the cart-pole problem. The cart-pole also known as an inverted pendulum with a center of gravity above its pivot point. It is unstable and falls over but can be controlled by moving the cart.

The goal of the problem is to keep the pole balanced by moving the cart left or right by applying appropriate forces to the pivot point. Fig. 8 demonstrates the simplest implementation of the cart-pole task.

This implementation made on the base of gym library. It is a collection of test problems. We can use them to work out our reinforcement learning algorithms. These environments have a shared interface that allows writing the general algorithms.

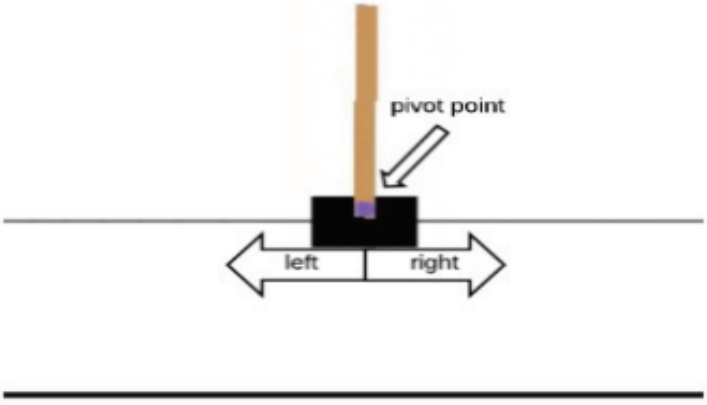


Fig. 8. A typical graphical representation of the cart-pole task

Investigation of the received algorithm showed its high stability while maintaining the high speed of learning. In fig. 9 we can show an example of comparison of learning speed using asynchronous parallel method and parallel one-step Q - method. The x-axis shows training time in hours and the y-axis shows the average score.

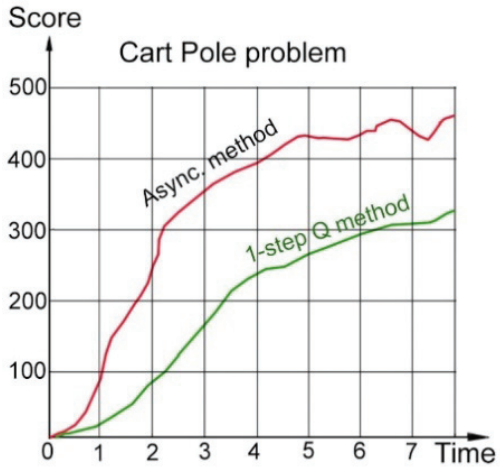


Fig. 9. Comparison of asynchronous and single-step parallel methods

Our asynchronous method shows significant speedups from using greater numbers of parallel actors.

Conclusions. In connection with the development of robotics, reinforcement learning is highly relevant as it is an important part of artificial intelligence technology. The main problem for reinforcement learning of convergent multilayer networks remains the stability of the method, which significantly influences the learning speed. The paper proposes an asynchronous parallel method for reinforcement learning, which allows increasing the speed of learning in comparison with one-step parallel methods.

References

1. Mnih V., Kavukcuoglu K., Silver D., et al.–Nature, 2015.–vol.518, №7540.– P.529-533

2. Van Hasselt H., Guez A., Silver D. Deep reinforcement learning with double q-learning, 2015.– *preprint arXiv:1509.06461*.

3. Nair A., Srinivasan P., Blackwell S., et al. – Massively Parallel Methods for Deep Reinforcement Learning, 2015. – *arXiv:1507.04296v2.*

4. Grounds M., Kudenko D. Parallel reinforcement learning with linear function approximation // *Proceedings of the 5th, 6th and 7th European Conference on Adaptive and Learning Agents and Multi-agent Systems: Adaptation and Multi-agent Learning*. – Springer-Verlag: 2008, P. 60–74.

5. Mnih V., Badia A., Mirza M., et, al. Asynchronous Methods for Deep Reinforcement Learning, 2016. – arXiv: 1602.01783 [cs.LG]

6. Sutton, R. and Barto, A. *Reinforcement Learning: an Introduction*. MIT Press, 1998. – 548 p.

Довідка про автора

Новотарський Михайло Анатолійович – професор, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Novotarskyi Mykhailo – professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”. E-mail: novot@ukr.net

Розширена анотація

Сучасні підходи до створення алгоритмів у сфері штучного інтелекту включають як створення нових, так і модернізацію відомих методів навчання без учителя, зокрема, навчання з підкріпленням. Водночас перспективним засобом реалізації штучного інтелекту є глибокі нейронні мережі. Проте застосування алгоритмів навчання з підкріпленням до згаданих мереж викликає певні проблеми, обумовлені принципово нестійкими процесами навчання. Основним напрямком підвищення стійкості процесу навчання є використання агентом попередньо набутого досвіду з метою коригування наступних дій. Слід відмітити, що такий підхід має свої негативні сторони, оскільки вимагає додаткової пам’яті та додаткових обчислень з розрахунку на одну активність агента. Тому основний напрямок подальшого розвитку даних алгоритмів лежить у сфері розпаралелювання.

У даній роботі розглянуто метод глибокого навчання з підкріпленням, який ґрунтується на застосуванні асинхронних паралельних алгоритмів при реалізації градієнтного спуску. На основі запропонованого методу побудовано актор-критик алгоритм, який характеризується більшим стабілізуючим ефектом при навчанні у порівнянні з існуючими паралельними методами. Крім того, запропонований підхід дозволяє розпаралелювати процес навчання шляхом застосування властивості багатоядерності сучасних комп’ютерів.

Section 1. SEC (Security of computer systems and networks. Fault-tolerant distributed computing)

UDC 004.8

Larysa Doroshenko,
Oleksandr Markovskyi, Andrii Honchar

ORGANIZATION OF RESERVATION AND RECONSTRUCTION OF DATA

Лариса Дорошенко,
Олександр Марковський, Андрій Гончар

ОРГАНІЗАЦІЯ РЕЗЕРВУВАННЯ ТА ВІДНОВЛЕННЯ ДАНИХ

In the article the method of additional packets formation during the information transfer on the Internet is offered. Also suggested is the way to use additional packages to recover lost or corrupted packets. Linear redundant codes are used to recover. An example of using the proposed method is given.

Key words: erasure codes, information packets, data packet reconstruction, linear codes.

Fig. 0. Tabl.: 1. Bibl.: 4.

У статті запропоновано метод формування додаткових пакетів при передачі інформації в Інтернет. Також запропоновано спосіб використання додаткових пакетів для відновлення втрачених або пошкоджених при передачі основних пакетів. Для відновлення використовуються лінійні надлишкові коди. Наведено приклад застосування запропонованого методу.

Ключові слова: відновлюючі коди, інформаційні пакети, відновлення пакетів даних, лінійні коди.

Рис. 0. Табл.: 1. Бібл.: 4.

Target setting. In the modern world the role of the Internet is dominant among the means of data exchange. Since there is a probability of loss, damage or delayed arrival of data packets over a global network, there is a need to create new methods and tools to solve these problems [1].

Actual scientific researches and issues analysis. Today the following technologies are used to reconstruct lost data during transmission: partial or complete duplication of data and their re-transmission, usage of the Reed-Solomon correction codes (the disadvantage is the reduction of the data to the solution of the nonlinear equations system, which is time-consuming), the usage of CRC-codes [2, 3]. The last two technologies perform error correction at the symbol level, which doesn`t allow the recovery of lost packets during transmission [4].

Not investigated parts of the general subject. At present great attention is paid to erasure codes that work at the packet level. The main principle is the transfer of basic packages along with additional ones, which partially contain the information that is in the main packets. With the help of erasure codes. It is possible to reconstruct lost or damaged packets with a certain probability. However, the erasure codes that are used do not provide high data reconstruction efficiency.

Research objective. The purpose of the article is to investigate the dependence of the lost data packets reconstruction on the number of main and additional data packets. Analyze the extermination results and form on the basis of the analysis the method of data packets reconstruction with the highest probability.

Principal statements. The information packet is divided into n blocks, each of which is transmitted separately. The block, with the number j, where $j \in \{1,..,n\}$, consists of a sequence of words: $a_{j1},a_{j2},\ldots,a_{jm}$. In the process of transmission in case of loss or distortion of block data with numbers q and r, where $q,r \in \{1,\ldots,n\}$ it is necessary to restore the words $a_{q1},a_{q2},\ldots,a_{qm}$ and $a_{r1},a_{r2},\ldots,a_{rm}$ from information, stored in n-2 main and k backup blocks. To ensure the possibility of simple data recovery when one carrier is lost, it is proposed in the first backup block to store the sum of the module 2 of all the same words of the main blocks:

$$\forall j = 1,..., n : s_{1j} = \bigoplus_{i=1}^{n} a_{ji} \quad (1)$$

Considering the task of reconstruction the q and r main units of the words $a_{q1},a_{q2},\ldots,a_{qm}$ and $a_{r1},a_{r2},\ldots,a_{rm}$, their values can be found as a result of solving systems of two 30oolean equations of the form:

$$\begin{cases} a_{qj} + a_{rj} = z_{1j} \\ a_{qj} = z_{2j} \end{cases} \text{ or } \begin{cases} a_{qj} + a_{rj} = z_{1j} \\ a_{rj} = z_{3j} \end{cases}, \ \forall j \in \{1,\ldots, n\}, \quad (2)$$

where the '+' symbol is denoted by a binary add operation in module 2. The first equation in systems (2) is transformed from (1), that is, the data stored in the first data backup block are used to obtain it. Obtaining the first equation of systems (2) is much more complicated, due to the fact that the a priori values of q and r are unknown. It is obvious that the second equation of system (2) can be obtained from the system of linear 30oolean equations, which for any values of q and r contains an equation in which only a_{q1} or only a_{r1} is included in the term. If n is a degree 2, then the example of such a system may look like:

$$\begin{cases} a_{1j} + a_{2j} + ... + a_{m/2,j} = y_j \\ a_{1j} + a_{2j} + ... + a_{m/4,j} + a_{m/2+1,j} + ... + a_{3m/4,j} = y_2 \\ \ldots \\ a_{1j} + a_{3j} + a_{5j} + ... + a_{m-1,j} = y_{\log_2 n,j} \end{cases}, \forall j \in \{1,...,n\} . \quad (3)$$

In the general case of arbitrary value n, the number of equations of the system of the form (3) is equal to the nearest whole, equal to or greater than $\log_2 n$: $\lceil \log_2 n \rceil$. In order to reconstruct the information packet when any two blocks (main or backup) are lost, it is necessary to save in the same backup unit the sum of the two basic words of the same name (1), the amounts determined by the system (3) in the backup blocks and in one more backup unit duplicate the last information block.

Thus, if each word from a pair of blocks that is lost or damaged during data transfer is reconstructed independently, then the number of backup units (packets) of the media is:

$$kp = 2 + \lceil \log_2 n \rceil.$$

General structure. The reconstruction of the lost "pack" of information packets is proposed to be organized as follows. Assuming that the lost packets belong to one, for the definiteness of the i-th block, and have sequence numbers from j to (j + l) in the block, where $j \in \{1,2,\ldots,h\cdot(l\text{-}1)\}$, then it is quite obvious that the number of lost packets in each of the columns of the matrix M does not exceed one. Accordingly, the recovery of p_{ij}, p_{ij+1}, …, $p_{i(j+l)}$ packets is performed by calculating the amount by module 2 of the received packs columns and the corresponding backup code. Formally, the recovery procedure is described by the formula:

$$p_{ij} = C_{i,j \bmod l} \oplus \bigoplus_{q=0, q \neq j \, div l}^{h-1} p_{i,q\cdot l + j \bmod l}. \quad (4)$$

The proposed method is actually specialized for the dominant type of information packets loss when broadcasting video information transmission in peer-to-peer networks. The analysis showed that the main reason for such losses is the exclusion of the network node, through which the main stream of video information is transmitted. Accordingly, the dominant type of the continuity violation of delivery to subscribers of video information is the loss of a group of consecutive packages that form a "pack."

Testing. The analysis of the data presented in table 1 indicates that the proposed approach provides the reconstruction possibility of the loss of blocks of large multiplicity with a probability close to one for data blocks of real length. In this case, the number of operations required to determine the positions of blocks, distorted during the transfer does not exceed $\log_2 m$. This means that the error correction of a large multiplicity can almost be done at the rate of data transmission. Situations, in which the correction is impossible, can be easily detected by analyzing the differences between the components of the control code. In this case, the errors can be corrected by re-transmission, but the probability of this, as seen from table 1, is sufficiently small considering the actual probability of damage of the blocks during the transfer.

Table 1

Experimental probability of data recovery

The number of main data blocks	*Number of damaged blocks*	*The number of backup data blocks*			
		2	*3*	*4*	*5*
6	2	0.28	0.67	0.87	1
	3	0	0.34	0.52	0.86
	4	0	0	0.37	0.47
8	2	0.32	0.72	0.90	1
	3	0	0.42	0.63	0.82
	4	0	0	0.47	0.59
10	2	0.39	0.76	0.92	1
	3	0	0.44	0.64	0.79
	4	0	0	0.49	0.62

Thus, as a result of the conducted research, the theoretical substantiation and development of the method of data packet reconstruction during their transmission in global computer networks on the basis of linear redundant codes, which are executed in the form of reserving packets, is carried out. Development of a method of formation by reserving packages and methods for reconstruction lost packets. The simulation carried out for the experimental study of the use of linear codes for the reconstruction of data packets in order to select the optimal characteristics for different types of channels and data transmission errors confirmed, in general, the results obtained theoretically.

Conclusion. As a result of the conducted research, the method of reservation and reconstruction of data packets for transmission in the global networks was proposed.

To achieve the goal, a method is proposed for the formation of additional packets and reconstruction lost, damaged or delayed critical time of the main data packets using them.

References

1. Tanenbaum A.S. Computer networks. Prentice Hall PTR.- 2016.-960 p.

2. Стіренко С.Г., Габінет А.В., Костенко Ю.В. Забезпечення безперервного відтворення потокового відео в однорангових мережах з використанням erasures кодів. // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: збірник наукових праць. – К.: «Век+», 2015. – № 62. – С. 105–110.

3. Leong D. Erasure coding for real-time streaming / D. Leong, T. Ho // Proceedng IEEE Int. Symposium Information Theory – ISIT-2012 . – 200.

4. Brinkmeier M., Fischer M., Grau S., Schaefer G., Strufe T. Methods for Improving Resilience in Communication Networks and P2P Overlays. PIK // Praxis der Informationsverarbeitung und Kommunikation 32, 2009.

Autors

Doroshenko Larysa – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: larysad785@gmail.com

Дорошенко Лариса Юріївна – студентка, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Markovskyi Oleksandr – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: markovskyy@i.ua

Марковський Олександр Петрович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Honchar Andrii – student, Department of Mobile and video information technology, State University of Telecommunications.

E-Mail: andrew.gonchar98@gmail.com

Гончар Андрій Анатолійович – студент, кафедра Мобільних та відеоінформаційних технологій, Державний Університет Телекомунікацій.

РОЗШИРЕНА АНОТАЦІЯ

Л. Ю. Дорошенко,
О. П. Марковський, А. А. Гончар

ОРГАНІЗАЦІЯ РЕЗЕРВУВАННЯ ТА ВІДНОВЛЕННЯ ДАНИХ

Актуальність теми дослідження. У сучасному світі спостерігається зростання домінуючої ролі Інтернету в процесі передачі інформації, що вимагає розробки нових засобів та методів для забезпечення надійності пересилки пакетів даних.

Постановка проблеми. При передачі інформації через глобальні мережі можлива втрата, пошкодження або запізнення отримання (по настанню критичного моменту часу) пакетів даних. Використання запропонованого методу підвищує вірогідність відновлення втраченої інформації під час передачі.

Аналіз останніх досліджень і публікацій. На сьогоднішній день використовується декілька технологій для відновлення втраченої інформації. Найпростішою є часткове або повне дублювання втрачених пакетів даних і повторна їх передача, використання коригуючих кодів Ріда-Соломона та CRC-кодів для виправлення помилок на рівні символів, використання відновлюючих кодів.

Виділення недосліджених частин загальної проблеми. Стаття присвячена формування методу побудови додаткових пакетів даних при передачі інформації та аналізу залежності надійності передачі даних від способу формування додаткових пакетів. Найбільшу увагу приділено розробці методу формування додаткових пакетів для найбільшої вірогідності відновлення втрачених пакетів.

Постановка завдання. Завданням є розробити метод формування додаткових пакетів, який би дав можливість найбільш ефективно відновлювати втрачені та пошкоджені основні пакети даних.

Викладення основного матеріалу. Проведено дослідження та аналіз ймовірностей відновлення втрачених пакетів в залежності від кількості посилання додаткових пакетів та кількості втрачених пакетів. Визначено спосіб формування додаткових пакетів для найбільшої ймовірності відновлення втрачених пакетів.

Висновки. Досліджено залежності відновлення пакетів даних від способу резервування даних та кількості додаткових пакетів. Проаналізовано експериментальні дані та виділено найкращі результати, на основі яких сформовано запропонований метод.

Ключові слова: відновлюючі коди, інформаційні пакети, відновлення пакетів даних, лінійні коди.

UDC 004.8

Anna Doroshenko,
Oleksandr Markovskyi

ACCELERATION OF BOOLEAN TRANSFORMATIONS NONLINEARITY TESTING FOR CRYPTOGRAPHIC ALGORITHMS

Анна Дорошенко,
Олександр Марковський

ПРИСКОРЕННЯ ТЕСТУВАННЯ НЕЛІНІЙНОСТІ БУЛЕВИХ ПЕРЕТВОРЕНЬ КРИПТОГРАФІЧНИХ АЛГОРИТМІВ

The method proposed in this article allows to significantly accelerate the testing of balanced 35oolean transformations nonlinearity by successive reconstruction of the nearest by Hamming distance linear function to a given nonlinear function. The results of experimental simulation, which prove the effectiveness of the proposed method, are presented.

Key words: Boolean transformations nonlinearity, linear cryptanalysis, cryptographic algorithms testing, cryptoresisting measuring.

Fig.: 1. Tabl. 0. Bibl.: 4.

Запропонований у статті метод дозволяє значно прискорити тестування нелінійності балансних булевих перетворень шляхом послідовного відновлення найближчої за відстанню Геммінга лінійної функції до заданої нелінійної функції. Наведено результати експериментального моделювання, що доводять ефективність запропонованого методу.

Ключові слова: нелінійність булевих перетворень, лінійний криптоаналіз, тестування криптографічних алгоритмів, вимірювання криптостійкості.

Рис.: 1. Табл. 0. Бібл.: 4.

Target setting. The automatic construction of cryptographic algorithms systems, the stages of which are the Boolean transformations generation and the determination of their cryptostability, are becoming increasingly popular in recent years. Cloud technologies make the large amount of information processing and complex calculations performing possible. Therefore, development of new cryptographic algorithms with increased cryptoresistance has become a current topic [1, 2].

Actual scientific researches and issues analysis. Currently only for the narrow class of Boolean functions there are nonlinearity evaluation methods which do not use a brute force. Thus, the problem of the nonlinearity evaluation has an

exponential complexity, depending on the number of variables n. Majority of existing methods designed for arbitrary Boolean functions achieved the acceleration of nonlinearity evaluation by either narrowing the problem or defining a nonlinearity with a predetermined error [3].

Not investigated parts of general subject. Existing methods do not take into account such features of Boolean transformations, which are used in cryptographic protection algorithms in practice, as balancedness. And as a result, these methods cannot meet the current requirements for testing cryptographic algorithms.

Research objective. The objective of this paper is to propose and investigate a new method that will allow to organise more effective cryptostability testing of data protection algorithms based on Boolean transformations by accelerating the nonlinearity determination.

Principal statements. Nonlinearity is the mandatory property of irreversible Boolean transformations [4]. Such Boolean transformation is represented by a system of nonlinear balanced Boolean functions $f_1(x_1,x_2,\ldots x_n)$, $f_2(x_1,x_2,\ldots x_n)$, … $f_k(x_1,x_2,\ldots x_n)$ from n variables, $\forall i \in \{1,2,...n\} : x_i \in \{0,1\}$. Testing the Boolean transformation nonlinearity consists in the nonlinearity evaluation of each of these nonlinear Boolean functions. Nonlinearity testing of Boolean function $f(x_1,x_2,\ldots x_n)$ is based on its Hamming distance $HD(f, g)$ to a linear function $g(x_1,x_2,\ldots x_n)$: $g(x_1, x_2,...x_n) = a_0 \oplus a_1 x_1 \oplus a_2 x_2 \oplus ... \oplus a_n x_n$, where $\forall j \in \{1,2,...n\} : a_j \in \{0,1\}$. The Hamming distance $HD(f, g)$ is determined by the number of input sets $x_1,x_2,\ldots x_n$ on which the function $f(x_1,x_2,\ldots x_n)$ and the linear function $g(x_1,x_2,\ldots x_n)$ take different values (Hamming weight of the functions exclusive disjunction), namely, the Hamming distance $HD(f, g)$ can be determined by the formula:

$$HD(f, g) = \sum (f(x_1, x_2,...x_n) \oplus g(x_1, x_2,...x_n)). \quad (1)$$

Balanced Boolean function $f(x_1,x_2,\ldots x_n)$ from n variables is a function whose weight is equal to:

$$HW(f) = 2^{n-1} \quad (2)$$

Linear Boolean function is a function that does not contain the product of variables in Zhegalkin polynomial form.

In turn, the nonlinearity $NL(f)$ of the Boolean function $f(x_1,x_2,\ldots x_n)$ is the minimal Hamming distance, namely, the Hamming distance to the nearest linear function. The nonlinearity $NL(f)$ can be determined by the formula:

$$NL(f) = \min HD(f, g). \quad (3)$$

To ensure the cryptoresistance of Boolean transformations against cracks using linear cryptanalysis, they must have the most possible nonlinearity.

General structure. The linear approximation constructing principle, which is basic for the proposed method, is the use of the probability p_i of changing the function value while the inverting of the i-th variable x_i of the Boolean function $f(x_1,x_2,\ldots x_n)$. The probability p_i can be determined by the formula:

$$p_i = \frac{1}{2^n} \cdot \sum_{X \in \Omega} f(X) \oplus f(X \oplus C_i), \quad (4)$$

where Ω is the set of all 2^n possible vectors X, $C_i=\{c_{i1},c_{i2},\ldots,c_{in}\}$ is an n-bit binary vector whose i-th component is equal to 1 and all others are 0: $\forall l \in \{1,..,i\text{-}1,i\text{+}1,\ldots,n\}$: $c_{il}=0$, $c_{ii}=1$.

Analysing the vector $P=\{p_1, p_2,..., p_n\}$ of probability values, which of the coefficients $a_1,a_2,\ldots,a_n$ in the linear equation will have the value of 1 can be assumed. In the process of analysis the values of the vector P in descending order in n steps, a set $\Theta_n(x_1, x_2,..., x_n)$ of all possible linear functions from $x_1,x_2,\ldots x_n$ which have the smallest Hamming distance to a given nonlinear Boolean function $f(x_1,x_2,\ldots x_n)$ is obtained.

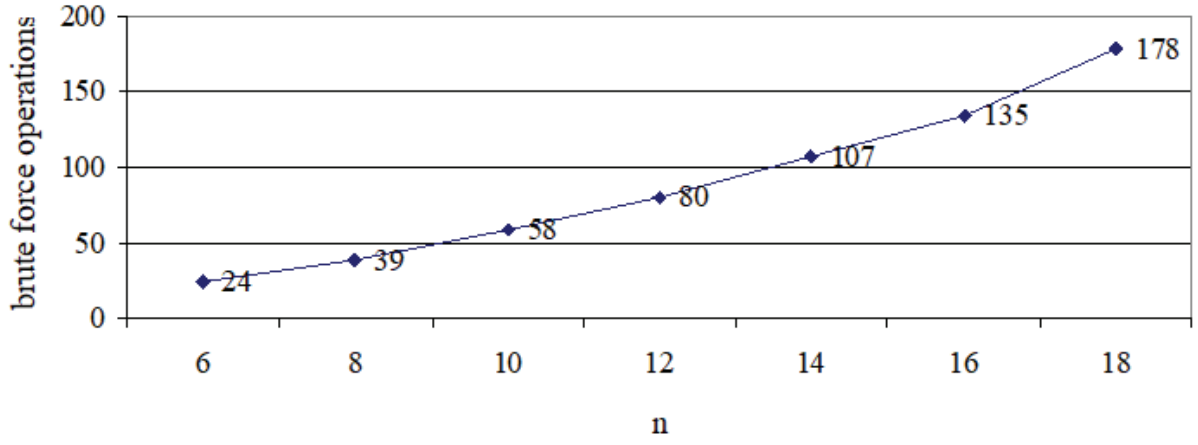
Testing. The purpose of experimental modelling was to analyse how the proposed method is effective in terms of the brute force operations number and how accurate the evaluated nonlinearity is.

In the experiments nonlinear Boolean functions with different number of variables have been tested. The graphs of the results are shown in Fig. 1 a) and b).

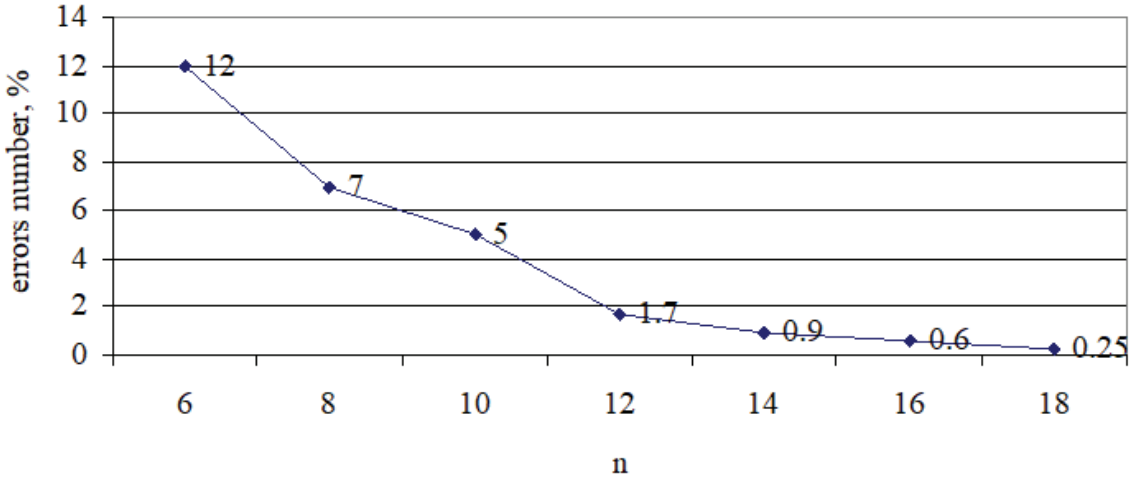
Received results illustrate that the proposed method, for example, for functions from 14 variables allows to accelerate computation approximately by $\frac{2^{14}}{107} \approx 153$ times.

Such an acceleration is achieved by the admissibility of the error in the nonlinearity evaluation.

As can be seen from Fig. 1 b), the proposed method extremely efficient for functions from a large number of variables, which is an advantage, since such functions are used in practice.



a)



b)

Fig. 1. The dependencies of a) the brute force operations number to evaluate the nonlinearity and b) the error magnitude from the variables number

Conclusion. The conducted experimental and theoretical studies have shown the effectiveness of the proposed method in testing modern algorithms of cryptographic data protection, functional basis of which is Boolean transformations.

To accelerate the construction of linear approximation, probabilities of changing the function values while the inverting of a particular variable of the Boolean function $f(x_1,x_2,\ldots x_n)$ are used, which is a peculiarity of the proposed method.

Application of the proposed method allows to provide better reliability and testing speed of wide class cryptographic algorithms.

References

5. Давиденко А.Н. Вероятностная оценка надежности реализации функций защиты информации // Моделювання та інформаційні технології: Зб.наук. праць.-Львів:НВМ ПТ УАТ.-2002.-Вип.14.-С.64-70.

6. Марковський О.П. Комбиаторный анализ булевых функций специальных классов для систем криптографической защиты информации // А.П. Марковский, Э.Р. Искаков, Г.В. Гарасимович // Збірник доповідей міжнародної науково-технічної конференції “The International Conference on Security, Fault Tolerance, Intelligence” (ICSFTI2018). – Київ, 10-11 травня 2018. – С.42-50.

7. Mesnager S. Bent Function: Fundamentals and Results / S. Mesnager // IEEE Trans. On Information Theory.-2016.- Vol.62, No. 7, , pp. 1825-1834.

8. Xiang C. A construction of linear codes from Boolean functions / C. Xiang, K.Feng, C.Taug // IEEE Trans. Inform.Theory, 2017.-Vol.63, № 1. – P. 167-176.

Autors

Doroshenko Anna – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: annadoroshenko03@gmail.com

Дорошенко Анна Юріївна – студентка, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Markovskyi Oleksandr – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: markovskyy@i.ua

Марковський Олександр Петрович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

А. Ю. Дорошенко,
О. П. Марковський

ПРИСКОРЕННЯ ТЕСТУВАННЯ НЕЛІНІЙНОСТІ БУЛЕВИХ ПЕРЕТВОРЕНЬ КРИПТОГРАФІЧНИХ АЛГОРИТМІВ

Актуальність теми дослідження. З появою хмарних технологій з’явилася можливість обробляти великий об’єм інформації та виконувати складні обчислення. Проте окрім очевидних переваг цих технологій існує вагомий недолік, пов’язаний із сумнівною безпекою даних та імовірністю їх пошкодження, викрадення або навіть знищення. Тому актуальним є розробка нових криптографічних алгоритмів з підвищеною криптостійкістю.

Постановка проблеми. Наразі існує тенденція збільшення кількості змінних булевих перетворень, які лежать в основі значної частини криптографічних алгоритмів. У зв’язку з цим постає проблема зростання часу та необхідних об’ємів ресурсів для тестування цих алгоритмів. Тому виникає нагальна потреба прискорення оцінювання криптостійкості алгоритмів захисту даних.

Аналіз останніх досліджень і публікацій. У більшості існуючих методів прискорення визначення нелінійності булевих перетворень досягається шляхом

звуженням поставленої задачі або визначенням нелінійності з наперед заданою похибкою.

Виділення недосліджених частин загальної проблеми. Існуючі методи не враховують такої особливості булевих перетворень, які використовуються в алгоритмах криптографічного захисту на практиці, як балансність. І як результат, ці методи не можуть задовольнити сучасні вимоги тестування криптографічних алгоритмів.

Постановка завдання. Завданням є запропонувати та дослідити метод, який дасть змогу більш ефективно проводити тестування криптостійкості алгоритмів захисту інформації, що мають за основу булеві перетворення, шляхом прискорення визначення їх нелінійності.

Викладення основного матеріалу. Проаналізовано недоліки існуючих методів визначення нелінійності. В основу запропонованого методу покладено концепції динамічного програмування, які дозволяють послідовно реконструювати лінійну апроксимацію за показниками імовірностей зміни значення заданої нелінійної булевої функції при інвертуванні конкретної змінної. Експериментальне моделювання показало високу ефективність запропонованого методу.

Висновки. Розроблено, теоретично обґрунтовано та досліджено метод підвищення ефективності тестування криптографічних алгоритмів шляхом пришвидшення визначення нелінійності булевих перетворень, які є основою цих алгоритмів. Наведено результати експериментального моделювання.

Ключові слова: нелінійність булевих перетворень, лінійний криптоаналіз, тестування криптографічних алгоритмів, вимірювання криптостійкості.

UDC 004.8

Igor Boyarshin,
Oleksandr Markovskyi

METHOD OF HASH TRANSFORMATIONS CONSTRUCTION FOR STRICT USER IDENTIFICATION

Ігор Бояршин,
Олександр Марковський

МЕТОД ПОБУДОВИ ХЕШ-ПЕРЕТВОРЕНЬ ДЛЯ СТРОГОЇ ІДЕНТИФІКАЦІЇ КОРИСТУВАЧІВ У СИСТЕМІ

The paper describes a new method of hash transformations construction for strict user identification. The key feature of this method is the utilization of a different algebraic basis, namely 41oolean functional transformers, which allows for a number of advantages compared to traditional approaches. The described method can generate for a given output key a range of unique input keys that satisfy the following rule: the given hash, once applied to such input key, yields the given output key.

Key words: cryptography, hash transformations, 41oolean functional transformers, strict identification.

Fig.: 1. Tabl.: 1. Bibl.: 5.

У статті описується новий метод побудови хеш-перетворень для строгої ідентифікації користувачів у системі. Особливістю методу є використання для його реалізації іншого алгебраїчного базису, а саме булевих функціональних перетворювачів, що дає ряд переваг у порівнянні з традиційними підходами. Метод дозволяє для заданого вихідного ключа згенерувати безліч унікальних вхідних ключів, що якщо до них застосувати задане хеш-перетворення, то буде отримано вихідний ключ.

Ключові слова: криптографія, хеш-перетворення, булеві функціональні перетворювачі, строга ідентифікація.

Рис.: 1. Табл.: 1. Бібл.: 5.

Target setting. With the rapid development of information technologies the demand for remote processing power and various services increased dramatically. As the majority of such systems is commercial, this requires for an efficient strict user identification algorithm to be developed. Such an algorithm must be both easy to use and provide sufficient security against attacks. The key question is thus to find a balance between the reliability of this algorithm and its ease of use.

Actual scientific researches and issues analysis. The analysis of known attacks on the identification systems showed that the most effective approach to withstand such attacks is to periodically re-identify the user in the system [1]. This implies that the method to be used for identification must have sufficient capabilities to provide enough sessions keys for the user to use when logging into the system [2].

Not investigated parts of the general subject. Although the matter of user identification in the system is not new, there have been few works addressing the usage of 42oolean functional transformations in it. The key advantage of such transformers is that they require much less computational power compared to other methods, and even more so with hardware implementation [3,4].

It is a known fact that 42oolean functional transformators are able to perform the computations in one third of the usual time. As a result, the choice of 42oolean functional transformations for the described task is obvious.

Research objective. The objective of this paper is to prove that the new method of hash transformations construction for strict user identification is viable and measure its performance, i.e. the amount of unique input keys x that the system can generate within the provided architecture for a given output key y [5].

As there are multiple ways to split the input vector into fragments, the resulting amount of generated vectors depends on it to some degree. That is why it is important to measure the performance for different splits of the key and find those that yield the most input keys.

Principal statements. The basic idea of the algorithm relies heavily on the underlying structure of the system, which is shown on Fig. 1 (for split of the key into 3 fragments). As can be seen from the figure, the system is comprised of multiple 42oolean functional transformers (each column corresponds to a single 42oolean functional transformer). An input key flows from top to bottom, layer by layer, resulting in an output key. The outputs of functional transformers are interconnected with each other by the means of XOR operator.

The essence of the algorithm is as follows: for a given output key y, keep filling the functional transformers with randomly-generated numbers from the bottom layer to the top layer, while meeting the rules of the underlying structure. Upon reaching the top layer a new input key is yielded. That concludes a single iteration I out of t total. Repeat the process until the 42oolean functional transformers become saturated. End by filling the remaining free slots of 42oolean functional transformers with randomly generated numbers.

If the algorithm is run for t iterations of the process, it results in t unique input keys x that all, once run through the system, yield the desired output key y. The proposed model structure provides sufficient non-linearity while keeping the overall performance of 42oolean functional transformators.

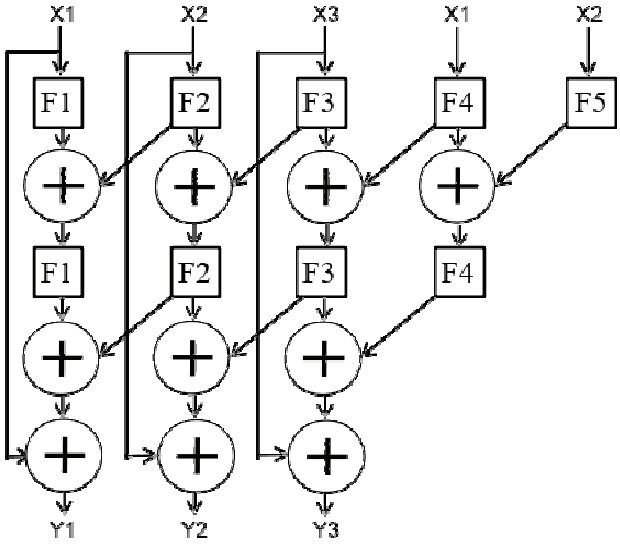


Fig. 1. The model structure for the split of key into 3 fragments

Testing. The purpose of testing is to find an approximate number of unique input keys *x* that the algorithm is able to generate for a given output key *y*. The testing was performed for different fragment size *k* and fragments amount *m*, so that *m*k=n*, where *n* is the size (bitness) of input and output keys. Table 1 gives the summary of the testing. As can be seen from the table, increasing the fragment size *k* by 1 results in doubled amount of generated keys.

Table 1

Amount of generated input keys for a given output key

Amount of fragments m	***Bitness of fragment k***			
	10	***11***	***12***	***13***
8	265	547	1102	2181
9	236	493	994	1988
10	221	452	909	1837
11	207	406	848	1702
12	193	379	789	1569

Conclusion. The study has proved that the new method based on 43oolean transformations is viable and can be used in user identification scenarios. It has been further shown that this method can generate sufficient number of unique input keys for a given output key. As the splitting of the key to be used inside the system can vary, a comprehensive testing was conducted to reveal dependencies between the amount of fragments and their bitness in regards to the amount of generated keys. The resulting performance of the method is a couple of magnitudes faster than that of other transformations based on a different underlying architecture, which confirms that the proposed solution can outperform existing solutions.

Future developments of the method could include experimenting with different system structures, as there are multiple ways functional transformers could be interconnected. Other variations of this structure could, for example, prove to be even more resilient to attacks or render even greater performance boost.

References

1. Широчин В. П., Мухин В. Е., Кулик А. В. Вопросы проектирования средств защиты информации в компьютерных системах и сетях. К.: 2000.- 111 с.

2. Захариудакис Лефтерис. Метод быстрой аутентификации удаленных пользователей на основе концепции “нулевых знаний” /Наукові записки Українського науково-дослідного інституту зв’язку. 2017.- № 1 (45).– С.109-117.

3. Захарченко Н. А., Топалова К. Н. Использование булевых преобразований для быстрой идентификации абонентов на основе концепции нулевых знаний. // Матеріали XII Міжнародної науково-технічної конференції ”Системний аналіз та інформаційні технології”.- К.:НТУУ ”КПІ”.-2010.- С.441.

4. Марковский А. П., Зюзя А. А., Шерстюк В. Д. Получение булевых преобразований специальных классов для построения эффективных алгоритмов защиты информации // Вісник Національного технічного університету України ”КПІ”. Інформатика, управління та обчислювальна техніка. К.,»ВЕК++»,- 2008.- № 49.- С.7-13.

5. Bardis N., Doukas N. Markovskyi O. A Method for strict remote user authentication using non-revesible Galois field transformations // Proceeding of IEEE Symposium on Computers and Communications. ISCC-2017. 3-6 July 2017. Heraclion, Crete, Greece. P.243-249.

Autors

Boyarshin Igor – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-Mail: igor.boyarshin@gmail.com

Бояршин Ігор Іванович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Markovskyi Oleksandr – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: markovskyy@i.ua

Марковський Олександр Петрович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

І. І. Бояршин,
О. П. Марковський

МЕТОД ПОБУДОВИ ХЕШ-ПЕРЕТВОРЕНЬ ДЛЯ СТРОГОЇ ІДЕНТИФІКАЦІЇ КОРИСТУВАЧІВ У СИСТЕМІ

Актуальність теми дослідження. З ростом популярності хмарних обчислень, що надають який-небудь сервіс або просто обчислювальні потужності, на передній план виходить проблема ідентифікації користувача у системі. Особливе місце серед можливих рішень посідає використання булевих нелінійних функціональних перетворень, що мають, по-перше, необхідну властивість незворотності, і по-друге, виконуються у декілька разів швидше за відомі аналоги. Ця робота присвячена опису нового методу побудови хеш-перетворень та його тестуванню.

Постановка проблеми. Завданням є надання такого методу ідентифікації користувача у системі, що був би водночас достатньо швидким, щоб система могла обробляти безліч користувачів за коротких проміжок часу, та з іншої сторони надавала можливість повторної ідентифікації користувача в цій системі для протистояння атакам.

Аналіз останніх досліджень і публікацій. Хоча проблема ідентифікації користувача в системі не є новою, використання в якості базису функціональних булевих перетворень є відносно новим. Основною перевагою такого базису у методах ідентифікації є швидкість їх роботи у порівнянні з іншими базисами, а також можливість зручної апаратної реалізації. Аналіз відомих атак на системи ідентифікації виявив, що найкращим способом протидії їм є повторна ідентифікація користувача в системі через деяких час.

Виділення недосліджених частин загальної проблеми. В цій статі описується новий метод побудови хеш-перетворень для строгої ідентифікації користувача в системі, що базується на булевих функціональних перетвореннях, а також тестування його роботи.

Постановка завдання. Для заданого вихідного ключа у необхідно згенерувати якомога більше унікальних вхідних ключів x, що якщо їх пропустити через систему (нелінійне булеве перетворення), то буде отримано заданий вихідний ключ y.

Викладення основного матеріалу. Побудована та проаналізована система, створена за описаним методом. Результати тестування показали, що побудована за цим методом система генерує достатньо велику кількість унікальних вхідних ключів користувача, а також виконується у декілька разів швидше за відомі аналоги на іншому алгебраїчному базисі.

Висновки. Новий метод строгої ідентифікації користувача в системі добре себе показує та дає задовільні результати з точки зору кількості згенерованих вхідних ключів та швидкості роботи. Таким чином, доведена можливість та доцільність його використання в системах ідентифікації користувачів.

Ключові слова: криптографія, хеш-перетворення, булеві функціональні перетворювачі, строга ідентифікація.

UDC 004.056

Dmytro Pylypiuk,
Oleksii Aleschenko

AUTHENTICATION METHODS IN WEB APPLICATIONS

Дмитро Пилип'юк,
Олексій Алещенко

СПОСОБИ АВТЕНТИФІКАЦІЇ КОРИСТУВАЧА У WEB-ДОДАТКАХ

The article is considering different methods of user authentication in web applications.

Key words: authentication, web application.
Fig.:5. Tabl. 0. Bibl. 4

У статті розглядається, як web-додатки проводять автентифікацію користувачів, використовуючи різні методи.

Ключові слова: автентифікація, web-додаток.
Рис.:5. Табл. 0. Бібл.:4.

The relevance of the research topic.

The topic of user authentication is widespread in the sphere of web application development and has become increasingly relevant in recent years, since the volume of private data transmitted over the Internet increases over time.

Target setting. Nowadays we use a certain amount of different applications and web resources, that make life easier. However, in order to receive a range of services, we provide private information about ourselves: phone number, address, email, bank card numbers, etc. Each user would prefer that such information be kept confidential and nobody other than the user could access it. This problem has begun development of various mechanisms of authorization and authentication of users. Let's consider how modern services conduct authentication of users and leave access to private data only to the user himself.

Actual scientific researches and issues analysis.

The source [1] addresses the issues of user authentication and describes the authentication protocols that are common at this time.

In article [2], a method is proposed for authenticating users of computer systems, resistant to spying attack, based on a graphic password and a gesture (move) selected by the user, similar to the movements of chess pieces.

The opportunity of two-factor authentication usage in the control systems and access management on the basis of Quick Response codes with one-time passwords is analyzed in the work [3].

Uninvestigated parts of general matters defining.

There is no system that uses an authentication method that is completely safe. Each authorization system can be attacked and allow hackers to steal data.

The research objective.

The purpose of this article is to analyze the most popular methods of user authentication.

The statement of basic materials.

1. Password authentication.

This method consists in the fact that the user must provide the system with a pair of login / password that was specified during registration for successful identification / authentication. This pair is specified when creating a user account on the system. There are standard password authentication protocols that can be applied in web applications.

1.a. HTTP authentication.

This protocol is described in HTTP 1.0 / 1.1 and is applicable in the corporate sphere. The principle of work is as follows:

1. When accessing an unauthorized user, the server returns the "401 Unauthorized" HTTP status and adds the "WWW-Authenticate" header with the specified parameters and the authentication scheme.

2. When receiving such an answer from the server, the browser automatically displays the form of entering the necessary parameters that the user can enter in order to access the resource.

3. In all subsequent user requests for this web resource, the browser automatically adds the HTTP header "Authorization", which transfers the data specified by the client when authorizing.

4. The server authenticates the user according to the data from this header.

Http authentication has several different schemas that differ in security:

1. Basic - the simplest, the parameters are transmitted in the header in unencrypted form. Relatively safe when using HTTPS.

2. Digest - is a schema where a server sends a unique "nonce" value, and the browser sends the MD5 hash to a user's password that was calculated using the "nonce" value. A more secure schema than Basic, but may be striked by the "man-in-the middle" attacks. Also, this scheme is not designed to use modern hash functions to store passwords on the server.

3. NTLM or Windows authentication - as well as Digest, based on the challenge-response principle, in which the user password is transmitted in encrypted form. Not an HTTP standard, but is supported by most browsers and servers. It is mainly used to authenticate Windows Active Directory users in web applications. Sensitive to "pass-the-hash" attacks.

It's worth noting that when using HTTP authentication, the user does not have the standard ability to exit the web application, except to close all browser windows.

2.Certificate authentication.

The certificate is a set of attributes that identifies the user and is signed by the certificate authority (CA). CA acts as an intermediary, which guarantees the authenticity of certificates. Also, the certificate is cryptographically associated with a private key, which is stored by the certificate owner and confirms the fact of possession of the certificate.

On the client side, the certificate may be stored along with the private key in the operating system, in the browser, in the file on the physical device. The private key is additionally protected by the password.

Web applications traditionally use certificates X.509. Authentication with such certificates occurs at the time of connection to the server and is part of the SSL / TLS protocol.

During authentication, the server performs validation based on the following rules:

1. Certificate signed by CA.
2. The certificate has not expired.
3. The certificate shall not be withdrawn by the relevant CA.

After a successful authentication, the web application can execute an authorized request based on the certificate parameters.

Using certificates is a much more reliable way than password authentication. In the process of authentication, a digital signature is created, the presence of which proves the fact of using the private key. However, problems with the distribution and support of certificates make this method of authentication unpredictable in the sphere of information technology.

3.Token authentication.

Tokens are created by the server, signed by a secret key and passed to the client, who in the future uses a token for authentication. There are several standards for web tokens. Consider the most common - JWT - JSON Web Token - a standard for creating access tokens based on the JSON format.

The JWT Token consists of three parts:

1. Header - specifies the information needed to describe the token itself (encryption algorithm, token type, type of content).
2. Payload - is a set of fields where user information (name, level of access, role) is specified.
3. Signature - is generated using encryption algorithms and is calculated based on the first two token blocks.

Tokens are divided into 2 types and perform important roles in the authentication of users of the client-server application:

– Access token is a token that gives its owner access to secure server resources. Usually, it has a short life time and contains additional information.

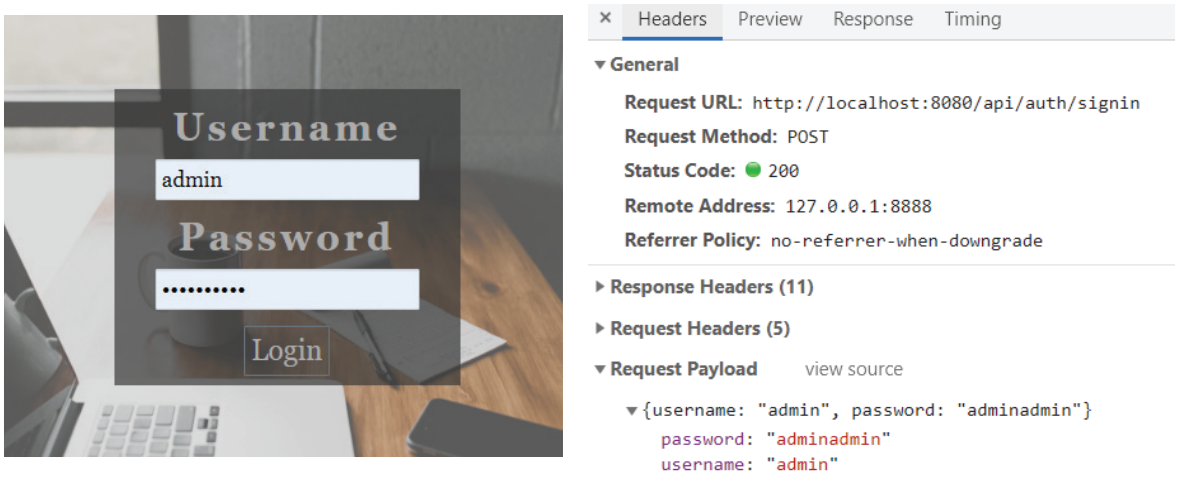
– Refresh token - this token allows clients to request new access tokens after their lifetime ends. These tokens are issued for a long time.

Using tokens in client-server applications:

1. The client is initially authenticated.
2. If the authentication was successful, then the server sends access and refresh tokens to the client.
3. During subsequent queries to the server, the client uses the access token. The server validates validity and provides access to resources.
4. If the access token is not valid, then the client sends a refresh token, in response the server will update both tokens.
5. If the refresh token is not valid then the client must pass the initial authentication process again.

Experiments. It was created its own authorization and authentication system based on access tokens. The tokens, among other things, contain information about the role of the user, which affects the reaction of the system when interacting with it. Tokens are also used as a password encryption mechanism. They contain a hash for the user password, which provides an alternative way to save it in a secure form.

Login frame of the created system and authentication query details are shown in the figure 1. There are user name and password in request payload. There is token and other technical information in the response (see fig. 2). After login user can see landing page in the figure 3.



A B

Fig.1. A – login frame, B – headers of authentication query

Conclusions. This article describes methods of user authenticate in web applications that can be considered fundamental. Among all of these mechanisms, attention was focused on token authentication. This method is more reliable than all of the above. Access Tokens is a much safer mechanism than HTTP authentication, since JWT is almost non-attackable and allows you to store encrypted data in a database. Also, tokens are more practical than certificates, because they exist only during the application works and they need not be maintained as certificates. Tokens are also the only authentication mechanism that allows you to build an SSO (Single Sign-On)

system, where one application allows you to switch to another without re-authentication (like Gmail and YouTube).

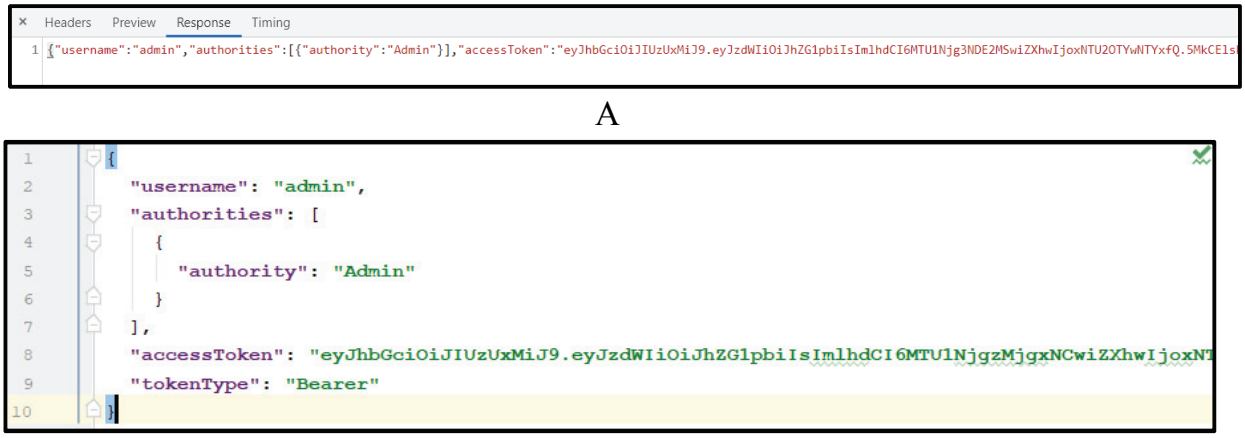


Fig.2. A – authentication response in browser tool view,
B – formatted authentication response

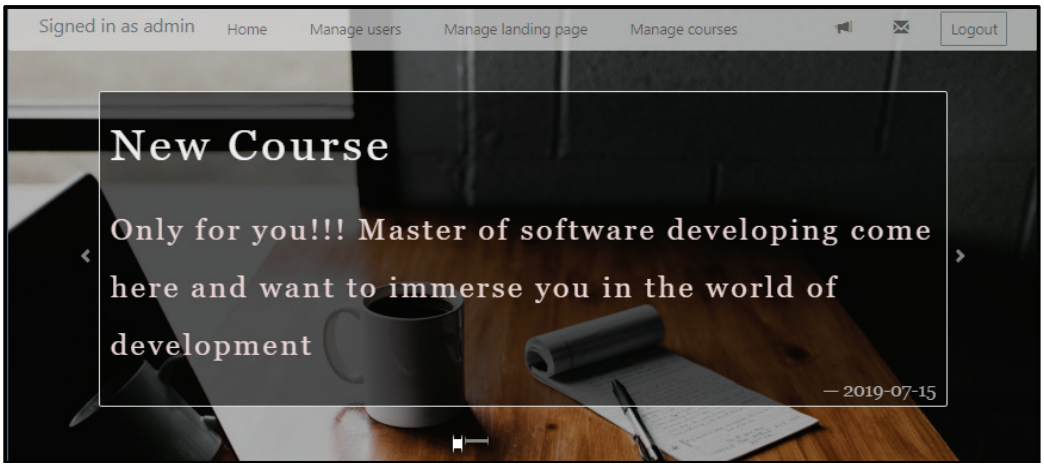


Fig.3. Landing page of the system after login

References

1. Молдовян А. А., Молдовян Д. Н., Левина А. Б. (2016). *Протоколы аутентификации с нулевым разглашением секрета.* (pp. 3-29).

2. Яковлев В. А., Архипов В. В., (2014). *Аутентификация пользователей на основе устойчивого к подсматриваниям графического пароля «Шахматы».* (pp. 25-35).

3. A. Y. Iskhakov (2013). *Two-Factor Authentication System based on QR-Codes.* (pp. 97 – 101).

4. *«Обзор способов и протоколов аутентификации в веб-приложениях»* [Electronic source] (2015) – Access mode: https://habr.com/ru/company/ /dataart/blog/262817/.

Authors

Pylypiuk Dmytro – bachelor student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.
E-mail: pylypyuk.dmytro@gmail.com

Пилип’юк Дмитро Олександрович – студент ІІІ курсу, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Aleshchenko Oleksii – senior lecturer, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".
E-mail: alexey.aleshchenko@gmai.com
Алещенко Олексій Вадимович – старший викладач, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Д. О. Пилип’юк,
О. В. Алещенко

СПОСОБИ АВТЕНТИФІКАЦІЇ КОРИСТУВАЧА У WEB-ДОДАТКАХ

Актуальність теми дослідження. Тема автентифікації користувачів є поширеною в сфері розробки web-додатків і в останні роки стає все більш актуальною, оскільки об`єм приватних даних, що передаються через Інтернет, збільшуються з часом.

Постановка проблеми. В сучасному світі ми користуємось певною кількістю застосунків та онлайн ресурсів, які спрощують наше повсякденне життя. Однак для того, щоб отримувати певний спектр послуг, ми надаємо приватну інформацію про себе: номер телефону, адресу, електронну пошту, номери банківських карт тощо. Кожен користувач волів би, щоб така інформація залишалась конфіденційною і ніхто крім самого користувача не міг мати доступ до неї. Дана проблема дала початок розвитку різних механізмів авторизації та автентифікації користувачів. Розглянемо, як сучасні сервіси проводять автентифікацію користувачів і залишають доступ до приватних даних лише самому користувачу.

Аналіз останніх досліджень і публікацій. Протягом останніх років з'являється все більше статей, що зосереджують увагу на нових протоколах і механізмах автентифікації користувача. Окрім автентифікації в Інтернеті, стає популярною тема біометричної автентифікації.

Виділення недосліджених частин загальної проблеми. Немає ні одної системи автентифікації, що була би повністю безпечною. Кожна система може бути вражена атаками і це дозволить хакерам викрасти дані.

Постановка завдання. Метою даної статті є аналіз найбільш популярних методів автентифікації користувачів.

Викладення основного матеріалу. Проведено аналіз трьох методів автентифікації користувачів. Розглянуто принцип їх роботи, переваги використання та недоліки, слабкі місця. Наведений приклад застосування авторизації за допомогою токенів на готовому програмному продукті.

Висновки. В даній статті було розглянуто способи автентифікації користувачів у web-додатках, які можна вважати фундаментальними. Серед усіх зазначених механізмів, увагу було зосереджено на автентифікації через токени. Даний спосіб є надійнішим, ніж усі вище зазначені. Токени доступу є набагато безпечнішим механізмом, ніж HTTP-автентифікація, оскільки JWT майже не підлягає атакам і дозволяє зберігати зашифровані дані у базі даних. Також токени більш практичні з сертифікати, оскільки вони існують лише під час роботи додатку і їх не потрібно підтримувати, як сертифікати. Також токени це єдиний механізм автентифікації, який дозволяє побудувати SSO (Single Sign-On) систему, де один додаток дозволяє перейти в інший без повторної автентифікації (наприклад Gmail і YouTube).

Ключові слова: автентифікація, web-додаток.

UDC 004.056

Roman Bozhok,
Oleksii Aleshchenko

WEB APPLICATION SECURITY

Роман Божок,
Олексій Алещенко

БЕЗПЕКА WEB-ЗАСТОСУНКУ

The article discusses the security issue of a web application. As a research, the site is used. Testing is carried out at the expense of external independent resources.

Key words: OWASP, website, threat, XSS, security.
Fig.: 4. Tabl. 0. Bibl.: 11.

У статті розглядається питання безпеки web-застосунку. В якості дослідження використовується сайт. Тестування виконується за рахунок зовнішніх незалежних ресурсів.

Ключові слова: OWASP, сайт, загроза, XSS, безпека.
Рис.: 4. Табл. 0. Бібл.: 11.

Target setting. The relevance of the security problems of WEB-applications is that they use confidential information, as well as the company's business processes.

Issues analysis. The vast majority of external attacks on corporate information systems are aimed precisely at the vulnerability of web applications.

Actual scientific researches. In recent years, the topic of application security has filled a lot with hype, there are many articles, discussions, and a search for solutions. To solve these problems, an international project on the security of WEB-applications (OWASP) [1] was formed. Problems and their solutions using OWASP are described in more detail in article [2]. Also, in article [3] a certified method is described that will help developers to minimize the occurrence of holes in the program.

Uninvestigated parts of general matters defining. There is no unified protection against all threats and security tools are developing more slowly than methods of application attacks.

The research objective. Investigate the types of threats and check for these threats' web application.

The statement of basic materials. A WEB application is a client-server application, where the client is a browser that displays the user interface, generates

requests to the server, and processes responses from it. And the server part is a WEB-server that processes customer requests. The interaction between the client and the server, as a rule, is carried out via the HTTP protocol [4]. The architecture of WEB applications has three levels [5], which are shown in Figure 1.

Fig. 1. Web application architecture [6]

In connection with the rapid growth of the popularity of information technologies, recommendations have emerged among developers to ensure the security of WEB applications, which resulted in a project called: The Open Web Application Security Project (OWASP).

OWASP is an open source WEB application security project that includes corporations, educational organizations and individual developers who together form articles, recommendations and tutorials that are freely available and recommended when developing WEB applications.

OWASP Top 10 Application Security Risks – 2017 [7]

- A1:2017-Injection
- A2:2017-Broken Authentication
- A3:2017-Sensitive Data Exposure
- A4:2017-XML External Entities (XXE)
- A5:2017-Broken Access Control
- A6:2017-Security Misconfiguration
- A7:2017-Cross-Site Scripting (XSS)
- A8:2017-Insecure Deserialization.
- A9:2017-Using Components with Known Vulnerabilities.
- A10:2017-Insufficient Logging and Monitoring.

Experiments. We are exploring the site [8] of the "Department of Computing Engineering", National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" for safety.

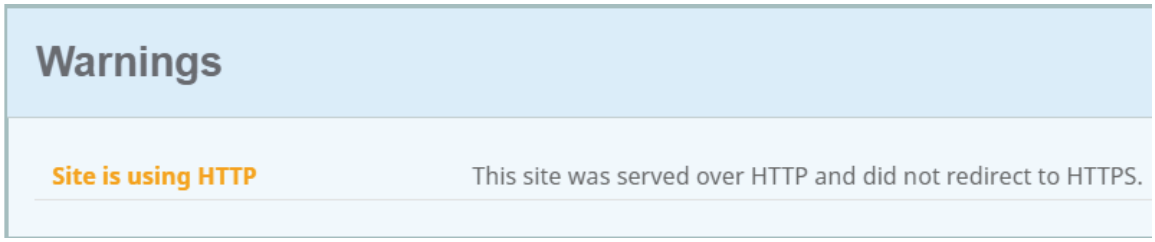


Fig.2. Test from the SecurityHeaders.io service [9]

Security misconfiguration is the most commonly seen issue. This is commonly a result of insecure default configurations, incomplete or ad hoc configurations, open cloud

storage, misconfigured HTTP headers, and verbose error messages containing sensitive information. Not only must all operating systems, frameworks, libraries, and applications be securely configured, but they must be patched/upgraded in a timely fashion.

Redirection	✗	-20	Does not redirect to an HTTPS site
Referrer Policy	–	0	Referrer-Policy header not implemented (optional)
Subresource Integrity	–	0	Subresource Integrity (SRI) not implemented, but all scripts are loaded from a similar origin
X-Content-Type-Options	✗	-5	X-Content-Type-Options header not implemented
X-Frame-Options	✗	-20	X-Frame-Options (XFO) header not implemented
X-XSS-Protection	✗	-10	X-XSS-Protection header not implemented

Fig.3. Test from the Observatory by Mozilla service [10]

Redirections. Sites that listen on port 80 should only redirect to the same resource on HTTPS. Once the redirection has occurred, HSTS should ensure that all future attempts go to the site via HTTP are instead sent directly to the secure site.

Referrer Policy. When a user navigates to a site via a hyperlink or a website loads an external resource, browsers inform the destination site of the origin of the requests through the use of the HTTP Referrer (sic) header.

X-Content-Type-Options is a header supported by Internet Explorer, Chrome and Firefox 50+ that tells it not to load scripts and stylesheets unless the server indicates the correct MIME type. Without this header, these browsers can incorrectly detect files as scripts and stylesheets, leading to XSS attacks

X-Frame-Options is an HTTP header that allows sites control over how your site may be framed within an iframe. Clickjacking is a practical attack that allows malicious sites to trick users into clicking links on your site even though they may appear to not be on your site at all.

X-XSS-Protection is a feature of Internet Explorer and Chrome that stops pages from loading when they detect reflected cross-site scripting (XSS) attacks.

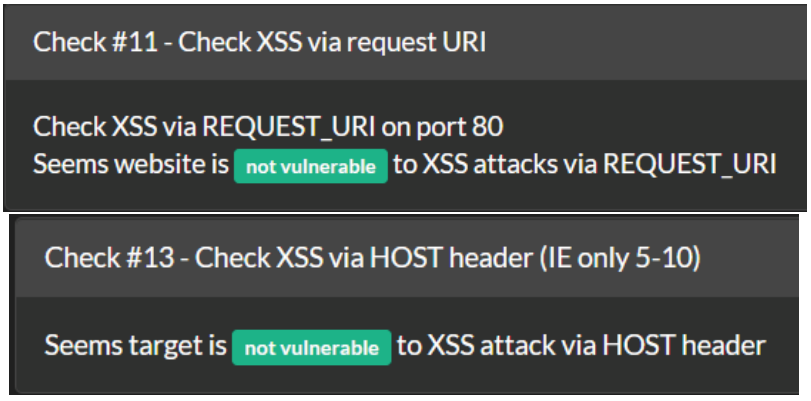


Fig.4. Test from the service One button scan [11]

XSS flaws occur whenever an application includes untrusted data in a new web page without proper validation or escaping, or updates an existing web page with user-supplied data using a browser API that can create HTML or JavaScript. XSS allows attackers to execute scripts in the victim's browser which can hijack user sessions, deface web sites, or redirect the user to malicious sites.

Conclusions of experiments. This web application is not properly secure. According to the test results, the average safety rating is “F”.

Conclusions. The article reviewed a list of popular threats. The site of the department was also checked for safety. From the research it is clear that the site is not reliable and needs to be improved. OWASP can help find and fix flaws.

References

1. OWASP. The Open Web Application Security Project. [Електронний ресурс]. — Режим доступу: https://www.owasp.org/index.php/Main_Page

2. БЕЗОПАСНОСТЬ ВЕБ-ПРИЛОЖЕНИЙ. [Електронний ресурс] / Королев О.Л., Лукьянова М.А.// Международная научно-практическая конференция "Проблемы информационной безопасности". — 2016. — №6. — С.166-167. — Режим доступу до журналу: http://ieu.cfuv.ru/sites/default/files/2018-12/sbornik-trudov-2-megd-konf-problemy-inform-bezopasn-2016.pdf#page=166

3. Разработка типовой методики анализа уязвимости в веб-приложениях при проведении сертификационных испытаний по требованиям безопасности информации. [Електронний ресурс] /Баранов А. В., Федичев А.В. // Вопросы кибербезопасности. — 2016. — №2(15). — С.2-8 — Режим доступу до журналу : https://cyberleninka.ru/article/v/razrabotka-tipovoy-metodiki-analiza-uyazvimostey-v-veb-prilozheniyah-pri-provedenii-sertifikatsionnyh-ispytaniy-po-trebovaniyam

4. Таненбаум Э. Компьютерные сети. Пятое издание / Компьютерные сети. 5-е изд. — СПб.: Питер. — 2012. — С. 724-726

5. Пьюривал С. Основы разработки веб-приложений / С. Пьюривал — СПб.: Питер. — 2015. —272с.

6. ОБЗОР УГРОЗ БЕЗОПАСНОСТИ WEB-ПРИЛОЖЕНИЙ. [Електронний ресурс] / Елисеев Н.А., Федоров С.А., Антонов О.Д. // Вопросы технических наук в свете современных исследований: сб. ст. по матер. V-VI междунар. науч.-практ. конф. — 2018. — № 1(4). — С. 18-23. Режим доступу до журналу: https://sibac.info/conf/technolog y/v/95451

7. OWASP Top 10 - 2017 The Ten Most Critical Web Application Security Risks. [Електронний ресурс]. — Режим доступу: https://www.owasp.org/images/ /7/72/OWASP_Top_10-2017_%28en%29.pdf.pdf

8. КАФЕДРА ОБЧИСЛЮВАНОЇ ТЕХНІКИ [Електронний ресурс]. — Режим доступу: http://comsys.kpi.ua/

9. Security Headers [Електронний ресурс]. — Режим доступу: https://securityheaders.com/?q=http%3A%2F%2Fcomsys.kpi.ua%2F&followRedirects=on

10. Mozilla Observatory [Електронний ресурс]. — Режим доступу: https://observatory.mozilla.org/analyze/comsys.kpi.ua

11. Scan #32710 for comsys.kpi.ua from Mon, 29 Apr 2019 21:18:06 +0300 [Електронний ресурс]. — Режим доступу: https://sergeybelove.ru/one-button-scan/result/378e03b41dbeb49d656496b34593c0049728804f/

Autors

Bozhok Roman – bachelor student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: romabos98@gmail.com

Божок Роман Юрійович – студент III курсу, кафедра обчислюваної техніки, Національний технічний університет України ”Київський політехнічний інститут імені Ігоря Сікорського”.

Aleshchenko Oleksii – senior lecture, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: alexey.aleshchenko@gmail.com

Алещенко Олексій Вадимович – старший викладач, кафедра обчислюваної техніки, Національний технічний університет України ”Київський політехнічний інститут імені Ігоря Сікорського”.

РОЗШИРЕНА АНОТАЦІЯ

Роман Божок,
Олексій Алещенко

БЕЗПЕКА ВЕБ ЗАСТОСУНКУ

Актуальність теми дослідження. Актуальність проблем безпеки веб-застосунків представляється в тому що в них використовується конфіденційна інформація, а також здійснюються бізнес-процеси компанії. Дана робота присвячена захисту та тестуванню саме сайту кафедри, оскільки кожна учбова структура повинна мати добрий захист.

Постановка проблеми. Переважна більшість зовнішніх атак на корпоративні інформаційні системи націлені саме на уразливості веб застосунків.

Аналіз останніх досліджень і публікацій. Протягом останніх років з'являється все більше статей присвячених захисту веб застосунку, зокрема, завдяки появі нових методів був сформований міжнародний проект по забезпеченню безпеки веб застосунків (OWASP). Проте підходи до пошуку відкритих частин для в злому не можливо вивчити досконало так як немає єдиного захисту.

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена вивченню та аналізу запропонованих підходів для пошуку загроз, зокрема на прикладі застосунку одного учбового закладу. Немає єдиного захисту від усіх загроз і безпека не розвивається з великою швидкістю.

Постановка завдання. Завданням є дослідити типи загроз і перевірити ці загрози на веб застосунку.

Викладення основного матеріалу. Проведено аналіз загроз та тестування веб застосунку. Описано рейтинг популярних та актуальних способів в злому сайту. З тестування застосунку видно що він не надійний.

Висновок. Проаналізовано веб застосунок на порядок загроз за допомогою різних незалежних веб ресурсів. Підхід показав себе добре та показав вразливість сайту. Наведені результати експериментів та аналіз загроз.

Ключові слова: OWASP, сайт, загроза, XSS, безпека.

UDC 004.056

Kostiantyn Minkov,
Viktor Selivanov, Artem Volokyta

PROTECTION SYSTEM OF MICROSERVICE SYSTEMS

Костянтин Міньков,
Віктор Селіванов, Артем Волокита

СИСТЕМА ЗАХИСТУ МІКРОСЕРВІСНИХ СИСТЕМ

The article provides an overview of the security system for microservices. As methods used by this system, tools such as HTTPS, JWT, OAuth2, RM, TOTP were considered. The implementation is performed using the Java programming language.

Key words: microservice systems, security systems, distributed systems, authentication, role model.

Fig.: 1. Tabl. 0. Bibl.: 12.

У статті виконується огляд системи захисту для мікросервісних систем. У якості методів, якими оперує дана система, були розглянуті засоби, такі як HTTPS, JWT, OAuth2, RM, TOTP. Реалізація виконана за допомогою мови програмування Java.

Ключові слова: мікросервісні системи, системи захисту, розподілені системи, автентифікація, рольова модель.

Рис.: 1. Табл. 0. Бібл.: 12.

Relevance of research topic. Due to the rapid introduction of microservice systems among various e-commerce companies (Netflix, Amazon, Hailo), there is a need for the means for their effective protection.

Analysis of recent research and publications. Despite the large number of works and studies devoted to the systems of protection, the protection of the microsystems itself isn't quite detailed in the literature.

Identification of unexplored parts of the general problem. Integration of known methods of protection into a coherent system. Creating a system of protection focused on microservices.

Setting objectives. The purpose of this work is to systematize known methods of protection and their consolidation to create a unified system in the context of microservice systems.

The statement of basic materials. Given the rapid development and globalization of modern business, there has been a demand for the development and

support of large systems for automating business processes in large companies, thus making their digital transformation. Such systems are most often implemented in the form of monolithic systems [1]. Over time, such systems grow more and more and can become difficult to maintain, deploy and develop by a large team. The introduction of small changes often leeds to the need to re-plan the entire application. With the development of cloud technologies, the issue is about scaling such systems, since monolithic applications usually offer a large number of services, some of which are used more often than others. This leeds to constant financial overheads for the maintenance of such software products [2].

Microservices have become the solution to the above problems. The microservice pattern suggests dividing system A into a plurality of small services μS (μS_1, μS_2, μS_n), each offering a subset of services S (S_1, S_2, S_x) provided by program A. Each microservice is developed and tested by the team of developers μT_i. Each microservice is developed using the independent code bases and the μT_i team, which is also responsible for deploying, scaling, operating and upgrading the micro-service on IaaS / PaaS solutions in a cloud environment. This approach allows to simplify the scaling, deployment and making necessary changes to individual parts of the system [3].

However, it should be noted that the microservice technology is not without some disadvantages. Such an approach complicates the architecture of the system, there is a complex network model of interaction between its components, considering that the number of services can reach several hundreds (Figure 1). Delay in the network, fault tolerance, message transformation, network reliability, asynchronousity, versioning of different subsystems, changes in loads within a particular version of applications - common issues in such systems [4].

The security challenge caused by the complexity of the network is the ever-increasing difficulty in monitoring, auditing and analyzing the functioning of the entire program. Since microservices are often deployed in a cloud environment that the application owners do not control, it is difficult for them to imagine the overall view of the entire application. Thus, attackers can use this complexity to launch attacks on applications. Another security issue is related to the trust among distributed microservices. An individual microservice can be compromised and controlled by an attacker. For example, an attacker can take advantage of the vulnerability in the microservice facing the public user and increase his privileges on the virtual machine on which the microservice operates. As a result, individual microservices can become unreliable [5]. Also, there is a question of authentication of individual system services among themselves.

Since, as noted above, the system is distributed, and communication is most often performed using the REST architectural style, one of the first methods to help secure the system is to configure the HTTPS protocol accross of the system [7]. It should be noted that the use of keys generated in certification center is not necessary for all system services. This should only be done for client-facing applications. For

internal services there is enough to have self-signed certificate, which must be added to the keystore on the virtual machines of other microservices.

Fig. 1. Scheme of Heptio microsystem [6]

One of the standard ways to protect the system is to introduce the role model for users. This model simplifies the setup of checks in the system and builds a clear access model that fits well into the hierarchical model of the company's departments. Mathematically, such a model is described using formula 1 [8].

$$permissions(s) = \bigcup_{r \in roles(s)} \{p_i | (p_i, r) \in PA\} \quad (1)$$

OAuth2 is an open standard that allows third-party applications to have limited access to the HTTP service on behalf of the owner of the resource.

This standard solves the problem of the need to store a password in the client system, since compromising a third-party program will compromise the end-user password and all data protected by this password. Instead of using the resource owner credentials to access secure resources, the client receives an access token, a string that denotes a specific area, lifetime, and other attributes of access. Access Tokens are issued to third-party clients by the authorization server with the permission of the owner of the resource. The client uses a token to access secure resources hosted on the resource server [9].

Oauth2 introduces a clear limitation of the rights and duration of access to a secured resource by the client, since the token is temporary and contains only a set of specific permissions.

As an add-on to OAuth2, another one is used - JWT. In the standard

implementation of OAuth2, a random string is used as a token. It has some disadvantages, namely:

1. At each request to one of the protected microservices, an additional request to the SSO is required to confirm the validity of the token, which increases the load on the network and introduces an additional delay to the response time.

2. The token does not have any information about the user or the authentication system.

To resolve this issue, a RFC 7519 standard was introduced, which defines a JSON token containing a payload (user, custom role, SSO, time before disabling the token, etc.), and a digital signature generated with the private key of the authentication system based on the body of the request using RSA or ECDSA. Thus, any service having a public key can check the token without any additional queries and get information about the user directly from the token [10].

A known problem in developing security systems is the storage of user passwords in a relational database. Storing passwords in the usual way is equivalent to writing them on a digital paper. If an attacker gets access to the database and steals a password table, then he will be able to access each user's account.

The safest way to store a password is to hash it.

The problem with the SHA family of algorithms is that they were designed in such a way as to have a higher computing speed. Rapid calculations mean faster brute-force attacks. An example of adaptive functions that can compensate increasing computer power is the bcrypt algorithm. Also, the algorithm is resistant to attack using rainbow tables, because it uses a random tape (salt) that is added to the hash of the password while storing.

One of the standards that helps protect a user from losing a password is the two-factor authentication of the TOTP protocol.

One-time passwords are often better than stronger authentication forms, such as public key infrastructure (PKI) or biometric data, since this method does not require the installation of any client software on a user's computer [11].

The HOTP algorithm is based on the HMAC-SHA-1 algorithm and is adapted to increase the counter value size representing the message in the HMAC calculation. TOTP is a time version of this algorithm, where the value T, derived from the time and time step, replaces the counter C in the HOTP calculation (2) [12].

$$TOTP\ (K,T)\ =\ Truncate(HMAC-SHA-1\ (K,T)) \qquad (2)$$

As we combine all methods described above we will receive a security system which meets the goals defined for this work.

Conclusions. The result of the design and development is a system that includes the methods in the main part, and can be used to protect the micro-service systems.

References

1. Matt C. Digital Transformation Strategies / C. Matt, T. Hess, A. Benlian. // Business & Information Systems Engineering. – 2015. – №57. – С. 339–343.

2. Newman S. Building Microservices / Sam Newman. – Sebastopol, CA: O'Reilly Media, Inc., 2015. – 472 с.

3. Evaluating the monolithic and the microservice architecturepattern to deployweb applications in the cloud [Електронний ресурс] / V.Mario, O. Garcés, H. Castro,S. Gil.–2015.–Режим доступу до ресурсу: https://www.researchgate.net/publication/304317852_Evaluating_the_monolithic_and_the_microservice_architecture_pattern_to_deploy_web_applications_in_the_cloud.

4. Wootton B. Microservices-Not A Free Lunch! [Електронний ресурс]/Benjamin Wootton.–2014.–Режим доступу до ресурсу: http://highscalability.com/blog/2014/4/8/microservices-not-a-free-lunch.html.

5. Y. Sun, S. Nanda and T. Jaeger,"Security-as-a-Service for Microservices-Based Cloud Applications," 2015 IEEE 7th International Conference on CloudComputing Technology and Science (CloudCom), Vancouver, BC, 2015, pp. 50-57.

6. Mirko Novakovic. Introducing Dynamic Focus for Application Performance Management [Електронний ресурс] / Mirko Novakovic. – 2017. – Режим доступу до ресурсу: https://www.instana.com/blog/introducing-dynamic-focus-application-performance-management/.

7. E. Rescorla. The Transport Layer Security (TLS) Protocol Version 1.3 / E. Rescorla. – Mozilla: IETF, 2018. – 160 с.

8. В. Л. Цирлов. Основы информационной безопасности автоматизированных систем / В.Л. Цирлов., 2008. – 119 с.

9. D. Hardt, Ed. The OAuth 2.0 Authorization Framework / D. Hardt, Ed.., 2012. – 76 с.

10. Sebastián E. Peyrott. The JWT Handbook / Sebastián E. Peyrott., 2017. – 85 с.

11. TOTP: Time-Based One-Time Password Algorithm / D. M'Raihi, S. Machani, M. Pei та ін.]., 2011. – 16 с.

12. HOTP: An HMAC-Based One-Time Password Algorithm / D. M'Raihi, M. Bellare, F. Hoornaert та ін., 2005. - 37 с.

Autors

Minkov Kostiantyn – student of Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: k.minkov-2017@kpi.ua

Міньков Костянтин Павлович – студент кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Viktor Selivanov – associate professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

Селіванов Віктор Левович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Volokyta Artem – associate professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: artem.volokita@kpi.ua

Волокита Артем Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

К. П. Міньков,
В. Л.Селіванов, А. М. Волокита

СИСТЕМА ЗАХИСТУ МІКРОСЕРВІСНИХ СИСТЕМ

Актуальність теми дослідження. Завдяки швидкому впровадженню систем мікросервісу між різними компаніями електронної комерції (Netflix, Amazon, Hailo) виникла потреба в засобах їх ефективного захисту.

Аналіз останніх досліджень і публікацій. Незважаючи на велику кількість робіт і досліджень, присвячених системам захисту, захист самих мікросистем досить докладно описана в літературі.

Виявлення недосліджених частин загальної проблеми. Інтеграція відомих методів захисту в когерентну систему. Створення системи захисту, орієнтованої на мікросервіси.

Цілі дослідження. Метою даної роботи є систематизація відомих методів захисту та їх консолідації для створення єдиної системи в контексті мікросистемних систем.

Викладення основного матеріалу. Визначено причини появи та розвитку систем мікросервісу, їх переваги та базові уразливості. В якості методів безпеки розглядаються такі стандарти, як HTTPS, JWT, OAuth2, RM, TOTP. Результатом є система, реалізована в Java відповідно до стандартів програмування.

Висновки. Результатом розробки та розробки є система, що включає в себе основні методи і може бути використана для захисту систем мікросервісу.

UDC 004.056

Aksyonenko Ilya,
Pavlo Rehida

APPLICATIONS OF SEQUENCE-TO-SEQUENCE AUTOENCODER NETWORKS IN REQUEST ANOMALY DETECTION

Аксьоненко Ілля,
Павло Регіда

ЗАСТОСУВАННЯ SEQUENCE-TO-SEQUENCE AUTOENCODER НЕЙРОННИХ МЕРЕЖ ДЛЯ РОЗПІЗНАВАННЯ АНОМАЛЬНИХ ЗАПИТІВ

This paper provides an insight into utilizing machine learning techniques to improve web application firewall (WAF) performance. A brief overview of existing techniques is provided, and a solution is proposed to optimize security breach alerts and anomaly detection capabilities of WAF software. An existing seq2seq autoencoder architecture is applied to solve the problem of efficient attack detection in WAF software.

Key words: WAF, LSTM, seq2seq, autoencoder.
Fig.: 4. Tabl. 0. Bibl. 0.

У статті розглядається використання технологій машинного навчання для підвищення ефективності web application firewall (WAF). На основі існуючої архітектури та рішень пропонується новий метод розпізнавання аномалій та атак. Нейронна мережа з архітектурою seq2seq використовується для вирішення задачі ефективного розпізнавання атак у послідовностях символів.

Ключові слова: WAF, LSTM, seq2seq, autoencoder.
Рис.: 4. Табл. 0. Бібл. 0.

Relevance analysis. WAF usage is becoming a staple in securing web applications, being required by industry-leading data protection standards, such as PCI DSS. Most of the provided solutions, however, rely on blacklisting [1,2] malicious requests either via regular expressions or attack fingerprints.

Although effective to some extent, these protection methods can be bypassed, which has happened in the past, as in the example [3]. Moreover, they by design can only prevent against known attacks, as they do not have measures past basic heuristics to prevent attacks previously unknown to them. Thus, a more general approach based on generic anomaly detection was proposed [4], and our method is an extension of this idea.

WAF general architecture. According to OWASP, a web application firewall (WAF) is an application firewall for HTTP applications. It applies a set of rules to an HTTP conversation. [5] Generally, these rules cover common attacks such as cross-site scripting (XSS) and SQL injection. As a WAF can be considered a filtering reverse proxy, its simplified and generalized data flow diagram could be visualized as follows:

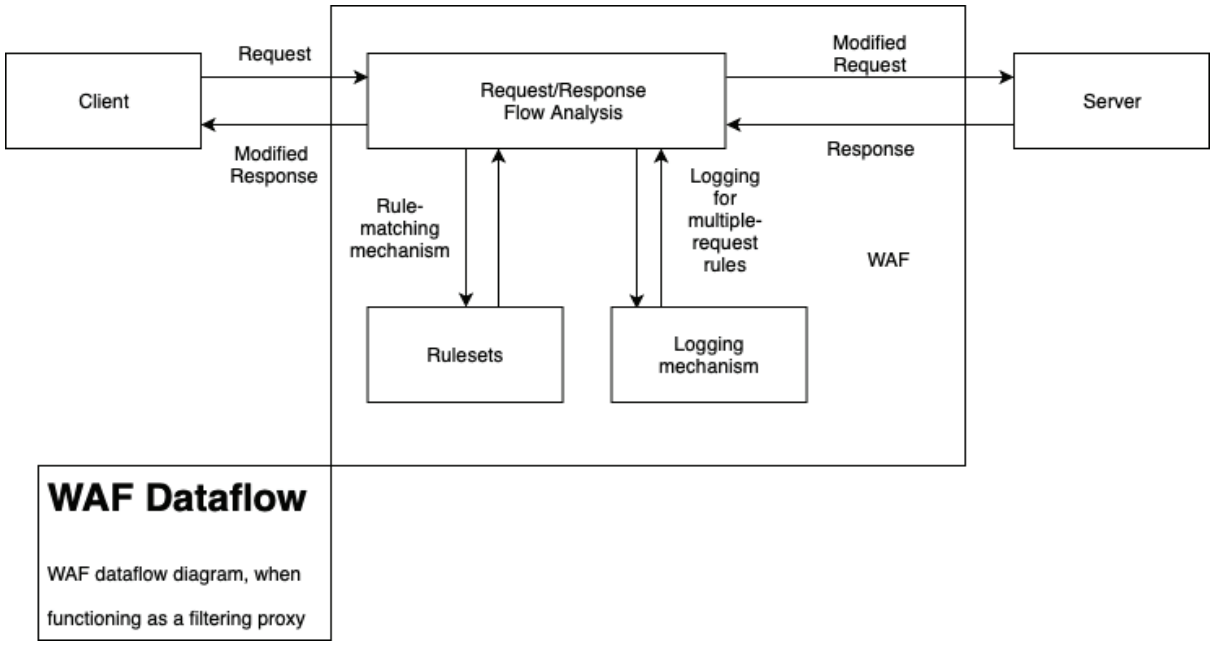


Fig.1. Generic WAF dataflow diagram

The proposed solution does not change this basic data flow, however, regular-expression based rulesets are replaced with a seq2seq autoencoder neural network that learns on previous trusted user requests. This also allows to implement our solution on top of existing software, combining machine learning-based anomaly detection with a vast range of existing attack fingerprints and rulesets.

Overview of existing solutions. This approach is mostly based on the work described above [3], as it pioneered the method. The idea is to implement a sequence to sequence autoencoder neural network with the following configuration:

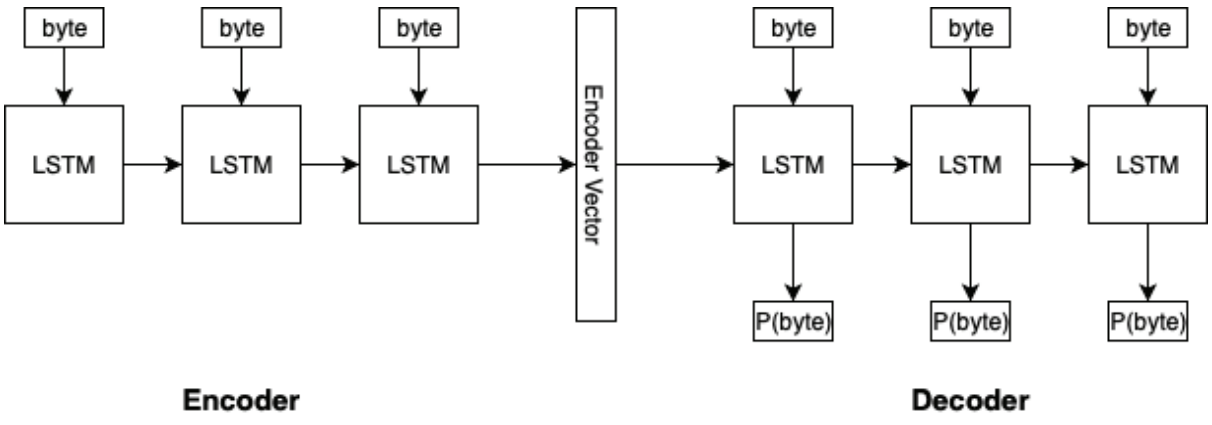


Fig. 2. Simplified LSTM network architecture

Figure 2 illustrating how the sequence is processed by the encoder and how the encoder internal state is used to initialize the decoder. The decoder output is used to determine estimated probabilities for sequence symbols, with anomalous parts of the sequence having significantly lower probabilities.

The network consists of the encoder and decoder LSTM networks [6], both trained with the exact same dataset of legitimate requests to the application. The internal state of the encoder, a vector of fixed length, is then used to initialize the decoder. The decoder is then used to determine the probability of the next symbol in the sequence, as it is trained to reconstruct known sequences from the input vector. Thus, if the request is anomalous, the probabilities of symbols in the anomalous (previously unseen) part of the request are significantly lower than average, allowing to highlight possible payloads for vulnerability exploitation. [4]

As described in the article above, the detection process can be visualized by the following example:

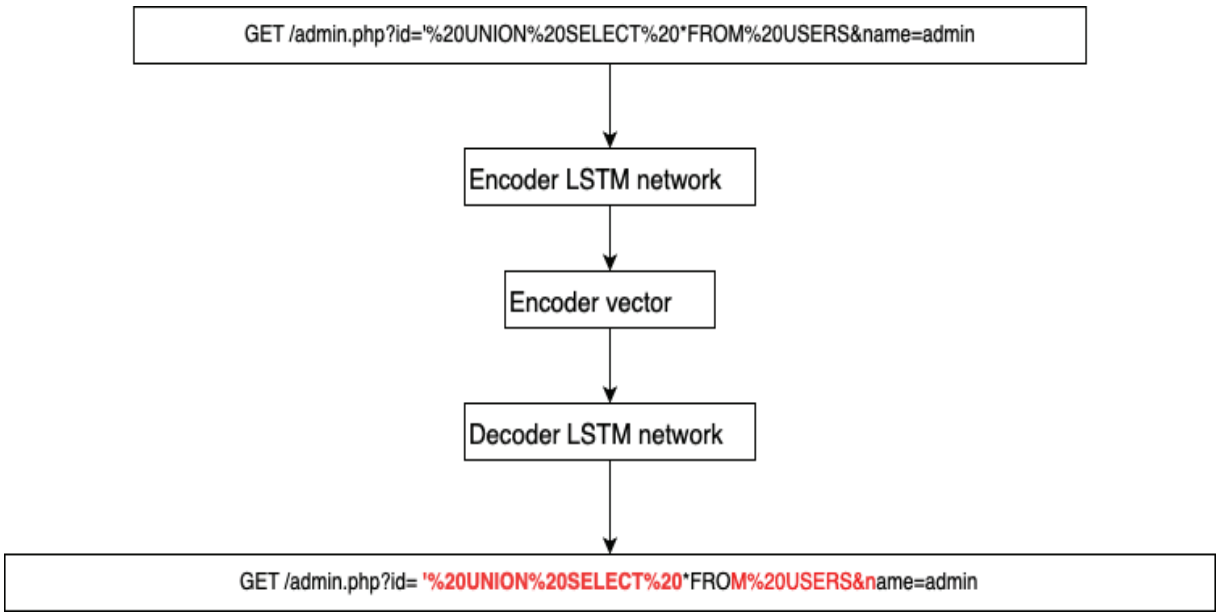


Fig.3. An implemented method to detect HTTP request anomalies.

The example details that the process does not give exact results, only estimating probability thresholds for an anomalous sequence.

Our approach. However, the approach described in the article can be optimized. The algorithm suggested by Alexandra Murzina, Irina Stepanyuk, Fedor Sakharov, and Arseny Reutov trains both networks on the whole request string, and thus, more time is required to allow the network to recognise HTTP request patterns.

Instead, we propose field-based anomaly detection, which adds an extra step to the detection flow. Instead of training one anomaly detection seq2seq autoencoder network, we propose to add a layer that separates HTTP form field values and supplies them both as training data and input for the several detectors, identical in structure. Field variable names can be also whitelisted, as the list of legitimate request fields is

known before the implementation of the WAF. With this approach, it is possible to achieve faster and more specific training, as each network will learn to only reproduce values of a single field.

With our solution, the example above is solved in another way:

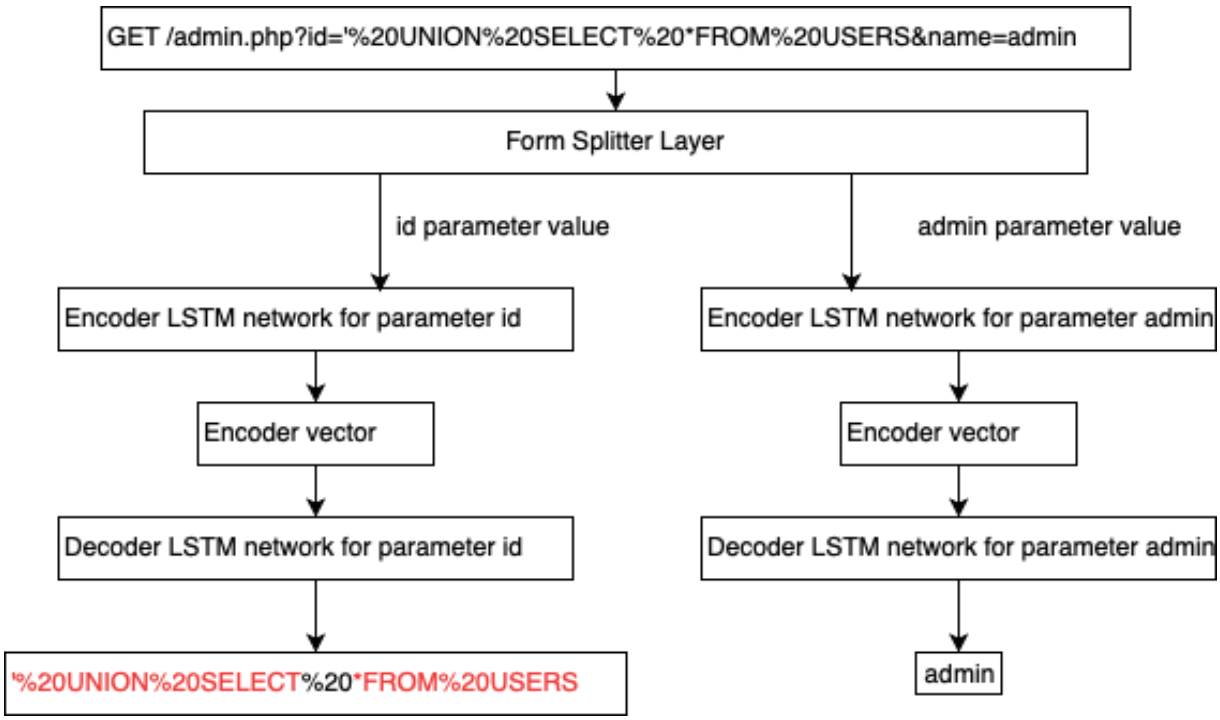


Fig. 3. An example of request processing with our approach

Although the method is still the same, the request-splitting approach allows to pinpoint the form field where the injection takes place.

In general, the algorithm proposed gives up multiple-parameter anomaly detection in favor of faster training and better flexibility (as when parameter submission form is changed, the trained networks can be reused to validate form values).

It could be best applied together with conventional attack-detection methods. As our method allows for localized detection, as opposed to highlighting anomalous parts of the request string, it is possible to use it as a trigger for a conventional WAF mechanism to use an extended set of heuristics to check an anomalous request.

Also, the approach could be used as a standalone solution, alerting the security personnel and automatically blocking users that send a high number of potentially anomalous requests in each timeframe.

Additionally, it can be used only as a monitoring solution, inspecting the mirrored traffic to/from an application and raising alerts based on detected anomalies.

Conclusions. In general, the idea proposed is a tradeoff between multi-value anomalies and faster, more specific detection. This theoretical approach can be the basis to implement practical machine learning in the field of web application firewall applications.

References

1. Nginx WAF. https://docs.nginx.com/nginx-waf. Accessed 12 May 2019
2. Mod_security documentation. http://modsecurity.org/rules.html. Accessed 12 May 2019
3. Web Application Firewall (WAF) Evasion Techniques #2. https://medium.com/secjuice/web-application-firewall-waf-evasion-techniques-2-125995f3e7b0. Accessed 12 May 2019
4. A. Murzina, I. Stepanyuk, F. Sakharov, A. Reutov. Detecting Web Attacks with a Seq2Seq Autoencoder. https://habr.com/en/company/pt/blog/441030/. Accessed 12 May 2019
5. Open Web Application Security Project documentation. https://www.owasp.org/index.php /Web_Application_Firewall. Accessed 12 May 2019
6. A Gentle Introduction to LSTM Autoencoders. https://machinelearningmastery.com/lstm-autoencoders/. Accessed 12 May 2019

Довідка про авторів

Аксьоненко Ілля Олегович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

E-mail: ilya.aksyonenko@gmail.com

Aksyonenko Ilya – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

Регіда Павло Генадійович – асистент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

E-mail: pavel.regida@gmail.com

Rehida Pavlo – Assistant Professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

РОЗШИРЕНА АНОТАЦІЯ

Аксьоненко Ілля,
Павло Регіда

ЗАСТОСУВАННЯ SEQUENCE-TO-SEQUENCE AUTOENCODER НЕЙРОННИХ МЕРЕЖ ДЛЯ РОЗПІЗНАВАННЯ АНОМАЛЬНИХ ЗАПИТІВ

Актуальність теми дослідження. Проблема розпізнавання веб-атак стає більш актуальною в останні дні у зв'язку зі зростаючою долею застосунків, що використовують веб-технології. Таким чином, проблемою є створення універсального способу розпізнавання атак, яка не буде базуватися на фільтрах. Дана робота присвячена проблемі розпізнавання веб-атак як аномалій за допомогою нейронних мереж.

Постановка проблеми. Неефективність та недоліки існуючих систем розпізнавання веб-атак,

Аналіз останніх досліджень і публікацій. Робота побудована на і є поширенням ідеї використаня seq2seq мереж для розпізнавання аномалій. Ідея була представлена у статті A. Murzina, I. Stepanyuk, F. Sakharov, A. Reutov: *Detecting Web Attacks with a Seq2Seq Autoencoder.*

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена розробці додаткового шару обробки даних до нейронної мережі для оптимізації певних характеристик системи.

Постановка завдання. Завданням є створити теоретичну архітектуру, що базується на існуючих моделях, але використовує підхід поділу вхідних даних до обробки них мережею.

Викладення основного матеріалу. Проведено аналіз seq2seq підходу до розпізнавання веб-атак, запропоновано новий метод розділення даних по параметрам запиту.

Висновки. За результатами аналізу було зроблено висновок, що запропонований підхід має переваги для певних сценаріїв застосування і може використовувватися разом з існуючими методами захисту.

Ключові слова: WAF, LSTM, seq2seq, autoencoder.

UDC 004.7

Oleksandr Honcharenko,
Artem Volokyta, Heorhii Loutskii

FAULT-TOLERANT TOPOLOGIES SYNTHESIS BASED ON EXCESS CODE USIGN THE LATIN SQUARE

Олександр Гончаренко,
Артем Волокита, Георгій Луцький

СИНТЕЗ ВІДМОВОСТІЙКИХ ТОПОЛОГІЙ НА ОСНОВІ НАДЛИШКОВОГО КОДУВАННЯ ЗА ДОПОМОГОЮ ЛАТИНСЬКОГО КВАДРАТУ

The article discusses the method of synthesis of fault-tolerant topologies using a Latin square and excess encoding of nodes numbers. Ways to fill the square were considered, several topologies were synthesized and their characteristics were analyzed. The usages of redundancy have been analyzed.

Key words: fault tolerance, excess code, Latin square
Fig.: 5. Tabl.: 5. Bibl.: 6.

У статті розглядається метод синтезу відмовостійких топологій за допомогою латинського квадрату з використанням надлишкового кодування номерів вершин. Було розглянуто способи заповнення квадрату, синтезовано кілька топологій і виконано аналіз характеристик. Проаналізовано можливості використання надлишковості.

Ключові слова: відмовостійкість, надлишковий код, латинський квадрат
Рис.: 5. Табл.: 5. Бібл.: 6.

Urgency of the research. In modern world the distributed computing is an important branch of the development of computer technology. Improving of their fault-tolerance is a one of most important tasks. One of perspective methods of its solution is a method of fault-tolerant topologies synthesis, that allow hardwarely provide a high level of systems fault-tolerance. In article discusses the method of fault-tolerant topologies synthesis with excess code 0/1/-1 and Latin square.

Target setting. A topological structure is an important part of distributed computing system. A lot of parameters depends from it, including fault-tolerance. There are some methods of fault-tolerant topologies synthesis, one of these is a using of an excess code 0/1/-1 in nodes numbers encoding. This article proposes a method of fault-tolerant topologies synthesis with Latin square and excess encoding.

Actual scientific researches and issues analysis. Now a topologies synthesis with Latin square was good considered. Also method of topologies synthesis with excess encoding was proposed [1], the synthesis of fault-tolerant versions of hypercube and quasi-quantum topologies was completed. In previous publications was perfomed an analysis of main advantages and disadvantages of these topologies, proposed methods of using redundancy to improving of fault-tolerance.

Uninvestigated parts of general matters defining. In last time only a synthesis of fault-tolerant topologies based on codes transformations was consider, a using of Latin square and other methods of synthesis without explicit nodes codes transformations wasn't considered.

The research objective. The purpose of the research is a consideration and analysis of possibilities of using Latin square for topologies synthesis with excess encoding of node's numbers, creating examples of these topologies and its analysis.

The statement of basic materials. The problem of synthesis with the square in its usual form is as follows: there are Latin square. In first column – numbers of nodes, for which the definition of neighbors is carried out. The order in which the nodes in the column are listed can be any, it is important that the code of each node is encountered only once. Next, the sequence written in the first column is cyclically shifted up and written to the second column. This doing repeats for next columns analogically until the square been filled. All cols except first are the possible neighbors of nodes, written in first column. After the square creating from possible neighbors in any way, several columns may be selected and a topology is constructed on them.

Basic definitions. The excess code 0/1/-1 has the same parameters as the usual binary, but contains an additional digit -1 denoted by the letter T [1]. The main feature of this code is that one and the same number has several possible views in this code. That's what makes it excess.

Topologies that use this code as a basis have high fault tolerance. This is ensured, first, by the presence of nodes with the same number, and secondly, using routing trees to bypass failures in the system. But, as a rule, the disadvantage of such topologies is too high a power or diameter. In synthesis based on transformations, this is due to the fact that the transformation must take into account an additional number -1, which increases the number of results. For example, on Exchange transformation it leads to the appearance of several results. Exchange with the lowest digit of code 010 gets not only 011, but also 01T. In the framework of transformations-based synthesis paradigm nothing can be done. An alternative to this is a synthesis without transformations.

An example of topology synthesis using Latin square method. In order to construct a Latin square, it is necessary and sufficient to determine the sequence of node's codes entries in the first column. For example, tables 1 and 2 illustrates the Latin squares, constructed for natural sequence and for Gray's code.

To perform the synthesis of the topology, you must select columns in square. For example, select cols 1 and 3. For first square this means, that neighbors of node

000 are nodes 001 and 011, neighbors of 001 – 010 and 100, etc. For second square neighbors of node 000 are nodes 001 and 010, for 001 – 011 and 110, etc.

Table 1

Latin square for 3-bits binary code, written in natural sequence

№	*Node*	*Possible neighbors*						
		1	*2*	*3*	*4*	*5*	*6*	*7*
0	**000**	001	010	011	100	101	110	111
1	**001**	010	011	100	101	110	111	000
2	**010**	011	100	101	110	111	000	001
3	**011**	100	101	110	111	000	001	010
4	**100**	101	110	111	000	001	010	011
5	**101**	110	111	000	001	010	011	100
6	**110**	111	000	001	010	011	100	101
7	**111**	000	001	010	011	100	101	110

Table 2

Latin square for 3-bits binary code, written in Gray's code sequence

№	*Node*	*Possible neighbors*						
		1	*2*	*3*	*4*	*5*	*6*	*7*
0	**000**	001	011	010	110	111	101	100
1	**001**	011	010	110	111	101	100	000
2	**011**	010	110	111	101	100	000	001
3	**010**	110	111	101	100	000	001	011
4	**110**	111	101	100	000	001	011	010
5	**111**	101	100	000	001	011	010	110
6	**101**	100	000	001	011	010	110	111
7	**100**	000	001	011	010	110	111	101

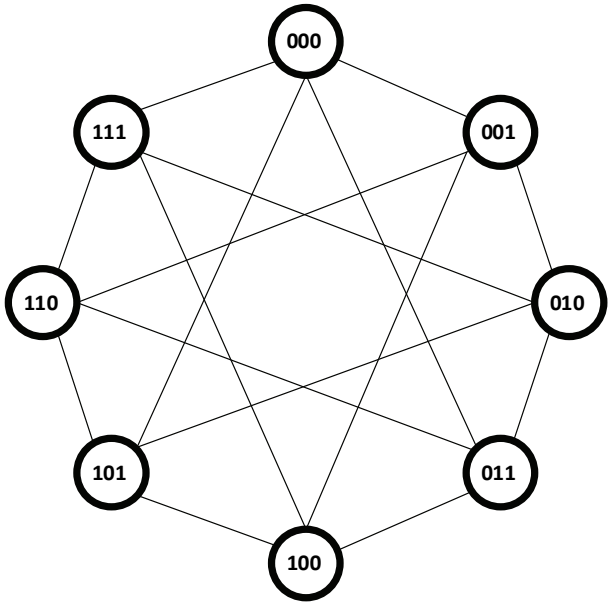


Fig 1. Topology, built with first Latin square with selected cols 1 and 3

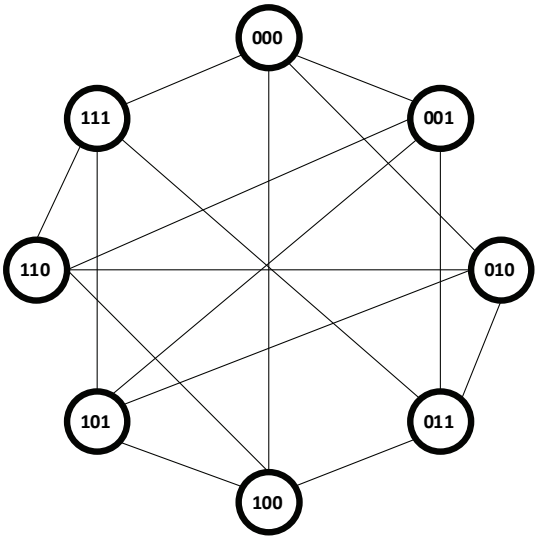


Fig 2. Topology, built with second Latin square with selected cols 1 and 3

The most interesting thing in this synthesis is that both topologies have power 4. Theoretically, in some cases, the power of the synthesized topology can be less than 4, but not more: in 2 columns opposite each node 2 neighbors and the node itself in each of the two columns opposite some other occurs only once.

Also the interesting thing is that obtained topology depends not only from selected columns, but also from sequence in first column.

Also, there is an interesting property in second topology: 3 links of every node analogical to links in hypercube, and 4th link can be received by the way of all code's bits' inversion. It means that routing methods, similar for routing in hypercube, can be used for this topology. For comparison, in table 3 the second topology's system of links described in terms of functional dependencies between node's codes.

Table 3

System of links in second topology in terms of codes transformation

Node	***Exchange***			
	1st bit	***2nd bit***	***3rd bit***	***All bits***
000	100	010	001	111
001	101	011	000	110
010	110	000	011	101
011	111	001	010	100
100	000	110	101	011
101	001	111	100	010
110	010	100	111	001
111	011	101	110	000

Topology synthesis for excess code. In essence, synthesis occurs similarly, but there are differences. There is no natural sequence of excess codes, because some numbers in it repeat with different code, as result, all codes can't be located so that the

value of each next be more than 1. It isn't critical for synthesis, but it can be important for filling the square, because, as it was being described above, properties of topology depend from square's filling. For synthesis use one of possible pseudo-natural sequence for 2-bits excess code. Table 4 illustrates Latin square for this sequence.

Table 4

Latin square for pseudo-natural sequence of the excess codes

№	*node*	*Neighbors*							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
-3	**TT**	T0	T1	0T	00	01	1T	10	11
-2	**T0**	T1	0T	00	01	1T	10	11	TT
-1	**T1**	0T	00	01	1T	10	11	TT	T0
-1	**0T**	00	01	1T	10	11	TT	T0	T1
0	**00**	01	1T	10	11	TT	T0	T1	0T
1	**01**	1T	10	11	TT	T0	T1	0T	00
1	**1T**	10	11	TT	T0	T1	0T	00	01
2	**10**	11	TT	T0	T1	0T	00	01	1T
3	**11**	TT	T0	T1	0T	00	01	1T	10

For example, also select columns 1 та 3. The blue dotted line highlights the connections that pass between nodes with identical numbers.

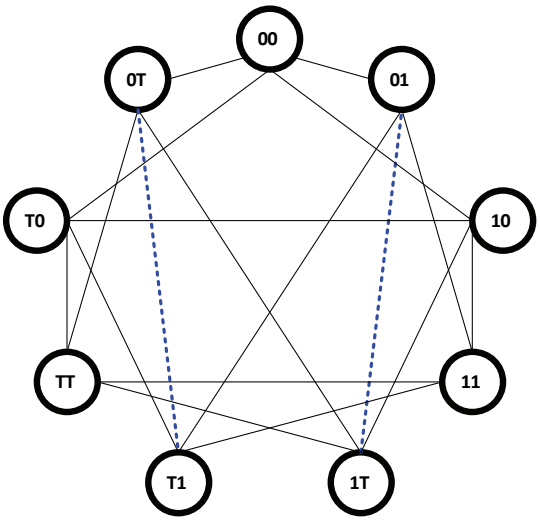


Fig 3. Topology, built with Latin square with a pseudo-natural sequence of excess codes

There is interesting property in this topology: all connections between nodes except 0T – T1, 01 – 1T, 11 – TT complete between nodes with codes, that differs by one bit. In allows in some cases use the routing algorithms, that similar to routing in hypercube.

Using of redundancy. In terms of using the redundancy obtained topology interesting is that has connections between nodes with same number. The reason for

these connections is to select column 1 for the synthesis of the topology: since in sequence these nodes are adjacent, then they are, in this way, adjacent in the topology. The presence of these links allows clustering.

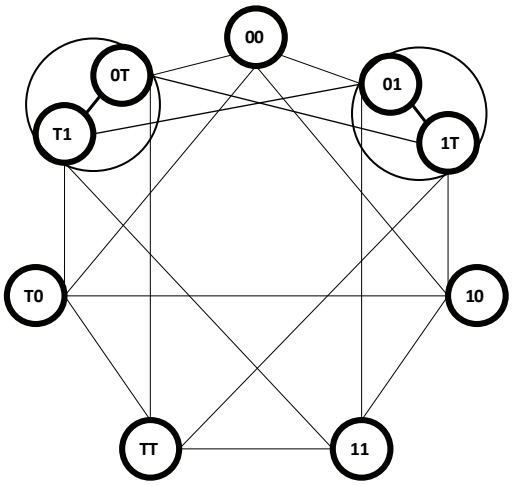


Fig 4. Topology clustering

As you can see, clusters have not only links inside, but also links between. So, even on case of fail one node of cluster this link not be broken. Moreover, if for nodes in cluster somehow provide possibility to intercept packets, that appointed to other node in cluster, every node can be substitute other in case of failure and to reroute packets, that were sent through failed node

Routing in redundant topology. As in the case of hypercube and de Bruyn topology, routing trees can be used in this topology too. It allows use all advantages of this routing method: bypass the failures and avoid locks.

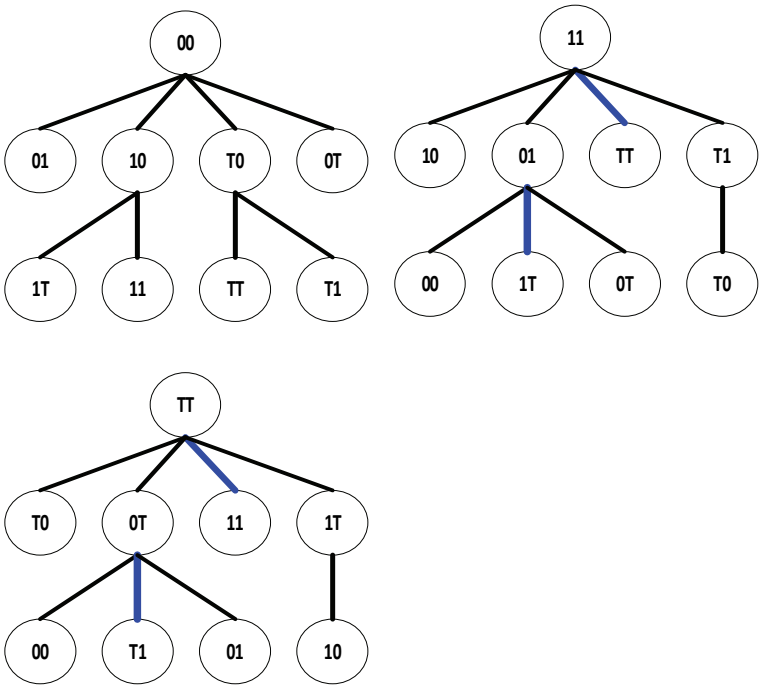


Fig 5. Routing trees for obtained topology

Comparison of the characteristics of the obtained topology. Was completed the comparison between obtained topology with same topology, based on usual binary code and with classic hypercube. In table 5 the results of this comparison are presented.

Table 5

Topologies comparison

Topology	***Classic Hypercube***			***Latin square with natural codes sequence***			***Redundant Latin square***					
Selected cols	-			1, 3			1, 3 (example above)			1, N/3		
Count of bits (N)	3	5	8	3	5	8	2	3	5	2	3	5
Count of nodes	8	32	256	8	32	256	9	27	243	9	27	243
Diameter	3	5	8	2	6	44	2	5	40	2	5	40
Power	3	5	8	4	4	4	4	4	4	4	4	4
Diameter *Power	9	25	64	8	24	176	8	20	160	8	20	160
Count of edges	12	80	1024	16	64	512	18	56	486	18	56	486

Topology	Redundant Latin square					
Selected cols	1, 3, 6, … , N/3			1, 3, 9, …, 3^{N-1}		
Count of bits (N)	2	3	5	2	3	5
Count of nodes	9	27	243	9	27	243
Diameter	2	3	5	2	3	5
Power	4	8	54	4	6	10
Diameter *Power	8	24	270	8	18	50
Count of edges	18	108	6561	18	81	1215

Conclusions. In article proposed method of fault-tolerant topologies synthesis, based on using excess code, with Latin square. Considered the clustering as method of fault-tolerance improving, received the routing trees, that can be used in this topology. Analyzed the characteristics of redundant topologies, that can be received by using Latin square, their comparison with each other was carried out.

A main advantage of method is that in the process of synthesis it is possible to directly determine the desired power of topology. Also, sequence in basis of Latin square allows to delegate some features to topology, that also can be used. Defined, that in these topologies is possible to use routing, based on trees, that allows bypass failures and provides a high level of fault-tolerance. Another advantage of the method is that all the nodes of the resulting topology often have the same power.

But there are disadvantages too. First, obtained topologies not always can be described through transformations with node's codes. This limits the ability to use some routing methods. At second, there are difficulty with defining diameter of topology before synthesis.

There are some ways to improve the method. Firstly, a features of topology depends by sequence in first column of Latin square. Theoretically, it possible to give some features to topology through choosing of sequence. The selected columns are important too. As the comparison showed, changing the numbers and the count of selected columns can be manipulated by power and diameter.

References

1. Goncharenko Olexandr, Pavlo Rehida, Artem Volokyta, Heorhii Loutskii, and Vu Duc Thinh: Routing Method Based on the Excess Code for Fault Tolerant Clusters with InfiniBand. Advances in Intelligent Systems and Computing, vol. 938, pp. 335-345. Springer, Heidelberg (2019)

2. Washington N., Perros H.: Performance Analysis of Traffic-Groomed Optical Networks Employing Alternate Routing Techniques. Lecture Notes in Computer Science, vol. 4516, pp. 1048-1059. Springer, Berlin, Heidelberg (2007).

3. Hu, Z., Mukhin, V., Kornaga, Y., Volokyta, A., & Herasymenko, O. The scheduler for distributed computer systems based on the network centric approach to resources control. In: Proceedings of the 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017, pp. 518-523(2017).

4. Emanouilidis, E and Bell, R. Latin squares and their inverses. Math. Gaz., vol 88(511), pp. 127–128 (2004)

5. Richard J.Cole, Bruce M.Maggs, Ramesh K.Sitaraman, On the Benefit of Supporting Virtual Channels in Wormhole Routers, Journal of Computer and System Sciences, vol. 62(1), pp 152-177 (2001)

6. Ian M. Wanless. Cycle Switches in Latin Squares, Graphs and Combinatorics, vol. 20(4), pp 545-570 (2004).

Autors

Olexandr Goncharenko – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" (Solomenskiy district, ave. Pobedy, 37, 03056, Kyiv, Ukraine).

E-mail: alexandr.ik97@ukr.net

Гончаренко Олександр Олексійович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (Солом'янський район, пр-т Перемоги, 37, м. Київ, 03056, Україна).

Волокита Артем Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Volokyta Artem – associate professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: artem.volokita@kpi.ua

Луцький Георгій Михайлович – професор, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Loutskii Heorhii – professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

РОЗШИРЕНА АНОТАЦІЯ

Олександр Гончаренко,
Артем Волокита, Георгій Луцький

СИНТЕЗ ВІДМОВОСТІЙКИХ ТОПОЛОГІЙ НА ОСНОВІ НАДЛИШКОВОГО КОДУВАННЯ ЗА ДОПОМОГОЮ ЛАТИНСЬКОГО КВАДРАТУ

Актуальність теми дослідження. В сучасному світі розподілені обчислення є важливою галуззю розвитку обчислювальної техніки. Підвищення їх відмовостійкості - одна із найважливіших задач. Доволі перспективним методом її вирішення є синтез відмовостійкої топології для системи. В даній статті розглядається метод синтезу відмовостійких топологій з використанням надлишкового коду 0/1/-1 та латинського квадрату.

Постановка проблеми. Топологічна структура є важливою частиною розподіленої обчислювальної системи. Від неї залежить маса параметрів, включаючи відмовостійкість. Є кілька способів синтезу відмовостійких топологій, одним із яких є використання надлишкового кодування 0/1/-1 в нумерації вершин. В даній статті пропонується метод синтезу відмовостійких топологій за допомогою латинського квадрату і надлишкового кодування.

Аналіз останніх досліджень і публікацій. На даний момент загальновідомим методом є синтез топологій за допомогою латинського

квадрату. Також розроблено метод синтезу топологій з використанням надлишкового кодування [1], виконано синтез відмовостійких версій гіперкуба та квазі-квантової топології. В попередніх публікаціях було виконано аналіз основних переваг та недоліків таких топологій, запропоновано методи використання надлишковості для збільшення відмовостійкості.

Виділення недосліджених частин загальної проблеми. До цього часу розглядався лише синтез відмовостійких топологій на основі перетворень, не розглянутим залишається використання латинського квадрату та інших методів створення топологій, які не використовують явних перетворень коду вершин для отримання сусідів кожної вершини.

Постановка завдання. Завданням є розгляд та аналіз можливостей використання латинського квадрату для синтезу топологій з надлишковим кодуванням номерів вершин, синтез таких топологій та їх аналіз.

Викладення основного матеріалу. Задача синтезу за допомогою квадрату у звичайному вигляді полягає в наступному: є латинський квадрат. В першому стовбці – номери вершин, для яких проводиться визначення сусідів. Порядок, в якому перелічуються вершини в стовбці, може бути будь-яким, важливо, щоб код кожної вершини зустрічався в ньому лише раз. Наступний стовбець визначається так: послідовність, що записана в першому стовбці, циклічно зсувається вгору і записується в другий стовбець. Для наступних стовбців аналогічно, і так до тих пір поки квадрат не буде заповнено. Всі стовбці, крім першого, - можливі сусіди вершин, записаних в першому стовбці. Після формування латинського квадрату із можливих сусідів будь-яким чином обирається кілька стовбців і по ним будується топологія. Основна відмінність запропонованого методу – в використанні надлишкового коду для формування квадрату.

Висновки. Виділено основні особливості запропонованого методу. Виконано синтез топології для прикладу, показано застосування надлишковості для підвищення відмовостійкості. Проведено аналіз характеристик топологій, що можуть бути синтезовані з використанням запропонованого методу, проаналізовано основні переваги та недоліки методу, висунуто пропозиції щодо його покращення.

Ключові слова: відмовостійкість, надлишковий код, латинський квадрат

Section 2. RT (Internet of Things, Real-Time Systems)

UDC 004.315

Dmytro Obozniy, Kateryna Poshtatska,
Valentyna Tkachenko, Oleksandr Verba

RECONFIGURABLE MATH COPROCESSOR ON FPGA

Обозний Дмитро, Поштацька Катерина,
Валентина Ткаченко, Олександр Верба

МАТЕМАТИЧНИЙ СПІВПРОЦЕСОР НА ПЛІС З МОЖЛИВІСТЮ РЕКОНФІГУРАЦІЇ

The article deals with the development of a specialized calculator constructed on the basis of FPGA, for calculating arithmetic-logic functions. The calculator is proposed for integration with the reconfigurable processor, for the implementation of the concept of accelerating the implementation of programs, by hardware realization of the time-critical functional core.

Key words: FPGA, reconfigurable processor, hardware implementation of functional core, mathematical coprocessor, acceleration of computations.

Fig.: 1. Tabl. 0. Bibl. 0.

У статті розглядається розробка спеціалізованого обчислювача, побудованого на базі ПЛІС, для обчислення арифметико–логічних функцій. Обчислювач запропоновано для інтеграції з реконфігуровним процесором, для реалізації концепції прискорення виконання програм, шляхом апаратної реалізації критичних до часу виконання функціональних ядер.

Ключові слова: ПЛІС, реконфігуровний процесор, апаратна реалізація функціональних ядер, математичний співпроцесор, прискорення обчислень.

Рис.: 1. Табл. 0. Бібл. 0.

Relevance of research topic. Today it is very actual to search for new architectural solutions for improving the efficiency of computer systems. This trend is becoming more and more relevant, considering that extensive means of accelerating computations have reached their limitations [1]. On the other hand, the amount of information that is currently being used by high-performance computer systems today is growing extremely rapidly and will continue to grow. Artificial intelligence, which the world now cares about, puts on computer systems a task that can further become overwhelming for modern computers [2]. Therefore, one of the common solutions to

the problem of increasing the efficiency is the hardware acceleration of critical functional nuclei by the use of modern programmable logic integrated circuits (FPGAs). The peculiarity of the modern element base is the ultra-high degree of integration and the possibility of dynamic reconfiguration, which allows the implementation of flexible, high-speed architectural solutions of any complexity that can dynamically adapt to the requirements of solvable problems [3]. The acquisition by Intel of one of the largest companies of FPGA developer Atlera also confirms the relevance and feasibility of the direction of hardware upgrading of computer systems and, in particular, processor cores.

Actual scientific researches and issues analysis. A number of papers [4, 5, 6] present classic solutions for increasing the efficiency of work at the expense of acceleration multiplication. In these works, the authors investigate the logical and hardware methods of accelerating multiplication. Logical methods can reduce the computational time due to more efficient multiplication algorithms, in particular due to the use of redundant numerical systems and systems with a basis of more than two [7]. In hardware methods, the emphasis is on the schematic reduction of computation and addition time [8].

In work [9] the efficiency of using a mathematical coprocessor, which, expansion of the capabilities of the central processor, is implemented as a separate functional module. Such a coprocessor was used to perform complex mathematical calculations, eliminating the central processor from a large number of tasks. According to Intel, a mathematical coprocessor reduces the execution time of mathematical operations, such as multiplication, division and elevation to an extent of 80 percent or more [10].

Today, all Intel and AMD processors, starting with 486DX, have a built-in math coprocessor and do not require a separate coprocessor (except Intel486SX). Although in first-generation computers (i80386, i80486), the mathematical coprocessor module was installed on the motherboard as a separate chip. The built-in SIMD-based extension coprocessor allows data parallelism to be provided. The Intel Advanced Vector Extensions extension provides a set of SIMD instructions for floating-point data processing in groups of 256 bits in length. The Intel MIC co-processors include a 512-bit instruction set.

Using programmable logic integrated circuits allows you to increase the set of instructions for processing data to the bit, limited to only the number of incoming contacts.

There are currently two of the most powerful companies involved in the production of ASICs and Altera and Xilinx programmable integrated circuits (FPGAs). The main products are programmable chips, as well as services for converting projects under FPGA to ASIC for mass production. Companies also develop software for embedded software for FPGAs, as well as compilers under the core of their own processor processors. In order to increase processor performance in

2015, Altera was acquired by Intel and is now expected to release new products in symbiosis of these two manufacturers [11].

The research objective. Within the framework of the modern eco-system of the electronic industry, which is organized according to the model of Fabless, today the world library of complex-functional IP-cores (Silicon IP) is rapidly improving. Developers of Silicon IP compete in optimizing performance and using hardware resources. Today's topical issue is the optimization of hardware solutions to accelerate the implementation of the transaction and energy efficiency. Such blocks are used for the design of computing systems, systems-on-chip, digital devices, reconfigurable processors.

In order to solve the actual problem of efficiency increase, the following tasks are set out in the article:

- develop a functional block model (IP-core) for performing basic mathematical and logical operations for integration into a reconfigurable processor core;
- to investigate the time of propagation of signals during execution of arithmetic and logical operations;
- To investigate the efficiency of the IP multiplier nucleus, implemented on the basis of the multiplication acceleration method in comparison with the Altera library core IP.

The statement of basic materials. The article proposes the use of a coprocessor of the matrix multiplier called "Bo-Wooley" for the implementation of the multiplication operation. The structure of this matrix multiplier includes 2 types of blocks, which differ in the presence after the block OR the block NO. These blocks consist of the block OR, the adder, the block NO (depending on the type) and has 4 inputs and 2 outputs. Also, on the last layer there are adders. If necessary, it is possible to expand the grid matrix according to the needs of the bit.

Figure 1 depicts ALU that performs functions: multiplication, division, addition, subtraction, AND, OR, XOR, NOT. The device has a digit of 4 carries out operations with a sign.

At the input of the coprocessor are the arguments A and B, as well as the operation code F. All operations blocks perform appropriate actions, the results of which are sent to the multiplexer. The multiplexer (MUX) outputs the result of an operation that was specified by the operation code F.

Simulation of the work of the coprocessor. The project is developed by Altera's Quartus II CAD. The correctness of the data was verified by simulating a time chart. In the simulation, the Multiplexer (MUX) tested outputs the result of the operation that was specified by the operation code. The value of the delay of signals in the chains of the block synthesized using the mega-function LPM_MULT and the block synthesized using the Bo-Wool algorithm. The block time performance by the Bo-Wool method proved to be the best (12.889 ns) for the LPM_MULT (13.437 ns) block with four-bit input data.

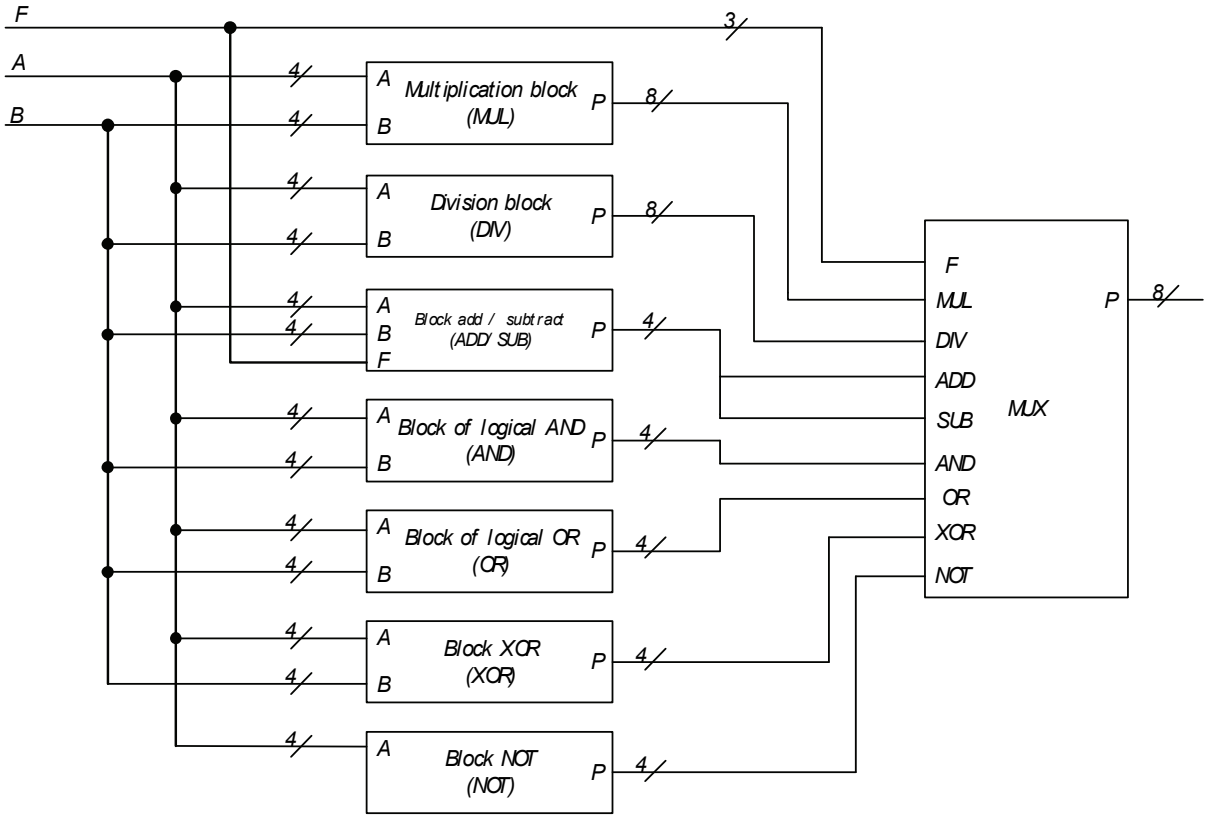


Fig. 1. The overall structure of the reconfigurable processor

Conclusions. Multistep arithmetic operations require a considerable time to execute at the software level and can significantly reduce system performance. The proposed implementation of computing units on an FPGA contributes to increased productivity by reducing the time spent doing the calculations and the possibility of dynamic reconfiguration from the bit rate of words processed to change the functionality in accordance with the requirements of solved tasks. The proposed solution allows you to implement operations of adding, multiplication and floating-point division on a coprocessor that is part of the computing system.

The main requirements for the development of new operating devices is the acceleration of operations. This was achieved during the implementation of the task in question.

The simulation results of the coprocessor showed a decrease in the multiplication time compared to the software implementation of the calculations by means of microcontrollers company Altera.

According to the results of the time analysis, the delay of the developed mathematical coprocessor is 12,889 ns, which is 0.548 ns less, compared with the calculator Altera (13,437 ns).

The main advantage of simulating a coprocessor is scalability. This means that the four-bit modules that are modeled and demonstrated in the article can be n * 4 arguments. Accordingly, with increasing the bit rate indicators will provide even better results.

The prospect for the development of work is the development of new units that perform other arithmetic operations (square root, elevation to degree). Also, as a further extension, it is expedient to develop a unit of firmware control.

References

1. Angepat H., FPGA–Accelerated Simulation of Computer Systems / H. Angepat, D. Chiou, E. Chung, J. Hoe. – IEEE, 2014. – 125–126.

2. Zhang Z., FPGA–oriented moving target defense against security threats from malicious FPGA tools/ Z. Zhang ; Q. Yu ; L. Njilla ; C. Kamhoua// 2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST). – IEEE,2018. – P 5–6.

3. N. Shylashree. FPGA implementation of high speed scalar multiplication for ECC in GF(p)/ N. Shylashree, V. Sridhar// TENCON 2015 – 2015 IEEE Region 10 Conference, (Macao, China, 1–4 Nov. 2015) / IEEE, ISBN: 978–1–4799–8641–5. – P 8–9.

4. Braun, E. Digital Computer Design, Logic Circuitry, Synthesis//Academic Press, New York, 1963. – P 324–326.

5. Lehman, M. High–speed Digital Multiplication, IRE Transaction on Electronic Computers// Vol. EC–6–6, № 3, 1957. – P 6.

6. Taylor F., Multiplier policies for digital signal processing/ F. Taylor. – IEEE ASSP Magazine, vol. 7, Jan. 1990, – P. 6–20.

7. Booth, A., A Signed Binary Multiplication Technique, Quart. J. Mech. Appl. Math. 4, part 2/ 1951, –P. 236–240.

8. Chopade S. Performance analysis of vedic multiplication technique using FPGA/ S. S. Chopade, Rama Mehta// 2015 IEEE Bombay Section Symposium (IBSS), (Mumbai, India, 10–11 Sept. 2015)/ IEEE, ISBN: 978–1–4673–9542–7, – P. 489–490.

9. Al–Wattar A. Efficient On–line Hardware/Software Task Scheduling for Dynamic Run–time Reconfigurable Systems / A. Al–Wattar, S. Areibi, F. Saffih // Proceeding in 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW). – IEEE, 2012. – P 401 – 406.

10. Gonzalez I. Dynamically Reconfigurable Coprocessors in FPGA–based Embedded Systems: Doctor of Science in Electronics Thesis : 21.03.2006 / I. Gonzalez; Universidad Autónoma de Madrid. – Madrid, 2006. P — 56 p.

11. Baugh, C. R., Wooley, B. A., A Two's Complement Parallel Array Multiplication Algorithm, IEEE Transactions on Computers, C–22, Dec. 1973, – P.1045–1047.

Autors

Poshtatska Kateryna –student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: poshtatskayakatirina@gmail.com

Поштацька Катерина Володимирівна – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Obozniy Dmytro –student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: DObozniy@gmail.com

Обозний Дмитро Миколайович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Tkachenko Valentyna – Associate Professor, Candidate of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: tkavalivas@gmail.com

Ткаченко Валентина Василівна – доцент, кандидат технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Верба Олександр – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Verba Oleksandr – Associate Professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: olverba@gmail.com

UDC 004.315

Обозний Дмитро, Поштацька Катерина,
Ткаченко Валентина, Верба Олександр

МАТЕМАТИЧНИЙ СПІВПРОЦЕСОР НА ПЛІС З МОЖЛИВІСТЮ РЕКОНФІГУРАЦІЇ

Анотація. Розроблено спеціалізований обчислювач для інтеграції з реконфігуровним процесором, який дозволяє прискорити виконання програм, шляхом апаратної реалізації критичних до часу виконання функціональних ядер. Запропонована модель спеціалізованого обчислювача на ПЛІС для виконання базових математичних та логічних операцій.

Вступ. В сучасному світі великої актуальності набув пошук нових архітектурних рішень для підвищення ефективності комп'ютерних систем. Ця тенденція стає все більше актуальною, зважаючи на те, що екстенсивні засоби прискорення обчислень досягли своїх граничних можливостей .

Мета роботи. Збільшити продуктивність обчислювальних блоків на ПЛІС за рахунок зменшення часу виконання обчислень та можливості динамічної реконфігурації від розрядності оброблюваних слів до зміни функціональних можливостей відповідно до вимог вирішуваних задач.

Постановка проблеми. Для вирішення актуальної проблеми підвищення ефективності у статті поставлені наступні завдання:

– розробити модель функціонального блока (IP–ядро) для виконання базових математичних та логічних операцій для інтеграції у реконфігуровне процесорне ядро;

– дослідити час розповсюдження сигналів під час виконання арифметичних та логічних операції;

– дослідити ефективність IP–ядра помножувача, що реалізований на базі методу прискорення множення у порівнянні з бібліотечним IP–ядром компанії Altera.

Основна частина роботи. В роботі запропоновано апаратурний співпроцесор на ПЛІС. Для реалізації операції множення розроблено матричний множник, що носить назву «Бо-Вулі В структуру розробленого матричного множника входить 2 типи блоків, що відрізняються наявністю після блоку АБО блоку НІ. Ці блоки складаються з блоку АБО, суматора, блоку НІ (в залежності від виду) та має 4 входи та 2 виходи. Також на останньому шарі знаходяться суматори. За необхідності можливо розширювати сітку матриці згідно з потребами розрядності.

Проект розроблений в САПР Quartus II компанії Altera. Коректність даних була перевірена за допомогою моделювання часової діаграми. При моделюванні були перевірені ульліплексор (MUX) видає на вихід результат операції, яку було

задано кодом операції F.значення затримки сигналів в ланцюгах блоку синтезованого за допомогою мега-функції LPM_MULT та блоку синтезованого за допомогою алгоритма Бо–Вулі.

Висновки. Розроблений математичний співпроцесор для виконання базових математичних та логічних операцій. Прискорення виконання операції множення досягнуто за рахунок зменшення часу виконання обчислень та можливості динамічної реконфігурації від розрядності оброблюваних слів до зміни функціональних можливостей відповідно до вимог вирішуваних задач.

Результати моделювання роботи співпроцесора показали прискорення виконання операції множення у порівнянні з відомою мегафункцію LPM_MULT компанії Altera. За результатами часового аналізу, час формування результату засобами розробленого математичного співпроцесора складає 12,889 нс, що на 0,548 нс менше, порівняно з відомим обчислювачем (13,437 нс). Використання методу «Бо-Вулі» для реалізації множення дозволило побудувати масштабований співпроцесор на ПЛІС. За результатами експериментів отримано, що при збільшенні розрядності показники швидкодії роботи співпроцесора збільшуються.

Перспективою до розвитку роботи є розробка нових блоків, що виконують інші арифметичні операції (квадратний корінь, піднесення до степеню). Також, в якості подальшого продовження, доцільним є розробка блоку мікропрограмного управління.

UDC 004.7

Victor Petrov,
Iryna Klymenko, Oleksandr Verba

METHOD TO IMPROVE EFFICIENCY OF MANUFACTURING ACTIVITY BY INTERNET OF THINGS TECHNOLOGY

The questions of increasing the efficiency of production activity by implementing the Internet of Things technology at the enterprise are considered. The method to use the Internet of Things and GPS technology to control and monitor the quality of personnel performance is proposed. Software in which the proposed method is implemented has been developed. The proposed tools of the Internet of Things can be integrated into the activities of enterprises. Using smartphones with the Android operating system has minimized the resources spent on achieving the effect of the proposed tools.

Keywords: Internet of Things, GPS, control, increase of efficiency

1. Relevance of the topic of research. In recent years, in various spheres of human activity such a notion as the Internet of Things (IOT) is gaining popularity. This technology allows you to display different devices in an open network so that they can interact with each other. IoT is also the ability to connect devices without human intervention, and most importantly - a large amount of data that generates and assembles devices, which can then be analyzed in order to be used in future for various needs, for example - increasing the comfort and business decision making.

2. Substantiation of research problems. At this stage of mankind's existence, population growth and technology improvement, the question arises of creating a productive way of correctly setting up diverse tasks and controlling their implementation. The Internet of things is the best suited for achieving such goals and their research - it will allow collecting and analyzing information that can be used already to meet the goals.

The article suggests the use of internet of things on the example of one of the enterprises that faced the problem of setting tasks and monitoring their implementation. The selected enterprise uses obsolete inefficient technologies in this matter, which use too much material resources and do not allow to compete normally in their market with other enterprises.

Having reviewed the enterprise and its methods of achieving the task, it was concluded that the emphasis should be placed on controlling the fulfillment of the tasks by the workers - this is an urgent question in this enterprise, as well as laying the foundation for technologies and algorithms that will allow to perform qualitative statement of tasks. That is why it was chosen to use tracking to solve the problem as a basis.

3. Analysis of recent research and publications. A well-known example of the implementation of tracking control issues on the basis of IOT is the control of the movement of diverse objects. These may be containers for sea transportation, vehicle

tracking, etc. For these purposes, trackers and sensors are typically used in a well-protected solid case and have a battery as a power source.

In the case of container transport, special equipment is usually used, which is created specifically for such needs. BlackBerry Radar from BlackBerry, which specializes in container tracking, can be considered as a representative of this class [2]. This is a device for tracking and container logistics. It is not large (292mm x 93mm x 42mm), it has a solid, well-protected mechanical damaged case, which has a number of sensors installed to monitor the condition of the container and objects that can be located and should be installed in the middle of the container. There are also universal solutions. As an example, you can consider the universal product ARUBA [3]. This product differs from the past only with its dimensions and is designed to be more for use within a single enterprise or production. Its main features are compactness, high reliability, low battery power consumption and a very long battery life of 36 months.

There are also products that are designed for logistics and are used for controlling the movement of vehicles such as: wagon carrying containers or tracking of urban transport. These include Atmel GPS Asset Tracking from Microchip. [4]

The main disadvantage of the above products can be considered that they carry out control of transportation by object, and in our case, control should be carried out by the subjects, that is, workers of the enterprise. Also, the disadvantages include the non-universality and dimensions. But it should be noted that these devices also have interesting advantages, for example - the ability to work in the city.

The next step was to focus on devices that can perform human-tracking. An example is Smart IoT People Tracking. [5] This development is very versatile and does not have the disadvantages described above. It allows you to perform quality control in many areas of human activity and use the technology of the internet of things to the fullest. But such an approach does not allow to solve all the questions and problems raised, namely the task setting for the workers of the enterprise.

After reviewing the solutions presented and a clearer understanding, attention was drawn to how other companies managed this. The best representative can be considered a whole stratum of modern enterprises such as Uber. They use, and for tracking, and for task statements - mobile smartphones. But their kind of activity is aimed at services in the field of taxi.

After fully understanding and analyzing the technologies that are being used and by which one can solve the problem, the following product was introduced: "Where are you?". [6] A versatile product that has both control and task functions.

Having studied this product in detail, attention was paid to some of the drawbacks and nuances. This product is secured to a specific area and will not be able to work except for it. Also, this product does not use independent development, and Navixy API, which in turn can cause some errors in the software. And most importantly, the product does not use the internet of things technology, which, in its turn, would help to carry out work more qualitatively and add many other useful features to the product in the future.

4. Uninvestigated parts of general matters defining. Thus, with regard to the existing means of control at enterprises, the traditional problem of the present can be considered as outdated traditional solutions. These tools are often autonomous and specific, they require a lot of expenses for their support and improvement according to the modern requirements of users. Modern world trends require the integration of a wide range of enterprises and production process control systems in the IOT to maximize the effect of their activities. Therefore, the research done in the article is relevant, modern and expedient.

5. Target setting. Not based on the review of existing solutions and analysis of the shortcomings of modern enterprise management systems, the following task has been put forward and fulfilled. Software for equipping the workers of the enterprise has been developed. Devices that perform the main task - these are smart phones workers. This solution was facilitated by the fact that the product should be widely used by personnel that is constantly changing. That is why there is no need to use specialized equipment, which in its turn is rather cumbersome, does not meet all the goals set. The task of each device is to recognize and establish the exact geolocation of the user in the city environment, in the shops of wholesale and retail trade. Unlike the existing tools, it is proposed to use algorithms that accurately identify a location with a low error, index user when they are next to each other and the ability to add other functions related to personnel management and task setting.

The main features and functionalities aimed at solving this task include the use of both GPS and GSM smartphone modules, which contributes to precise geolocation in a city environment that is used for commercial purposes, which in turn allows you to use logic and algorithms for other tasks, which can create a market. Also, the development provides a high saving of resources in technical terms. Saving resources is to develop only software, and use as a device - a smartphone user.

6. The statement of basic materials. Because it was chosen to use mobile phones to achieve the goal - product development is implemented in the Java programming language and has the look of the Android application. This allows you to use a wide range of technical features and capabilities of this operating system, and the technical resources of the devices themselves, such as built-in GSM and GPS modules in order to accurately determine the exact location of the user.

To achieve the user location goal, the Location-based service was used, which in turn uses the gsm technology to recognize and can use wi-fi for greater accuracy.

The solution to the problem is implemented on Android 8.1 (Oreo) because in my opinion - it is a sufficiently proven version of the operating system, which came out in 2017 and will be supported by the largest number of smartphones, which in turn will make the application as widely available and effective in distribution.

As mentioned above, the gsm module will be used to improve and accurately position my product in addition to gps. The algorithm of the work is that at first the program will use the GPS module of the device and in case when it will be impossible to do this - will use the GSM module. This algorithm is depicted in Fig. 1

The chosen operating system allowed adding to the product a useful functionality for setting tasks and having feedback from users. An example is the use of forms, chat, and easy communication with managers.

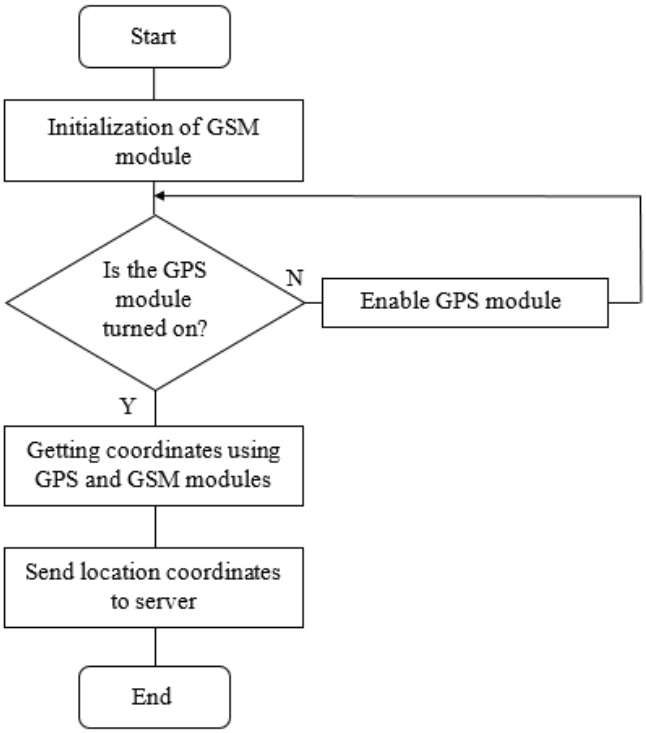


Fig. 1. Schematic representation of gps and gsm modules

There was also a permanent work with the server, from which the user can receive instructions and perform the tasks in real time. The instructions and tasks are created in a special form, which simplifies its sending from the server and makes the interface to the user more intuitive. At the same time, the implementation is executed in the opposite direction - the user has the opportunity to form and send to the server the result of their work in a convenient form. This allows the enterprise to get rid of unnecessary misunderstandings.

In this mode, a high employee benefit rate is monitored and monitored, and he is reminded that he must return to work if he leaves his place of work or moves away from that department or in other emergencies. And most importantly - with the help of using special forms in setting tasks it becomes possible to make the most of the resource that we provide the internet of things, which will make the work of the company more efficient.

7. Conclusions. The article was devoted to questions of improving the efficiency of production activities using technology internet of things, was tasked with the example of the chosen company - to implement this question. For this purpose, an overview of existing decisions was made, and the setting of specific objectives, goals and solutions to current problems in the world on the issue. Having obtained and analyzed the received information, it was specified and set the task of solving the problem of using mobile smartphones with the combination of technology internet of things.

References

1. AIN.UA (2017) Available at: https://ain.ua/special/what-is-iot/ (accessed 10.04.2019).

2. Asset Monitoring Devices Purpose-Built For Transportation (2018) Available at: https://www.blackberry.com/us/en/products/blackberry-radar/tracking-devices (accessed 12.04.2019).

3. Asset Tracking with Aruba Tags (2018) Available at: https://www.aruba-networks.com/products/location-services/aruba-tags/?source=sem&gclid=Cj0KCQjwsZHPBRClARIsAC-VMPDuRxZkSJRDq6gfOnzzV5jp4N_1e9NZGbih-Uqhp4f5xx6QE5MN-fQaAv6OEALw_wcB (accessed 12.04.2019).

4. Smart Connected Secure (2018) Available at: https://www.microchip.com/ /design-centers/internet-of-things (accessed 15.04.2019).

5. Smart IoT People Tracking (2018)– Available at: http://www.hitssolutions.com/smart-iot-worker-location-tracker/ accessed 20.04.2019).

6. GPS-трекер for Android (2018) Available at: https://www.gdemoi.ru /app/tracker/android/ (accessed 20.04.2019).

Autors

Petrov Victor – master student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: thevityun@gmail.com

Петров Віктор – магістр, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Klymenko Iryna – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: ikliryna@gmail.com

Клименко Ірина – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Verba Oleksandr – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: olverba@gmail.com

Верба Олександр – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

EXTENDED ANNOTATION

Petrov Victor,
Klymenko Iryna, Verba Oleksandr

METHOD TO IMPROVE EFFICIENCY OF MANUFACTURING ACTIVITY BY INTERNET OF THINGS TECHNOLOGY

Relevance of research topic. In recent years, popularity in various spheres of human activity is gaining such a notion as the Internet of Things (IOT). This technology allows you to display different devices in an open network so that they can interact with each other. IoT is also the ability to connect devices without human intervention, and most importantly - a large amount of data that generates and collects devices, which can then be analyzed in order to be used in future for various needs, for example - increasing the comfort and business decision making.

Formulation of the problem. Absence of cheap tools for improving the efficiency of production activities of employees using technology Internet of Things.

Analysis of recent research and publications. A well-known example of the use of IOT technology for improving efficiency is the special equipment used in tracking, for example, marine containers. Also, the technology is used for tracking diverse transport or small-sized cargoes between enterprises. But this technology is not used to track people - employees of enterprises.

Setting objectives. Since the market is to some extent not using the technology of the IOT in improving the efficiency of production activities of employees, and after analyzing existing similar solutions, it was decided to create a mobile application by which the worker's work statistics are collected, his tracking is going to be. The statistics are transmitted to the server where the analysis is carried out and the conclusions are executed.

Presentation of the main material. Created application for use on mobile smartphones of enterprise workers for collecting information about their location and transferring this information for analysis on the server. This approach uses the logic of the IOT and allows using the information obtained to improve the efficiency of production activities.

Conclusions. A new approach was developed to use the IOT technology in a trekking process for the employees of the company using their mobile phones. This approach has allowed to improve production efficiency more than 2 times.

Keywords: Internet of Things, GPS, control, increase of efficiency

UDC 004.7

Kruk Yaroslav, Kulakov Yurii

ENERGY EFFICIENCY IN WIRELESS NETWORKS OF INTERNET OF THINGS

Крук Ярослав, Кулаков Юрій

ЕНЕРГОЗБЕРЕЖЕННЯ В БЕЗДРОТОВИХ МЕРЕЖАХ ПРИСТРОЇІВ ІНТЕРНЕТУ РЕЧЕЙ

Annotation. This article is devoted to the analysis and design of energy-efficient work and communication model of Internet of Things devices, which connected by a wireless network. The sleep-wakeup cycles taken as a base of the energy efficiency. The degree of filling node`s buffer and the number of sleep-wakeup cycles without data transmission are taken as correction metrics of node sleep duration. The routing performed by considering the shortest paths and node loading.

Keywords: energy efficiency, wireless asynchronous networks, scaling, Internet of Things, routing.

Fig.: 0. Tabl.: 0. Bibl.: 3.

Анотація. Дана стаття присвячена аналізу та проектуванню енергоефективної моделі роботи та комунікації пристроїв інтернету речей, що поєднані бездротовою мережею. За основу енергозбереження взято цикли переходу пристроїв у режими сну та пробудження. В якості метрик корекції тривалості сну використовується ступінь заповнення буферу вузла та кількість циклів сну-пробудження без передачі даних. Маршрутизація відбувається по найкоротшим шляхам з врахуванням завантаженості.

Ключові слова: енергозбереження, бездротові асинхронні мережі, масштабування, інтернет речей, маршрутизація.

Рис.: 0. Табл.: 0. Бібл.: 3.

Relevance of research topic. Due to rapid development of mobile devices of internet of things, the topic of energy efficiency becomes very important to the spheres of life, where autonomous lifetime and scalability are the critical characteristics. Because of the limitations in these characteristics, there is a slowdown of implementation in certain spheres of life, for example, agricultural industry.

Analysis of resent research and publications. There are plenty of methods of work lifetime prolongation of wireless sensors – radio optimisation, aggregation and intermediate data processing, duty cycles schemes, energy-efficient routing and

charging methods [1]. Only one of these methods is the most energy-effective and which configured by a software – duty cycles schemes. In [2] clusterization method is used along with duty cycle schemes where a duty node is chosen in each cluster, which receives messages from other nodes within the cluster for some period of time and then sends this data further to the sink and control node (base station). The disadvantage of this method is low scalability because with node distance increase the clusterization ability decreases.

Target setting. The best method of energy saving in wireless networks of internet of things devices is switching nodes to a sleep state which is characterized by ultra-low power consumption. The switching to a sleep state means disconnect from the network, which makes the further communication impossible. Thus, a periodical wakeup is required for reconnection. In this case, each node with a certain periodicity can generate data retrieved from the connected sensors. Thus, this scheme of work nodes communication includes sleep/wakeup cycles with asynchronous communication, which ensured by buffer presence and work duty cycling.

The statement of basic materials

The determination of cycle duration. Immediately after awakening, the node activates the radio channel and broadcasts the beacon message so the other nodes in the covered region can find out about its presence and active status in the network. The time needed for beacon message transmission is denoted as T_B. So for receiving of the beacon message, an interval of time T_{ACK} is needed to eliminate the possibility of missing a reply to a message and which is 2-3 times bigger than T_B correspondingly. Besides that, the active mode also involves the work with connected sensors which takes T_W and communication time T_{RTX} itself. Since the process of sensing is periodical, its frequency can be denoted as ω and the maximum sleep time will be $T_{Smax} = 1/ \omega$ correspondingly. Hence, we get a formula of effective time of sleep:

$$T_S = \frac{1}{\omega} - (T_B + T_{ACK} + T_W + T_{RTX}) \quad (1)$$

The given formula (1) doesn't resolve the synchronization problem (the presence of 2 and more nodes in access zone for the same time slot). So, with a long time work there's a high probability of local buffer overflow and hence the unavailability to transmit the data. Thus, (1) is suitable only for the initial setting sleep/wakeup cycles right after routing tables initialization on each node. So it means that it's required to change sleep duration dynamically depending on nodes state. As such metric the node`s buffer filling degree was taken. The more data in the buffer, the faster this data should be transmitted to the subsequent nodes in order to avoid data accumulation and buffer overflow. This can be expressed as follows:

$$T_{S_i} = T_S * \left(1 - \frac{n}{C}\right) \quad (2)$$

Where, n – data size in buffer, c – buffer capacity, T_S – maximum sleep duration, T_{S_i}- sleep duration for the next cycle. Due to (2) the problem with buffer overflow was resolved but the same time may be a case where buffer is not replenished by new messages (for example, because of sensors failure or their complete absence) and the condition of decreasing of sleep time never happens. Therefore, a new metric should be presented, such as the number of cycles *m* during which the message was stuck at the buffer without being transmitted. Let's introduce the K_i coefficient for which value the next sleep duration will be divided:

$$T_{S_i} = \frac{T_S * \left(1 - \frac{n}{C}\right)}{K_i} \tag{3}$$

Where K_i is determined as:

$$\begin{cases} K_0 = 1, \text{для } m = 0 \\ K_i = K_{i-1} * (1 + m * k_b), \text{ для } m \geq 1 \end{cases} \tag{4}$$

In (4) the k_b defines the coefficient of message importance (priority) and is constrained: $0 < k_b \leq 1$. Thus, the longer message stays in the buffer the less time the node will sleep so there's bigger probability to detect ready for transmission neighbor nodes in access range.

Routing. At the first glance, routing in asynchronous networks with periodically sleeping nodes is a difficult problem. Since duty cycles guarantees with delay though, the existence of connection between two and more nodes the problem can be simplified for finding the shortest path. During network initialization, there is construction of routing tables for each V_i node from the W set. In this case each node calculates and saves in memory the cheaper path cost to each node in the netwok which can be achieved using Dijkstra's algorithm. Taking into account asynchrony, the same time receiving node choice algorithm requires adjustments in both selection principle itself and the routing table. So each node stores the information about paths costs from itself to other nodes, but the same time it has similar records for neighboring nodes that are available. This is due to the fact that during the optimal path selection, the node through which the path lies can be in a sleeping state. And in fact the routing is as follows: select the available nodes, which have path cost less than the transmitting node and select the less costing node taking into account load balancing.

Balancing. In order to avoid nodes overhead which make an optimal path it's a normal practice to balance the loading by the mean of these nodes which are less involved in data transmission. The node load can be expressed through the degree of buffer filling:

$$\beta = \frac{n}{C} \tag{5}$$

Correspondingly, the receiver node selection will be as follows:

$$N_i = \min\left(h_i + \sqrt{h_i} * \beta\right) \tag{6}$$

Where, h_i – path cost from i-neighbor node to the destination node and N_i – is a number of the selected receiver node during the routing task.

Conclusions. Using the proposed integration method of nodes operation mode the significant reduced energy consumption can be achieved due to a sleep mode. In spite of the periodicity of wakeup state, the nodes data transmission ability was saved by dynamic sleep/wakeup cycles reconfiguration. This method allows you to reduce node density on some area, improve the coverage correspondingly without increasing energy consumption, and improve the scalability of the system.

References

1. Tifenn Rault. Energy-efficiency in wireless sensor networks [Електронний ресурс] // Sorbonne Universit´es, Universit´e de Technologie de Compi`egne, CNRS, Heudiasyc UMR 7253 – 2017. – Режим доступу до ресурсу: https://tel.archives-ouvertes.fr/tel-01470489/document

2. L. Kokila, Dr. D .Saraswady. Energy Efficient Routing Protocol for SelfAdaptive Sleep/Wake-Up Scheduling Approach in WSN [Електронний ресурс] // International Journal for Research in Applied Science & Engineering Technology (IJRASET) – 2018. – Режим доступу до ресурсу: https://www.ijraset.com/ fileserve.php?FID=16835

3. splitface - Алгоритм Дейкстры. Поиск оптимальных маршрутов на графе [Електронний ресурс] // habr – 2011. – Режим доступу до ресурсу: https://habr.com/ru/post/111361/

Autors

Kruk Yaroslav – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Politechnic Institute”.

E-mail: uakruk@ukr.net

Крук Ярослав Ігорович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Kulakov Yurii – Doctor of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Politechnic Institute”.

E-mail: ya.kulakov@gmai.com

Кулаков Юрій Олексійович – доктор технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

EXTENDED ANNOTATION

Kruk Yaroslav,
Kulakov Yurii

ENERGY EFFICIENCY IN WIRELESS NETWORKS OF INTERNET OF THINGS

Relevance of research topic. Due to rapid development of mobile devices of internet of things, the topic of energy efficiency becomes very important to the spheres of life, where autonomous lifetime and scalability are the critical characteristics. Because of the limitations in these characteristics, there is a slowdown of implementation in certain spheres of life.

Analysis of resent research and publications. There are plenty of methods of work lifetime prolongation of wireless sensors – radio optimisation, aggregation and intermediate data processing, duty cycles schemes, energy-efficient routing, charging methods and clusterization method.

Uninvestigated parts of general matters defining. The requirements in hardware change configuration for most of methods which result in more costs of the system. The methods that can be implemented with a software have problems with data integrity, delivery and scalability.

Target setting. The best method of energy saving in wireless networks of internet of things devices is switching nodes to a sleep state which is characterized by ultra-low power consumption. Thus, this scheme of work nodes communication includes sleep/wakeup cycles with asynchronous communication, which ensured by buffer presence and work duty cycling.

Найкращим можливим способом збереження енергії в бездротових мережах пристроїв інтернету речей є переведення вузлів у режим сну, який характеризується наднизьким енергоспоживанням. Таким чином, дана схема роботи та комунікації вузлів передбачає чергування циклів сну та робочого стану вузлів з асинхронною комунікацією, яка забезпечується наявністю буфера у вузлів та циклічністю роботи.

The statement of basic materials. The sleep cycle duration is determined dynamically depending on the node and the network state. The required data sensing rate, time required for sending the beacon message, time requeired for listening for incoming messages, data transmission time and sensors work time are taken into account. In this case, the sleep duration is adjusted depending on buffer load as well as the time messages spent in the buffer in the dimension of cycles sizes and importance of these messages. The routing is based on the shortest path principle taking into account the load of the nodes.

Conclusions. Using the proposed integration method of nodes operation mode

the significant reduced energy consumption can be achieved due to a sleep mode. This method allows you to reduce node density on some area, improve the coverage correspondingly without increasing energy consumption, and improve the scalability of the system.

Keywords: energy efficiency, wireless asynchronous networks, scaling, Internet of Things, routing.

UDC 004.7

Yevheniia Zubrych,
Oleksandr Podrubailo

GENERATION OF THE SHORTEST ROUTE BASED ON THE VISIBILITY OF THE INTERMEDIATE POINTS

Євгенія Зубрич,
Олександр Подрубайло

ПОБУДОВА НАЙКОРОТШОГО МАРШРУТУ НА КАРТІ З УРАХУВАННЯМ ОБЛАСТЕЙ ВИДИМОСТІ ПРОМІЖНИХ ПУНКТІВ

This paper analyzes the existing routing solutions and identifies their disadvantages. Based on the obtained results, an algorithm for constructing the shortest route on the map is proposed.

Key words: route planning, tourist routes, area of object's visibility.
Fig.: 3. Tabl.: 1. Bibl.: 9.

У даній роботі було проаналізовано існуючі рішення для географічної маршрутизації, виявлено їх недоліки. Розроблено алгоритм для побудови маршруту з урахуванням областей видимості проміжних пунктів.

Ключові слова: планування маршрутів, туристичні маршрути, області видимості об'єктів.
Рис.: 3. Табл.: 1. Бібл.: 9.

Relevance of research topic. Today, automated systems for building and visualizing routes are used in many industries. Route planners have become a part of everyday life. Tourism is the most urgent area. The positive impact of information technology on the dynamics of the local and international tourist flow has led to the transformation of the tourism industry from a service-oriented to a diversified field of activity aimed at meeting the diverse needs of millions of individual tourists [1].

The real-time planners used for tourist purposes are especially in demand.

Formulation of the problem. Existing algorithms are focused at finding the fastest and shortest route. However, using systems that generate only the shortest route, the tourist risks not to see many of showplaces in the particular area.

Analysis of recent research and publications. With the growing demand for navigation systems and route planners, the number of researches in this area has also increased. There are many systems that built the shortest path between the two points.

Most of them can be divided in two main types of existing routing applications: offline, or onboard planners and online or web planners. Numerous applications of both types exist, this chapter will only name a few very popular ones and several that are in some way related to this project.

TomTom is a large international company offering stand-alone navigation devices. Their devices are some of the most popular, mainly due to the intuitive interface, speed and accuracy of route calculations. Devices can count on routes for traveling on a car, bike or on foot. Unfortunately, their routing algorithm and data sources are not open to developers. TomTom has also recently released an online version of its route planner, but this version does not have many features that offers a built-in version [2]. Among other things, it lacks the ability to plan routes by bike or on foot. Also, neither an online version nor a built-in version has a Ukrainian localization.

Google has launched its own routing service, called Google Maps [3]. This is a very fast free service. Obviously, Google performs some preprocessing or caching to make their routing service so fast, but the details and routing algorithm remain secret. Google Maps is available in Ukrainian, but the Google Maps API does not allow to modify the algorithm or consider the priorities of the objects when constructing the route.

Via Michelin is an online routing service based on the maps of the well-known Michelin roadmap issuer [4]. The service is free and works fast, but again, the information about the algorithms is confidential. Via Michelin is not available in Ukrainian, builds only the shortest routes and does not always work exactly for cycling routes.

YourNavigation.org is a demo website for the YOURS project. The purpose of the project is to create a routing website based on OpenStreetMap (OSM) data using other open source applications. It uses an open source routing mechanism called Gosmore [5]. The service is not very fast, and its website mentions that Gosmore is not intended to generate routes longer than 200km. Planner gives the ability to choose between two types of route: the shortest or fastest, and one of the nine modes of transport, but it is only available in English.

OpenRouteService.org is another online service that uses OSM data [6]. As well as YourNavigation.org, this is a non-profit service. The service uses the A * algorithm and is slower on long routes than other analogues. OpenRouteService.org is available in Ukrainian, can configure the types of roads, the type of route (fastest or shortest), the type of vehicle and its type of fuel. However, as well as other similar services, it does not take into account the types of intermediate points of the route and does not allow them to be configured as a priority when constructing a route relative to the user's preferences.

The results of comparing the five routing services are shown in Table 1. Unfortunately for the best performing applications the routing algorithms were confidential. The two that did reveal their routing algorithm both used A*. It is obvious that non-commercial routing services tend to stick with less complex algorithms, to limit the time that needs to be invested. Commercial applications have a lot depending on the performance of their service and can afford to invest more time

and money in order to increase performance. It was found that most of the existing solutions do not have the Ukrainian language version and none of them supports the construction of tourist routes.

Selection of unexplored parts of the general problem. Nowadays, systems that can generate route based on the types of objects on the map are unexplored and not investigated. However, it would be useful to modify the route in such a way that it was, on the one hand, short, on the other hand, covered as many objects of the given type as possible.

Table 1

Comparison of existing services

Name	***Algorithm***	***Data source***	***Selecting start and destination***	***Supports Ukrainian language***	***Supports priorities of intermediat e points***
TomTom	*confidential*	Own data	Addresses	No	No
Google Maps	*confidential*	Own data	Both addresses and select on a map	Yes	No
Via Michelin	*confidential*	Michelin maps	Both addresses and select on a map	No	No
YourNavig ation.org	A*	OSM	Both addresses and select on a map	No	No
Open Route Service.org	A*	OSM	Both addresses and select on a map	Yes	No

Setting objectives. Thus, the actual task is to develop a system that when constructing a route will consider its length and the number of objects covered by the given type. Also, when calculating the route, it is necessary to reflect the limits of visibility of each of the points, because it would be enough that the path would be passed by objects, and not through them.

Presentation of the main material

1. The task of finding the shortest path

The problem of finding a route on a map can be represented in a simple form by using a graph, as in Figure 1. To find the shortest path from, e.g., A to H, one would need to find the collection of edges that connect A to H through any other nodes, for which the sum of the weights is as low as possible. Throughout this paper, the length of the shortest path from some node s to some node t will be referred to as the distance from s to t.

Consider the weighted graph $G = (\mathrm{V}, \mathrm{E})$, where V – is a finite set of nodes and E - is a set of edges between these nodes. Number of nodes $|V|$ we denote as n , and the number of edges $|E|$ - m. Each edge e has a weight $w\ (e)$. The path is determined by the sequence of nodes $(v_1, \dots, v_k)$, for which $(v_i, v_{i+1}) \in E$ and $1 < i < k$. A node connected to a specific node v with some edge is called a neighbor v.

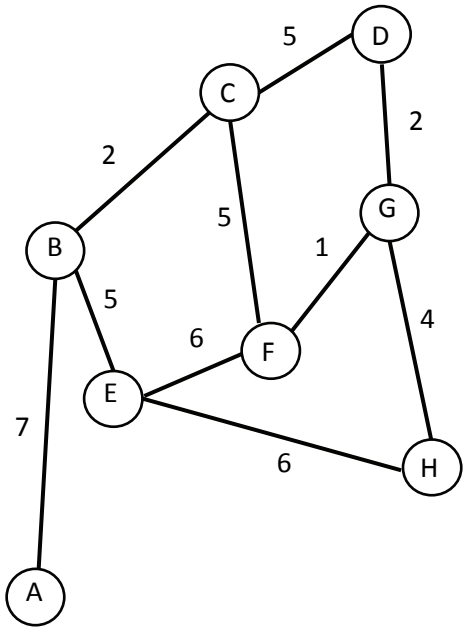


Fig. 1. Weighted graph

If the starting node is a vertex $s \in V$ and a finite node $t \in V$ then the shortest path is defined as the path $(s, \ldots, t)$, which has the minimum sum of the weights of all edges in the path. The length of the shortest path from s to node v is defined as $g\ (v)$ and is also called the distance from s to v.

2. Algorithm A*

When traveling to a certain destination, it usually does not make sense to look for a path in the opposite direction. Therefore, an algorithm that prefers the vertices in the direction to the destination appeared and first visits them, unlike the Dijkstra algorithm, which searches in all directions of space. The algorithm does not search in the direction of the target node. Hart, Nilsson and Raphael [7] introduce the A* algorithm, which adds heuristics to Dijkstra, making it more directional to the final vertex. Instead of the weight of node v, we use the estimate of the shortest path $\hat{g}(v)$ from the initial vertex to the finite, which runs through the node v. For this purpose, a function (1) is introduced that represents the shortest path from s to t passing through v, with $\hat{g}(v)$ being the distance from s to v, and $h(v)$ is the distance from v to t.

$$f(v) = g(v) + h(v) \tag{1}$$

Also, estimates of the values of functions were introduced:

$$\hat{f}(v) = \hat{g}(v) + \hat{h}(v) \tag{2}$$

The estimate of the distance from s to v, $\hat{g}(v)$ is determined in the same way as in the Dijkstra algorithm, it is the shortest path from s to v, found at the current iteration. Estimation of the distance from v to t, $\hat{h}(v)$ is determined heuristively. This function can be any function and often its definition is a separate task. Euclidean

distance from v to t is most often used as heuristic function for route planning. The heuristics should be admissible, that is, they do not have to overestimate the cost of the route; the estimate of the path should be in the range $[0; k]$ where k is the actual distance, and the motoonon, ie, for each vertex v and adjacent vertex $v`$ it must be inequality (3), where $c(v, v`)$ is the actual distance between v and $v`$:

$$h(v) \leq c(v, v`) + h(v`) \tag{3}$$

3. Calculation of the weight of the edge

In order to solve the problem, it is proposed to introduce the following definition of the function $\dot{w}(e)$:

$$\dot{w}(e) = k(e) * w(e), \tag{4}$$

where $k(e)$ – the function which defines the value of the coefficient k for each edge e.

The value of k should be based on types of points, which areas of visibility cover that edge e. Moreover $k(e) \in (0,1]$, accordingly $\dot{w}(e) \in (0, w(e)]$. Therefore, it will be possible to take into account the type of intermediate point when constructing the route and to achieve an increase in the probability of entering the priority object to the resulting route, because at identical distances, the vertex, the edge to which passes through the scope of the priority object will have a smaller value $\dot{w}(e)$. It means that when choosing a vertex v from the set X such that $\hat{g}(v) := \min_{u \in X} \hat{g}(u)$ [9], where $X :=$ set of nodes v for which the path from s to v is not defined, the vertex with the smallest $\dot{w}(e)$ will be selected.

For each edge e_i passing through the visibility of the vertex v, $\dot{w}(e_i)$ will be equal to the edge weight, multiplied by the coefficient $k(v)$: $\dot{w}(e_i) = w(e_i) * k(v)$, and for others edges $\dot{w}(e_i) = w(e_i)$, since for them $k = 1$. The calculated value $\dot{w}(e_i)$ can then be used in the algorithms for constructing the route instead of the distance between the vertices that connect the edge e_i, that is, its length.

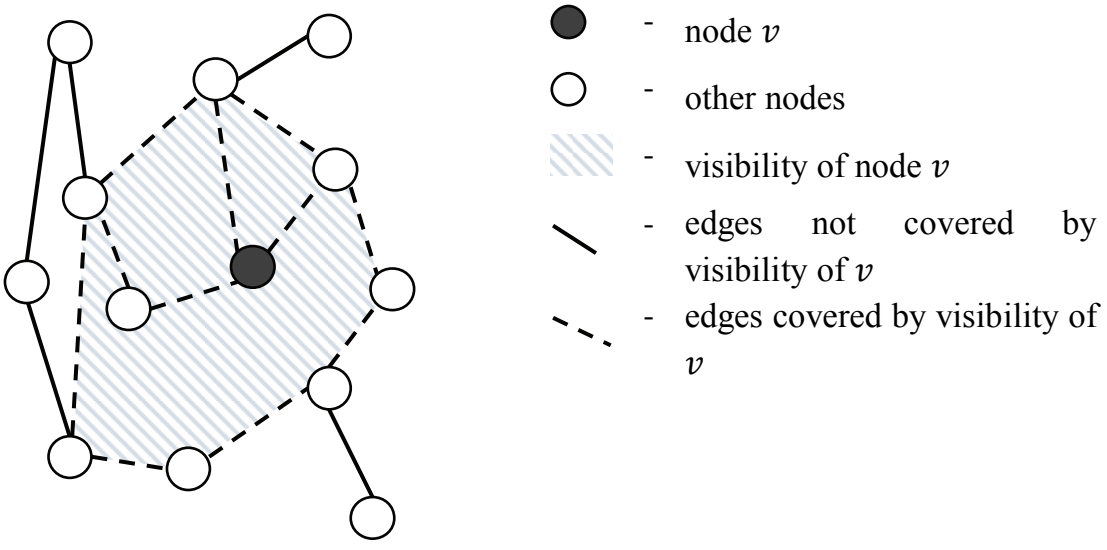


Fig. 2. Visibility of the node

4. Calculation of the object's visibility on the map

When calculating the route, it would be necessary to determine the scope of visibility of each of the priority objects, that is, the set of points from which this object will be visible on the locality. This is especially important in the context of a dense building of the city, because the neighboring buildings can overlap the view of the object.

As an input to the system it is better to use OSM maps in XML format. All map objects in OSM XML format are divided into 2 types: <node>, which consist only one point and are described by parameters such as latitude, longitude and id (node id), as well as ways (<way >), which can consist of 2 or more dots (up to 2000). The way may be open or closed. These types of objects are interconnected by relations (<relation>). Buildings, or rather, their boundaries are usually described by means of closed way. This means that it is possible to get the coordinates of the boundaries of buildings located within a radius from point x (lat, lon) from OSM XML maps.

By the method of ray tracing [8] it is possible to construct a polygon, which would determine the scope of the object. The point of the start of all rays may be the center of the object, but for the sake of greater accuracy, it is necessary to take two more distant points from each other that are at the boundary of the object, for which the polygon is calculated, because some buildings may be too long.

Figure 3 shows an example of constructing a polygon of an visibility area. In order to construct a polygon, the point x of the rays is selected (in the picture - the center of the object). A bounding circle is constructed with radius r, which is the maximum distance from which the object can be seen. After that n rays with start in x are throwed in all directions.

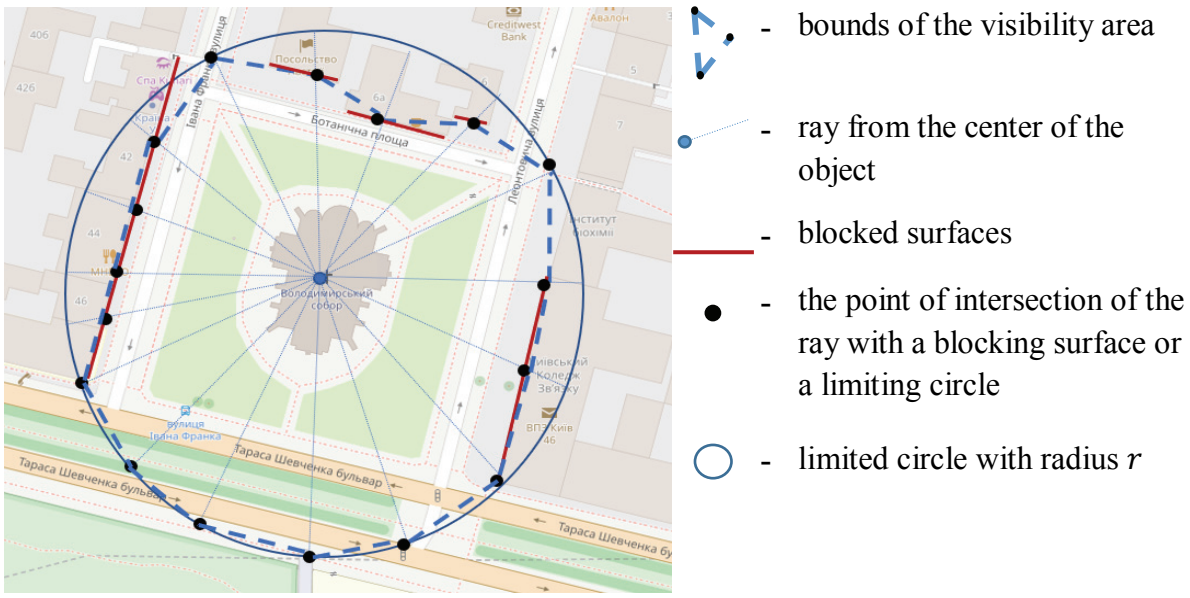


Fig. 3. Calculation of the object's visibility

Each of the rays intersects either the limiting circle or the blocking surface. The blocking surface mean the first boundary of the building, which occurs on the path of the ray. A plurality of intersection points forms a polygon, or the area of visibility.

5. Calculation of the coefficient *k*

Most objects in OSM XML maps have metadata other than coordinates and IDs. For example, the <amenity> tag may be one of the most useful for identifying an object type. This tag is used to designate public infrastructure objects: banks, schools, hospitals, cinemas, theaters, fountains, etc. Also, such tags as <*tourism*>, <*building*> with value "*architecture*" would be useful in systems oriented on tourist routes building. If it needed the route to go through parks and green areas, it be useful to find all objects marked with the <*leisure*> tag with the value of "*park*".

Once all the priority objects are found, it is necessary to determine the level of their importance, since the monuments are different, and have varying degrees of influence on the route. For example, for matching with each of the conditions you can count points: +3 for the tag <*historic*>, +4 for <*leisure*> with the value of "*park*", +2 for a link to Wikipedia with an article about this object, +1 for each additional tag <*name*>.

The sum of the points should be converted into a coefficient with a value within (0; 1] and matched to each visibility area with the value of the factor of the object around which this area was constructed.

Conclusions. This paper analyzes the existing routing solutions and identifies their disadvantages. Based on the obtained results, an algorithm for constructing the shortest route on the map is proposed. Using this algorithm to modify Dijkstra and A* algorithms, it's possible to create a system for constructing tourist routes.

References

1. Мельниченко С. «Інформаційні технології в туризмі: теоретичні та практичні аспекти» / Мельниченко С., Запоріжжя: Вісник Запорізького національного університету №2(6). – 2010. – С.129.

2. Marcin Wojnarski «TomTom Traffic Prediction for Intelligent GPS Navigation» / M. Wojnarski, IEEE International Conference on Data Mining Workshops – 2010 – C.20-21.

3. Gabriel Svennerberg "Google Maps API 3" / G. Svennerberg, Apress – 2010 – C.157-160.

4. ViaMichelin Navigation. User Manual [Electronic source] / ViaMichelin, URL: http://enav.download.viamichelin.com/nav/tel/manuels/gbr/User_Manual_GBR_VMN_New_Edition_v7.pdf (Accessed: 01.05.2019).

5. YourNavigation.org About [Electronic source] / YOURS, URL: https://wiki.openstreetmap.org/wiki/YOURS/ (Accessed: 01.05.2019).

6. Sebastian Schmitz «New Applications based on collaborative geodata – the case of Routing» / Sebastian Schmitz, Proceedings of XXVIII INCA international congress on collaborative mapping and space technology – 2008 – C.1-7.

7. Hart, Peter E., Nils J. Nilsson, and Bertram Raphael. «A formal basis for the heuristic determination of minimum cost paths» / Hart, Peter E., Nils J. Nilsson, and Bertram Raphael., IEEE transactions on Systems Science and Cybernetics 4.2 – 1968 – C.100-107.

8. Spencer G.H., M.V.R.K. Murty «General ray-tracing procedure» / Spencer G.H., M.V.R.K. Murty, JOSA 52.6 – 1962 – C.672-678.

9. Knuth, Donald E. «A generalization of Dijkstra's algorithm» / Knuth, Donald E, Information Processing Letters 6.1 – 1977 – C.1-5.

Autors

Oleksandr Podrubailo – research assistant, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: alex.podrubailo@gmail.com

Подрубайло Олександр Олександрович – асистент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Yevheniia Zubrych – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: evg.zubrich@gmail.com

Зубрич Євгенія Сергіївна – студентка, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Євгенія Зубрич,
Олександр Подрубайло

ПОБУДОВА НАЙКОРОТШОГО МАРШРУТУ НА КАРТІ З УРАХУВАННЯМ ОБЛАСТЕЙ ВИДИМОСТІ ПРОМІЖНИХ ПУНКТІВ

Актуальність теми дослідження. Сьогодні автоматизовані системи побудови та візуалізації маршрутів використовуються у багатьох галузях, а планувальники маршрутів стали невід'ємною частиною повсякденного життя. Особливо затребуваними стали в наші дні і продовжують активно розвиватися системи планування маршрутів, що працюють в режимі реального часу.

Постановка проблеми. Існуючі алгоритми здебільшого направлені на пошук найшвидшого та найкоротшого маршруту. Проте використовуючи системи, що генерують лише найкоротший маршрут турист ризикує не побачити частину пам'яток даної місцевості.

Аналіз останніх досліджень і розробок. З ростом попиту на навігаційні системи та планувальників маршрутів, збільшилась і кількість досліджень у цій галузі. Насамперед з'явилось чимало систем, що будують найкоротший шлях між двома точками. У даній роботі було проаналізовано існуючі рішення для географічної маршрутизації, виявлено такі їх недоліки як відсутність підтримки української мови та побудови туристичних маршрутів.

Виділення недосліджених частин загальної проблеми. На даний момент недослідженими та нереалізованими є системи які б при побудові найкоротшого маршруту враховували типи об'єктів, що є на мапі, та класифікували їх за тематикою.

Постановка завдання. Актуальною є розробка системи, що при побудові маршруту буде враховувати його довжину та кількість охоплених об'єктів заданого типу. Також при розрахунку маршруту необхідно врахувати межі видимості кожного з пунктів, адже достатньо, щоб шлях проходив повз об'єкти, а не крізь них.

Викладення основного матеріалу. У даній роботі будо запропоновано алгоритм побудови найкоротшого шляху, що базується на областях видимості проміжних пунктів. Визначено поняття області видимості, розроблено метод її розрахунку на основі координат об'єкта, відносно якого ця область будується та координат будівель, що знаходяться не далі, ніж деяка відстань r від центру об'єкта.

Висновки. Було проаналізовано існуючі рішення побудови найкоротших маршрутів та виявлено їх недоліки. На основі отриманих результатів було запропоновано алгоритм побудови найкоротших маршрутів. Використовуючи даний алгоритм для модифікації алгоритмів Дейкстри та A*, можливе створення системи побудови туристичних маршрутів.

Ключові слова: планування маршрутів, туристичні маршрути, області видимості об'єктів.

UDC 004.7

Heorhii Loutskii,
Andrii Dolgolenko, Oleksandr Dolgolenko

METHOD OF SIMPLIFICATION OF COMPUTATIONS WITH A FLOATING POINT IN THE SUPERSCALAR PROCESSOR

Георгій Луцький,
Андрій Долголенко, Олександр Долголенко

СПОСІБ СПРОЩЕННЯ ОБЧИСЛЕНЬ З ПЛАВАЮЧОЮ КРАПКОЮ В СУПЕРСКАЛЯРНОМУ ПРОЦЕСОРІ

This article describes results of development of the approach to building fast operational device of adding-subtracting a long sequence of floating-point numbers with dynamic branching of work at the level of RISCs, which without additional software complications, will ensure the law of associativity when performing addition of sequence of positive numbers. This paper describes the functional circuit of such operational device which does not require for its work elements of firmware control. The operational device can be implemented for SF and F formats of floating-point numbers. For other formats such implementation of the operational device is more reasonable to base on an algorithm similar to the Kahan 's algorithm.

Keywords: Operational Device, RISC Operations, Floating-Point, Law of Associativity, Kahan's Algorithm.

Fig.: 1. Tabl.: 1. Bibl.: 8.

У статті розглядається підхід до побудови швидкого операційного пристрою додавання-віднімання довгої послідовності чисел з плаваючою крапкою, що без додаткових програмних ускладнень забезпечить виконання закону асоціативності при складанні послідовності додатних чисел. Описана функціональна схема такого пристрою, котра не потребує для своєї роботи елементів мікропрограмного керування. Показано, що операційний пристрій за цією схемою може бути реалізованим для половинного та одинарного форматів представлення чисел з плаваючою крапкою. Для старших форматів представлення чисел з плаваючою крапкою реалізація подібного операційного пристрою виглядає більш розумною на основі алгоритму, подібного доалгоритму Кехена.

Ключові слова: операційний пристрій, скорочений набір операцій, плаваюча крапка, закон асоціативності, алгоритм Кехена.

Рис.: 1. Табл.: 1. Бібл.: 8.

Target setting. When constructing cores of most of the modern microprocessors with the x86-64 architecture, OoOE (Out-of-Order Execution) technology, based on the implementation of a Restricted Data Flow (RDF) [1,2], is used. Such microprocessors are

called superscalar microprocessors [3], or microprocessors with the CISC-RISC (CISC-outside RISC-inside) architecture, where: CISC - Complex Instruction Set Computing (full set of operations of x86-64 architecture), RISC - Reduced Instruction Set Computing (short set of operations implemented by a number of microprocessor operating devices). RISC is also called: uop, micro-ops, μops, or similar terms.

In the process of operating the cores of such microprocessors, a number of CISC commands, currently active in the flow of commands, are simultaneously decoded into a plurality of RISC operations. RISC implementation planning is performed according to the RDF architecture, based on the readiness to execute RISC operations operands. Prior to RISCs obtaining the ready to execute status, they are placed in the reserve station cells [4-5]. RISC, which became ready to execute, can be transmitted from cell reservation station to a free operating device that can execute it. Thus, in modern microprocessors, dynamic parallelism is organized at the level of RISCs.

When forming CISC flows – commands that operate with floating point operands, both programmers and developers of optimizing compilers must take into account features of specific implementations of arithmetic with floating-point [6]. As a consequence of these features, for example, floating-point arithmetic does not perform standard mathematical laws such as commutative and associative [7] and it is possible to do so that the calculated answers almost entirely consist of "noise" [7].

For example, multiplication and division operations do not greatly increase the relative error, but subtracting almost equal quantities can significantly increase it. One of the consequences of the possible unreliability of the addition operation sequence of floating-point numbers is a violation of the law of associativity: $(u + v) + w \neq u + (v + w)$ for some u, v, w. The distributive law that binds the operations × and +: $u \times (v + w) \neq (u \times v) + (u \times w)$ may also be violated. Performing addition and subtraction operations sequence numbers even with fixed-point is known as [7] left-associative. This means that operations in such an arithmetic expression must be strictly executed from left to right. Even guidelines for programmers are developed, that contain recommendations for the organization of computing with fewer errors [7]. For example, if you want to add a long sequence of positive numbers [7], you should first sort them out and perform operations starting with the smallest numbers.

Analysis of these guidelines shows that it is often difficult to carry out such rules for the programmers, for example, because unknown values of variables, because the need for pre-sorting of numbers by size, etc. The implementation of such guidelines by the compiler at the stages of preparation for computing is also difficult for the same reasons. In addition, the implementation of these rules, such as the need to change the order of filing operands when performing addition sequence of positive numbers with floating-point, creates the complexity for organizing parallel calculations.

To reduce the error of adding-subtracting a long sequence of floating-point numbers, Kahan 's algorithm, which is also known as compensatory summation, is used [8]. Reducing the error is achieved by introducing an additional variable to store a

growing amount of error. With compensating summation, the worst error does not depend on the number of operands, so a large number of operand values can be combined with an error that depends only on the accuracy of the floating-point representation format. But according to this algorithm, each operation of addition-subtraction is transformed into 4 operations of addition-subtraction type and 4 assignment operations.

The research objective. The purpose of this project is to develop the approach to building fast operational device of adding-subtracting a long sequence of floating-point numbers for dynamic branching of work at the level of RISCs, which without additional software complications, will ensure the law of associativity when performing addition sequence of positive numbers.

The approach to creation of the operational device. Ensure the implementation of the associativity rule for the addition operation over a sequence of positive of floating-point numbers without additional software complications is possible if you perform computation of all intermediate results without losing significant bits.

Consider the possibility of constructing an operational device (OD) for adding-subtracting a sequence of floating-point numbers, which performs the operation $o = o \pm x$ at each step of the calculation, where: x is the next operand of the sequence that can be taken at each step of the calculation on OD inputs for processing; o - an intermediate result of addition-subtraction of a sequence of numbers. To increase the accuracy of the o in the OD will be calculated with an accuracy that is limited only by the range of order change and the accuracy of the representation of the mantissa of the processed format [6]. At the beginning of computing a new sequence of numbers o will be zeroed. Simultaneously with the calculation of the new value of o, its previous value will be converted to the processed numeric format with the rounding to the nearest [6] and output to the OD outputs. Suppose that at the OD input, according to standard [6], each floating-point number x has the form: $x = f_x \cdot 2^{ex}$, where: f_x is a n-bit normalized ($1 \le |f_x| < 2$, with $x \neq 0$) fractional part of the number x (mantissa); e_x - number order (unsigned integer from interval $[e_{max}, 0]$); f_x and e_x are represented by a direct binary code. The floating-point number has two characters: a sign number ***(sign)***, displayed in a separate bit; the order sign, is displayed by the ***bias*** of the order [6].

The schematic design of OD addition-subtraction operations over a sequence of floating-point numbers with increased accuracy of execution is shown in Figure 1. His work is as follows.

Filing of the operands is one at a ***clock*** (see fig.1) of work of the OD. So, on the i-th clock of the work on the inputs of the OD, a regular operand (number x) is given. In this case, the control node 1 is its transformation from a processed floating-point number format [6] to the internal range of processing numbers r, where $r = e_{max}+1+n$ and is the number of binary digits in the representation of a fixed-point number. Namely, the input f_x is given by a normalized mantissa of the operand, the input e_x is the order of the operand, the input $sign_x$ - sign of the operand. With the arrival of the front edge of the ***clock*** pulse on the input ***clock*** data are recorded, respectively, in the

registers of the mantissa ***RG*** f_x, the order ***RG*** e_x and the trigger control ***Tg*** $sign_x$. At the same time, if the number x is the initial operand of the new sequence of numbers, the value of ***RG*** f_o is reset using the ***reset*** signal.

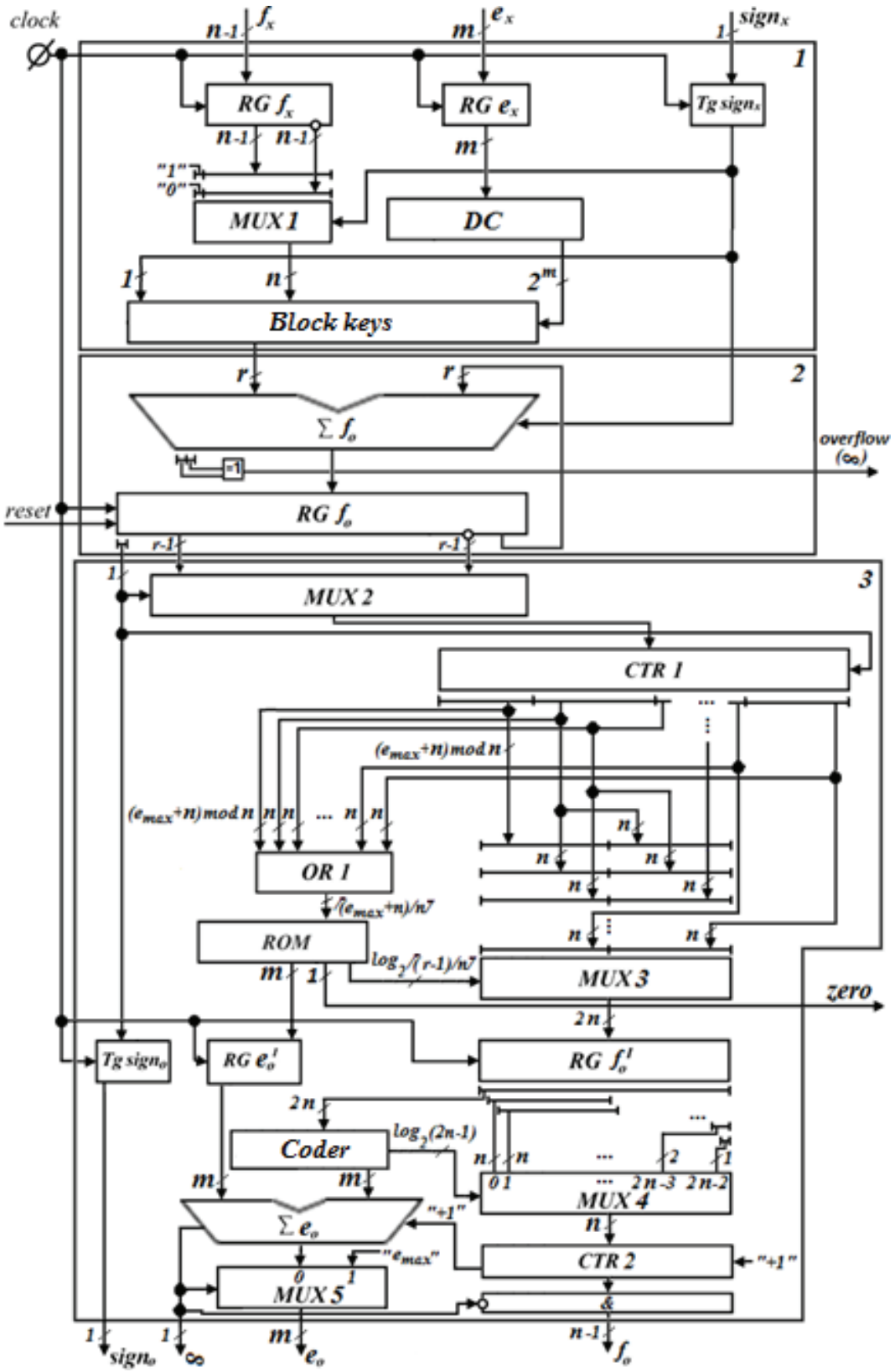


Fig. 1. Functional operating device (OD) addition-subtraction operations scheme

With the ***MUX 1*** multiplexer and depending on the $sign_x$ value, the transfer of ***RG*** f_x from the mantissa f_x code to the ***Block keys*** input with the recovery of the hidden bit ***MSB*** [6] and the $sign_x$ is carried out. Depending on the value of the order e_x, on one of the outputs of the ***DC*** decoder, a signal is generated that provides the input of the ***Block keys*** to the first input of the adder $\sum f_o$. As a result of this transfer, the arithmetic left shift of the inverse code f_x is carried out with ***MSB*** and $sign_x$ on e_x bits, and all other lower bits of the first input of the adder $\sum f_o$ are filled with the value of the $sign_x$, which, together with the submission of the $sign_x$ also to the input of the junior carry input adder $\sum f_o$ provides formation supplementary code x, brought to the r range.

On the $(i+1)$-th step of the OD's work on its inputs, a regular operand of a computable sequence can be taken. At the same time, in the register ***RG*** f_o in the node 2 summation from the outputs $\sum f_o$, the result of the summation of the previous operand will be recorded, and in the trigger ***Tg*** $sign_o$ and registers: ***RG*** e_o^I and ***RG*** f_o^I in the node 3, the result will be written, respectively: the sign, order and mantissa are partially normalized the previous value of o (from the summation of the operand that could be taken at the inputs of the OD on the $(i-1)$-th step of his work).

Thus, OD represents a conveyor information converter consisting of three segments. At each step of the OD's work on its inputs, a regular value of the number x of the executed operation $o = o + x$, in the general case of different microprocessor cores (from the core that captured the OD clock cycle) can be taken. If any of the cores need to perform the operation $o = o - x$, the $sign_x$ value applied to the OD inputs must be changed to the opposite.

Simultaneously with the summation on $\sum f_o$ of the number x with the accumulated amount f_o, which flows from the register of ***RG*** f_o to the second input $\sum f_o$, an ***overflow*** signal is generated (as the sum of modulo 2 of the two highest bits $\sum f_o$), which indicates about the output of a new value of f_o from the range r. In the node 3 forming result transfers the previous accumulated amount of f_o to an intermediate representation in a floating-point form with a $2n$-digit binary code of the mantissa of the f_o^I result in the forward code and the m-bit value of the order corresponding to the f_o^I mantissa. To do this, first, the f_o value from ***RG*** f_o, using the ***MUX 2*** multiplexer and the ***CTR 1*** counter, is translated into a straight-line code, and from its representation the $sign_o$ is deleted, which is written to the ***Tg*** $sign_o$ at the next cycle of the work of the OD. Then with the help of a group of ***OR1*** formed address entry to permanent memory ***ROM*** (actually the older nonzero group of binary digits is found in f_o). The ***OR1*** group consisting of $\lceil (e_{max}+n)/n \rceil$ n-inputs elements ***"OR"*** (the senior element in the group has $(e_{max}+n) mod n$ inputs. At this address, from the first ***ROM*** outputs, the value of the order e_o^I is read, which corresponds to the finding of the higher significant digit f_o on the lower-rank position of the older non-zero binary digits in f_o. This read-out value of the order e_o^I is written to ***RG*** e_o^I at the next time the OD works. From the second output ***ROM*** the zero address reads a ***zero*** signal, which indicates that $f_o = 0$. In this case, from the first ***ROM*** outputs, the zero value of the

order of e_o^I is read [6]. From the third ***ROM*** outputs at the address formed in ***OR 1***, the control information is read out to the ***MUX 3*** multiplexer. This information provides the passage from the outputs ***CTR 1*** through ***MUX 3*** to the ***RG*** f_o^I inputs only of the highest non-zero binary digits group f_o and the discharges group that is next after by her. At the next cycle OU work, these two groups of digits are written to ***RG*** f_o^I.

In the next work of the OD normalizing the mantissa of the result f_o is achieved by removing the hidden bit, its rounding to the nearest, and making the corresponding correction to the value e_o. To do this, using the encoder, depending on the number of zeros to the first significant bit in f_o, the control information is generated by the ***MUX 4*** multiplexer. This information provides the passage from the outputs of ***RG*** f_o^I through ***MUX 4*** to the inputs of the ***CTR 2*** counter ***n*** of the highest significant bits f_o, shifted left to ***k*** bits, where ***k*** is the number of zeros to the first significant bit in f_o. In this case, the first significant bit in f_o on the inputs of the ***CTR 2*** counter is not transmitted, which provides for the removal of the hidden bit. At the same time, at other encoder outputs, generates correction an order of e_o equal to $(e_{max}+n) mod n-k$ is formed, if the first significant bit in the direct code f_o was found in the older group of digits, or $n-k$ in all other groups of digits. With the help of the $\sum e_o$ adder, this correction is added to the value e_o^I that arrives at the other inputs of the adder from ***RG*** e_o^I and is summed up with the value of the carry input entry of the adder coming from the counter ***CTR 2*** and indicates the ***overflow*** of the mantissa f_o as a result of its rounding to the nearest [6]. The result of this so that the value of the mantissa f_o, as well as without the ***CTR 2*** overflow signal, will be normalized, rounded and deleted with a hidden bit and is output from the outputs of ***CTR 2*** through a group of ***(n-1)*** 2-inputs elements ***"AND"*** to output f_o OD. Thus, from the outputs $\sum e_o$ through the input ***0*** multiplexer ***MUX 5*** output e_o OD issued value of the order of the intermediate result. When the ***carry*** signal from most significant bit $\sum e_o$ appearing, indicating the overflow e_o i.e. $e_o > e_{max}$, according to [6], the output of f_o OD gives a zero value, and the output e_o OD, through the input 1 of the multiplexer ***MUX 5***, gives the value of e_{max}.

Table 1 summarizes the values of the bits of the main OD blocks for different formats of representation of floating-point numbers, according to [6].

Conclusions. The paper describes the possibility of constructing an operational device for addition-subtraction of a sequence of floating-point numbers with an accuracy that is limited only to the range of order change and the accuracy of representation of the mantissa of the processed format.

This approach to constructing the operational device of addition-subtraction of a sequence of floating-point numbers looks very promising due to the simplification of the computational process from the programmer's point of view, because it ensures implementation of the associativity law when performing addition of sequences of positive numbers, without additional complications required on the software level.

The paper describes the functional circuits of such an operational device, which does not require elements of firmware control for its work. As it can be seen from the

table 1, the operational device, under the current technological level of components development, can be implemented for SF and F formats of floating-point numbers [6]. For other formats such implementation of operational device is more reasonable to base on an algorithm similar to the Kahan 's algorithm.

Table 1

The values of the bits of the main OD blocks for different formats of representation of floating-point numbers

Main OD blocks	***SF***	***F***	***DF***	***DEF***	***QF***
RG f_x	10	23	52	64	112
RG e_x, RG e_o^I, $\sum e_o$, MUX 5	5	8	11	15	15
MUX 1, MUX 4, CTR 2	11	24	53	65	113
RG f_o , $\sum f_o$	43	280	2101	32833	32881
MUX 2, CTR 1	42	279	2100	32832	32880
OR 1	4	12	40	506	291
ROM	16 x 8	4K x 13	T x 18	2^{506} x 25	2^{291} x 25
MUX 3, RG f_o^I	22	48	106	130	226

References

1. Y. Patt, W. Hwu, et al, Experiments with HPS, a Restricted Data Flow Micro architecture for High Performance Computers, Digest of Papers, COMPCON 86, (March 1986), pp. 254-258.

2. M. Simone, A. Essen, A. Ike, A. Krishnamoorthy, T. Maruyama, N. Patkar, M. Ramaswami, M. Shebanow, V. Thirumalaiswamy, D. Tovey (1995). Implementation trade-offs in using a restricted data flow architecture in a high performance RISC microprocessor. New York. pp. 151-162.

3. John L. Hennessy, David A. Patterson. Computer Architecture. A Cuantitative Approach, USA, Morgan Kaufmann, 2012 – 497 p.+ add-ins.

4. Kanter, David (November 13, 2012). "Intel's Haswell CPU Microarchitecture" (http://www.realworldtech.com/ haswell-cpu/).

5. J.Shen, M. Lipasti. Modern Processor Design: Fundamentals of Superscalar Processors. Waveland Press, 2013- 642 p.

6. IEEE 754: Standard for Binary Floating-Point Arithmetic / 3 april 2014. – URL: http://grouper.ieee.org/groups/754/.

7. What Every Computer Scientist Should Know About Floating-Point Arithmetic. – URL: https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/ 02Numerics/Double/paper.pdf.

8. Higham, Nicholas J. Accuracy and Stability of Numerical Algorithms. SIAM, 2002, pp. 110–123.

Autors

Heorhii Loutskii – professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: georgijluckij80@gmail.com

Луцький Георгій Михайлович – професор, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Andrii Dolgolenko – PhD student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: andrew@ncube.co.uk

Долголенко Андрій Олександрович – аспірант, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Oleksandr Dolgolenko – associate professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: aleks.dolgolenko@gmail.com

Долголенко Олександр Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Георгій Луцький,
Андрій Долголенко, Олександр Долголенко

СПОСІБ СПРОЩЕННЯ ОБЧИСЛЕНЬ З ПЛАВАЮЧОЮ КРАПКОЮ В СУПЕРСКАЛЯРНОМУ ПРОЦЕСОРІ

Актуальність теми дослідження. При побудові ядер більшості сучасних мікропроцесорів з архітектурою *x86-64* використовується технологія *OoOE (Out-of-Order Execution)*, що заснована на реалізації обмеженої архітектури потоку даних (*Restricted Data Flow (RDF*)). Такі мікропроцесори отримали назву суперскалярних мікропроцесорів, або мікропроцесорів з архітектурою *CISC-RISC («CISC-outside RISC-inside»),* де: *CISC - Complex Instruction Set Computing*, *RISC - Reduced instruction set computing* (скорочений набір команд, що реалізується множиною операційних пристроїв ядра мікропроцесора).

Постановка проблеми. При формуванні потоків *CISC* – команд, що оперують операндами з плаваючою крапкою, як програмістам так і розробникам оптимізуючих компіляторів доводиться враховувати особливості реалізації арифметики з плаваючою крапкою. В наслідок цих особливостей, наприклад, для арифметики з плаваючою крапкою не виконуються стандартні математичні закони, такі як комутативний та асоціативний і неважко так невдало провести обчислення, щоб їх відповіді майже цілком складалися із "шуму".

Аналіз останніх досліджень і публікацій. Протягом останніх років, побутова ядер суперскалярних мікропроцесорів ґрунтується на тому, що деяка кількість *CISC* – команд, активного в даний момент потоку команд, одночасно декодується на множину *RISC*– операцій. Планування виконання *RISC* – операцій здійснюється відповідно до архітектури *RDF*, на підставі готовності до виконання операндів *RISC* – операцій. До набування *RISC* – операціями стану готовності до виконання, вони розміщуються в комірках станції резервування. *RIS*C – операція, котра набула стану готовності, може бути переданою з комірки станції резервування в вільний операційний пристрій, що може її виконати. Таким чином в ядрах сучасних мікропроцесорах організовується динамічний паралелізм на рівні *RISC* – операцій (їх також називають: *uop, micro-ops, μops*, або подібними термінами).

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена вивченню та аналізу підходу до побудови більш точного операційного пристрою суматора-віднімача послідовності чисел з плаваючою крапкою.

Постановка завдання. Завданням є розробка швидкодіючого операційного пристрою суматора-віднімача з плаваючою крапкою який може бути використаними для динамічного розгалуження роботи суперскалярного ядра на рівні *RISC* – операцій і при цьому, без додаткових програмних ускладнень, забезпечить виконання закону комутативності для довгої послідовності додатних чисел.

Викладення основного матеріалу. Розглянуто підхід до побудови швидкого операційного пристрою додавання-віднімання довгої послідовності чисел з плаваючою крапкою, що без додаткових програмних ускладнень забезпечить виконання закону асоціативності при складанні послідовності додатних чисел. Описана функціональна схема такого пристрою, котра не потребує для своєї роботи елементів мікропрограмного керування.

Висновки. Операційний пристрій за розглянутою схемою може бути реалізованим для половинного та одинарного форматів представлення чисел з плаваючою крапкою. Для старших форматів представлення чисел з плаваючою крапкою реалізація подібного операційного пристрою виглядає більш доцільною на основі використання алгоритму, подібного до алгоритму Кехена.

Ключові слова: операційний пристрій, скорочений набір операцій, плаваюча крапка, закон асоціативності, алгоритм Кехена.

UDC 004.315

Valerii Zhabin, Valentina Zhabina

EFFICIENCY IMPROVEMENT OF DIVISION OPERATION REALIZATION IN ON-LINE MODE

There is considered a method of numbers division, which enables to overlap realization of digit-by-digit input of operands, their processing and digit-by-digit output of the result starting with high-order digits. It allows reduction of the necessary FPGA hardware resource during realization of parallel systems with direct connection between calculation units by means of data transfer in serial code. There is provided partial overlapping of calculations that gives a potential opportunity to accelerate the computation process for realization of the series of operations connected by data. The method enables to add functionalities to the division operation in redundant symmetrical number system.

Key words: division operation, on-line computations, redundant number system, dependent operations overlap.

Tabl.: 1. Bibl.: 8.

Relevance of the research topic. For parallel systems the time of task solution depends not only on parallel algorithm branches processing time but also on time expenditures for information interchange between calculation units (CU). The use of systems with direct connection by means of data between CUs (DCS) permits to reduce the information interchange time [1-4]. Data transfer time reduction, as well as hardware implementation simplification is a relevant objective that requires additional researches, including development of efficient arithmetic operation realization methods.

Problem statement. When building systems based on FPGA, the efficient use of the microcircuit resource (internal functional blocks and external pins) is an important problem [5, 6]. By the example of the division operation there is considered an opportunity to solve the above-mentioned problem reducing number of connections between system units by means of data transfer in serial code.

Analysis of resent researches and publications. In parallel arithmetic the division methods require availability of all digits of numerator and denominator before calculation starts. It does not allow overlapping the processes of digit-by-digit input of operands and the process of result digits calculation, so realization of computations in on-line mode is impossible. To overlap the above-mentioned processes, division methods of quasi-parallel arithmetic were developed [7, 8]. In this case the calculations are realized starting with high-order digits and redundant number systems are used to avoid carries to higher number positions. Nevertheless, the well-known methods place restrictions to the form of operands representation, namely to their range and sign.

Singling out of unexplored parts of the general problem. Methods of division in on-line mode require further research. Adding functionalities to methods including possibility to process numbers with any sign, as well as to extend the range of operands representation is an important task to be solved to increase the data processing efficiency.

Task definition. The purpose of the research is to increase the efficiency of division operation realization in on-line mode by adding functionalities to the methods.

Statement of the main material. For division of numbers we will use a redundant symmetrical number system with the base $k = 2$ and the digits {-1,0,1}. In DCS the units are connected directly according to data flow graph of the task. It enables to interchange data between CUs without main memory use and, thus, to accelerate computations.

At the time of execution of series of operations connected by data the CUs work in the following way. At the each step CU accepts one operand digit from the previous CU and forms one digit of the result for the next CU. The operation is realized starting with the high-order digits. The formed result digit does not need any correction as in the redundant number system there are no carries to result high-order digits.

In the on-line mode the result digits are formed with the delay of p cycles, i.e. for division the CU executes binary operation $Z = 2^{-p} \cdot X / Y$, in which the operands are n-digit fractional numbers and the result has $m = n + p$ digits:

$$X = \sum_{i=1}^{n} x_i 2^{-i},\ Y = \sum_{i=1}^{n} y_i 2^{-i},\ Z = \sum_{i=1}^{m} z_i 2^{-i}. \quad (1)$$

Let's suggest that $0 \le |X| < |Y|$ та $2^{-r} \le |Y| < 1$ (r is an integer number, not greater than n). It is obvious, that in this case we get $|Z^*| < 1$ and $|Z| < 2^{-p}$. With X_i, Y_i, Z_i we define numbers that have only i high-order digits.

To get n digits of the function $Z^* = X / Y$, it is necessary to form $m=n+p$ digits of the function $Z = 2^{-p} X / Y$.

The condition of the symmetrical chopping of the function Z in i-th cycle of calculation is the following

$$Z_i - 2^{-i-1} \le 2^{-p} X_i / Y_i < Z_i + 2^{-i-1}. \quad (2)$$

If in each cycle the expression (2) is fulfilled, then after m cycles we will get a function with the measure of inaccuracy that is not greater than 2^{-m-1} in absolute value. In fact, for $i = m$ the expression (2) is the condition of chopping $Z = 2^{-p} X_m / Y_m$ to m digits. During calculation process in i-th cycle the measure of inaccuracy of the result will not be greater than the half of the weight of the digit that is formed.

Let's define the following designation

$$R_i = (2^{-p} X_i - Z_i Y_i) 2^i . \quad (3)$$

At that, the condition (2) can be defined as follows

$$-2^{-1} Y_i \le R_i < 2^{-1} Y_i . \quad (4)$$

Let's suggest that in $(i-1)$- th cycle the condition (4) is fulfilled, i.e.

$$-2^{-1} Y_{i-1} \le R_{i-1} < 2^{-1} Y_{i-1} . \quad (5)$$

We should find the minimal delay p in forming function Z^* digits, for which the condition (2) will be fulfilled in each cycle.

Using (1), (3) and (5) we get

$$R_i = 2R_{i-1} + 2^{-p} x_i - Z_{i-1} y_i - z_i Y_i . \quad (6)$$

Designating

$$H_i = 2R_{i-1} + 2^{-p} x_i - Z_{i-1} y_i , \quad (7)$$

we get

$$R_i = H_i - z_i Y_i . \quad (8)$$

In this case, taking into account (8) the expression (4) can be represented as follows

$$(z_i - 2^{-1}) Y_i \le H_i < (z_i + 2^{-1}) . \quad (9)$$

Using the formula (9), we define the rule for selection of quotient digit from the set {-1, 0, 1} in i -th cycle:

$$z_i = \begin{cases} -1, \ if \ -\frac{3}{2} Y_i \le H_i < -\frac{1}{2} Y_i ; \\ 0, \ if \ -\frac{1}{2} Y_i \le H_i < \frac{1}{2} Y_i ; \\ 1, \ if \ \frac{1}{2} Y_i \le H_i < \frac{3}{2} Y_i . \end{cases} \quad (10)$$

In order to get the correct result, it is necessary to keep H_i values within the interval $-\frac{3}{2} Y_i \le H_i < \frac{3}{2} Y$. Selecting p values it is possible to provide the fulfillment of the following expressions:

$$H_{i\max} < \frac{3}{2} Y_i , \ H_{i\min} \ge -\frac{3}{2} Y_i . \quad (11)$$

Taking into account the formulas (4), (6) and the range of numbers representation we can prove that the expressions (11) are correct when the minimal integer value of the delay is the following: $p = 3$.

Algorithm of i-th cycle execution (where $i = 1,..., m$) is boiled down to the following calculations.

Using the formula (7), we find the intermediate value H_i ($H_0 = 0$), then according to the rule (11) the digit of the result z_i is formed. For the next cycle, the value of R_i is calculated by the formula (8).

The process of division of 5-digit fractional operands $X = 0.101\bar{1}0$ and $Y = 0.1101\bar{1}$ is shown in the table 1.

Table 1

Values of variables during process of number division

i	x_i	y_i	Y_i	H_i	*Fulfilled condition*	z_i	Z_i	R_i
1	1	1	0.1	0.0010000	-1/2 $Y_1 < H_1 <$ 1/2 Y_1	0	0.0	0.0010000
2	0	1	0.11	0.0100000	-1/2 $Y_2 < H_2 <$ 1/2 Y_2	0	0.00	0,0100000
3	1	0	0.110	0.1010000	1/2 $Y_3 < H_3 <$ 3/2 Y_3	1	0.001	-0,0010000
4	-1	1	0.1101	-0.1000000	-3/2 $Y_4 < H_4 <$ -1/2 Y_4	-1	0.0001	0.0101000
5	0	-1	0.11001	0.1011000	1/2 $Y_5 < H_5 <$ 3/2 Y_5	1	0.00011	-0.0001100
6	-	-	0.11001	-0.0011000	-1/2 $Y_5 < H_6 <$ 1/2 Y_6	0	0.000110	-0.0011000
7	-	-	0.11001	-0.0110000	-1/2 $Y_7 < H_7 <$ 1/2 Y_7	0	0.0001100	-0.0110000
8	-	-	0.11001	-0.1100000	-3/2 $Y_8 < H_8 <$ -1/2 Y_8	-1	0.00010111	0.0000100

After $m = n + p = 5 + 3 = 8$ cycles are completed, the following results represented in the canonical binary number system are formed:

$$Z = 2^{-3} X / Y = 0.00010111 \; ; \; Z^* = X / Y = 0.10111 .$$

The measure of inaccuracy of each function calculation is not greater than the half of the weight of the low-order digit of the result.

Conclusions. There was proposed an algorithm for number division in on-line mode that stipulates the use of redundant symmetrical number system with the digits {-1, 0, 1} and the base $k = 2$.

The increase of division operation realization effectiveness in comparison with the well-known methods was achieved due to adding functionalities including the capability to process numbers with different signs, as well as to extend the range of number representation.

The operation realization is digit-by-digit, starts with high-order digits and allows overlapping of input, processing and output of digits. Also, it allows overlapping of connected operations execution in different units, i.e. execution of series of connected operations in concurrency mode at the level of digits processing, and enables to accelerate computations.

The on-line calculations enable to transfer data between system calculation units in the serial code. It reduces the scope of necessary FPGA functionalities and commutation resources (functional cells and input-output cells). Moreover, in this mode the energy consumption is also reduced and the system reliability is increased.

References

1. Zhabin V.I. Design of High-Speed Specialized Computers for the Realization of Many-Place Expressions / V. I. Zhabin, V. I. Korneichuk, V. P. Tarasenko // Automatic Control and Computer Sciences. – 1981. – vol. 15, no 6. – P. 15-18.

2. Zhabin V. I. Computation of Rational Functions for Many Arguments / V.I.Zhabin, V.I.Korneichuk, V.P.Tarasenko // Automation and Remote Control. – 1978. – vol. 38, no 12. – P. 1864-1871.

3. Дичка И. А. Совмещение зависимых операций на уровне обработки разрядов операндов / И.А.Дичка, В.В.Жабина // Искусственный интеллект. – 2008. – №3. – С. 649-654.

4. Жабин В. И. Выполнение последовательностей зависимых операций в режиме совмещения / В. И. Жабин // Вісник Національного технічного університету України “КПІ”. „Інформатика, управління та обчислювальна техніка”. – 2007. – №46. – С. 226-233

5. Максфилд К. Проектирование на ПЛИС. Архитектура, средства и методы / К. Максфилд. – М.: Издательский дом «Додэка-XXI», 2007. – 408 с.

6. Жабин В.И. Реализация неавтономных вычислений в избыточных системах счисления на ПЛИС / В.И. Жабин, В.В. Жабіна, А.В. Скоріченко // Вісник НТУУ "КПІ". Інформатика, управління та обчислювальна техніка. – 2016. – №64. – С. 150-155.

7. Zhabin V.I. Division Method for Numbers in Redundant Representation / V.I.Zhabin, V.I.Korneichuk, V.P.Tarasenko // Control and Computer Sciences. – 1978. – vol. 38, no 12. – P. 1 864-1871.

8. Жабин В.И. Реализация операции деления при поразрядном вводе и выводе информации / В.И.Жабин // Проблеми інформатизації та управління: Зб. наук. праць. – К.: НАУ, 2007. – Вип. 2 (20). – С. 65-71.

Information about authors

Valerii Zhabin – professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: viz.kpi@gmail.com

Valentina Zhabina – assistant professor, Department of Computer Systems Software, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: val.zhabina2@gmail.com

РОЗШИРЕНА АНОТАЦІЯ

Валерій Жабін, Валентина Жабіна

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ РЕАЛІЗАЦІЇ ОПЕРАЦІЇ ДІЛЕННЯ В НЕАВТОНОМНОМУ РЕЖИМІ

Актуальність теми дослідження. Зростаюча складність задач обробки інформації в реальному часі обумовлює застосування паралельних обчислювальних систем. Час рішення задач в таких системах залежить не тільки від часу виконання паралельних гілок алгоритмів, але і від витрат часу на обмін інформацією між обчислювальними модулями. Зменшити витрати часу на обмін даними дозволяє використання паралельних систем з безпосередніми зв'язками між модулями системи за рахунок пересилання даних послідовним кодом. Підвищення ефективності таких систем є актуальним завданням, що вимагає додаткових досліджень, в тому числі розробки ефективних методів виконання арифметичних операцій.

Постановка проблеми. При побудові систем на базі ПЛІС важливою проблемою є економне використання ресурсу мікросхем (внутрішніх функціональних блоків і зовнішніх виводів) з метою зменшення числа мікросхем для реалізації обчислювальної системи. На прикладі операції ділення у роботі розглядається можливість вирішення зазначеної проблеми за рахунок зменшення кількості зв'язків між вузлами системи шляхом пересилання даних послідовним кодом.

Аналіз останніх досліджень і публікацій. Методи ділення в паралельній арифметиці потребують наявності всіх розрядів діленого та дільника перед початком обчислень. Це не дозволяє суміщувати процеси порозрядного введення операндів та формування розрядів результату. Для суміщення вказаних процесів розроблені методи ділення квазіпаралельної арифметики. Обчислення в цьому випадку виконуються в неавтономному (on-line) режимі зі старших розрядів операндів, а для вилучення переносів в старші розряди використовують надлишкові системи числення. Операційні модулі в системі з'єднуються відповідно потоковому графу задачі. При виконанні ланцюжків залежних за даними операцій одержаний в i-му циклі розряд результату в одному модулі може в $(i+1)$-му циклі в іншому модулі оброблятися як черговий розряд операнду. Це забезпечує часткове суміщення операцій на рівні обробки розрядів операндів.

Виділення недосліджених частин загальної проблеми. Методи ділення в неавтономному режимі потребують подальшого дослідження. Для підвищення ефективності обробки даних в системах реального часу важливим завданням є підвищення функціональних можливостей методів, в тому числі, забезпечення обробки чисел з довільними знаками та розширення діапазону подання операндів.

Постановка завдання. Метою роботи є підвищення ефективності виконання операції ділення чисел в неавтономному режимі шляхом розширення вказаних вище функціональних можливостей методів.

Викладення основного матеріалу. Запропоновано алгоритм ділення чисел в неавтономному режимі з використанням надлишкової симетричної системи числення з цифрами {-1, 0, 1} та основою $k = 2$. Метод дозволяє суміщення процесів порозрядного введення операндів, їх обробки та порозрядного виведення результату. Це дає можливість реалізувати паралельну обробку на розрядному рівні залежних за даними операцій, що дозволяє прискорити обчислення. Показано приклад ділення чисел.

Висновки. Досягнуто підвищення ефективності виконання операції ділення відносно відомих методів за рахунок розширення функціональних можливостей, в тому числі, забезпечення обробки чисел з довільними знаками та збільшення діапазону подання операндів.

Завдяки пересиланню даних між вузлами системи послідовним кодом зменшуються необхідні функціональні та комутаційні ресурси ПЛІС (функціональні комірки та комірки введення-виведення). В свою чергу, зменшується енергоспоживання та підвищується надійність систем.

Розглядається метод ділення чисел в неавтономному режимі, що дозволяє суміщення процесів порозрядного введення операндів, їх обробки та порозрядного виведення результату, починаючи зі старших розрядів. Це дає можливість зменшити необхідний апаратний ресурс ПЛІС при реалізації паралельних систем з безпосередніми зв'язками між обчислювальним модулями за рахунок пересилання даних послідовним кодом. При виконанні послідовності залежних за даними операцій забезпечується часткове суміщення виконання операцій, що дає потенційну можливість для прискорення обчислень. Метод розширює функціональні можливості операції ділення чисел в надлишковій симетричній системі числення.

Ключові слова: операція ділення, неавтономні обчислення, надлишкові системи числення, суміщення залежних операцій.

Автори

Жабін Валерій Іванович – професор, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

E-mail: viz.kpi@gmail.com

Жабіна Валентина Валеріївна – доцент, кафедра програмного забезпечення комп'ютерних систем, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

E-mail: val.zhabina2@gmail.com

РОЗШИРЕНА АНОТАЦІЯ

Valerii Zhabin, Valentuibna Zhabinba

EFFICIENCY IMPROVEMENT OF DIVISION OPERATION REALIZATION IN ON-LINE MODE

Relevance of the research topic. For parallel systems the time of task solution depends not only on parallel algorithm branches processing time but also on time expenditures for information interchange between calculation units. Data transfer time reduction, as well as hardware implementation simplification is a relevant objective that requires additional researches, including development of efficient arithmetic operation realization methods.

Problem statement. When building systems based on FPGA, the efficient use of the microcircuit resource (internal functional blocks and external pins) is an important problem. By the example of the division operation there is considered an opportunity to solve the above-mentioned problem reducing number of connections between system units by means of data transfer in serial code.

Analysis of resent researches and publications. In parallel arithmetic the division methods require availability of all digits of numerator and denominator before calculation starts. It does not allow overlapping the processes of digit-by-digit input of operands and the process of result digits calculation, so realization of computations in on-line mode is impossible. To overlap the above-mentioned processes, division methods of quasi-parallel arithmetic were developed. In this case the calculations are realized starting with high-order digits and redundant number systems are used to avoid carries to higher number positions. Nevertheless, the well-known methods place restrictions to the form of operands representation, namely to their range and sign.

Singling out of unexplored parts of the general problem. Methods of division in on-line mode require further research. Adding functionalities to methods including possibility to process numbers with any sign, as well as to extend the range of operands representation is an important task to be solved to increase the data processing efficiency.

Task definition. The purpose of the research is to increase the efficiency of division operation realization in on-line mode by adding functionalities to the methods.

Statement of the main material. There is proposed a method of numbers division in on-line mode, which enables to overlap realization of digit-by-digit input of operands, their processing and digit-by-digit output of the result starting with high-order digits. It enables to realize parallel processing of operations connected by data at the level of operand digits that allows acceleration of data transfer. Unlike the well-known methods, this one gives an opportunity to realize division of numbers with any sign. Moreover, the range of data representation is extended.

UDC 004.055

Artem Volokyta,
Artem Kaplunov, Oleksandr Pospishnyi

THE PROBLEM OF RESOURCE SEARCH IN DISTRIBUTED COMPUTING SYSTEMS

Артем Волокита,
Артем Каплунов, Олександр Поспішний

ПРОБЛЕМА ПОШУКУ РЕСУРСІВ В РОЗПОДІЛЕНИХ ОБЧИСЛЮВАЛЬНИХ СИСТЕМАХ

Formalism of resource records requests creates significant difficulties for users due to the high complexity of the Grid-resources information model. This leads to errors and inaccuracies in the resource requests, threatening the efficiency of dispatching.

The results of studies of the Grid model's imitation is presented in this work. With its help the aforementioned errors' influence on the systems efficiency is researched.

Keywords: Grid-system, executive resource, semantics attribute, problematic user's orientation, erroneous requests.

Tabl.: 1. Fig.: 1. Bibl.: 5.

Формалізми записів ресурсних запитів, що використовуються на практиці, через високу складності інформаційної моделі Grid-ресурсів, створюють істотні труднощі для користувачів системи і сприяють появі помилок і неточностей в ресурсних запитах, ставлячи під загрозу ефективність диспетчеризації.

У статті представлені результати дослідження імітаційної моделі Grid, за допомогою якої вивчено вплив вищевказаних помилок на ефективність роботи системи.

Ключові слова: Grid-система, виконавчий ресурс, семантика атрибутів, проблемна орієнтація користувача, помилкові запити.

Табл.: 1. Рис.: 1. Бібл.: 5.

Problem formulation. The information of mechanisms for resources finding in the Grid systems is systematized. The active development of Grid technology in the field of distributed computing systems was taken into account. The collected information is summarized in the table 1.

The processes such as search and composition of resources for solving the problem in the Grid system, should be transparent to the user as much as possible. The

resources that maximally fulfill all user's requirements and are able to solve the task set by one, must be allocated by the system automatically, with minimal participation by the user. He should only formulate the tasks and requirements for the resources.

Analysis of recent scientific researches and publications. In modern Grid-systems the user is forced to actively participate in the process. He is guided by the information about the resources provided to him and independently selects them. At the same time, he must have a good understanding of his task's requirements and the information model of the resources he has access to. Moreover, the user must have an excellent command of the language, which is necessary to impose requirements of the task to the resources of the system.

Table 1

Search for resources in distributed computing systems

Software	***Inform. model***	***Language***	***Possibility to expand***	***Complexity of inform. models***	***Checking queries***	***Subjective orientation***
ARC	ARC, MDS, GLUE	xRSL, JDL	no	medium	absent	no
GT 4	GLUE, CIM	RSL	no	high	absent	no
gLite	GLUE	JDL	no	high	absent	no
Legion	no	MESSIAHS	yes	low	absent	partial
Condor	no	ClassAd	yes	low	absent	yes

Specification of uninvestigated parts of general matter. The situation is complicated by the use of various syntactically incompatible schemes while the resources describing. As a result, we need strict agreements between resource providers and users about the names of attributes and their possible meanings. Allocation of descriptions only at the syntactic level, regardless the semantics of attributes and their values, as well as the need to coordinate the set of attributes among all the participants, makes such systems inflexible and difficult to expand.

Thus, the most important shortcomings of the resources' search and selection in Grid systems can be distinguished:

1. High complexity of relevant information models. The information models used to represent and search resources in Grid have a rather complex structure and use a large number of attributes in order to represent all aspects and details of the hardware and software components of the system. The study and application of such schemes creates tangible difficulties for new users, as well as creates favorable conditions for the occurrence of errors in the preparation of queries and incorrect interpretations in the meaning of attributes.

2. Lack of validation of the resource request. The combination of the above

defects leads to a large number of false resource requests from users. Unfortunately, all Grid systems used today do not validate such requests, which leads to idle time computational resources, erroneous assignments and loss of user's working time.

3. Insufficient flexibility of search mechanisms. As a rule, users and virtual organizations are deprived of the opportunity to expand or modify the resource information model and adapt it to their specific needs, while resource search engines are based on a strict syntactic mapping of attribute values for the resource and a query, which is not a sufficiently flexible nor an efficient method.

All of these disadvantages considerably complicate the process of using Grid-systems in practice and establish a high entry bar for new users.

Problem statement. Considering the collective nature of the Grid systems functioning, the necessary mechanism for quality service provision is an automatic dispatching. This is a set of management actions that optimize the whole system, carrying out operative planning and redistribution of running tasks on the basis of specified resource queries.

Unfortunately, the formalities of resource queries records used in practice create significant difficulties for users of the system and contribute to the occurrence of errors and inaccuracies in the resource queries, due to the high complexity of the informational Grid-resources model. This endangers the effectiveness of dispatching. At the same time, the two-level of management organization in Grid-systems repeatedly amplifies the negative effects that arise as a result of such errors.

Let's take as a basis the following calculation of the typical job execution time:

$$t = t_w + t_s + t_p + t_q + t_b + t_c, \quad (1)$$

where t_w - time of preparation and transfer of tasks to the controller; t_s - is the time needed to handle the task of the controller and its auxiliary modules; the search and selection of executive resources; t_p - the time for the task transfer to the subsystem management executive resource; t_q - waiting time in the queue of the local batch processing system; t_b - time of task execution on computing nodes; t_c - the time of the task results transfer to a user.

The statement of basic materials. An error requesting an inefficient resource is considered. It arises as a result of an executive resource query that can successfully complete a task, but increases its execution time due to a non-optimal selection criterion:

$$t^{III} = t_w + t_s + t_p + t_q + \varphi t_b + t_c, \quad (2)$$

where φ is the coefficient of efficiency decrease, $\varphi \in R$, $0 < \varphi < 1$.

The results of the simulation Grid model study are presented. With a help of that, the influence of the above errors on the efficiency of the system has been studied. According to the data collected, if the presence of false queries is 25% from the total (with requests of inefficient resources), the flow of tasks performed by the system decreases by approximately 15%.

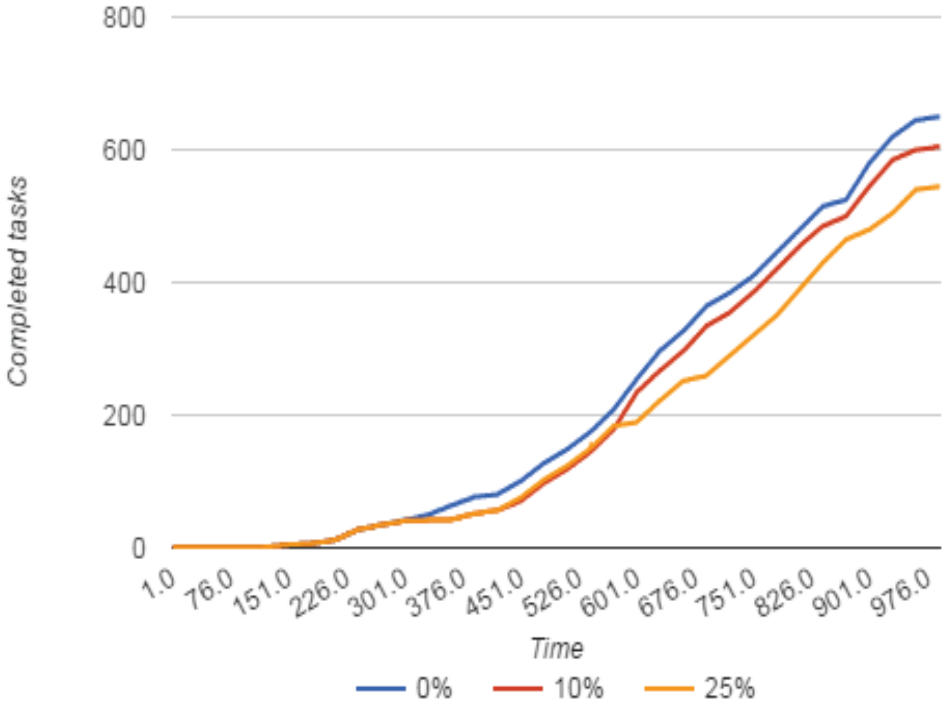


Fig. 1. The influence of an error requesting an inefficient resource on the performance of the Grid system

Conclusion. Due to the disadvantages described above, a high threshold is created that needs to be overcome by new users in order to familiarize themselves with the Grid computing industry. This happens due to the high complexity and inclination to error in resource retrieval that creates significant difficulties for users without experience of Grid system usage.

The further development of Grid technologies will lead to a multiplication of resources and their diversity, which will exacerbate the existing problem. In this regard, the development of new methods that will address these shortcomings and improve the process of finding resources in the Grid system is an actual topic of scientific and practical interest.

References

1. Tangmunarunkit H. Ontology-based resource matching in the Grid – the Grid meets the semantic web / H. Tangmunarunkit, S. Decker, C. Kesselman // The Semantic Web (ISWC). – 2003. – PP.706-721.

2. Ambrosi E. A Description Logic based Grid Inferential Monitoring and Discovery Framework / E. Ambrosi, M. Bianchi, C. Gaibisso, G. Gambosi // Proceedings of the 2005 International Conference on Grid Computing and Applications. – 2005. – PP. 18-23.

3. Parkin M. The knowledge of the grid: A grid ontology / M. Parkin, S. Van den Burghe, O. Corcho, D. Snelling, J. Brooke // Proceedings of the 6th Cracow Grid Workshop. – 2006. – PP. 658-662.

4. Said M. S-MDS: Semantic Monitoring and Discovery System for the Grid / M. Said, I. Kojima // Journal of Grid Computing. – No.7(2). – 2008. – PP.205-224.

5. Xing W. An ActOn-based semantic information service for Grids / W. Xing, Wei, O. Corcho, C. Goble, M.D. Dikaiakos // Future Generation Computer Systems. – No. 26(3). – 2010. – PP. 324-336.

Autors

Artem Volokyta – PhD in Technical Sciences, Associate Professor, Department of Computer Engineering, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

E-mail: artem.volokita@kpi.ua.

Артем Миколайович Волокита – кандидат технічних наук, доцент, Кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Artem Kaplunov – Master of Technical Sciences, student, Department of Computer Engineering, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

E-mail: art.kaplunov@gmail.com

Артем Володимирович Каплунов – магістр технічних наук, студент, Кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (пр. Перемоги, 37, корпус 18, м. Київ, 03056, Україна).

Pospishnyi Oleksandr – Master of Technical Sciences, Department of Computer Engineering graduate, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

Поспішний Олександр Сергійович – випускник, Кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

А. М. Волокита,
А. В. Каплунов, О. С. Поспішний

ПРОБЛЕМА ПОШУКУ РЕСУРСІВ В РОЗПОДІЛЕНИХ ОБЧИСЛЮВАЛЬНИХ СИСТЕМАХ

Актуальність теми дослідження. Враховуючи активний розвиток Grid-технології у галузі розподілених обчислювальних систем, все більш гостро постає питання ефективності механізмів пошуку ресурсів в Grid-системах.

Постановка проблеми. Подальший розвиток Grid-технологій неминуче призведе до багаторазового збільшення кількості ресурсів та їх різноманітності, що спричинить загострення наявної проблеми виникнення помилок і неточностей у ресурсних запитах.

Аналіз останніх досліджень і публікацій. Протягом останніх років з'являється все більше статей присвячених оптимізації підбору ресурсів у Grid-середовищі. Проте підбір ресурсів в більшості із них базується на простому синтаксичному порівнянні значень атрибутів, присутніх в описі ресурсу, зі значеннями, які зазначені в запиті користувача.

Виділення недосліджених частин загальної проблеми. Формалізми записів ресурсних запитів, що використовуються на практиці, через високу складності інформаційної моделі Grid-ресурсів, створюють істотні труднощі для користувачів системи і сприяють появі помилок і неточностей в ресурсних запитах, ставлячи під загрозу ефективність диспетчеризації.

Постановка завдання. Завданням є виділення та формальне представлення основних типів помилок, що виникають у ресурсних запитах, та дослідження їх впливу на зменшення потоку виконуваних системою завдань.

Викладення основного матеріалу. На основі аналізу літературних джерел були формалізовані основні типи помилкових запитів ресурсів. Було досліджено їх вплив на ефективність роботи системи.

Висновки. У статті представлені результати дослідження імітаційної моделі Grid, за допомогою якої вивчено вплив вищевказаних помилок на ефективність роботи системи. Наведені результати експериментів та хі аналіз.

Ключові слова: Grid-система, виконавчий ресурс, семантика атрибутів, проблемна орієнтація користувача, помилкові запити.

UDC 004.7

Tiku Vladislav,
Prokopovych Oleksandr,
Kulakov Yuriy

REVIEW OF IOT SYSTEMS BASED ON EDGE COMPUTING

Тіку Владислав,
Прокопович Олександр,
Кулаков Юрій

ОГЛЯД СИСТЕМ IOT НА ОСНОВІ ПЕРИФЕРІЙНОГО ОБЧИСЛЕННЯ

This article reviews optimization of processing data in Internet of Things systems via edge computing concept.

Keywords: gateway, IoT, edge computing.
Tabl.: 0. Fig.: 2. Bibl.: 5.

Анотація. У даній роботі розглядається актуальність та проблематика оптимізації обробки інформації в системах інтернету речей за допомогою методу периферійних обчислень.

Ключові слова: шлюз, інтернет речей, edge computing.

Relevance of research topic. The popularity of Internet of things directly affects the number of calculations and increase Internet traffic, so there is a need to optimize the operation of data centers and reduce the cost of servicing IoT systems.

Formulation of the problem. Most IoT-systems transmit data to a cloud, which is accompanied by a large flow of traffic through it and the creation of additional load on the server. The cost of leasing a server directly depends on the amount of traffic coming into the cloud, and also increases the risk of system vulnerability.

Analysis of recent researches and publications. The modern look at the distributed systems in the IoT consists of reducing the traffic between the cloud and the place where the data is collected, which increases the security of such systems. Cloud-based organizations usually receive a lot of data that is not valid and not used, but companies try to avoid deleting this information.[1]

One of the studies was about creation of a face recognition platform and transfer it to Edge Computing and compared to the cloud, the response time was from 900 to 169 ms. Moreover the unloading of the cloud can reduce the power consumption by 30-40% [2].

Nowadys, edge computing mainly serves as the reception, data storage, filtering and data transfer to cloud systems. But you need to understand that edge computing will not supplant the cloud in any case.

Selection of unexplored parts of the general problem. In addition to edge computing, there are other technologies that can solve the problem, such as distributed systems, although the method of edge computing partially uses distributed systems, so here is not all so unambiguous.

Setting objectives. Explore the concept of optimizing IoT-systems with edge computing, review the advantages and disadvantages of this method.

Presentation of the main material. Connecting IoT devices directly to data centers is accompanied by an increase in load and creates an excess of traffic.

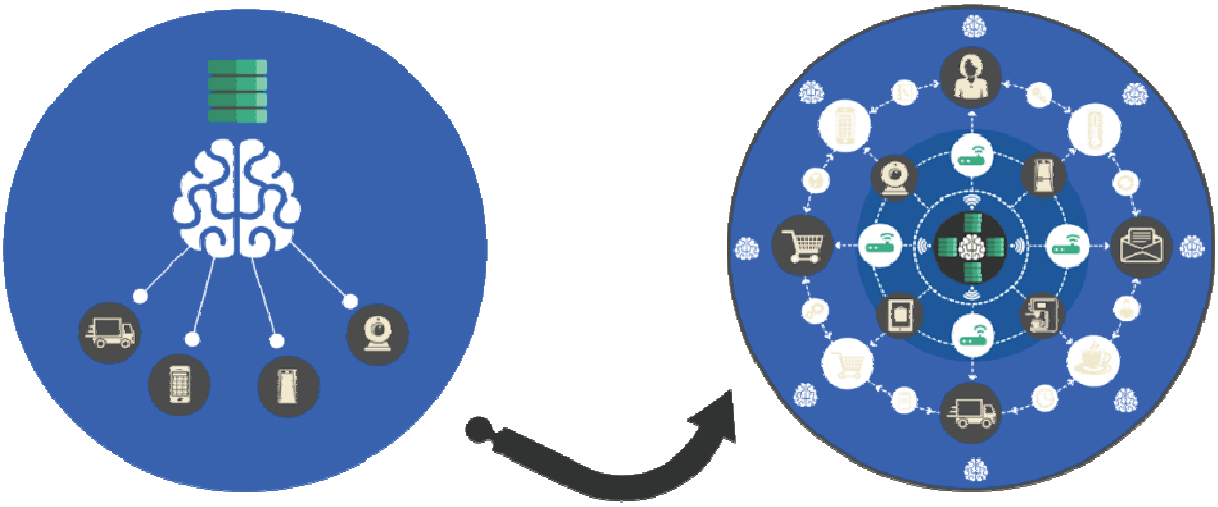


Fig. 1. General representation [5]

An effective solution is to transfer processing of raw information from IoT-devices to local devices (peripherals), and transmitting to the cloud only processed and serialized information.

Peripheral calculations can be divided into three categories:

- Calculations that are conducted only on the periphery. Typically, such systems are isolated from the network or connect to it in the case of transmission of already processed data. Such architectures are created for security systems or when the amount of raw information is too large or not available during processing.
- Distribution of computing between the cloud and the periphery. This category is the most widespread because it allows you to get all the benefits of these systems.
- Local edge computing. This is a network of gateways located in one local system.

One of such peripheral information processing device may be a single-board computer, such as Intel Edison, Intel Galileo, RaspberryPi, which can act as a gateway. This approach is not a novelty: single-board computers are the best solution for data transfer and processing on the Internet, and large businesses create cluster systems of hardware routers for their own purposes.

To solve the tasks, a single-payment computer must have a relatively small list of properties, namely:

- Reliable operating system. With the ability to collect your own distribution that will meet the task [3].
- Support for containerization tools, such as Docker.
- Remote control of the gateway.
- Ability to process, validate the data, and synchronize with other IoT devices and peripherals.
- Software for connecting new devices using the API.

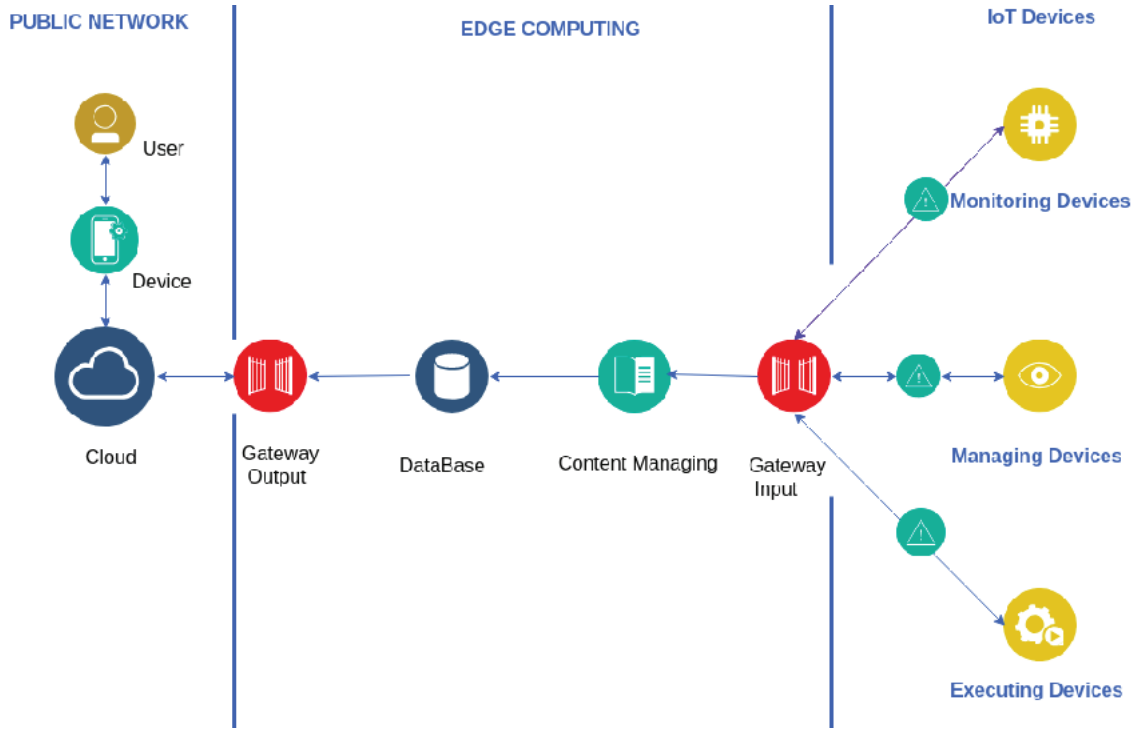


Fig. 2. The typical scheme of Edge Computing

Implementation of autopilots and smart city systems (traffic lights, parking places, roads) makes it impossible to transfer all computations to the cloud, since such situations require immediate solution on the spot, so the gateway can be an instrument that will help avoid such problems.

Moreover, many companies are interested in IoT systems and offer solutions for organization of remote systems, such as WebThings Mozilla and Intel IoT Gateway, so deployment of the network with edge computing becomes even more affordable with every year, and the range of functionalities is constantly increasing[4].

The disadvantages include a large number of edge devices, which complicates administration and maintenance than a centralized network with multiple servers, so horizontal scaling of edge computing systems needs to take into account this moment.

Conclusions. In conclusion, use of systems built on the edge computing method is going to distribute the load on data-center, organize and structure data transmitting between devices, but at the same time with the horizontal scaling of such system grows its needs.

References

1. Edge Computing – Calculating at the edge of the network. [Online]. Available: https://www.ionos.com/digitalguide/server/know-how/edgecomputing/?kwk= 614192440

2. Weisong Shi. Edge Computing: Vision and Challenges. /Weisong Shi, Fellow, IEEE, Jie Cao, Student Member, IEEE, Quan Zhang, Student Member// IEEE, Youhuizi Li, and Lanyu Xu. IEEE INTERNET OF THINGS JOURNAL, VOL. 3, NO. 5, OCTOBER 2016

3. What is edge computing? [Online]. Available: https://www.cloudflare.com/learning/serverless/glossary/what-is-edge-computing/

4. M. Armbrust, "A view of cloud computing," Commun. ACM, vol. 53, no. 4, pp.50–58, Apr.2010.[Online].Available: http://doi.acm.org/10.1145/1721654.1721672

5. IoT Edge Computing Software. [Online]. Available: https://www.post scapes.com/iot-edge-computing-software/ Pic 1. https://www.postscapes.com /wp-content/uploads/2018/05/unnamed-2.png

Authors

Tiku Vladislav - 3rd year student of the Department of Computing Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Prokopovich Alexander - 2nd year student of the Department of Computing Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Yuriy Kulakov - Professor, ScD of Computer Science, Department of Computing Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

РОЗШИРЕНА АНОТАЦІЯ

Тіку Владислав,
Прокопович Олександр,
Кулаков Юрій

ОГЛЯД СИСТЕМ IOT НА ОСНОВІ ПЕРИФЕРІЙНОГО ОБЧИСЛЕННЯ

Актуальність теми дослідження. Популярність Інтернету Речей безпосередньо впливає на кількість обчислень та збільшення інтернет трафіку. В IoT та Cloud Computing, виникає потреба оптимізації роботи датацентрів та зменшення вартості обслуговування IoT-систем.

Постановка проблеми. Більшість IoT-систем передають дані в хмару, що супроводжується великим потоком трафіку через неї та створенням додаткового навантаження на сервер.

Вартість оренди сервера напряму залежить від кількості трафіка, що надходить до хмари, а також збільшує ризик вразливості системи. Тому варто розглянути технологію Edge Computing, як засіб вирішення цих проблем.

Аналіз останніх досліджень і публікацій. Сучасний погляд на розподілені системи в IoT зводиться до зменшення трафіку між хмарою та місцем, де виконується збір даних, що збільшує безпеку таких систем. Організації, що працюють з хмарами, зазвичай отримують багато данних, що не є валідними та не використовуються, але компанії намагаються уникати видалення цієї інформації.[1]

Одним із досліджень було створення платформи розпізнавання облич та перенесення її на Edge Computing. В порівнянні з хмарою, час відгуку склав від 900 до 169 мс, також розвантаження хмари може зменшити електроспоживання на 30-40 %.

Виділення недосліджених частин загальної проблеми. Окрім граничних обчислень існують інші технології, що здатні вирішити поставлену проблему, наприклад метод розподілених систем. Розподілення Edge Computing на під системи, в залежності від задач. Розподілення навантажень між Хмарою та переферією.

Постановка завдання. Дослідити концепцію оптимізації IoT-систем за допомогою периферійного обчислення, розглянути переваги та недоліки данного метод.

Викладення основного матеріалу. Було проаналізовано способи організації переферійних обчислень, та поділено на критерії. Виділено основні властивості одноплатних комп'ютерів та недоліки.

Висновки. Отже, використання системи, побудованої на методі граничного обчислення, розподілить навантаження на датацентр, організує та структурує дані, які передаються між пристроями, але в той же час з розширенням системи росте і її потреба в обслуговуванні.

Section 3. AI (Intelligent Systems. Machine learning, big data)

UDC 004.93.12

Bandurin Vladyslav, Pavlo Rehida,
Victor Steshyn, Boldak Andriy, Artem Volokyta

AUTOMATION OF THE PROCESSING OF CERTIFICATES OF ENTRANTS THROUGH COMPUTER VISION

Бандурін Владислав, Павло Регіда,
Віктор Стешин, Болдак Андрій, Артем Волокита

АВТОМАТИЗАЦІЯ ОБРОБКИ АТЕСТАТІВ АБІТУРІЄНТІВ ЗА ДОПОМОГОЮ КОМП'ЮТЕРНОГО ЗОРУ

The paper deals with the problem of automatic processing of documents with the help of computer vision.

Key words: Computer Vision, Google Cloud Vision, Text recognition.
Tabl.: 0. Fig.: 2. Bibl.: 3.

У статті розглядається задача автоматичної обробки документів за допомогою комп'ютерного зору.

Ключові слова: Computer Vision, Google Cloud Vision, Text recognition.
Табл.: 0. Рис.: 2. Бібл.: 3.

Actual scientific researches and issues analysis. In 2018, during the admission company, the members of the admissions committee faced the problem of manual checking and counting the average score of the certificates of applicants. Every day there were up to 500 certificates a day and a large number of man-hours spent on their routine processing.

There was an idea to automate this process with the help of neural networks and algorithms of machine learning.

Formulation of the problem. When solving the problem of automation of this production process you can distinguish two main problems:

• The program required high precision, due to the specific application area of the application.

• There are few grand datasets in Ukrainian to study the model in open access.

The research objective. In recent years, you can see a significant leap in the development of text recognition systems. Among the leaders in this field are ABBYY FineReader, Tesseract, Microsoft Office Document Imaging, Google Cloud Vision. All of them are based on neural networks. However, it should be noted that some of them either do not support Ukrainian or recognize Ukrainian text with great error.

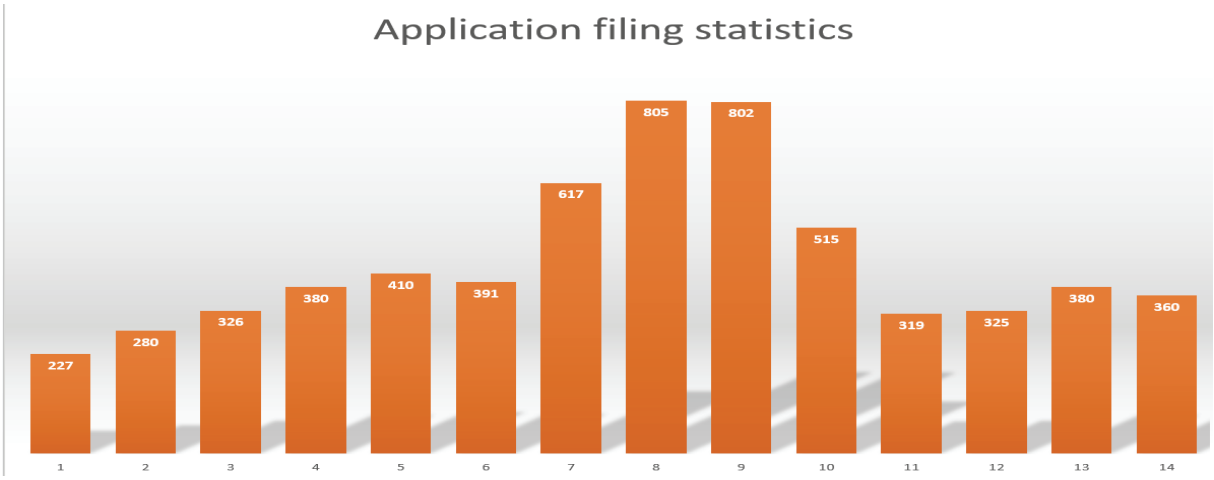


Fig. 1. Application filing statistics by day

Unexplored parts of the general problem. The decision to automate the processing of certificates in Ukraine has not yet been presented. With the problem of the ineffective use of human resources during the entrance company, all the higher educational institutions of Ukraine simultaneously faced. Therefore, there was a need to create such a service.

The research objective. Creating a service, which will take the image of the student's certificate, will work with Ukrainian words and recognize them in the image, as well as to calculate the average mark of the certificate and check the necessary information indicated on the certificate.

The statement of basic materials. The image of the certificates received by the admission committee was of a different quality. Very often the photo was taken on a bad camera and / or in poor lighting conditions. The grade for the discipline was written in a word. Also, in different secondary schools somewhat different format of certificates, number of educational disciplines. The share of print and handwriting certificates was 70% and 30%, respectively. All this caused considerable difficulties in manual and automatic data processing.

Due to the lack of quality datasets for training the neural network for the recognition of Ukrainian text in the image, it was decided to use a ready-made solution for this task. Namely: Google Cloud Vision API. This service, compared to others, gives greater accuracy in recognizing Ukrainian text, has a convenient API and is very cheap.

The Google Cloud Vision API accepts the image on the Google Cloud Input, at the output, giving text, its position in the image, and its language.

Sometimes, due to noise and poor image quality, in the recognized words there were one or two incorrectly recognized letters (for example, ten instead of ten). To solve this problem, Levenshtein distance was used between the recognized word and the dictionary. The threshold of similarity was given, in which the words are considered identical. Also, the position of words in the image was taken into account in order to distinguish grades for discipline from everything else.

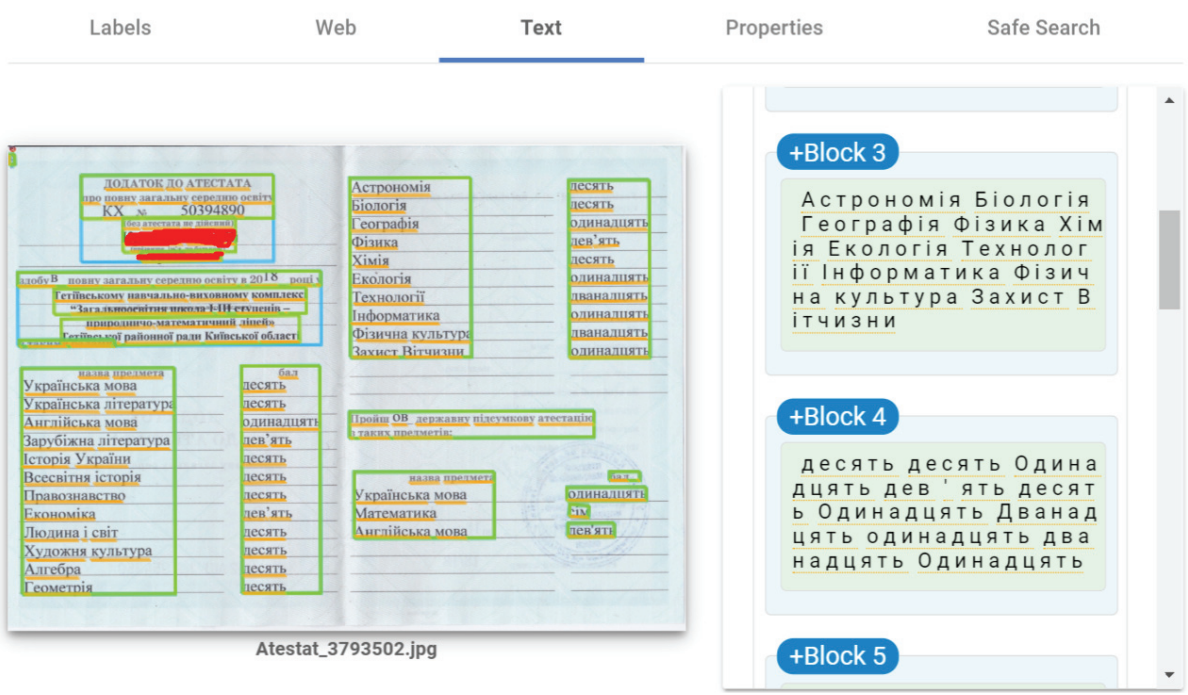


Fig. 2. An example of a certificate and display of the Google Cloud Vision API

In order to consider the algorithm to be reliable, handwriting certificates were filtered and hand-crafted. But even in this case, the efficiency of the work of the admissions committee was increased by more than 60%.

After completion of the text recognition phase, the average value of all evaluations was algorithmically determined. They checked all the necessary data about the entrant, with those that he noted. If the data do not match, then the program drew attention to the work of the admissions committee.

The results of the verification were entered into the table exel.

Conclusions. A program that was able to recognize text in high or medium quality images, with or without noise, was developed. The application is capable of recognizing English, Ukrainian and Russian alphabets, upper and lower case letters.

The accuracy of print recognition was 99%, errors were only in cases of very dark and blurred images. The accuracy of the recognition of handwritten Ukrainian text was about 60-70%.

References

1. Cloud Vision API Documentation https://cloud.google.com/vision/docs/

2. The Levenshtein Distance Algorithm https://dzone.com/articles/the-levenshtein-algorithm-1

3. Text recognition methods https://habr.com/post/220077/

Autors

Bandurin Vladislav Yuriyovich – 3rd year student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

E-mail: vladyslavbandurin@gmail.com

Бандурін Владислав Юрійович – студент 3 курсу, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

Rehida Pavlo – assistant professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: pavel.regida@gmail.com

Регіда Павло Генадійович – асистент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Victor Steshyn – assistant professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: studio.webmarker@gmail.com

Стешин Віктор Васильович – асистент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Boldak Andriy - associate professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: boldak.andrey@gmail.com

Болдак Андрій Олександрович - доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Volokyta Artem – associate professor, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".
E-mail: artem.volokita@kpi.ua

Волокита Артем Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Бандурін Владислав, Павло Регіда,
Віктор Стешин, Артем Волокита, Болдак Андрій

АВТОМАТИЗАЦІЯ ОБРОБКИ АТЕСТАТІВ АБІТУРІЄНТІВ ЗА ДОПОМОГОЮ КОМП'ЮТЕРНОГО ЗОРУ

Актуальність теми дослідження. У 2018 році, під час вступної компанії, члени приймальної комісії зіткнулися з проблемою ручної перевірки та підрахунку середнього балу атестатів абітурієнтів. Щодня надходило до 500 атестатів в день і витрачалося велика кількість людино-годин на їх рутинну обробку. Виникла ідея автоматизувати цей процес за допомогою нейронних мереж і алгоритмів машинного навчання.

Постановка проблеми. При вирішенні завдання автоматизації даного виробничого процесу можна виділити 2 головні проблеми. Від програми була потрібна висока точність роботи, в зв'язку зі специфікою області застосування програми. Великих датасетів українською мовою для навчання моделі у відкритому доступі майже немає.

Аналіз останніх досліджень і публікацій. Серед лідерів в галузі розпізнавання тексту можна зазначити: ABBYY FineReader, Tesseract, Microsoft Office Document Imaging, Google Cloud Vision. Всі вони засновані на нейронних мережах. Однак, варто зауважити, що деякі з них або не підтримують українську, або розпізнають український текст з великою похибкою.

Виділення недосліджених частин загальної проблеми. Рішення щодо автоматизації обробки атестатів в Україні ще не було представлено. З проблемою не ефективного використання людських ресурсів під час вступної компанії стикнулися одночасно всі вищі навчальні заклади України. Тому була потреба у створенні такого сервісу.

Постановка завдання. Створення сервісу, який з буде автоматично обробляти зображення атестату абітурієнта.

Викладення основного матеріалу. Проведено аналіз методів розпізнавання тексту та сервісів, які надають послуги по розпізнаванню тексту. Наведено аргументи чому використовувати Google Vision API найбільш раціонально для вирішення даної задачі. Також, наведені приклади результатів розпізнавання тексту за допомогою Google Vision API.

Висновки. Проаналізовано результати роботи програми та визначено відсоток приблизного зростання продуктивності труда у випадку використання програми.

Ключові слова: Computer Vision, Google Cloud Vision, Text recognition.

Yehor Zakupin, Valery Pavlov

CREATING TRANSLATORS OF HIGH-LEVEL PROGRAMMING LANGUAGES

Єгор Закупін, Валерій Павлов

СТВОРЕННЯ ПЕРЕКЛАДАЧІВ МОВ ПРОГРАМУВАННЯ ВИСОКОГО РІВНЯ

The paper deals with the creation of a program-translator for high-level programming languages. Analyzed the principles of building such applications and existing solutions. The program takes as a basis the structure of the translator, using as the input and output language - a high-level language. The program also allows you to choose the input and output languages.

Key words: programming language, translator.
Fig.: 4. Tabl.: 0. Bibl.: 8.

У статті розглядається питання створення програми-перекладача для мов програмування високого рівня. Проаналізовані принципи побудови подібних програм та існуючі рішення. Програма бере за основу структуру транслятору, використовуючи в якості вхідної та вихідної мови – мову високого рівня. Також програма надає змогу вибору вхідної та вихідної мови.

Ключові слова: мова програмування, перекладач.
Рис.: 4. Табл.:0. Бібл.: 8.

Relevance of research topic. A steady increase in the number of high-level programming languages can create significant problems when creating unified software. This necessitates the creation of an interpreter to facilitate and accelerate the development of software through the use of existing modules, regardless of the programming language in which they were written.

Target setting. The lack of programs that allow you to flexibly select incoming and outgoing languages, as well as the unreliability of existing solutions due to the lack of analysis of input text, which greatly complicates the use of such software.

Actual scientific researches and issues analysis. Typically, the translation of programs from one programming language to another is based on the theory of constructing compilers [1, 2, 3], but for the case when both the input and the output languages are high-level languages only cross-compilation is considered. There is a fairly small number of publications devoted to the problem of translation of programming languages. Whether they are solving local problems [4], or are based on outdated versions of programming languages [5], or are purely commercial projects,

where only demo versions are offered on a free basis [6, 7]. For example, you can specify "Java to C ++ Converter" as well as "ANSI / Turbo Pascal to C / C ++ converter", the main disadvantage of which is their strict attachment to the input and output languages. Also, in some such programs there is a lack of semantic verification systems.

Uninvestigated parts of general matters defining. The article proposes an analysis of the application of a new approach for translating between high-level programming languages. It is based on the use of pre-analysis of the text of the input language, as well as the use of external data to provide information on programming languages.

The research objective. The objective is to create a software application that allows you to transfer incoming text written in a high-level language, also at another high-level language. In this case, the program should conduct a preliminary analysis of the input text, and information on the structure of programming languages to take from external files.

The statement of basic materials. The solution method can be divided into two stages. In the first stage, the syntax of the programming language will be written, since the work of the translator depends on the form in which the syntax will be written. At the second stage, is developing a translator.

Method of writing language syntax. It is recommended to use the Backus - Naur form for language descriptions. In contrast to the meta-language of Chomsky or Chomsky-Schutzenberger, which were used in mathematical literature in the description of simple abstract languages, this meta-language was first used to describe the syntax of the real programming language Algol 60 [8]. Along with the new symbols of metacharacters, it used meaningful designations of non-terminals. This made the description of the language more vivid and allowed to continue to widely use this universal notation to describe the real languages of programming.

Developing a translator. Common properties and patterns are inherent in different programming languages, as well as translators from these languages. They have similar processes of converting the source text. In spite of the fact that the interaction of these processes can be organized in different ways, one can distinguish the functions, implementation of which leads to the same results. We call these functions the phases of the translation process. They determine the overall structure of the compiler, shown in Fig. 1.

It stands out:

1. The phase of lexical analysis.
2. The phase of syntax analysis, consisting of:

• recognition of syntactic structure;

• semantic parsing, in the process of which the work with tables is performed, the generation of an intermediate semantic representation or an object model of language.

3. The code generation phase, implementing:

• semantic analysis of the component of the intermediate representation or the object model of the language;

• Intermediate representation or object model translation into object code.

Along with the main phases of the translation process, additional phases are possible:

2a Phase of research and optimization of the interim report, consisting of:

2a.1. analysis of the correctness of the interim presentation;

2a.2. optimization of the interim presentation.

3a The phase of optimization of the object code.

In addition, we can select a single process for all phases to analyze and correct the errors that exist in the original source code of the program.

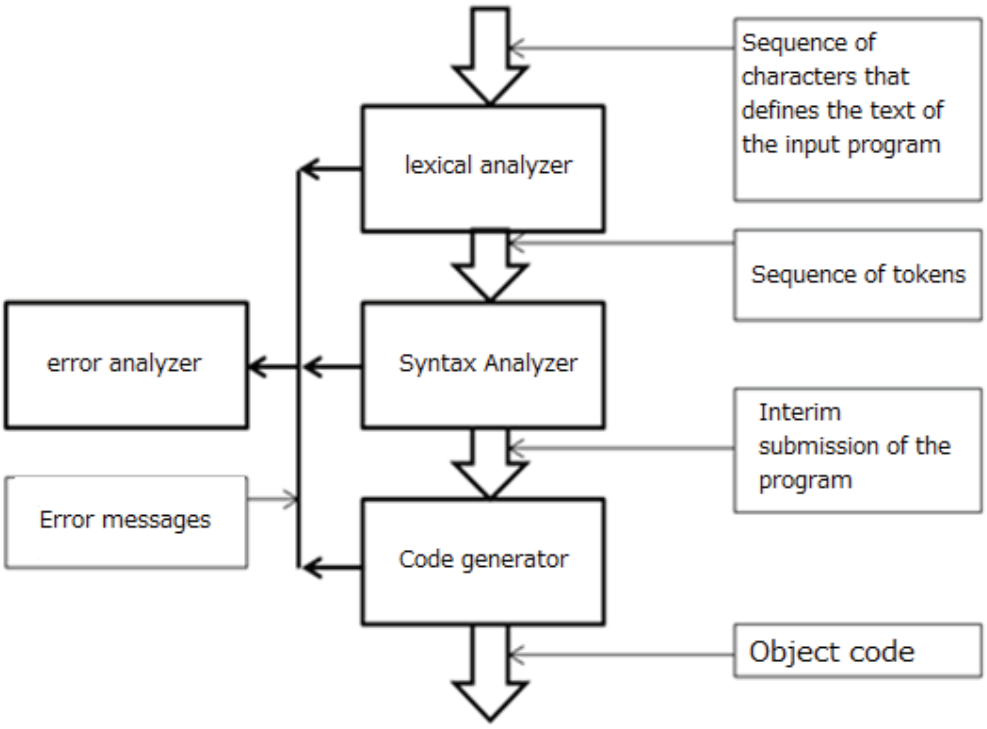


Fig. 1. Structure of the compiler

The syntax analyzer (Fig. 2.) performs the analysis of the input program using the received tokens, the construction of the syntactic structure of the program and semantic analysis with the formation of the object model of language. The object model represents a syntactic structure complemented by semantic links between existing concepts. These connections can be:

• references to variables, types of data, and procedure names that are placed in the names tables;

• links defining sequence of execution of commands;

• links defining the attachment of elements of the object model of language and others.

Thus, the parser is a fairly complex block of the translator. Therefore, it can be divided into the following components:

• recognizer;

• semantic analysis unit;

• an object model, or an intermediate representation, consisting of a table of names and a syntactic structure.

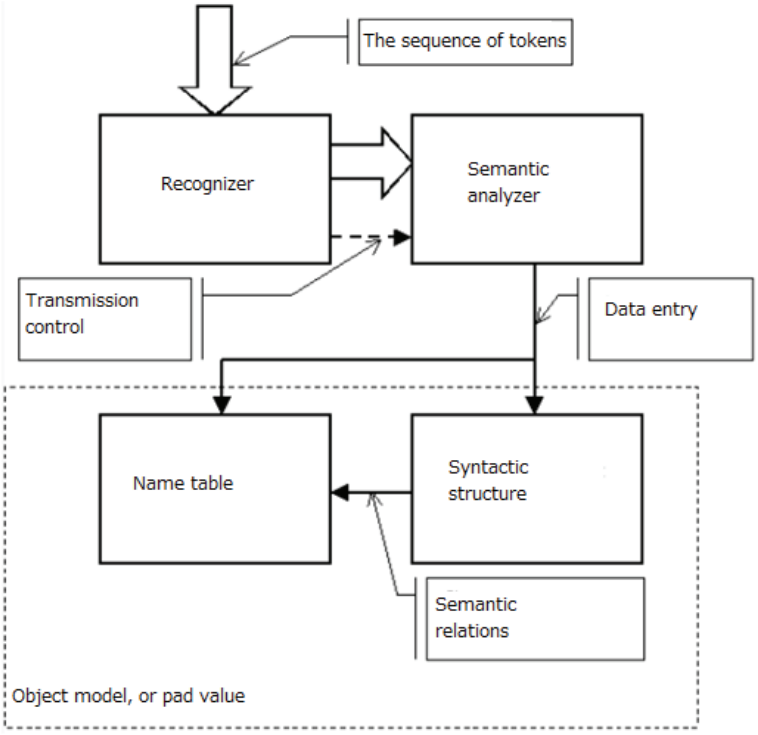


Fig. 2. The general scheme of syntax analyzer

Recognizer receives a chain of tokens and on its basis performs analysis in accordance with the rules used. Lexems, with successful parsing of rules, are transmitted to a semantic analyzer, which builds a names table and captures fragments of the syntactic structure. In addition, between the table name and syntactic structure recorded additional semantic relationships. As a result, the object model of the program is formed, freed from the binding to the syntax of the programming language. Quite often instead of a syntactic structure, it completely copies the hierarchy of objects of the language, creating its simplified analog, called intermediate representation.

Error analyzer receives error information that occurs in different blocks of the translator. Using the information he receives, he generates a message to the user. In addition, this unit may try to correct the error to continue the analysis further. It also relies on actions related to the correct completion of the program in the event that it is not possible to continue the subsequent translation.

The code generator builds the code based on an analysis of the object model or intermediate representation. The construction of the code is accompanied by additional

semantic analysis. At the stage of this analysis, the possibility of conversion is finally determined and effective options are selected. Code generation itself is the re-coding of some commands to others.

An important role in the algorithm is played by the code optimizer, which in this implementation is proposed to embed in the structure of the code generator. The role of the optimizer - before building the code, degrade it, to improve the compatibility of languages, and after the construction - optimize for a better result.

The importance of the previous decomposition of complex structures, on simpler (degradation), is the irreversibility of some transformations. For an example, let's take C ++, Java and Pascal. At first glance, the structure of these operators is very similar (Fig. 3), but if the Pascal language checks and increments (decrements) occur exclusively with the cycle variable in the beginning, in C ++ and Java, the cycle variable, verification and transformation may has nothing in common. Therefore, it makes sense to schedule such a cycle in languages C ++ and Java in the form:

for (<set initial conditions >; <condition>; <variable transform>)
{<body of the cycle}
in the form of:
<set initial conditions>;
while (<condition>) {
<body of the cycle >;
<variable transform>;
}

In this form, any C ++ and Java language cycle can be translated into Pascal without causing errors.

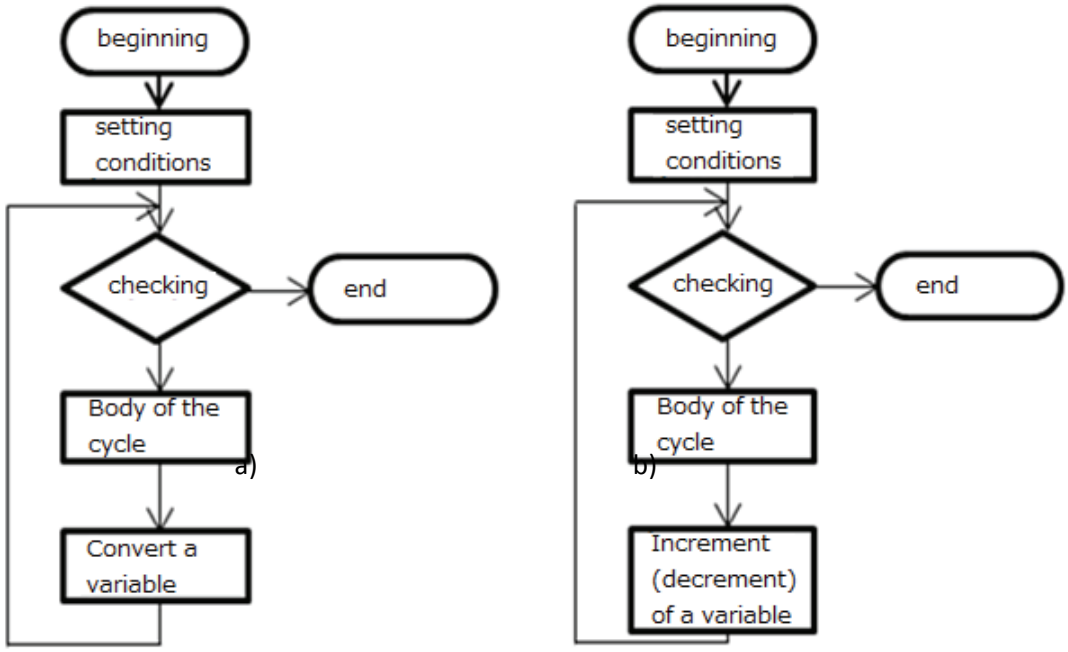


Fig. 3. The structure of cycle “for” a) in Java and C ++ languages; b) in Pascal

Due to such differences, there is a problem of the irreversibility of some transformations. As we see from Fig. 4. If there are only three languages, and one operator, the big part of the transformations can not be inverse. That is why the equally important part is the correct formation of relations between operators, and their structure at the stage of writing syntax languages.

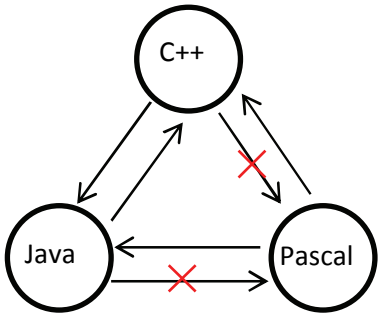


Fig. 4. Ability to convert operator for to other languages

Conclusions. The paper shows an approach to solving the problems of creating a high-level programming language translator. This is achieved through the use of a full-fledged translator model that includes an error detection system, as well as the use of language syntax descriptions and code optimization, which allows for the creation of correct relationships between similar high-level structures.

References

1. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. - М.: Мир, 1978.

2. Кауфман В. Ш. Языки программирования. Концепции и принципы. - М.: Радио и связь, 1993. - 432 с.

3. Льюис Ф., Розенкранц Д., Стринз Р. Теоретические основы проектирования компиляторов. - М.: Мир, 1979.

4. Brian Alliet. Complete Translation of Unsafe Native Code to Safe Bytecode. Rochester Institute of Technology. URL: http://www.megacz.com/berkeley/research/papers/nestedvm.ivme04.pdf. (дата звернення: 09.08.2009).

5. C2J Converter. URL: http://tech.novosoft-us.com/jsps/downloads.jsp. (дата звернення: 11.11.2001).

6. C to Java Translation. Migration Technology Systems. URL: https://www.mtsystems.com. (дата звернення: 01.02.2018)

7. Our Source Code Converters. Tangible Software Solutions Inc. URL: https://www.tangiblesoftwaresolutions.com/index.html. (дата звернення: 02.05.2019).

8. J. W. Backus. The Syntax and Semantics of the Proposed International Algebraic Language of the Zurich ACM-GAMM Conference. Proceedings of the International Conference on Information Processing, UNESCO, 1959, pp.125-132.

Authors

Pavlov Valery - associate professor, Ph.D., Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: pavlovvg@ukr.net.

Павлов Валерій Георгійович – доцент, к.т.н, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Zakupin Yehor – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: egoza343@gmail.com.

Закупін Єгор Олександрович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Єгор Закупін,
Валерій Павлов

СТВОРЕННЯ ПЕРЕКЛАДАЧІВ МОВ ПРОГРАМУВАННЯ ВИСОКОГО РІВНЯ

Актуальність теми дослідження. Постійне збільшення кількості мов програмування високого рівня може створювати суттєві проблеми при створенні уніфікованого програмного забезпечення. Таким чином виникає необхідність створення перекладача для полегшення та прискорення розробки програмного забезпечення за рахунок використання вже існуючих модулів в незалежності від мови програмування, на якій вони були написані.

Постановка проблеми. Відсутність програм, що дозволяли гнучко вибирати вхідні та вихідні мови, а також ненадійність існуючих рішень через відсутність аналізу вхідного тексту, що значно ускладнює процес використання подібного програмного забезпечення.

Аналіз останніх досліджень та публікацій. Зазвичай, переклад програм з однієї мови програмування на іншу базується на теорії побудови компіляторів, але для випадку, коли й вхідна, і вихідна мови є мовами високого рівня розглядається лише крос-компіляція. Існує досить невелика кількість публікацій, присвячених проблемі перекладу мов програмування. Існуючи рішення або

стосуються вирішання локальних задач, або спираються на застаріли версії мов програмування, або є суто комерційними проектами, де на безкоштовній основі пропонуються лише демоверсії.

Виділення недосліджених частин загальної проблеми. У статті пропонується аналіз застосування нового підходу для перекладу між мовами програмування високого рівня. Він ґрунтується на використанні попереднього аналізу тексту вхідної мови, а також використанні зовнішніх даних для надання інформації, щодо мов програмування.

Постановка завдання. Завданням є створити програмний додаток, що надає можливість перекладу вхідного тексту, написаного на однієї мові високого рівня, також на іншу мову високого рівня. При цьому програма має проводити попередній аналіз вхідного тексту, а інформацію щодо структури мов програмування брати з зовнішніх файлів.

Викладення основного матеріалу. Метод рішення можна розподілити на два етапи. На першому етапі буде проводиться запис синтаксису мови програмування, оскільки від того, в якому вигляді буде записаний синтаксис, залежить робота перекладача. На другому етапі проводиться розробка транслятору.

Висновки. В роботі показаний підхід до вирішення задачі створення перекладача мов програмування високого рівня. Це досягається за рахунок використання повноцінної моделі транслятора, що включає в себе систему виявлення помилок, а також використанням описів синтаксису мов та оптимізатору коду, що дозволяє створювати коректні співвідношення між подібними структурами мов високого рівня.

Ключові слова: мова програмування, перекладач.

UDC 004.934

Inna Humeniuk,
Olexander Markovskyi,
Olga Shevchenko

METHOD TO IMPROVE THE EFFICIENCY OF ELECTRONIC DICTIONARIES WITH CONTENT SEARCH

Інна Гуменюк,
Олександр Марковський,
Ольга Шевченко

СПОСІБ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ЕЛЕКТРОННИХ СЛОВНИКІВ З КОНТЕКСТНИМ ПОШУКОМ

In the article, the method of accelerating the work of electronic dictionaries of computer translation systems is offered. This method is based on using perfect hash-addressing and cryptographic transformations as a hash function.

Keywords: perfect hash addressing, electronic dictionary, cryptographic transformation.

Tabl.: 1. Fig.: 0. Bibl.: 5.

У статті запропоновано спосіб прискорення роботи електронних словників систем комп'ютерного перекладу, за допомогою організації пошуку на основі perfect хеш-адресації та використання криптографічних перетворень в якості хеш-функції.

Ключові слова: perfect хеш-адресація, електронний словник, криптографічне перетворення.

Табл.: 1. Рис.: 0. Бібл.: 5.

Target setting. The last two decades marked a rapid redistribution of centers of production and scientific activity. The countries of the East increasingly take leading positions in many areas of scientific, especially technological, researches.

These processes undoubtedly increase the size of scientific and technical information exchange between East and West. However, this process is inhibited by the language barrier, which is due to the huge linguistic difference between the languages of East and West. Traditional method of language acquisition by a wide range of industry specialists does not give the desired effect [1].

The most promising way to overcome the language barrier, in the context of the exchange of scientific-technical information, is the using better-advanced computer and computerized translation systems. Modern advanced computer translation technologies based on analyzing a large number of translation options, which requires

multiple access to electronic dictionaries. That dictates new requirements for the speed and efficiency of search in electronic dictionaries [2].

Thus, the scientific task of increasing the speed of search in electronic dictionaries is relevant and important for the current stage of the development of information technology.

Analysis of available solutions. To date, there are a number of approaches of organizations for searching in electronic dictionaries, oriented for using in computer translation systems.

There is part of the e-dictionaries based on the principles of databases [1]. The advantage of this approach is the ability to use existing technologies and software packages for working with databases. The main disadvantage is the low speed of search in terms of computer translation systems. E-dictionaries based on tree structures are the most widespread. They are represented by many variations and modifications of search trees [3, 4]. In the e-dictionaries of this class, the compromise between the search speed and the memory resources is resolved at an acceptable level. The number of requests to memory logarithmically depends on the number of words in the dictionary. But this speed is not enough in terms of modern computer translation technologies.

Potentially, the highest search speed is achieved with hash addressing. Until recently, its widespread use was constrained by the existing of collisions and a high level of redundancy memory utilization [5].

In modern conditions, the cost of hardware memory is reduced. As a result, the significance of the second disadvantage is reduced too. The most serious problem is collisions because resolving them requires complex mechanisms which need significant memory resources. It also does not allow to find wanted data by one access to memory.

Thus, existing electronic dictionaries do not provide the necessary speed of searching.

The research objective. The purpose of the research is to increase the speed of the contextual search in electronic dictionaries, for using them as a part of advanced computer translation systems.

The statement of basic materials. To achieve this goal the analysis of the features of the context search in electronic dictionaries has been performed. The Context Dictionary model consists of a search array of keywords and context data that included translation options and word subsets of context language constructs associated with a particular translation option [2].

Potentially, the fastest key search technology is hash addressing. It provides the independence of the search time from the size of the search array. Thus, only hash addressing is able to resolve the compromise between speed of search and size of e-dictionaries on an acceptable level for practical using.

The feature of the hash search in e-dictionaries is that words with the same root

or close to each other addressed in different areas of memory. This problem can be solved by the allocating roots or bases, using computer morpheme analysis [5] or stemming algorithms. The roots or bases are proposed to be used as keywords, and the word and its modifications are stored as context data.

To solve another problem - existing of collisions, it is proposed to use perfect hash addressing and, as a perfect hash function, symmetric cryptographic encryption algorithms, such as DES, AES. The cipher block key is used as a hash configuration code. Thus, the search keyword is introduced to the cipher block input, and the certain ciphertext or part of it is used as a hash address.

That all determine the feasibility of using a two-stage search to increase the efficiency of electronic dictionaries. In the first stage, the using of hash addressing for finding context data by keyword is proposed. The context search for the most relevant translation option takes place on the second stage. Technological realization of this process is carried out with the intelligent technologies of structural-linguistic and lexical-semantic computer analysis.

General structure. Implementation of the proposed approach provides that in the memory cell addressed by a hash transformation of a keyword, the address link to the certain contextual information in the hash memory is stored. Thus, filling in the hash memory is proposed to be carried out in two stages. At first, memory is reserving for primary address links. In the second stage, context data in size $w_1, w_2, \ldots, w_m$, of m keywords is recorded between m primary address links $a_1, a_2, \ldots, a_m$, in a way that for each word the primary address link and context data can be read to the cache by a minimum number of swap cycles. To store data it is necessary to divide memory for two parts: hash and overflow memory. The average value of each keyword's context information is w.

To implement this approach, it is proposed to use four formats of data organization in the memory cell, depending on the stored information:

• format A - free memory cells marked with a marker symbol M_1;

• format B - all bytes of this kind of memory cells are filled with context data;

• format C - for storing primary address links. The memory cells consist of three fields: the first - token M_2; the second - address link to the beginning of the context data of a certain word; the third - address of the last memory cell of the relevant context data.

• format D - for memory cells that contain address link to the rest context data of a certain keyword in the overflow memory. Memory cells of this format are marked by token M_3.

The size of the memory cells is determined by the format B, as $n/4 + 1$ byte. Tokens must be characters that are not used in the dictionary.

The context information of the dictionary words records in the ordering of their perfect hash-addresses and consists of the following action sequence:

1. The first byte of all hash cells is indicated according to the format A.

2. For all keywords, perfect hash addresses are calculated. The hash memory cells, which was addressed, mark according to the format C.

3. Let set j = 1, b = 2n + 1.

4. For the j keyword, a A-format memory cell searched starting with the address $a_j - v/2$.

5. Recording information is carried out in all bytes of a cell changing it to the format B. The address value increments: $a = a + 1$.

6. If the entire size of the j word context data has written to the hash, the address $a - 1$ is fixed in the third field of the memory cell a_j. Go to the step11.

7. If a addressed B-format memory cell and $a \leq a_j$, the address a is incremented by one: $a = a + 1$. Go to step 5

8. If $a = a_j + v/2$ or the memory cell at $a + 1$ is accorded to format B and at the same time $a > a_j$, the a memory cell is marked according to the format D. The current value of b is written to a memory cell. Go to step 10.

9. Go to step 5.

10. The rest of the j word context information is successively written to the overflow memory starting from the address b. The address of the last completed memory cell is written to the third field of the a_j memory cell, this value, incremented by one, is fixed in b.

11. The j increments: $j = j + 1$. If j is less than m, go to step 4.

12. Context information for all words has stored.

The first stage of searching is carried out in the following order:

1. The perfect hash address of keyword s is calculating: $a = h(s)$.

2. Block of memory cells, which size is v, is loaded to the cash from the hash memory, starting with the address $a - v/2$. The address of the first byte of the context information is q.

3. In the cache, the address of the begin s word context data from is hash memory is read at the address $q + (v/2)\cdot(n/4+1) + 1$ to the variable W. The address B, that is adress of the context data begin in the cache, is calculated as $B = q + (W - a + v/2) \cdot (n/4 + 1)$.

4. In the cache memory, the end context data address of the s word in the hash memory is read to the U variable. That is, the value at address $q + (v/2)\cdot(n/4 +1)$ is recorded to U. If $U < a + v/2$, the address E is calculated as: $E = q + (U - a + v/2) \cdot (n/4+1)$, else: $E = 0$.

5. For composing, i is setting as q, j— B.

6. In the cache, the byte addressed by j is forwarded into the byte, addressed by i. After that, both addresses are incremented: $i = i + 1$, $j = j + 1$.

7. If the byte addressed in the cache j contains the M_2 token, then $j = j + n/4 + 1$.

8. If $j - 1 = E$, go to step 12

9. If the byte addressed j contains an M_3 token, the rest of the s keyword

context data is in the overflow memory. From the address $j + 1$, the address of context data continuation is read to the variable D. The size of the continuation of context data is $r = U - D$. Go to step 11.

10. Go to step 6.

11. From the overflow memory to the cache, a block of r memory cells is loaded, starting with address D. g is the address of the first byte of the block in the cache. The end address of this block in the cache is calculated as $E = g + \mathrm{r} \cdot (n/4+1)$, j is set as g. Go to step 6.

12. The s keyword context data is found, loaded into the cache memory and is composed. The block address in the cache memory is q, the end address is $j - 1$.

Experiments. The effectiveness of the electronic dictionary, as components of computer translation systems, can be evaluated for speed of context search and level of memory utilization.

For modern computer systems, the search time $T_в$ is calculated as:

$T_в = h \cdot t_\rho + t_n$,

where t_ρ — the execution time of one swap cycle, t — the time of context search, h — the number of swap cycles.

An analysis of the computational processes on which the search for modern electronic dictionaries is based indicates that t_ρ is greater than t_n. That is, the speed of the dictionary can be evaluated by the average number of ρ swap cycles needed to access the keyword context data.

In addition to linguistic information, all electric dictionaries contain service data, by which access to keyword context data is made. This means that the size u of real dictionaries is always larger than the size y of actual linguistic information, herewith $y = w \cdot m$.The effectiveness of using memory can be estimated by comparing full size u of e-dictionaries, which save the same size of linguistic data.

For the experimental part of performance evaluation of the proposed e-dictionary organization, a software complex of the statistical simulation was developed. For the experiment, it was considered that $m = 1000$, $w = 450$ bytes. These parameters were determined by statistical researches of translated and explanatory dictionaries of computer terms.

The first cycle of research is aimed at detecting the influence of v on the ρ and t_ρ. However, experimental results showed no significant influence of v on ρ, so it allows to choose the value of v equal to w.

The main parameter that determines the effectiveness of the proposed vocabulary organization is α. The performed statistical research showed that the increase of value α reduces memory redundancy, but the value of ρ increases.

For example, for $\alpha = 0.013$, the $\rho = 1.45$. The number of hash memory cells can be calculated as m/α, which for this example is equal to $769 \cdot 10^3$. The memory cell size is 6 bytes, so the total amount of memory u is $8074 \cdot 10^3$ bytes.

Performance evaluation of the proposed electronic dictionary organization can be done by comparing with known developments.

In table 1 performance indicators of electronic dictionaries based on a tree with the storage of data in its nodes, based on a tree with separate storage addresses and context data, based on a hash search with collisions and separate storage of addresses and context data and proposed electronic dictionary organization, are shown.

Table 1

Performance evaluation of e-dictionary organizations

Performance indicator	***E-dictionaries based on:***			***Proposed e-dictionary based on a perfect hash addressing***
	a tree with the storage of data in its nodes	***a tree with separate storage addresses and context data***	***a hash search with collisions and separate storage of addresses and context data***	
ρ	13.29	5	2	1.45
$u \cdot 10^6$, байт	4.6	4.66	$4.686 \cdot 10^6$	8.074

The main advantage of the proposed organization of electronic dictionaries is the increasing speed of access to the keyword context data. The resulting effect is achieved by less efficient of using memory. Today hardware memory is becoming cheaper, so the proposed organization of electronic dictionaries is quite justified.

Conclusion. As a result of the research, a way to increase the speed of electronic dictionaries is proposed. It is based on the perfect hash addressing with taking into account the multilevel memory of modern computer systems.

To achieve the goal, the organization of recording and searching data in electronic dictionaries have developed. It can be used as components of high-speed intelligent computerized translation systems.

References

1. Марчук Ю. Н. Компьютерная лингвистика.- АСТ, Восток-Запад, 2007.-165 с.

2. Агапова Н. А. О принципах создания электронного словаря лингво-культорологического типа: к постановке проблемы / Н. А. Агапова, Н. Ф. Картофелева // Вестник Томского государственного университета. № 382 -2014.- С.6-10.

3. Кашеварова И. С. Электронный словарь как новый этап в развитии лексикографии // Молодой ученый. — 2010. — №10. — С. 145-147.

4. Марковский А. П. Интерактивно-шаблонный метод компьютерного перевода научно-технических публикаций / А. П. Марковский, О. Н. Шевченко, Фань Чуньлэй // Вісник Національного технічного університету України ”КПІ” Інформатика, управління та обчислювальна техніка, – Київ: ВЕК+ – 2013. – № 59. - С. 86-97.

5. Выдрин Д. В., Поляков В.Н. Реализация электронного словаря с использованием *n* грамм / Д.В. Выдрин, В.Н. Поляков // Штучний інтелект. № 4 – 2002. – С.180-183.

Autors

Humeniuk Inna – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: humeniuk.inna@gmail.com

Гуменюк Інна Олександрівна – студентка, кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Markovskyi Olexander – docent (Associated Professor), Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: markovskyy@i.ua

Марковський Олександр Петрович – доцент, кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Shevchenko Olga – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: ostl@ukr.net

Шевченко Ольга Миколаївна – student, кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Інна Гуменюк,
Олександр Марковський,
Ольга Шевченко

СПОСІБ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ЕЛЕКТРОННИХ СЛОВНИКІВ З КОНТЕКСТНИМ ПОШУКОМ

Актуальність теми дослідження. Останні два десятиліття ознаменували себе стрімким перерозподілом центрів виробничої і наукової діяльності. Країни Сходу дедалі частіше займають передові місця у багатьох галузях наукових, особливо технологічних досліджень, що призводить до збільшення інформаційного метаболізму між Сходом та Заходом. Основною перешкодою для обміну науковою технічною інформацією стає мовний бар'єр.

Найперспективнішим шляхом подолання мовного бар'єру є використання більш ефективних систем комп'ютерного перекладу.

Постановка проблеми. Наявні технології досягнення семантичної адекватності комп'ютерного перекладу базуються на різнорівневому аналізі альтернативних варіантів, що потребує багатократного звернення до електронних словників. Як наслідок, розвиток комп'ютерного перекладу ставить якісно нові вимоги до швидкості пошуку в електронних словниках.

Аналіз наявних рішень. На сьогоднішній день існує велика кількість способів організації електронних словників, найпоширенішими з яких є словники на основі: деревних структур, баз даних, хеш-пошуку з колізіями.

Постановка задачі. Мета досліджень полягає в підвищенні швидкості контекстного пошуку в електронних словниках, задля забезпечення ефективної роботи систем комп’ютерного перекладу.

Викладення основного матеріалу. Проведено теоретичні та експериментальні дослідження роботи електронних словників орієнтованих на використання в складі систем комп'ютерного перекладу. Визначено спосіб організації електронних словників, який забезпечує щонайменше в 2 рази швидший пошук в порівнянні з існуючими.

Висновки. В результаті проведених досліджень, запропоновано новий спосіб організації електронних словників систем комп'ютерного перекладу, який базується на perfect хеш-адресації.

Запропонована розробка забезпечує суттєве пришвидшення пошуку, що дозволяє використати її як компоненту швидкодіючих інтелектуалізованих систем комп'ютерного перекладу.

Ключові слова: perfect хеш-адресація, електронний словник, криптографічне перетворення.

UDC 004.414.28

Vadim Levkivskiy,
Andrii Boldak

APPROACH TO ORGANIZATION OF CLIENT SERVER INTERACTION FOR IMPLEMENTATION OF MODEL-VIEW-CONTROLLER PATTERN IN DISTRIBUTED SYSTEMS

Левківський Вадим,
Болдак Андрій

ВИКОРИСТАННЯ СЕРВІСІВ GOOGLE ДЛЯ ВИЯВЛЕННЯ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ТЕНДЕНЦІЙ.

In this article, the research is aimed at developing a system for analysing social facts in society by increasing the search queries for a given topic using Google services. As an example, the phenomenon of labour migration of Ukrainians to European countries has been explored based on the analysis of the Google Trends and Google News news charts.

Keywords: Google service, labour migration, trends, search term.

В даній статті проведенно дослідження спрямованне на передбачення певних соціальних явищ у суспільстві за рахунок збільшення пошукавих запитів по відповідній темі. А саме трудова міграція українців до країн Європи. Приведено графіки зацікавленості у пошуковому сервісі Google, відповідними країнами та подальші тенденції трудової міграції до них.

Ключові слова: трудова міграція, тенденції, пошуковий термін.

The relevance of research topic. Nowadays the methods of analytical processing of data allow us to study the formal models of facts, but the occurrence in the social sphere are usually determined informally. That is why, the development of approaches related to the construction of reflections of informal semantics in formal models is necessary for the application of analytical tools for analysing informal facts, which are social processes.

Target setting. The complexity of the analysis of social processes is related to their informal definition. Methods of detecting and predicting certain social facts have a high price. These methods are used mainly for the analysis of the economic aspect of society. Forecasting and analysing social facts is a non-standard use of Google services.

Actual scientific researches and issues analysis. More and more articles aimed at researching socio-economic facts, using data-mining methods, can be found

in broad access, at the moment. Scientific articles mainly cover the historical and political aspects of this phenomenon. Available articles describing the research, using Google Trends [4], economic facts (financial market) [1,5,6].

Uninvestigated parts of general matters defining. Investigation of cause-effect relationships of facts in the socio-economic sphere using the reflection of informal semantics in formal models and data-mining methods.

The research objective. To develop an approach to display informal semantics of social facts in quantitative models and to use methods of visual analysis [] and multivariate statistics [] for the analysis of causal relationships.

The statement of basic materials. The proposed approach to the study of social processes based on analytical processing of data is shown in Fig. 1.

From the informal description of the social phenomenon, the search tags used to derive quantitative models from Google services are distinguished. This data, along with data from other sources, is the source for data-mining procedures, which are aimed at analysing causation and building a prediction model.

Very useful for this is Google Trends (see Fig. 2). To obtain data that will be analysed together with the metric received, the news is used, from which we can take a description of facts in society.

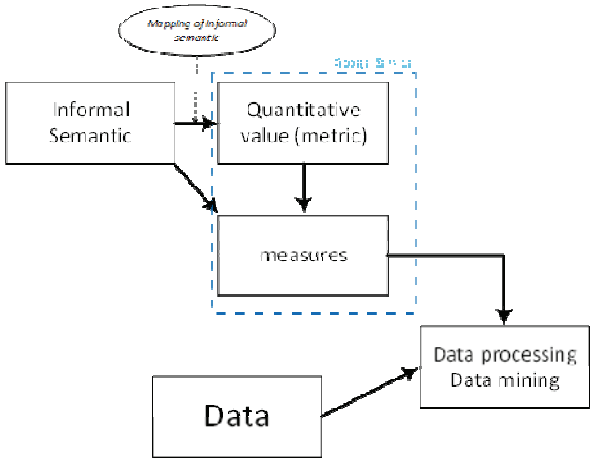


Fig. 1. Data Processing Scheme

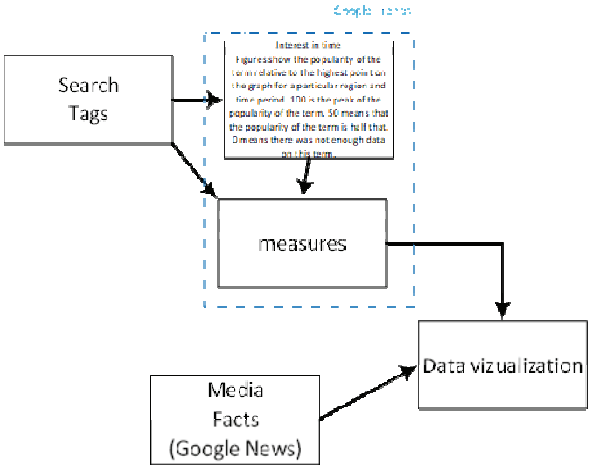


Fig. 2. Scheme of implementation

As an example of the application of the above-mentioned approach, the study of the processes of labour migration of Ukrainians was conducted. To this end, the link between interest in search terms and further social trends was analysed. Herewith were used Google Trends, Google News and the google-trends-api library. The frequency of seasonal increase in interest in the topics of work abroad, you can see from Fig. 3, Fig. 4. This season's increase comes from mid-December to mid-February of each year. This testifies to the preliminary preparation of Ukrainians to leave because the execution of relevant foreign documents takes up to 3 months [2]. Also, the charts show greater interest in Poland, which corresponds to more labour migration to this country.

Fig. 3. Schedule "Work in Poland" (red) and "Work in the Czech Republic" (blue) for the last 5 years

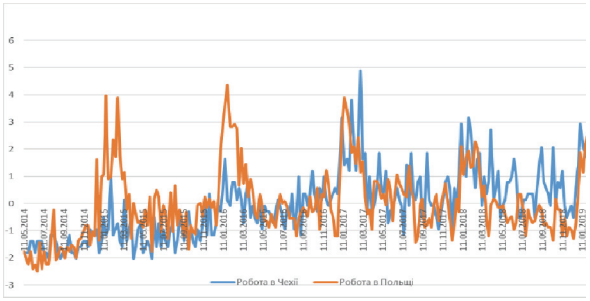


Fig. 4. The customized schedule of requests "Work in Poland and" Work in the Czech Republic "for the last 5 years

Behind the Fig. 3, you can see a change in the interest of a certain topic in percentage, over the last 5 years, from the largest of all the specified time levels. From Fig. 4 there is a change in interest in comparison with the average level, for each topic separately.

The appearance of this trend is due to the beginning of the orientation of Ukraine to Europe from 2014. This is evidenced by the relevant charts of interest in the search terms from July 2012 (Fig. 5). The jump in Poland's interest in 2015-2016 is correlated with record-breaking visa issuance at that time [3].

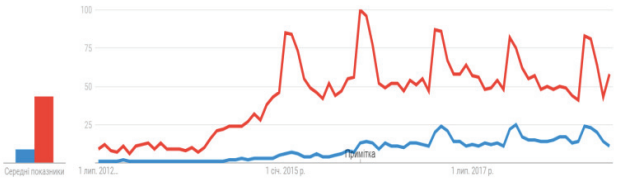


Fig. 5. Charts for interest rate search terms from July 2012

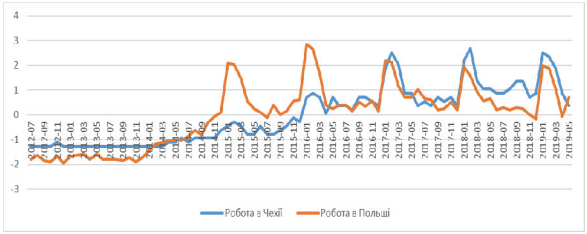


Fig. 6. Unrestricted interest rate schedules from July 2012

Also, by the level of similar queries as an example: "work in Poland for men", "work in Poland at the factory", "work in Poland for harvesting", and others; you can see the future distribution of workers by type of work, by the relevant groups (sex, single-seat or family) and the economic spheres of their employment.

A similar trend in the growth of search queries began to show in relation to other European countries as of 2018-2019 (Fig. 7, Fig. 8).

Fig. 7. Schedule of Interest "Work in Germany"

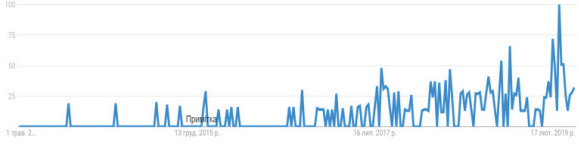


Fig. 8. Interest Graph "Work in Hungary"

Conclusions. The proposed approach of using Google services allows us to display the informal semantics of social facts into formal metrics that allow quantifying these facts. The advantage of this approach is to extend the set of quantitative data that can be processed using data-mining methods to analyse causation relationships and construct a prediction model.

References

1. Стаття «Complex dynamics of our economic life on different scales: insights from search engine query data». [Електроний ресурс] https://royalsocietypublishing.org//doi/full/10.1098/rsta.2010.0284.

2. Стаття «Що потрібно для поїздки за кордон». [Електроний ресурс] - http://lowcostavia.com.ua/travelguide-for-beginners/#travel1.

3. Стаття «В 2016 році українцям видали рекордну кількість віз в Польщу» [Електроний ресурс] - http://vsetutpl.com/v-2016-rotsi-ukrayintsyam-vydaly-rekordnu-kilkist-viz-v-polschu.

4. Сервіс Google Trends [Електроний ресурс] - https://trends.google.com//trends.

5. Эксперимент: Использование Google Trends для прогнозирования обвалов фондового рынка [Електроний ресурс] - https://habr.com/ru/company/iticapital/blog/279021.

6. Анализ маркетинговой информации на основе инструментария публичного web-приложения google trends [Електроний ресурс] - https://cyberleninka.ru/article/n/analiz-marketingovoy-informatsii-na-osnove-instrumentariya-publichnogo-web-prilozheniya-google-trends

Autors

Boldak Andrii – Candidate of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: boldak.andrey@gmail.com

Болдак Андрій Олександрович – кандидат технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Levkivskiy Vadim – Student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: vadlevkovskiy@gmail.com

Левківський Вадим Валерійович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Левківський Вадим,
Болдак Андрій

ВИКОРИСТАННЯ СЕРВІСІВ GOOGLE ДЛЯ ВИЯВЛЕННЯ СОЦІАЛЬНО-ЕКОНОМІЧНИХ ТЕНДЕНЦІЙ

Актуальність теми дослідження. Сьогодні методи аналітичного опрацювання даних дозволяють досліджувати формальні моделі явищ, але явища в соціальній сфері зазвичай визначаються неформально. Тому розробка підходів, пов'язаних з побудовою відображень неформальної семантики в формальні моделі, є необхідною для застосування засобів аналітичного опрацювання даних до аналізу неформальних явищ, якими є соціальні процеси.

Постановка проблеми. Складність аналізу соціальних процесів пов'язана з їх неформальним визначенням. Методи виявлення та передбачення певних соціальних явищ мають високу ціну. Дані методи використовуються здебільшого для аналізу економічного аспекту суспільства. Передбачення та аналіз суспільних явищ є нестандартним використанням сервісів Google.

Аналіз останніх досліджень та публікацій. Наразі у широкому доступі можна знайти все більше статей спрямованих на дослідження соціально-економічних явищ, за допомогою методів data-mining. Наукові статті переважно висвітлюють історичний та політичний аспект даних явищ. Наявні статті, які описують дослідження, за допомогою Google Trends [4], економічних явищ (фінансовий ринок) [1,5,6].

Виділення недосліджених частин загальної проблеми. Дослідження причино-наслідкових зв'язків явищ в соціально-економічній сфері з використанням відображення неформальної семантики у формальні моделі та методів data-mining.

Постановка завдання. Розробити підхід відображення неформальної семантики суспільних явищ у кількісні моделі та використання методів візуального аналізу і multivariate statistics для аналізу причинно-наслідкових зв'язків.

Викладення основного матеріалу. Запропонований підхід до дослідження соціальних процесів на основі аналітичного опрацювання даних. Показано, як можна пришвидшити деякі етепи моделі за допомогою існуючих api.

Висновки. Запропонований підхід використання сервісів Google дозволяє отримати відображення неформальної семантики соціальних явищ у формальні метрики, які дозволяють кількісно оцінювати ці явища. Перевагою такого підходу є розширення множини кількісних даних, які можуть опрацьовуватися за допомогою методів data-mining з метою аналізу причинно-наслідкових зв'язків та побудови prediction model.

Ключові слова: сервіси Google, трудова міграція, тенденції, пошуковий термін.

UDC 681.327

Victor Poriev

IMPROVING THE METHOD OF RUN LENGTH ENCODING

Віктор Порєв

ВДОСКОНАЛЕННЯ МЕТОДА КОДУВАННЯ ДОВЖИН ПОВТОРЕНЬ

The article analyzes modifications of the classical method of encoding repeat lengths for compression of raster data. As a result of the analysis, the usefulness of constructing optimal code sequences for fragments of raster images is noted. Proposals for organizing direct access have been made to accelerate the decoding of large-scale raster data fragments.

Key words: compression, prefix codes, run length encoding.
Fig.: 3. Tabl.: 0. Bibl.: 7.

В статті проаналізовані модифікації класичного метода кодування довжин повторень для компресії растрових даних. У результаті аналізу відзначається корисність побудування оптимальних кодових послідовностей для фрагментів растрових зображень. Зроблено пропозиції щодо організації прямого доступу для прискорення декодування фрагментів растрових даних великих обсягів.

Ключові слова: кодування довжин повторень, компресія, префіксні коди.
Рис.: 3. Табл.: 0. Бібл.: 7.

Relevance of the research topic. The growth in the amount of information processed by modern information systems prompts the search for effective forms of data storage and transmission. The improvement of methods of compression of information is actual.

Formulation of the problem. For information systems that store large volumes of information in raster formats, it is necessary to provide both a high degree of compression and a high decompression rate. To a large extent, these factors are contradictory.

Actual scientific researches and issues analysis. The method of compressing information based on encoding repeat lengths (RLE – *run length encoding*) has been known for a long time [1]. This is a very simple compression method, according to which each sequence of identical values is encoded by a pair (number of repetitions, values). This method has received wide popularity for recording images in a variety of file formats. The main known implementations of the RLE method are the PackBits method used in TIFF, TGA and others [2], as well as the version of the RLE method for the PCX file format [3].

Advantages of RLE:

- additional memory is not required (for example, for the dictionary)
- simplicity and high speed of unpacking (decoding)
- in simple implementations of the RLE method, the highest packaging speed is achieved
- the possibility of independent coding of individual lines, or other blocks, creates prerequisites for:
  - possibility of organizing quick direct access to any parts of the image
  - parallel (multi-thread) organization of coding-decoding

The disadvantage of known implementations of the RLE method is the small degree of compression. To increase the compression, some authors, for example [4], propose to combine the Huffman encoding with RLE [5].

In order to increase the RLE compression, authors [6, 7] proposed special prefix codes to represent the color values of raster images, as well as independent encoding of individual raster fragments for the optimal codes for these fragments. Such a version of the compression method is called RLE-БП. The developed adaptive encoder RLE-БП made it possible to increase compression by 1.5-2 times compared with the PackBits, PCX implementations while maintaining high decompression rates. This allowed it to compete with more powerful vocabulary LZ-like compression methods.

One of the advantages of the RLE method over vocabulary compression methods is that there is no need to accumulate a predefined decoding (dictionary). This allows you to encode independent raster fragments without losing compression that generates the ability to organize fast direct access to the desired parts of the image without decompressing the previous ones. Such opportunities are useful, in particular, in geographic information systems [6, 7].

Uninvestigated parts of general matters defining. The aspects of using RLE method together with methods of efficient bit sequence coding are not investigated.

The research objective. The main tasks are to search for encoding bit sequences for RLE that can provide both high compression rates and high decompression speed with minimal memory requirements.

The statement of basic materials. To estimate the number of bits needed to encode some image, we will write the following formula

$$Bits_{\text{image}} = Bits_{\text{pixel}} \times N_{\text{pixels}},$$

where: $Bits_{\text{pixel}}$ - number of bits per pixel, N_{pixels} - number of pixels in the image. For example, for a 256-color image ($Bits_{\text{pixel}} = 8$) of 1000 × 2000 pixels ($N_{\text{pixels}} = 2000000$), you need $8 \times 2000000 = 16000000$ bits.

You can write such an estimate for RLE encoding

$$Bits_{\text{RLE}} = M_{\text{bit single}} \times N_{\text{single}} + M_{\text{bit chains}} \times N_{\text{chains}},$$

where: $M_{\text{bit single}}$ - the number of bits needed to represent a single pixel, N_{single} - the

number of single pixels in the image, $M_{\text{bit chains}}$ - the number of bits needed to represent the pixel chain, N_{chains} - the number of pixel chains in the image.

For compression it is necessary that $Bits_{\text{RLE}} < Bits_{\text{image}}$. The amount of compression is determined, firstly, by the number of chains of identical pixels in a particular image, and secondly by the coding of chains and single pixels.

The number of bits per $M_{\text{bit chains}} = 16$ for the PackBits and PCX methods regardless of the particular image. For the RLE_БП method, the value of $M_{\text{bit chains}}$ can be substantially less than 16 and takes into account the features of a particular image.

The number of bits per single pixel depends on the image. For PCX and PackBits, the $M_{\text{bit single}}$ value is in the range of 8 to 16. For the RLE_БП method, the $M_{\text{bit single}}$ value may be less than 8.

General model structure. In general, it seems that for all known varieties of the RLE method, the following coding structure is used. Bitstream contains codewords. The first bit 0 of each codeword of them means that the next is the single pixel code or the code of the set of unique pixels (literals). If the first bit = 1, then the code of the chain of the same pixels is contained.

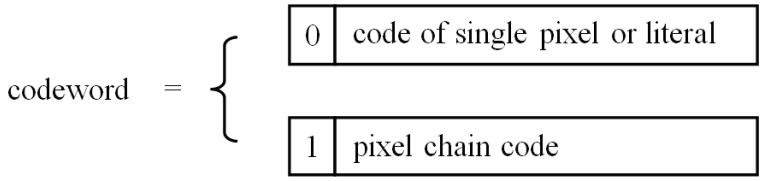


Fig. 1. Codeword structure

The single pixel code can be a normal binary code of the color index, or may be in the form of a prefix code, such as Huffman. There may be a combination of prefix and common codes used in RLE_БП, such as:

0 c . . c - m0 bits c
10 c . . c - m1 bits c
110 c. . . c - m2 bits c
1110 c. . . c - m3 bits c
1111 c. . . c - m4 bits c

where: c. . . c is the bits of the usual binary code of color indices. This RLE method improvement improves the compression of images that have many single pixels.

The pixel chain code must one way or another contain the chain length (n) and the pixel color index (s). For example, for the PackBits method, the pixel string code has the form: nnnnnnnccccccc, and for RLE_БП it is usually the color of the bits first and then the bits of length. One of the moments in improving the code structure in RLE_БП is the different number of bits of the chain code for different lengths of chains - for short lengths less bits, for longer ones. Some RLE_БП submethods use different versions of prefix codes for different lengths of pixel strings.

Another aspect of improving the RLE method is to control the length of the codewords using a plurality of encoding parameters. Such parameters can be, for

example, the number of bits of repeat length, the number of bits of color indices, types of code formats, etc. This allows each piece of the image to be written down by a code of minimum length, so that RLE_БП provides the highest degree of compression among known implementations of the RLE method.

Experiments. You can give an example of a specific image of 1125 × 1115, in which 74% of single pixels, 26% of the pixels in the chains. For him, Bitsimage is 10,003,500. This image is poorly compressed by all known versions of the RLE method. PCX encoding gives 10,821,000 bits, that is, even more so than Bitsimage. The value of $M_{\text{bit single}} = 9.43$, $M_{\text{bit chains}} = 16$. Encoding of PackBits gives 9710384 bits (compression is), $M_{\text{bit single}} = 8.2$, $M_{\text{bit chains}} = 16$. The RLE_БП method gives the greatest effect: 8590579 bits, $M_{\text{bit single}} = 7.69$, $M_{\text{bit chains}} = 9.93$.

In the following example, the image is better compressed, because the single pixels have only 8.4% of it, and the rest in the chains. The dimensions are 11680 × 10,000, which means Bitsimage = 934.4 million bits. Encoding PCX gives 192.0 million bits, $M_{\text{bit single}} = 10.32$, $M_{\text{bit chains}} = 16$. PackBits encoding gives 188.4 million bits, $M_{\text{bit single}} = 9.66$, $M_{\text{bit chains}} = 16$. The RLE_БП method here gives the biggest effect: 151.1 million bits, $M_{\text{bit single}} = 8.37$, $M_{\text{bit chains}} = 12.96$.

With regard to tests. For testing, you should use both real images and specially synthesized tests. As the simplest tests to assess the capabilities of RLE can recommend the image of vertical bands of different thicknesses of different colors. For example, the following tests:

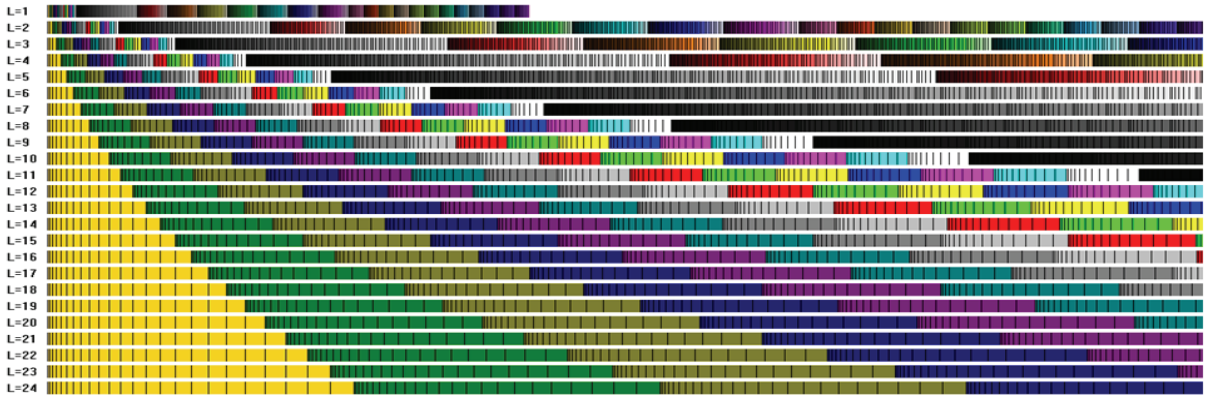


Fig. 2. A set of tests of vertical stripes of different widths

Fig. 2 shows 20 test variants for different maximum bandwidths (L). In each test, the set of stripes of all 256 colors, each color strip is separated by a black vertical line. On these tests, different versions of RLE show a significantly different degree of compression (Fig. 3).

Conclusions. The problems of improving the method of RLE based on the methods of optimal encoding of bit sequences of individual fragments of raster images are investigated. The comparative testing of realization of the modified method with known versions of RLE implementations is carried out.

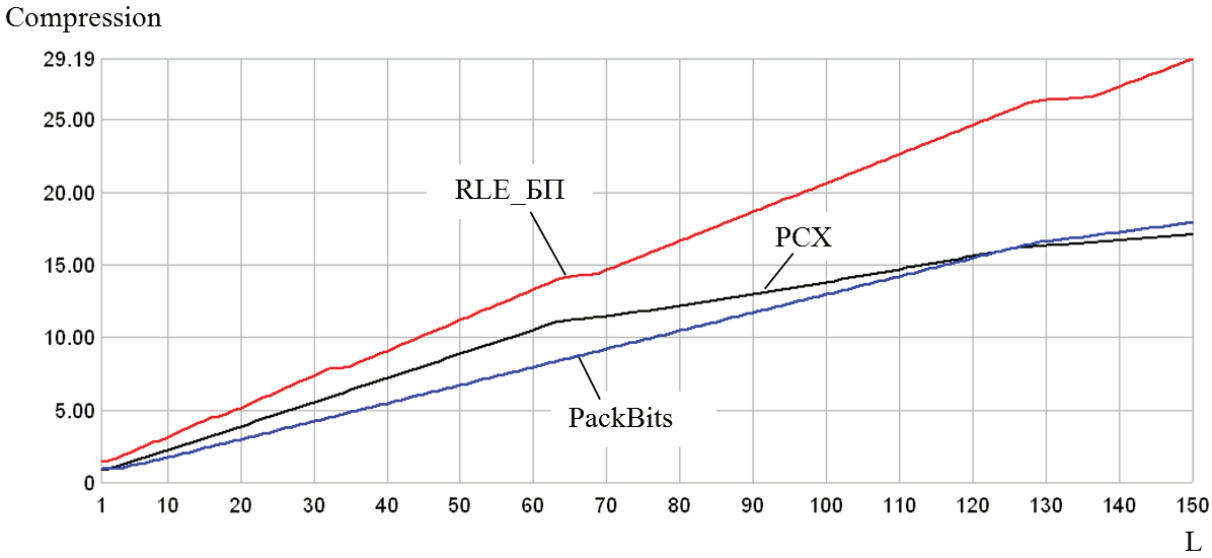


Fig. 3. The degree of compression in the tests for the width of the bands L = 1. . . 150

References

1. S.W.Golomb. Run-Length Encodings, IEEE Trans. Information Theory, 12:3 (1966) pp. 399-401.

2. PackBits. From Wikipedia, the free encyclopedia. URL: https://en.wikipedia.org /wiki/PackBits.

3. PCX File Format Summary. FileFormat.Info. URL: https://www.fileformat.info /format/pcx/egff.htm.

4. Bayadir Abbas Al-Himyari. Role of Run Length Encoding on Increasing Huffman Effect in Text Compression // Journal of Kerbala University, 2008, Vol. 6 No.2 Scientific. pp.199-204.

5. D.A. Huffman. A Method for the Construction of Minimum-Redundancy Codes. Proceedings of the IRE. 40 (1952): 1098–1101. doi:10.1109/JRPROC.1952. 273898.

6. Блинова Т.А., Порев В.Н. Некоторые способы кодирования растров в геоинформационных системах // Электронное моделирование. – 2008. – Т.30, №1. – С. 119-128

7. Blinova T., Porev V. Some Methods Of The Raster Encoding In Geographic Information Systems // Proc. int. conf. “CODATA`21”, Kyiv, 2008. – p.153.

Autors

Victor Poriev – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: v_porev@ukr.net

Порєв Віктор Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

В. М. Порєв

ВДОСКОНАЛЕННЯ МЕТОДА КОДУВАННЯ ДОВЖИН ПОВТОРЕНЬ

Актуальність теми дослідження. Зростання обсягів інформації, яку обробляють сучасні інформаційні системи спонукає шукати ефективні форми зберігання та передачі даних. Актуальним є вдосконалення методів компресії інформації.

Постановка проблеми. Для інформаційних систем, які зберігають великі обсяги інформації у растрових форматах, необхідно забезпечувати як високу ступінь компресії, так і високу швидкість декомпресії. Значною мірою ці фактори суперечливі.

Аналіз останніх досліджень і публікацій. Метод компресії інформації на основі кодування довжин повторень (RLE) відомий вже давно. Широку популярність цей метод отримав для запису зображень у різноманітних файлових форматах. Недоліком відомих реалізацій метода RLE є невелика компресія, проте інтерес до цього метода не зменшується. Дослідники шукають можливості його вдосконалити, щоб підвищити компресію і зберегти високу швидкодію.

Виділення недосліджених частин загальної проблеми. Недослідженими є аспекти використання метода RLE спільно зі способами ефективного кодування бітових послідовностей.

Постановка завдання. Основними завданнями є пошук процедур кодування бітових послідовностей для RLE, які здатні забезпечити одночасно і високу степінь компресії і високу швидкість декомпресії при мінімальних вимогах до пам'яті.

Викладення основного матеріалу. Запропонованої формулу для оцінки основних властивостей метода RLE, яку використано для порівняння різних версій реалізації цього метода. Визначено загальну модель, яка охоплює відомі різновиди реалізації метода RLE і дозволяє проаналізувати можливі модифікації метода. Запропоновані способи модифікації метода на основі оптимального кодування бітових послідовностей RLE з використанням префіксних кодів.. Виконане порівняльне тестування відомих версій реалізацій RLE та запропонованою автором, яка отримала назву RLE_БП.

Висновки. Досліджені питання вдосконалення метода RLE на основі способів оптимального кодування бітових послідовностей окремих фрагментів растрових зображень. Виконане порівняльне тестування реалізації модифікованого метода з відомими версіями реалізацій RLE.

Ключові слова: кодування довжин повторень, компресія, префіксні коди.

UDC 004.9

Bohdan Ivanishchev, Artem Volokyta,
Heorhii Loutskii, Vu Duc Thinh

INDOOR POSITIONING SYSTEM FOR DETERMINE COORDINATES OF OBJECTS IN SYSTEM'S SIDE WITHOUT RECEIVERS

Богдан Іваніщев, Артем Волокита,
Георгій Луцький, Ву Дук Тхінь

СИСТЕМА ПОЗИЦІЮВАННЯ У ЗАКРИТИХ ПРИМІЩЕННЯХ ДЛЯ ВИЗНАЧЕННЯ КООРДИНАТ ОБ'ЄКТІВ З БОКУ САМОЇ СИСТЕМИ БЕЗ ПРИЙМАЧІВ

У статті розглядається питання розробки системи позиціювання у закритих приміщеннях, за допомогою якої можна було б визначати координати об'єктів зі сторони самої системи без наявності приймачів у об'єктів та каналу зв'язку між системою та об'єктами.

Ключові слова: система позиціювання у закритих приміщеннях, безпровідні мережі, метод зваженого калібрування.

Рис.: 1. Табл.: 1. Бібл.: 3.

The paper deals with issue of development of indoor positioning system to determine coordinates of objects by control part of this system without receivers on objects.

Key words: indoor positioning system, wireless networks, fingerprinting weighting method.

Fig.: 1. Tabl.: 1. Bibl.: 3.

Topicality of research's topic. Positioning systems take the important place in modern life. They are used by common people and the companies. For example, on transport, in trade, in the industry and energetics.

Setting the research issue. In trade, in industry and in energetics indoor positioning systems are most demanded. Such systems are applied for creation of routes in buildings with internal walls. Therefore, there are high requirements to possible error of determination of coordinates by such systems. Acceptable error (distance between calculated and correct coordinates) must be no more than 5 meters.

Analysis of the last scientific researches and publications. At the moment there are developments of indoor positioning systems which are founded on different principles. These principles include using of recognition of optical images, magnetism, propagation of sound waves, pseudo GNSS (Global Navigation Satellite System), inertial systems, infrastructure of data transmission networks (Wi-Fi, Bluetooth). [1]

Defining of uninvestigated parts of general issue. The most indoor

positioning systems determine coordinates of objects by means of special receivers. Control centre of such systems can't know coordinates of objects without existence of data transmission channel between it and receivers. But it can be necessary in certain cases, for example in case of search of workers at the plant at the time of the accident.

Setting the research objective. The purpose of this paper is to develop indoor positioning system with control part which can determine coordinates of objects without receivers with them. The developed system should meet requirement to acceptable error that must be less than 5 meters.

Description of the developed indoor positioning system. The developed indoor positioning system is based on use of Wi-Fi networks and fingerprinting weighting method. The system includes group of Wi-Fi signal receivers which create a set of signal levels for each of transmitters. It is possible to determine transmitter coordinates by this set of signal levels and fingerprinting weighting method.

Fingerprinting weighting method. This method is divided into two phases, a calibration phase and a positioning phase [3]. During the calibration phase, the indoor area must be prepared by pre-measuring of signal level of several transmitters in different points that are called fingerprints

$$C_i = (c_{i1}, c_{i2}, ..., c_{ij}), \tag{1}$$

where i is number of calibration point, j is number of receiver. To compare relative, but not absolute values of signal levels, it is possible to expand the received set [2] as follows

$$CE_i = (CE_1, CE_2, ..., CE_n) = (c_{i1} - c_{i2}, ..., c_{i1} - c_{ij}, c_{i2} - c_{i3}, ...). \tag{2}$$

During the positioning phase, the calibration values

$$X = (x_1, ...x_j) \tag{3}$$

expand

$$XE = (XE_1, ...XE_n) = (x_1 - x_2, ..., x_1 - x_j, x_2 - x_3, ...) \tag{4}$$

and compare with the measured values from the receivers to determine coordinates of transmitter by calculating weights of calibration points:

$$W_i = \sum_{k=1}^{n} |XE_k - CE_k|. \tag{5}$$

Coordinates of transmitter are calculated as follows:

$$P = \frac{\sum_{i=1}^{m} \frac{1}{W_i} P_i}{\sum_{k=1}^{m} \frac{1}{W_k}}, \tag{6}$$

where P_i is coordinates of i calibration point.

Experiments. The developed indoor positioning system was tested in office building with plan which is represented in Fig. 1.

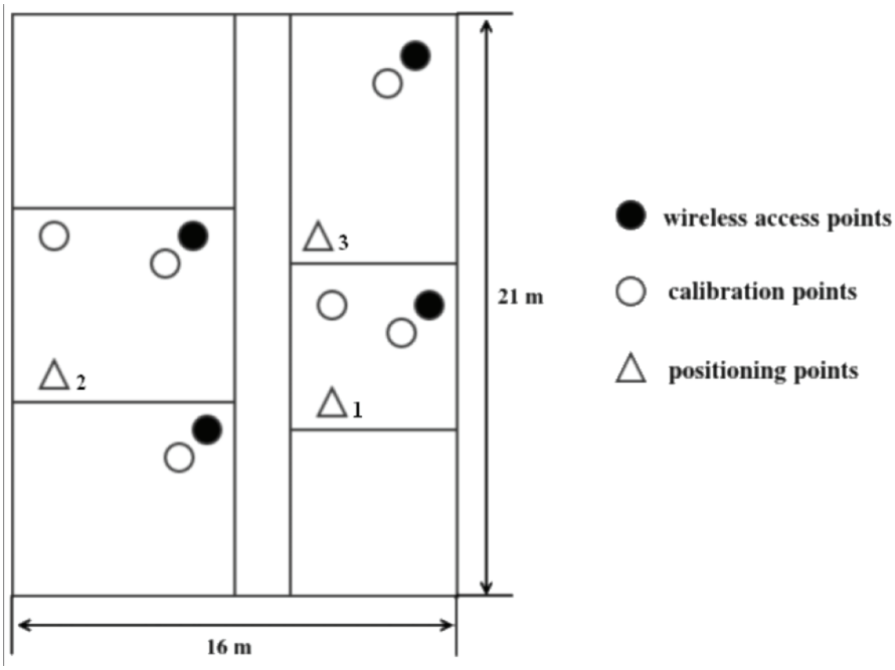


Fig. 1. Plan of office building for system testing

Result of system testing is represented in Table 1.

Table 1

Result of system testing

Number of positioning point	*Minimal error, m*	*Maximal error, m*	*Average error, m*
1	3.44	5.65	3.93
2	2.38	4.90	3.86
3	4.16	6.87	4.69

Conclusions. The indoor positioning system with control part which can determine coordinates of objects without receivers with them was developed. Average error received as a result of system testing meets requirements. The main directions for a further research are development of method of dynamic positioning of moving objects and research of other methods of remote position determination (triangulation, trilateration).

References

1. Curran K., Furey E., Lunney T., Santos J., Woods D., McCaughey A. (2011, May) *An evaluation of indoor location determination technologies.* In Journal of Location Based Services (pp. 61-78).

2. Eissfeller B., Gänsch D., Müller S., Teuber A. (2015) *Indoor Positioning Using Wireless LAN Radio Signals.*

3. Norlander A., Andersson P. (2012) *Indoor positioning using WLAN.* In degree project on computer engineering in Örebro University.

Autors

Ivanishchev Bohdan – PhD student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: callidus.iv@gmail.com.

Іваніщев Богдан Вячеславович - аспірант, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Volokyta Artem – associate professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: artem.volokita@kpi.ua.

Волокита Артем Миколайович – доцент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Луцький Георгій Михайлович – професор, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Loutskii Heorhii – professor, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

Ву Дик Тхінь – доцент, Факультет інформаційних технологій, Хошимінський університет харчової промисловості, В'єтнам.

Vu Duc Thinh – associate professor, Faculty of Information Technology, Ho Chi Minh City University of Food Industry, Vietnam.

E-mail: thinhvd@cntp.edu.vn

РОЗШИРЕНА АНОТАЦІЯ

Б. В. Іваніщев, А. М. Волокита,
Георгій Луцький, Ву Дик Тхінь

СИСТЕМА ПОЗИЦІЮВАННЯ У ЗАКРИТИХ ПРИМІЩЕННЯХ ДЛЯ ВИЗНАЧЕННЯ КООРДИНАТ ОБ'ЄКТІВ З БОКУ САМОЇ СИСТЕМИ БЕЗ ПРИЙМАЧІВ

Актуальність теми дослідження. Область застосування систем позиціювання постійно зростає. Сьогодні такі системи застосовуються як у повсякденному житті людей, так і в комерційних цілях: на транспорті, у промисловості та енергетиці, в торгівлі.

Постановка проблеми. Застосування систем позиціювання в комерційних цілях часто пов'язано з необхідністю позиціювання різних об'єктів у закритих приміщеннях (таких як торгові центри, заводські цехи, електростанції).

Аналіз останніх досліджень та публікацій. У даний момент розробки систем позиціювання у закритих приміщеннях базуються на різних принципах, один з яких використання безпровідних мереж.

Виділення недосліджених частин загальної проблеми. Більшість з наявних систем позиціювання у закритих приміщеннях не можуть визначати координати об'єктів без наявності спеціальних приймачів у самих об'єктів.

Постановка завдання. Завдання розробити систему позиціювання в закритих приміщеннях, яка не вимагала б наявності спеціальних приймачів у об’єктів, а також могла б визначати координати об'єктів зі свого боку.

Викладення основного матеріалу. Була розроблена система позиціювання у закритих приміщеннях, що базується на використанні безпровідної мережі Wi-Fi та методу зваженого калібрування. Тестування системи було проведено у закритому офісному приміщенні. Результати тестування проаналізовані.

Висновки. Тестування розробленої системи показало, що середнє відхилення визначення координат відповідає поставленим вимогам. Головні напрямки для подальшого дослідження — розробка методу динамічного позиціювання об'єктів, що рухаються, та дослідження інших методів віддаленого визначення координат (тріангуляція, трилатерація).

Ключові слова: система позиціювання у закритих приміщеннях, безпровідні мережі, метод зваженого калібрування.

Section 4. GN
(Global networks, grid and cloud systems)

UDC 004.8

Yuliia Chyzh, Simonenko Valeriy

SYSTEM FOR ACCOUNT AND STATISTICAL ANALYSIS OF THE DATA OF PATIENTS OF THE HOSPITAL

Юлія Чиж, Сімоненко Валерій

СИСТЕМА ОБЛІКУ ТА СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ ПАЦІЄНТІВ ЛІКАРНІ

The article describes the computer desktop program for automation of accounting of patients and statistical periodic reporting "Accounting of patients of the neurology department", developed to optimize the process of registration of patients.

Key words: database, client-server architecture, SQL, C++, statistical reporting
Fig.: 2. Tabl.: 0. Bibl.: 5.

У статті описана комп'ютерна десктопна програма для автоматизації ведення обліку пацієнтів та статистичної періодичної звітності «Облік пацієнтів відділення неврології», розроблена для оптимізації процесу реєстрації пацієнтів.

Ключові слова: база даних, клієнт-серверна архітектура, SQL, C++, статистична звітність
Рис.: 2. Табл.: 0. Бібл.: 5.

Introduction. The work of a neurologist is an important branch of medicine. The result of his work are diagnosis and further treatment of diseases of patients connected with the activities of the nervous system. That is why the high accuracy of diagnosis and quality of treatment is an important part of the neurologist's work.

After studying and analyzing the problems of this profession's scope, we concluded that the available programs that meet the needs of a doctor don't exist. Therefore, we decided to develop a computer program for optimizing the process of accounting and automation of statistical analysis of patients. The implementation of this system involves the facilitation of the neurologist's work; control over observance of protocols for the treatment of diseases according to the International Statistical Classification of Diseases; reducing the likelihood of errors in reporting and losing data about patients in the neurology department. The main advantage of implementation of the system is a reduction of the time for routine work with medical documentation, the registration of data on the receipt or discharge of patients, which, in its turn, increases the time for treating patients and provides better accuracy and quality of the result.

The relevance of the research topic. The relevance of the research work is to develop an effective tool in the form of a computer program that provides processing of medical records about patients and the creation of statistical reporting in order to improve the level of quality and accuracy of neurologist's treatment.

The aim of the research. The purpose of this work is to develop a computer desktop program that would take over certain functions of medical document circulation, reporting and statistics: registering patients with the preservation of a set of basic information for each of them, the formation of electronic reporting at specified intervals, etc.

Analysis of existing solutions. During solving the task, the analogs that are present in the Ukrainian market were investigated. In general, all existing solutions can be divided into two groups: too expensive for the state or municipal enterprises and software with limited functionality.

A bright example of the decision of the first group is the medical information system "Doctor Eleks". This software provides for the creation of electronic medical cards, receiption of reports, control of services and their payment. Backup copies of databases are available at an additional cost. The aforementioned product is more focused on private health care facilities. The cost of this product is quite high, which limits its distribution.

An example of decision of the second group is the Radiological Information System (IRIS). This program prevents time mismatches in patients receiving, contains anthropometric data of the patient which is necessary for proper implementation of radiological examination, improves the logistics of excepting patients. The main disadvantage of this software is the discrepancy of the existing functional to the needs of the client.

Another example of such software is the Medical Registry software package. This product provides the possibility of conducting electronic medical records, receiving periodic reports. The main disadvantage is the file-server architecture and the lack of customization of electronic cards and reports client's needs without the developer.

Problem statement. The main task of this work is to develop a computer desktop program "The accounting of patients of the neurology department" to implement the functions that were meant above.

The subject of the study is the relational database format SQL Server and the main components of the architecture of Microsoft SQL Server.

The object of the study is the database of electronic medical records documentation of National Children's Hospital "OKHMADYT" of the neurology department, the ways of its processing and the formation of the necessary reporting (with the use of main elements of the architecture Microsoft SQL Server).

One of the tasks of the work is to use the database of medical records for the registration of patients to receive the operational and periodic statistical reporting.

Materials and methods of research. The research material is the work of a

neurologist who makes diagnostic and further treatment of diseases connected with the activities of the nervous system.

Analysis of the results of work. The result of the research of the neurology department of the children's hospital staff and the analysis of the required accounting documentation is a system that optimizes the work of the department staff.

Applied computer program "The accounting of patients of the neurology department" (Fig. 1) is installed on any number of computers in the department, integrated into the local network. The database is installed on the server or one of the computers on the local network. The operators of this program can be a neurologist or a nurse who records patients. This application provides getting medical and statistical reporting.

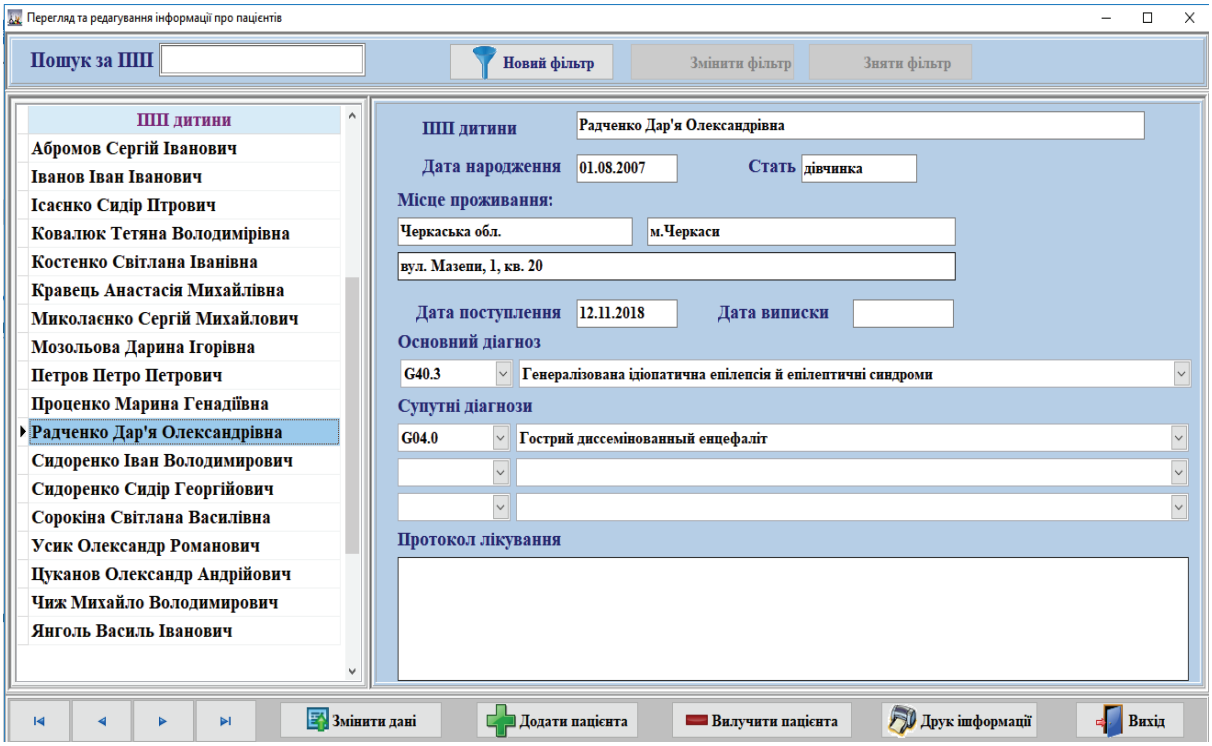


Fig. 1.

The program incarnates the following functions:

– patient records;

– filtering patients according to the selected criteria;

– saving basic information about each patient;

– dynamic management of the availability of the necessary diagnosis with ciphers according to the International Statistical Classification of Diseases and Treatment;

– formation of reporting for specified periods with the possibility of viewing and printing information according to certain indicators;

– visualization of the results of statistical analysis using diagrams of different types (Fig. 2).

The user of the program can do the following operations:

– creation / deletion / editing of a patient record;

– view information for each patient;

– search the patients by their full name or part of it

– filtering of patients by sex, region of residence, date of birth, date of admission, date of discharge, diagnosis;

– creation / deletion / editing of diagnostic data;

– print the following information:

• detailed report (a report containing complete information about all or screened patients);

• coverage of treatment by age category for the specified period with the division into articles;

• coverage of the treatment by diagnosis for the specified period with the distribution by age;

• coverage of treatment by region for the specified period by age distribution.

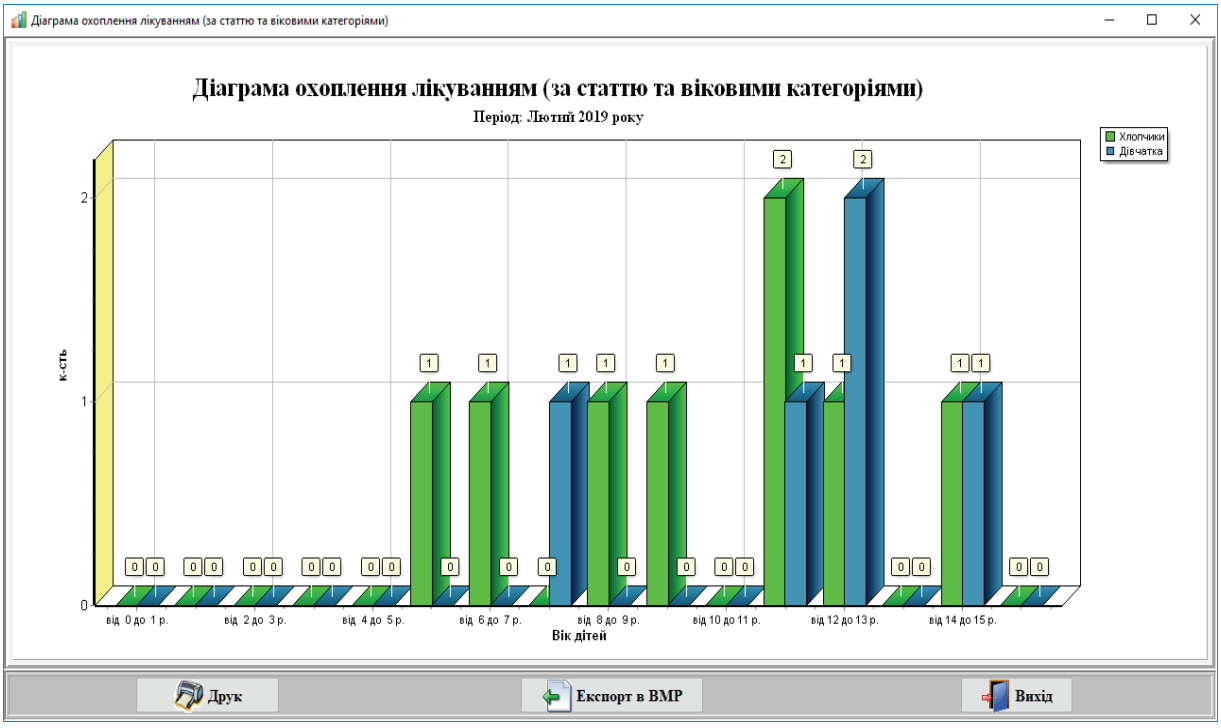


Fig. 2

Methods of solving the problem. For developing this system, the programming language C ++ and the programming environment of Embarcadero C++ Builder XE7 were used. An appropriate relational database SQL Server format has been developed for storing the data. Microsoft SQL Server 2012 Express was selected to work with the database.

Conclusions. The main result of the work is the program, which ensures optimization of the work of neurologists and children's hospital department staff.

The results of testing the program indicate the feasibility of its implementation.

References

1. International Statistical Classification of Diseases and Related Health Problems. 10th translation. P.1., ch.1, 2: Geneva : World Health Organization, 1998.

2. SQL Server Documentation. – [Electronic resource]. – Access mode: https://goo.gl/VrWkUY

3. Bruce Eckel Thinking in C++, Volume One: Introduction to Standard C++ (2nd Edition)/ Eckel Bruce.– : Prentice Hall, 2000. – 814 p.

4. Alan Beaulieu. Learning SQL. SECOND EDITION / Beaulieu Alan. – : O'Reilly Media, Inc, 2009. – 335 p.

5. В. С. Романчик, А. Е. Люлькин. Программирование в C++ Builder. – Минск, 2007. – 128 с.

Autors

Chyzh Yuliia – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: yuliia.chyzh@gmail.com

Чиж Юлія Михайлівна – студентка, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Simonenko Valeriy - Professor, ScD of Computer Science, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

Сімоненко Валерій Павлович - професор, доктор технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Ю. М. Чиж,
В. П. Сімоненко

СИСТЕМА ОБЛІКУ ТА СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ ПАЦІЄНТІВ ЛІКАРНІ

Актуальність теми дослідження. Актуальність науково-дослідницької роботи полягає в розробці дієвого інструменту у вигляді комп'ютерної програми, що забезпечує обробку облікової медичної документації про пацієнтів та створення статистичної звітності з метою покращення рівня якості та точності лікування лікаря-невролога.

Постановка проблеми. Відсутність програмного забезпечення, що задовольняло б потреби відділення лікарні.

Аналіз існуючих рішень. Під час виконання поставленого завдання було досліджено аналоги, що присутні на українському ринку. Загалом всі існуючі рішення можна поділити на дві групи: занадто дорогі для державного чи комунального підприємства та програмне забезпечення з обмеженої функціональністю.

Постановка завдання. Метою даної роботи є створення комп'ютерної десктопної програми «Облік пацієнтів відділення неврології» для реалізації вказаних вище функцій.

Одним із завдань роботи є застосування бази даних медичної документації для обліку пацієнтів для отримання оперативної та періодичної статистичної звітності.

Викладення основного матеріалу. Проведено дослідження реляційних баз даних формату SQL Server та основних компонентів архітектури Microsoft SQL Server. Об'єктом дослідження була база даних електронної медичної облікової документації відділення неврології дитячої лікарні «ОХМАДИТ», способи її обробки та формування необхідної звітності.

Для реалізації поставленої задачі використовувалась мова програмування C++. Кінцевий продукт впроваджено у відділенні неврології дитячої лікарні «ОХМАДИТ». Результат роботи програми оптимізує та спрощує роботу лікарів.

Висновки. Основним результатом роботи є програма, що забезпечує оптимізацію роботи лікарів-неврологів та персоналу відділення дитячої лікарні.

Результати проведеної апробації програми вказують на високу ефективність її впровадження.

Ключові слова: база даних, клієнт-серверна архітектура, SQL, C++, статистична звітність

UDC 004.414.28

Dmytro Korenko, Andrii Boldak

APPROACH TO ORGANIZATION OF CLIENT-SERVER INTERACTION FOR IMPLEMENTATION OF MODEL-VIEW-CONTROLLER PATTERN IN DISTRIBUTED SYSTEMS

The article considers the approach to organization of client-server interaction. This approach is used to implement the template model-view-controller in distributed systems. New concept organization WEB API implements ETL technology and allows you to move the controller functionality to the server side.

Keywords: client-server interaction, distributed systems, MVC, ETL technology.

Fig.: 1. Tabl.: 2.

У статті розглядається підхід до організації клієнт-серверної взаємодії. Даний підхід використовується для реалізації шаблону model-view-controller в розподілених системах. Нова концепція організації WEB API реалізує ETL-технологію та дає можливість перенести функціональність контролера на серверну сторону.

Ключові слова: клієнт-серверна взаємодія, розподілені системи, MVC, ETL-технологія,.

Fig.: 1. Fig.: 2.

Relevance of research topic. Modern WEB-applications are developed in accordance with the approach of SPA [1], first of all, those that apply processing large volumes of information, put forward fairly stringent requirements for the quality of communication channels. This is due to the fact that the generally accepted concepts of WEB API organization [2] are focused on the implementation of ELT-technology [3], the disadvantage of which is the need to transfer large volumes of intermediate data between a server whose API usually implements the CRUD-interface [4] model, and a client that implements a controller and a view as a MVVM template [5]. Studies related to the development of new concepts for the organization of the WEB API aimed at reducing the amount of intermediate data transmitted between the server and the client are relevant.

Target setting. The problem lies in the lack of concepts for organizing WEB APIs that implement ETL technology.

Actual scientific researches and issues analysis. Today, the generally accepted concepts of the WEB API organization are REST[6][7] and GraphQL[8], the essence of which is to implement the CRUD-interface on the server side.

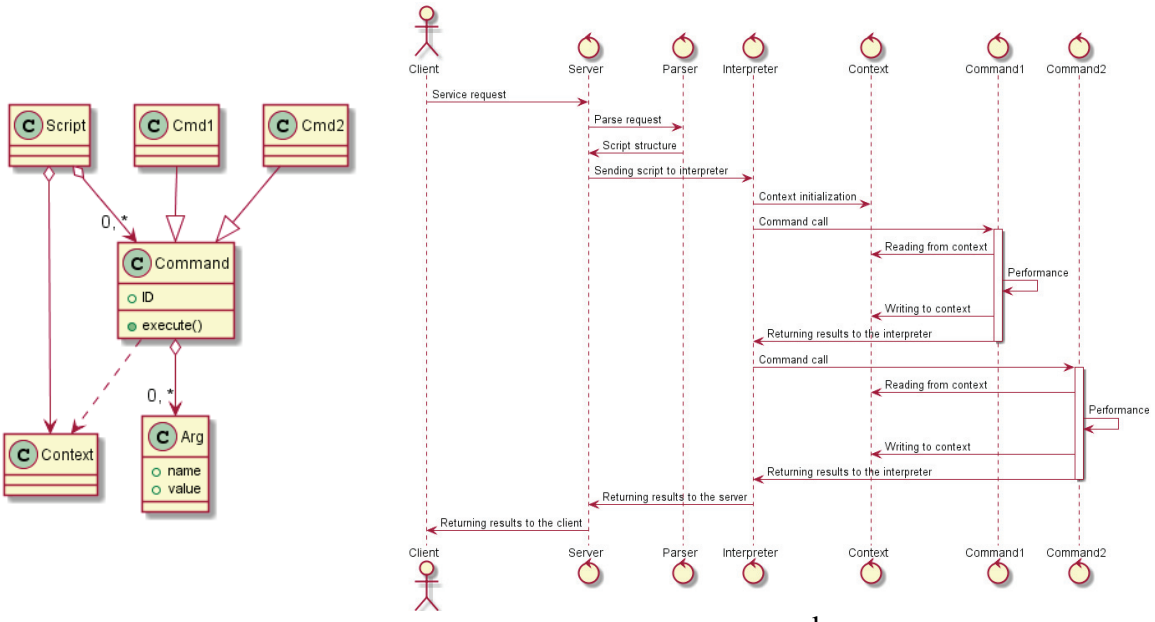
Uninvestigated parts of general matters defining. In this paper, we explore a new concept for the organization of the WEB API, which implements ETL technology and enables the functionality of the controller to be transferred to the server side.

The research objective. The task is to develop a new concept of client-server interaction, which involves transferring to the server side and executing on it a script that defines the necessary actions related to the implementation of the functionality of the controller.

The statement of basic materials. The proposed concept of WEB API organization is to use the POST request, which transmits the data structure (script), which defines the sequence of calls implemented on the server side methods (commands) (see fig.1.a). On the server side, there should be an interpreter (similar to an orchestrator in GraphQL) that implements a sequence of calls in a single scope of script context, as depicted in fig.1.b.

It is clear that the composition of the teams that are interpreted on the server side is determined by the purpose of the service. In terms of the functional purpose associated with the processing of data, in our view, the system of commands should include:

- data definition and data manipulation commands with data loading from external sources;
- collection filtering, collection joining, collection mapping and collection reducing commands;
- statistic commands such as calculation of statistics, principal component analysis, clustering etc.



a b

Fig.1. Script based WEB API: a. structure of script; b. request life cycle

The basis of the integral integrity of the distributed system is the implementation of the unified interface subsystems access to its resources. Remote Assignment Tools (RPCs) implemented at the level of underlying assets allow the delegation of data

processing to another resource that supports this protocol. In this case, the integration is implemented on the server side (fig.2.a), which can act as a proxy.

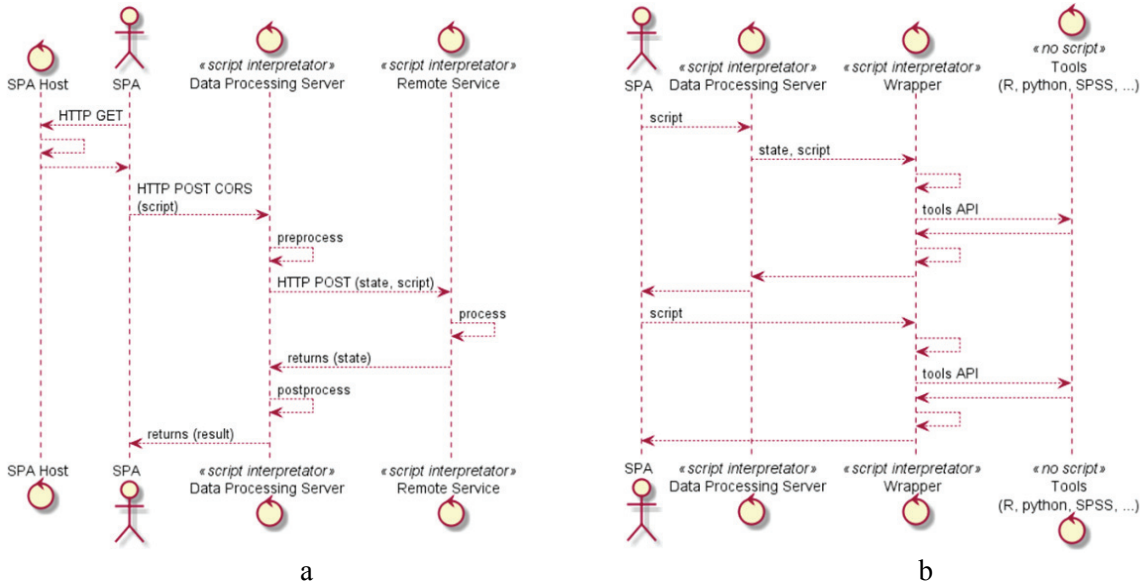


Fig.2. Integration of system resources: a. with RPC and CORS; b. with wrappers

Using Cross Origin Resource Sharing (CORS) [9] lies in the integration of resources such as one-page portal and data processing services, and enables the integration of client-side resources.

In case when there is a need to use resources that do not implement the above interface, it is advisable to use adapters (wrappers). This shell, as shown in fig.2.b, implements a unified interface for the subsystem and uses resource-specific software interfaces.

Thus, the proposed model implements a flexible scheme for organizing the interaction of various resources, including heterogeneous, both on the server side and on the client side.

Conclusions. The proposed concept of organization of the WEB API, in contrast to the known, such as REST and GraphQL, allows, in the framework of the SPA approach, to reduce the amount of intermediate data transferred between the server and the client by moving the implementation of actions related to the implementation of the controller on the server side. The implementation of such a concept involves the presence on the server side of a special software script interpretation. This approach to the organization of the WEB API has been successfully applied to the development of the data processing service [10][12], which is part of a distributed data processing system used by the World Data Center for Geoinformatics and Sustainable Development [11][13].

References

1. David Flanagan (2006). *JavaScript - The Definitive Guide*, 5th ed., *O'Reilly, Sebastopol, CA,* p.497.

2. E. Michael Maximilien, Ajith Ranabahu, Karthik Gomadam (Sep–Oct 2008). *An Online Platform for Web APIs and Service Mashups*. In Proceedings of the IEEE conference on computer vision and pattern recognition (Volume: 12 , Issue: 5).

3. Clive Skinner. (Jan 2018). *Using Redshift Spectrum to load data pipelines*. URL: https://www.dativa.com/using-amazon-redshift-spectrum-data-pipelines (Access time: 12.04.2019)

4. James Martin (1983). *Managing the Data-base Environment*. Englewood Cliffs, New Jersey: Prentice-Hall. p. 381. ISBN 0-135-50582-8.

5. John Gossman. "Introduction to Model/View/ViewModel pattern for building WPF apps ". URL:https://blogs.msdn.microsoft.com/johngossman/2005/10/ /08/introduction-to-modelviewviewmodel-pattern-for-building-wpf-apps (Access time: 08.04.2019)

6. Fielding, Roy (June 2014). *"Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, Section 4"*. IETF. Internet Engineering Task Force (IETF). RFC 7231. Retrieved 2018-02-14.

7. Fielding, Roy (June 2014). "Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content, Section 4". IETF. Internet Engineering Task Force (IETF). RFC 7231. Retrieved 2018-02-14.

8. Lee Byron (SEP 2015). *GraphQL: A data query language.* URL: https://code.fb.com/core-data/graphql-a-data-query-language (Access time: 08.04.2019)

9. Rick Anderson (2019). *Enable Cross-Origin Requests (CORS) in ASP.NET Core.* URL: https://docs.microsoft.com/en-us/aspnet/core/security/cors?view=aspnetcore-2.2 (Access time: 12.04.2019)

10. Andrii Boldak (Dec 2017). *Basic DJ Data Processing Server.* URL: https://github.com/boldak/dj-dps-server (Access time: 18.04.2019)

11. *World Data Center for Geoinformatics and Sustainable Development.* URL: http://wdc.org.ua/ (Access time: 18.04.2019)

12. Власов М. Д., Болдак А. О. Оптимізація конфігурацій розподілених інформаційних систем URL: http://sait.kpi.ua/media/filer_public/6e/80/6e804b3f-ae13-4899-b336-4daddbd45584/sait2018ebook.pdf (Access time: 18.04.2019)

13. Єфремов К. В., Болдак А. О. Предметно-орієнтована мова аналітичної обробки даних. Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. – К.: Век+, – 2012. – № 55. - с. 50-55.– 212 с

Autors

Boldak Andrii – Candidate of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: boldak.andrey@gmail.com

Болдак Андрій Олександрович – кандидат технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Korenko Dmytro – Student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: korenko.dima98@gmail.com

Коренко Дмитро Володимирович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

А. О. Болдак, Д. В. Коренко

ПІДХІД ДО ОРГАНІЗАЦІЇ КЛІЄНТ-СЕРВЕРНОЇ ВЗАЄМОДІЇ ДЛЯ РЕАЛІЗАЦІЇ ШАБЛОНУ MODEL-VIEW-CONTROLLER В РОЗПОДІЛЕНИХ СИСТЕМАХ

Актуальність теми дослідження. Сучасні WEB-застосунки розроблені у відповідності до підходу SPA, в першу чергу ті, що застосовують опрацювання великих обсягів інформації, висувають доволі жорсткі вимоги до якості каналів зв'язку. Це пов’язано з тим, що загально прийняті концепції організації WEB API є орієнтованими на реалізацію ELT-технології, недоліком якої є необхідність передачі великих обсягів проміжних даних між сервером, API якого, зазвичай, реалізує CRUD-інтерфейс моделі, та клієнтом, який реалізує контролер та зовнішній вигляд як шаблон MVVM.

Постановка проблеми. Проблема полягає у відсутності концепцій організації WEB API, які реалізують ETL-технологію.

Аналіз останніх досліджень і публікацій. Сьогодні загальноприйнятими концепціями організації WEB API є REST та GraphQL, сутність яких полягає у реалізації CRUD-інтерфейсу на серверній стороні.

Виділення недосліджених частин загальної проблеми. В цій роботі досліджуються нова концепція організації WEB API, яка реалізує ETL-технологію та дає можливість перенести функціональність контролера на серверну сторону.

Постановка завдання. Завдання полягає в розробці нової концепції клієнт-серверної взаємодії, яка передбачає передачу на серверну сторону та виконання на ній скрипту, що визначає необхідні дії, пов’язані з реалізацією функціональності контролера.

Викладення основного матеріалу. Запропонована концепція організації WEB API полягає у використанні POST-запиту, в якому передається структура даних (скрипт), що визначає послідовність викликів реалізованих на серверній стороні методів (команд). На серверній стороні повинен бути інтерпретатор (схожий на оркестратора в GraphQL), який реалізує послідовність викликів в єдиній області видимості змінних.

Висновки. Запропонована концепція організації WEB API, на відміну від відомих, дозволяє в рамках підходу SPA зменшити об’єм переданих між сервером і клієнтом проміжних даних за рахунок переміщення виконання дій, пов’язаних з реалізацією контролера на серверну сторону.

Ключові слова: клієнт-серверна взаємодія, розподілені системи, MVC, ETL-технологія.

UDC 004.72

Oleksii Cherevatenko, Yurii Kulakov

ANALYSIS OF TECHNOLOGIES OF SOFTWARE-DEFINED NETWORKS

Олексій Череватенко, Юрій Кулаков

АНАЛІЗ ТЕХНОЛОГІЙ РЕАЛІЗАЦІЇ МЕРЕЖ, ЩО ПРОГРАМНО КОНФІГУРУЮТЬСЯ

The article describes main modern technologies that implement the concept of software-defined networks (SDN) on different levels of management – OpenFlow, ONOS and CORD. An analysis of interconnectedness of these technologies has been conducted, innovations of each of the technologies has been studied. The future of the SDN concept has been predicted in the article.

Keywords: software-defined networks, OpenFlow, ONOS, CORD.
Fig.: 4. Tabl.: 0. Bibl.: 7.

У статті описуються основні сучасні технології, що реалізують концепцію програмно конфігурованих мереж (SDN) на різних рівнях управління мережею – OpenFlow, ONOS та CORD. Проводиться аналіз взаємопов'язаності цих технологій, вивчаються нововведення, що наявні у кожній з них. Прогнозується розвиток концепції SDN у майбутньому.

Ключові слова: програмно конфігуровані мережі, OpenFlow, ONOS, CORD.
Рис.: 4. Табл.: 0. Бібл.: 7.

Relevance of the research topic. The problem of scalability of computer networks in the XXI century has become global according to the big size of networks and variety of devices, so it requires automation of control. The concept of software-defined networks (SDN) can offer this and there are several interconnected multilevel implementations, which are OpenFlow, ONOS, and CORD, and they considered in this article.

Target setting. To date, there is no single laconic documentation that would explain the interconnection between the different levels of the software-defined network, and would describe the advantages and disadvantages of each solution in the context of the overall construction of the network.

Actual scientific researches and issues analysis. Currently, there are some studies about software-defined networks, but network equipment manufacturers have no single standard and opinion about how should the practical implementation of SDN

look like, and CORD technology is still not well understood by information technologies companies due to its novelty.

Uninvestigated parts of general matters defining. The article deals with the construction and interconnection of software-defined networks based on the CORD platform (including ONOS and OpenFlow at lower levels), which has not yet received general recognition in network technologies.

The research objective. The purpose of the article is to determine if it is technically and economically feasible to create a software-defined network based on the connection of OpenFlow-ONOS-CORD multilevel technologies, according to the analysis of the advantages and disadvantages of SDN.

The statement of basic materials. The principle of a software-defined network (SDN) consists of separating and managing the processes of transmitting traffic to the network.

This interpretation roots from the principle of building network devices (such as routers and switches) that implement three logical processes and have the appropriate hardware structure: dataplane, controlplane (regulation), and managementplane (administration). Typically, all these processes are monitored on each individual network element, which results in a large amount of time spent on network configuration and significant resource utilization. The principle of programmed configuration is the separation of two processes – administration and regulation – into a separate centralized system, which will use the software tools, to configure the entire network at once, and this will save time and increase traffic flow performance [1]. The network elements in SDN thus consist of only two components - the physical chips that are responsible for the ability to transmit traffic and the easiest forwarding table that will "carry" packets of traffic "through" the router or switch (Figure 1) [2].

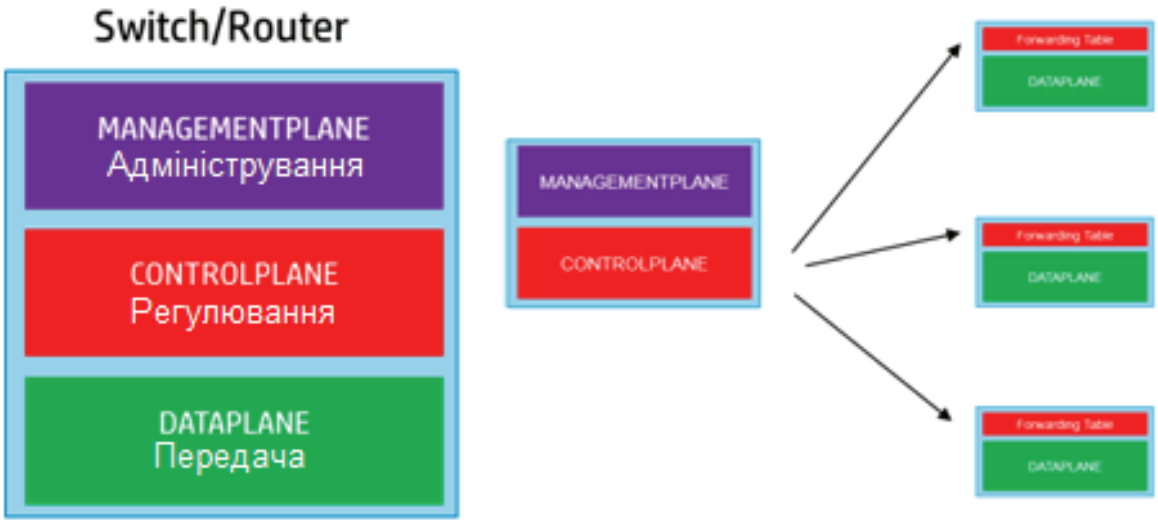


Figure 1. Separation of the levels of management of network devices in the SDN concept

The lowest level of the system under consideration is the open network protocol OpenFlow, which allows you to perform the adjustment not directly from the network device, but from the central controller of the network segment. In a basic implementation, OpenFlow connects a switch that supports OpenFlow (OpenFlow Switch, OFS) and an OpenFlow controller (OFC).

The OFS consists of two parts: a flow table that stores records received from the controller and a secure channel through which the OFS and the OFC communicate. Together with OpenFlow, some protocol such as SSL can be used in pair to provide more secure connection. After the controller software created a new command, the controller sends it to the OFS secure channel, and then adds to the flowchart as a new record. The commands received from OFC may be different - sending a frame to the port or IP address specified in the record title, dropping the frame, sending the frame back to the controller. The OFC itself can be various devices, including a personal computer (in this case, it will be the easiest for the programmer to make software adjustments), although usually a specialized server is selected as a controller. The main advantage of OpenFlow is the significant time savings required to reconfigure the system by establishing direct connections between the network elements [3]. The basic scheme of the SDN network using the OpenFlow protocol is shown in Figure 2.

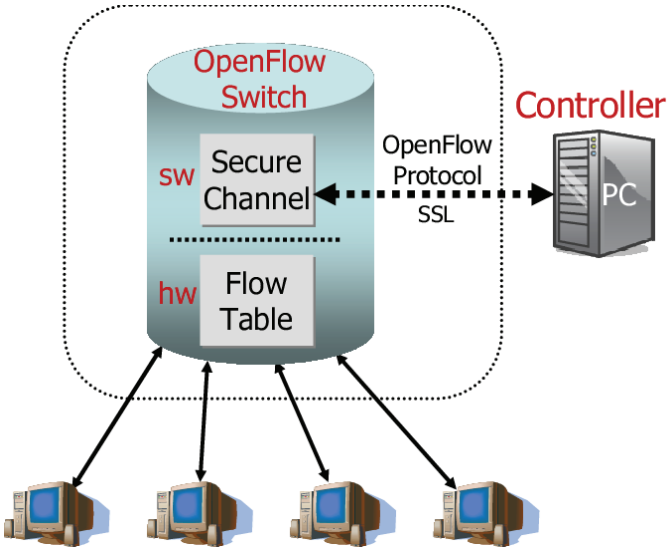


Figure 2. The basic structure of a simple SDN network,
the elements of which are connected through the OpenFlow protocol

The next level of the software-defined network is ONOS, which is an open operating system and is positioned as a network solution consisting of a service provider and its clients. ONOS is a linker for simpler SDNs, being a peculiar cluster for multiple nodes. The advantage of using ONOS as a system for a large network controller is the ability of this system to cope with system failure in case of a breach of the setting of individual nodes. In addition, the node where the addressing or connection error occurred may be promptly reconfigured by the software controller.

Considering that OpenFlow and other similar network protocols are integrated with ONOS will be incorrect, because OpenFlow is used only to establish a connection between network devices, but the ONOS operating system itself is provided with its own applications and software solutions. They are divided into tiers, each of which has its own software models (one of them interacts with OpenFlow) and can work independently. According to this level isolation, the system becomes versatile, because it is not tied to a single model or protocol and can interact at different levels with other network elements across the topology (Figure 3) [4].

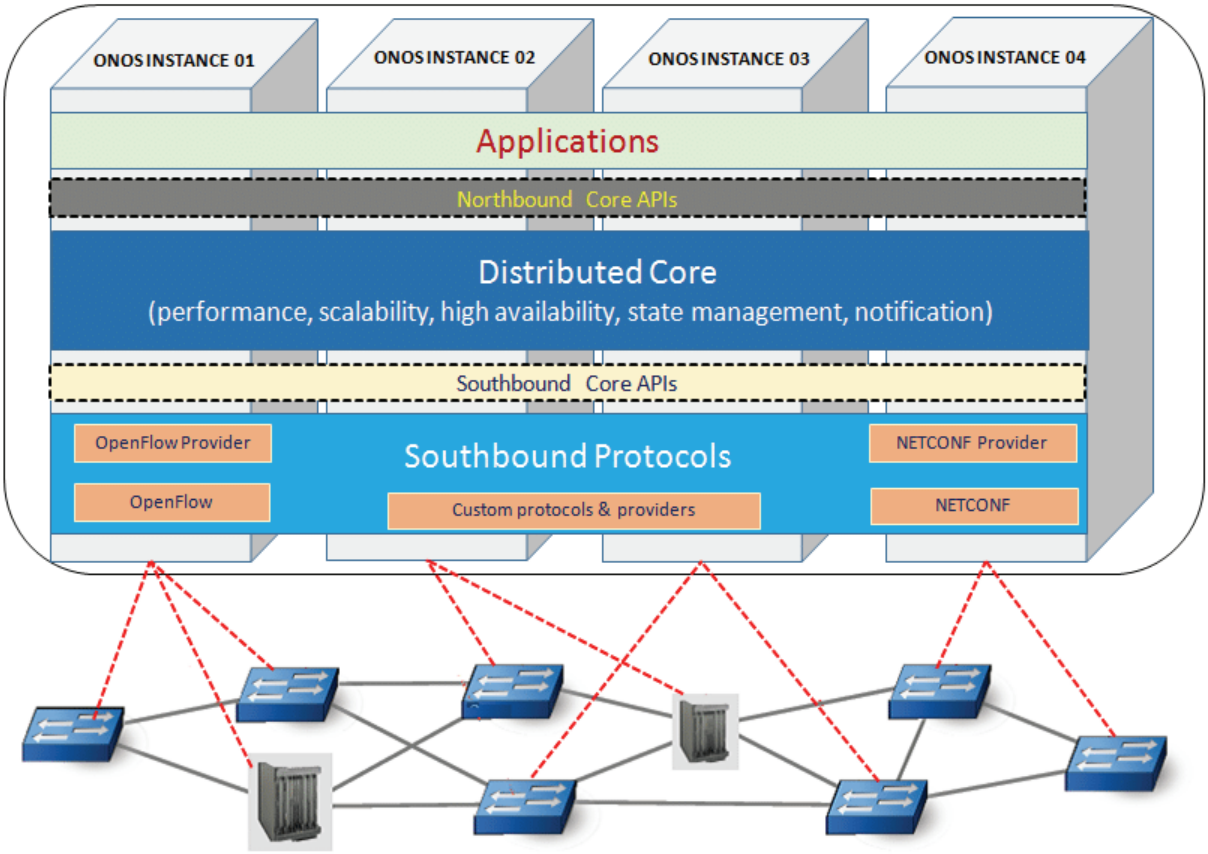


Figure 3. Dividing of the ONOS system to levels.
OpenFlow switches communicate via the OpenFlow Provider (bottom left)

Based on levels and models, ONOS has applications that allow the programmer to configure the network and control the traffic. Network topology information is stored in the control center device where the operating system is running. From the same center, you have the ability to connect directly to a particular device on the network and send it some data. This expands network control capabilities and simplifies application development and administration. Actual applications can be added or removed from the network dynamically without the need to stop traffic and disable or restart the network, which is a huge advantage for providers that require uninterrupted connection to subscribers [5].

Based on the CORD operating system, controllers are built using the CORD architectural solution, which is described below.

The highest level of the network described is the CORD technology, which is positioned as a solution for Internet service providers and implements the idea of a software-defined network on a large scale. It involves the creation of a single virtual management center that can process data dynamically, using cloud computing for this. This center communicates with intermediate elements of the network and end users. With the help of CORD, the service provider has the ability to set up a network that will be programmatically configured on remote computing power, and this provides the whole system with several important benefits - with automated control, the platform becomes simpler and faster (because it does not require human intervention, all configurations are made by the software), more flexible (because the program dynamically configures network elements, constantly synchronizing them and making necessary updates), while using cloud computing opens a perspective of decreasing the amount of physical hardware what brings tangible savings and reduce the technical complexity of the construction of the network.

CORD can be used to transfer data to three types of clients - mobile users, corporate and home users. Users are connected to intermediate data centers (one for each of the described types) and themselves are connected to a CORD controller that configures the entire network. The controller itself consists of many specialized controllers, each of which executes its own part of the software code - for example, one of these controllers can perform routing management, and another interconnection of different network levels. The CORD controller is connected to the server base, which stores the configuration of the system and performs operations that are sent to it by the controller. Additionally, it joins the core architecture of the CORD architecture, the configurable multiplexer, ROADM, which provides multiple inputs and one access output to the global network (Figure 4) [6].

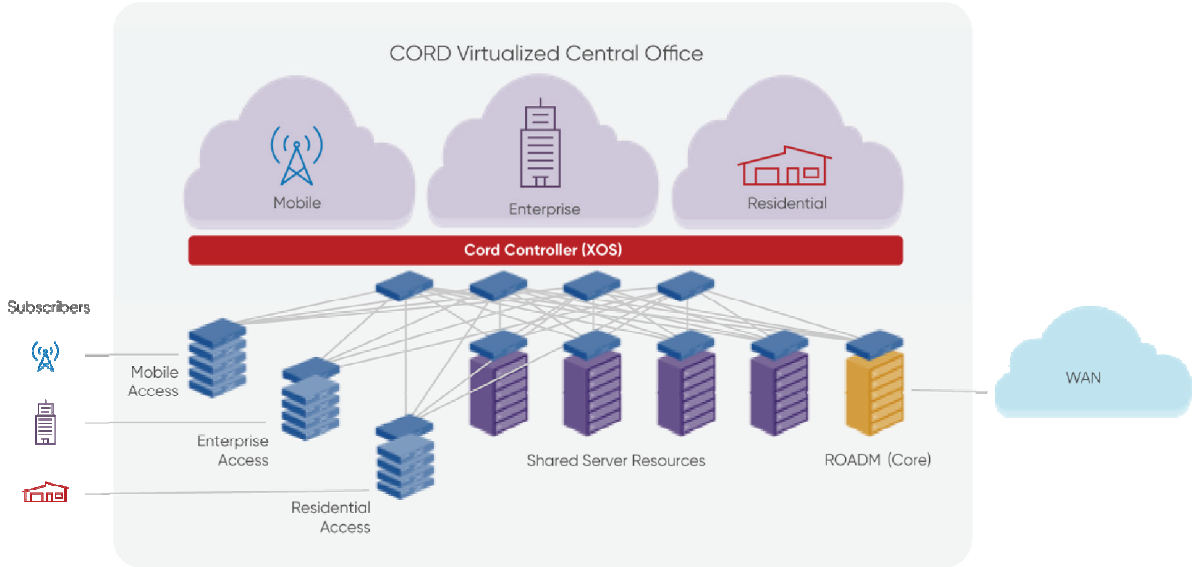


Figure 4. CORD hardware architecture

Although CORD is a large-scale solution, there are now concepts in which it is either an integral part or one of the possible solutions, for example, in the Blue Planet Platform, which offers very wide opportunities for setting up communication between the operator and users. However, these and other similar solutions are quite new and insufficiently tested so they will not be mentioned in this article.

Thus, a system that uses a general CORD-based plan, the main controllers of which are based on the ONOS operating system, which in turn relies on the OpenFlow protocol, which provides a connection to the network elements, is described. The obvious advantages of the system are the versatility described above, high scalability and stability of work, provided by automation of the administration and the exclusion of the human factor. Disadvantages of such a system until recently were the requirements for equipment and a specialist who need to be experienced with the software setup. However, today both minuses are less significant, because, for example, OpenFlow is supported in the vast majority of network devices from major manufacturers, and operators can use various devices as controllers, like not only specialized hardware, but also a fairly common, for example, a personal computer. The requirements for a specialist who need to be experienced are less necessary because of the possibilities of the ONOS system - the programs for network configuration in this system are created on rather simple commands of the Java language, therefore a qualified system administrator is enough effective to work with the controller and companies don't need a specialized programmer [7].

Conclusions. This article describes and analyzes key modern technologies that implement the concept of software-defined networks (SDN). Based on the advantages and expected disadvantages, the conclusion is made on the technical feasibility of creating OpenFlow-ONOS-CORD communication networks. Such a solution should be implemented among ISPs, as well as among diverse users, both private and corporate, this will allow companies to save material costs, since SDNs can be followed by much fewer specialists and allow users to use a stable and fast connection. According to the automation and high speed regulation and connectivity on SDN networks, it is sensibly to assume that this approach to network management will gain popularity in the future, as the number and size of networks are becoming larger and their administration needs more resources.

References

1. Kulakov, Y., Kohan A., Kopychko S. (2019). *Traffic Orchestration in Data Center Network Based on Software-Defined Networking Technology.* In Proceedings of the 2nd International Conference on Computer Science, Engineering and Education Applications (ICCSEEA2019) (pp. 228-237).

2. Лебеденко, Е. В. (2018). *Управление качеством обслуживания в системах информации на основе гистерезисного метода с двумя типами*

порогов. Диссертация на соискание учёной степени кандидата технических наук. Получено из: http://niiae.ru/components/com_chronoforms/uploads/Dissertation/ /20180427104156_%20.pdf

3. Koponen, T., Casado, M., Gude, N., Stribling, J. (2010). *Onix: A Distributed Control Platform for Large-scale Production Networks.* In Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (pp. 351-364).

4. Shah, S. A. R., Jaikar, A. (2016). *An adaptive load monitoring solution for logically centralized SDN controller*. (from 18th Asia-Pacific Network Operations and Management Symposium (APNOMS). Retrieved from: https://www.researchgate.net/ publication/309917884_An_adaptive_load_monitoring_solution_for_logically_centralized_SDN_controller

5. Lantz, B., O'Connor, B., Hart, J., Berde, P. and others. (2014). *ONOS: Towards an Open, Distributed SDN OS.* In the technical program of the ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN 2014). Retrieved from http://delivery.acm.org/10.1145/2630000/2620744/p1-berde.pdf?ip=176.37.26.167&id=2620744&acc=OPENTOC&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2EE994ED6114094BD1&_acm__=1556545733_366840876660aed76d8c4ea681cec66b

6. Open Networking Projects. (2017). *CORD.* Retrieved from https://www.opennetworking.org/cord/

7. Van der Meer, S., Grasa, E. (2016). *SDN Architectural Limitations: Towards a Full Software Network Vision.* IEEE Softwarization, A collection of short technical articles, May 2016. Retrieved from: https://sdn.ieee.org/newsletter/may-2016/sdn-architectural-limitations-towards-a-full-software-network-vision

Authors

Cherevatenko Oleksii – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: chereva@ukr.net

Череватенко Олексій Володимирович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Yurii Kulakov – professor, Doctor of Engineering Sciences, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: ya.kulakov@gmail.com

Кулаков Юрій Олексійович – професор, доктор технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

РОЗШИРЕНА АНОТАЦІЯ

Олексій Череватенко, Юрій Кулаков

АНАЛІЗ ТЕХНОЛОГІЙ РЕАЛІЗАЦІЇ МЕРЕЖ, ЩО ПРОГРАМНО КОНФІГУРУЮТЬСЯ

Актуальність теми дослідження. Проблема масштабованості комп'ютерних мереж у XXI столітті стала глобальною і вимагає автоматизації управління. Концепція програмно конфігурованих мереж (SDN) може запропонувати це і існує кілька взаємопов'язаних багаторівневих реалізацій, якими є OpenFlow, ONOS і CORD, які розглядаються в цій статті.

Постановка проблеми. До теперішнього часу не існує єдиної лаконічної документації, яка пояснювала б взаємозв’язок між різними рівнями мережі, що програмно конфігурується, та проводив би опис переваг і недоліків кожного з рішень у контексті загальної побудови мережі.

Аналіз останніх досліджень і публікацій. Наразі існує деяка кількість досліджень в області мереж, що програмно конфігуруються, а також технологій, що імплементують концепцію SDN. Однак серед виробників мережевого обладнання немає єдиного стандарту та бачення, як має виглядати практична реалізація SDN, а технології імплементації ще недостатньо вивчені компаніями сфери інформаційних технологій.

Виділення недосліджених частин загальної проблеми. У статті розглядається побудова та взаємозв’язок програмно конфігурованих мереж на базі платформи CORD, у яку входять на більш низьких рівнях система ONOS та протокол OpenFlow.

Постановка завдання. Завданням статті є визначити на основі аналізу переваг і недоліків, чи є технічно та економічно обґрунтованою побудова мережі, що програмно конфігурується, на базі зв’язки різнорівневих технологій OpenFlow–ONOS–CORD.

Викладення основного матеріалу. Описаний основний принцип SDN і технологій їх реалізації OpenFlow, ONOS і CORD. Проведено аналіз переваг і недоліків централізованої архітектури управління програмами для провайдера та клієнтів. Результати аналізу були достатньо інформативними та підтвердили очікувані висновки.

Висновки. У статті було описано та проаналізовано ключові сучасні технології, які реалізують концепцію програмно конфігурованих мереж (SDN). Зроблено висновок про технічну доцільність створення мереж з використанням зв’язки OpenFlow–ONOS–CORD. Припущено, що цей підхід буде набирати популярність у майбутньому.

Ключові слова: програмно конфігуровані мережі, OpenFlow, ONOS, CORD.

UDC 004.414.28

Andrii Boldak, Maksym Vlasov

METHOD OF IMPLEMENTATION OF SOFTWARE FOR SOLVING THE PROBLEM OF OPTIMIZATION OF CONFIGURATIONS OF THE DISTRIBUTED INFORMATION SYSTEM

The article examines how to implement software for solution of the problem of optimizing the configurations of the distributed information system. This method aims to obtain the optimal configuration in the form of a code from the input parameters.

Key words: distributed information system, DIS, optimization.

Relevance of research topic. Modern distributed information systems either already use cloud service providers, or soon begin to use them [1] [2]. Vivid examples of cloud providers are Amazon Web Services, Microsoft Azure, and the Google Cloud Platform [3]. Each of them has a lot of resources and services, with which one task can be performed even with a single provider using different services. In addition, every week new services appear or existing ones are improved, which makes it impossible to comprehend all interconnections and, as a result, obtain the optimal configuration of DIS using only limited human computing power [4].

Target setting. The problem is the lack of implementation or description of the implementation of the optimized distributed information systems.

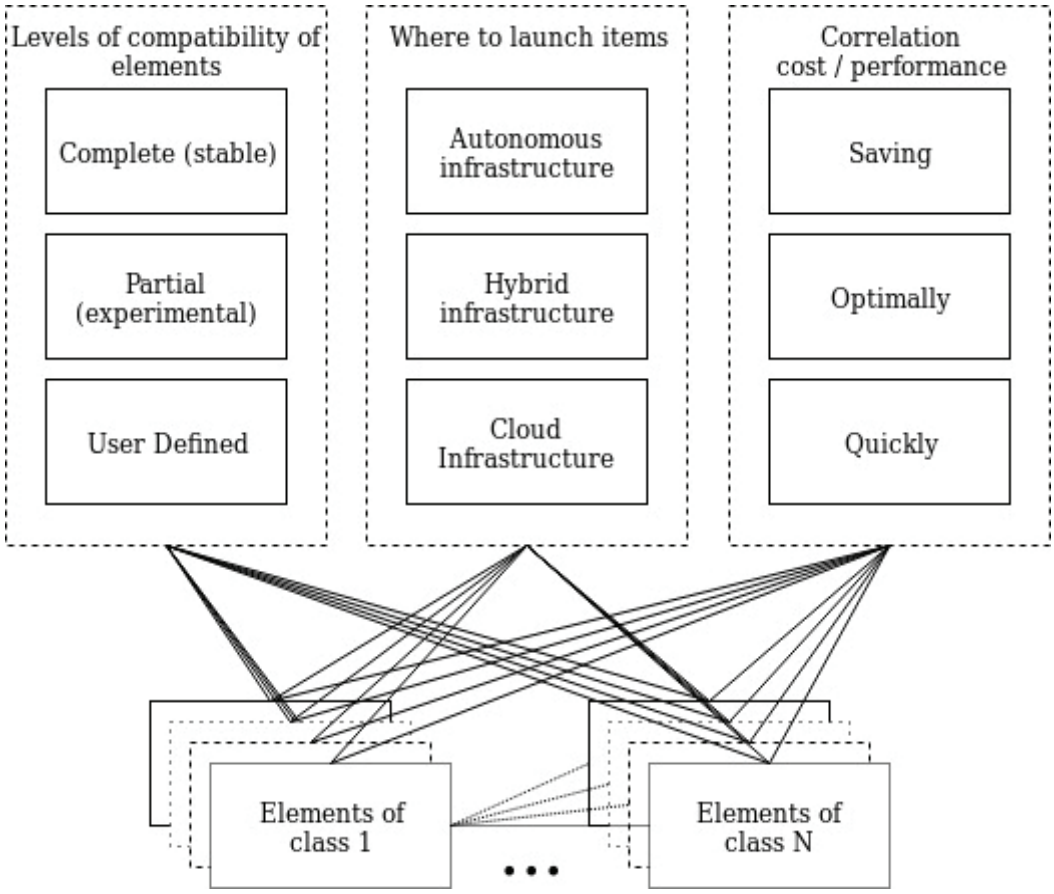
Actual scientific researches and issues analysis. Today, there is a necessary theoretical basis for optimizing DIS [5].

Uninvestigated parts of general matters defining. In this paper, we explore the method of implementing software tools to solve the RIS optimization problem.

The research objective. The task is to develop a concept for optimizing DIS configurations, which involves obtaining the description of applications and infrastructure in the form of a code [6], by which it is easy to deploy the optimal version of the distributed information system.

The statement of basic materials. The proposed concept of the implementation of software consists in finding in the informal description of the system provided by the user, the formal parts of the description of the elements, finding the coverage, which is best at the given time corresponds to the weight of the requirements [5] as shown in the picture 1. As a result of the system's operation, a descriptive system is described in HCL [7] and / or YAML [8], which are the language of the description in systems such as Terraform [9] and Docker-compose [10].

The structure of software looks like three linked modules, which are divided into functional: preprocessor, optimizer, postprocessor shownin the picture 2.



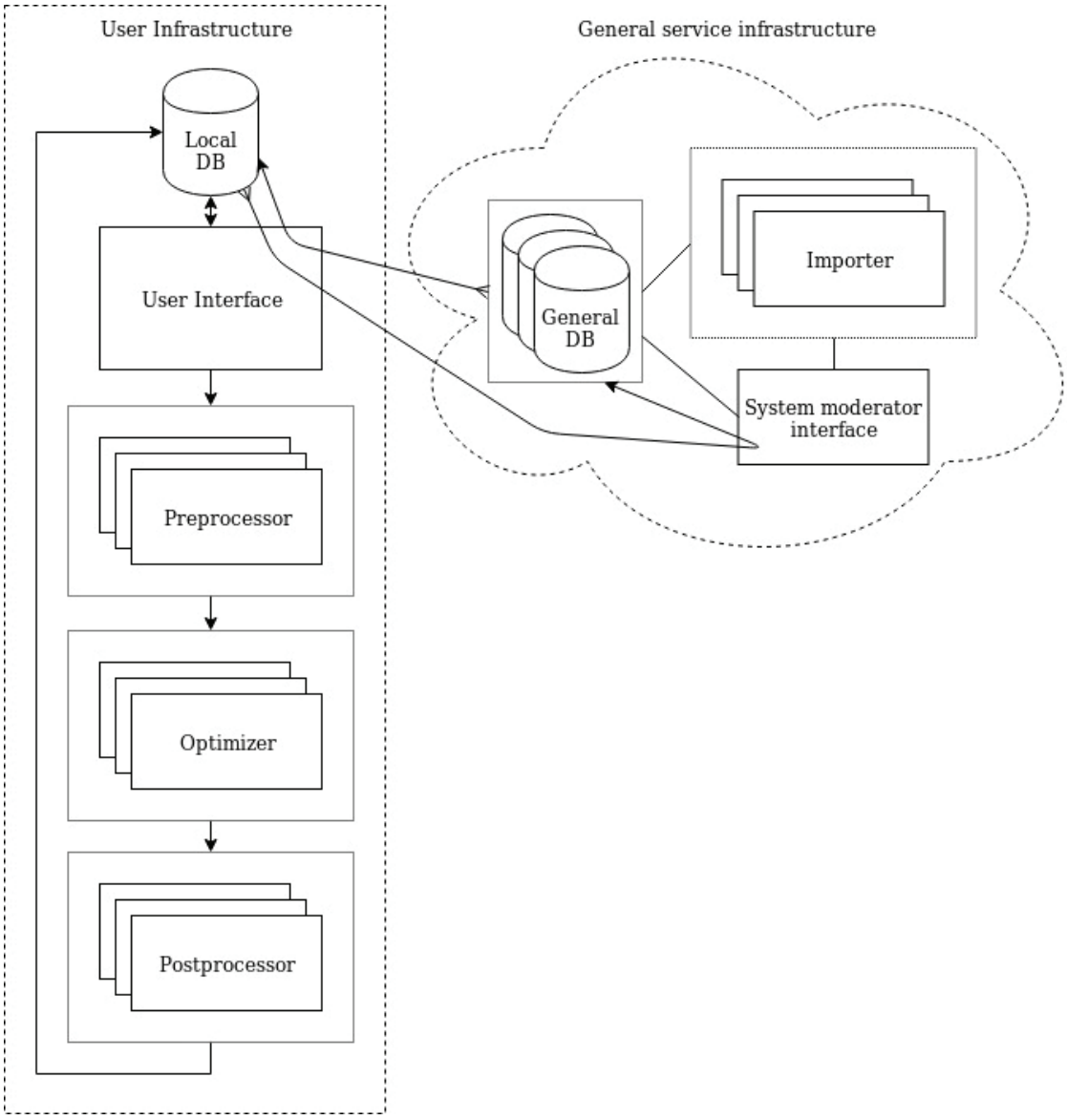
Picture 1. Block of tags

The preprocessor is responsible for data preparation. To find the formal description in the informal, first of all, it is proposed to use usecases that already have the necessary links with the formal description of the elements. To do this we use the concept of tags, and to fill formal descriptions with appropriate tags using a preprocessor with a minimum human participation - Multi-label classification [11].

The optimizer is responsible for finding the coverage and solving the optimization problem. For this purpose, already known algorithms and methods, such as ABC-analysis [12] and Quine-McCluskey [13] BOOM [14], Espresso and others, can be used. They will act as part of the formal description of the elements, looking for all possible matching combinations of elements.

The postprocessor selects the optimal parameters of each of the configurations based on the weight of each of the filters that were specified by the user and compares the results obtained. On the output, if possible, get several similar architectural projects, the user will be given the opportunity to compare the characteristics of each option and choose the one that more closely matches his goals.

At the time of choosing the desired configuration description, the user is given a code that describes its infrastructure dependencies, which allows the user to automatically deploy a system of any complexity.



Picture 2. System architecture

Conclusions. The proposed method of implementation of optimization DIS currently has no analogues and can dramatically increase the speed of development of prototype systems with the use of all the latest features and optimizations, which in turn allows the best practices to people who do not have the relevant knowledge, as well as adjust the decisions of professionals, pointing to shortcomings in their system.

References

1. Nik Simpson (Jun 2012). Comparing Data Center Costs With Public IaaS Cloud Services. URL: https://www.gartner.com/en/documents/2048419 (Access time: 28.04.2019)

2. Sanil Solanki, Michael Smith, Tomas Nielsen (Apr 2015). The Financial Case for Moving to the Cloud. URL: https://www.gartner.com/en/documents/3030826 (Access time: 28.04.2019)

3. Larry Dignan (Feb 2019). Top cloud providers 2019: AWS, Microsoft Azure, Google Cloud; IBM makes hybrid move; Salesforce dominates SaaS. URL: https://www.zdnet.com/article/top-cloud-providers-2019-aws-microsoft-azure-google-cloud-ibm-makes-hybrid-move-salesforce-dominates-saas/ (Access time: 29.04.2019)

4. Yuval Noah Harari (Sep 2014). Sapiens: A Brief History of Humankind, Chapter 10: The Scent of Money - https://www.ynharari.com/book/sapiens/ (Access time: 22.04.2019)

5. Власов М. Д., Болдак А. О. Оптимізація конфігурацій розподілених інформаційних систем URL: http://sait.kpi.ua/media/filer_public/6e/80/6e804b3f-ae13-4899-b336-4daddbd45584/sait2018ebook.pdf (Access time: 21.04.2019)

6. Kief Morris (Feb 2019). Introduction to Infrastructure Patterns URL: https://infrastructure-as-code.com/patterns (Access time: 27.04.2019)

7. Mitchell Hashimoto (Jul 2014). What is Hashicorp corporative language. URL: https://github.com/hashicorp/hcl/blob/master/README.md (Access time: 28.04.2019)

8. Ingy döt Net (Dec 2016). What is YAML. URL: https://yaml.org/ (Access time: 28.04.2019)

9. Mitchell Hashimoto (Jul 2014). What is Terraform. URL: https://github.com/hashicorp/terraform/blob/master/README.md (Access time: 28.04.2019)

10. Misty Linville (Sep 2016) Overview of Docker Compose. URL: https://docs.docker.com/compose/overview/ (Access time: 28.04.2019)

11. Kartik Nooney (Jun 2018). Deep dive into multi-label classification..! (With detailed Case Study). URL: https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff (Access time: 27.04.2019)

12. Ranganath Muttanna Singari (Feb 2014). Application of Selective Inventory Control Techniques for Cutting Tool Inventory Modeling and Inventory Reduction-A Case Study URL: https://www.researchgate.net/publication/273258286_Application_of_Selective_Inventory_Control_Techniques_for_Cutting_Tool_Inventory_Modeling_and_Inventory_Reduction-A_Case_Study (Access time: 27.04.2019)

13. Jiangbo Huang. Programing implementation of the Quine-McCluskey method for minimization of Boolean expression. URL: https://arxiv.org/ftp/arxiv/papers/1410/1410.1059.pdf (Access time: 27.04.2019)

14. PetrFiˇser, JanHlaviˇcka (Nov 2002). BOOM — A HEURISTIC BOOLEAN MINIMIZER. URL: http://www.cai.sk/ojs/index.php/cai/article/viewFile/450/356 (Access time: 25.04.2019)

Autors

Boldak Andrii – Candidate of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: boldak.andrey@gmail.com

Болдак Андрій Олександрович – кандидат технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Vlasov Maksym – Student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: m.vlasov@post.com

Власов Максим Дмитрович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

UDC 004.7

Dmytro Zhyzhko,
Simonenko Valery

THE ALGORITHM OF DYNAMIC DISTRIBUTION OF TASKS FOR A CLUSTER SYSTEM

Жижко Дмитро,
Сімоненко Валерій

АЛГОРИТМ ДИНАМІЧНОГО РОЗПОДІЛУ ЗАДАЧ ДЛЯ КЛАСТЕРНОЇ СИСТЕМИ

The article describes the principle of the algorithm for dynamic loading of cluster nodes and compares its work with existing solutions. Was analyzed it work and efficiency in various types of tasks, the ability to balance and distribute the flow of tasks between the system's kernels. A software model for testing was created. The results of the analysis are presented.

Key words: claster, distribution of tasks, planning algorithms.
Fig.: 6. Bibl.: 5.

У статті наведено принцип роботи алгоритму динамічного завантаження вузлів кластера та порівняно його роботи відносно існуючих рішень. Було проаналізовано його роботу та ефективність при різних типах задач, здатність балансувати та розподіляти потік завдань між ядрами системи. Створено програмну модель для тестування. Наведено результати аналізу.

Ключові слова: кластер, розподіл задач, алгоритми планування.
Рис.: 6. Бібл.: 5.

Relevance of research topic. In our time, computer technology is in almost all spheres of human activity. This, in turn, forces the specialists in the field of information and computer technologies to create new tools that would be able to perform tasks quickly and qualitatively. Every year the software becomes more flexible and perfect. Such programs consume more hardware resources. Initially, such a task was solved by an increase in clock speed of the processor, but now in most cases it is already impossible to do this because a certain limit is reached, or it is very expensive. An alternative is to use several identical processors that can perform different parts of one task in parallel. This approach makes it possible to significantly reduce the time of the task. Systems that have many computing modules are called clusters or grid systems. They have many advantages over noncommercial computers. The main advantage is that they allow you to implement parallel execution and

multitasking. Modern datacenters have thousands and millions of processors in their structure, which allows them to serve a large number of users. On the other hand, the end user is able to use these resources and not spend money on the purchase of personal machines [1, 2, 3].

Formulation of the problem. The increase in the number of computing units and the load on them on the one hand and the price of components on the other hand compel system software developers to create task schedulers that would allow them to achieve significant efficiency. The most important component of the task manager is the planning algorithm. For a few reasons, the purpose of this article is to develop such a planning algorithm that would enable a sufficient level of efficiency of the cluster system. An existing algorithm based on static scheduling gives average performance values of 70% - 90%, the value of which can be significantly reduced with considerable granularity of the computational task. The proposed algorithm is designed to solve this problem.

Analysis of recent research and publications. Since there has been a tendency in the last years to increase the number of cores in processors, as well as the growing popularity of technology, which provides the basis for scientific research on the topic of multi-threaded computing. A number of algorithms and strategies have been created that allow efficient allocation of resources of the cluster system. For an example, the DRF strategy can be called [4]. It makes it possible to distribute resources among users depending on their quotas. But this algorithm works at a higher level of abstraction, which does not allow it to influence the distribution of tasks between processor cores [5].

Identification of unexplored parts of the general problem. The main disadvantage of existing algorithms is that they have static scheduling using queues for each computing unit. With great granularity of tasks it leads to significant losses of efficiency. Also, the use of additional queues of tasks requires more system resources.

Setting objectives. The purpose of the research is to develop a new planning algorithm that would increase the efficiency of the cluster for different types of tasks.

Presentation of the main material. The algorithm of dynamic distribution of tasks for a cluster system.

For experiments and comparisons, a static planning algorithm will be considered. It is based on the following approach. The scheduler transmits a task that consists of n-th number of subtasks. We also have a computing system that includes k processors or cores. The algorithm will distribute all available subtasks between the kernels equally. Until all subtasks are executed, the system will not start the next task. Schematically, the operation of this algorithm can be seen in Fig. 1.

Its main disadvantage is that scheduling occurs only at the stage of cluster loading. If each of the subtasks has a different execution time, then after a certain period of time, one kernel will complete its work, while others do not. This results in lower efficiency. Also, the use of local queues for each core requires certain resources, and if, with this number of these cores are large, then we will have significant losses of RAM.

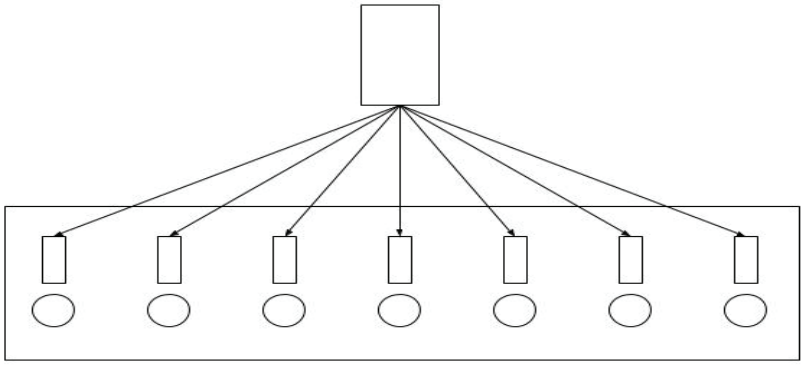


Fig. 1. The structure of the static planning algorithm

The proposed dynamic planning algorithm is devoid of these disadvantages. Planning occurs throughout the time of the task. It does not use local queues of subtasks, but has one common queue. The distribution of tasks occurs dynamically. This allows you to effectively distribute cluster resources. With this approach, the scheduler saves the queue of tasks locally. The kernels have no queues. When the subtask is completed, the scheduler gives the kernel a new task to execute. Schematically it can be depicted as in Fig. 2.

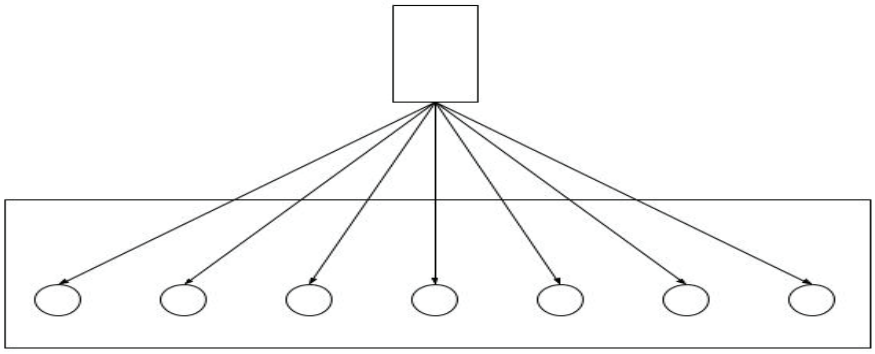


Fig. 2. The structure of the dynamic planning algorithm

To compare the algorithm's work, a number of experiments were carried out, among which measurements of efficiency at different types. In the first experiment, 1000 tasks were generated with execution times from 1 to 10. Efficiency was 0.9 and 0.99 for static and dynamic planning methods, respectively. This is explained by the fact that the static algorithm greatly depends on the granularity of the tasks, and the dynamic is devoid of this disadvantage.

For the second experiment, tasks with the same execution time were generated. The result of the planning for both algorithms was the same, and the efficiency reached 1.

In the third experiment, the dependence of efficiency on the number of processors for each of the algorithms will be analyzed. The results showed that, with increasing number of kernels, the efficiency of each algorithm decreases, but the performance of the second algorithm is better.

In the fourth experiment, the dependence of efficiency from the runtime range on each subtask was analyzed. The test system had 10 cores and 1000 subtasks. The results can be seen in Fig. 3 and fig. 4. Based on them, one can conclude that the efficiency of each of the algorithms does not depend on the range of the time of the task.

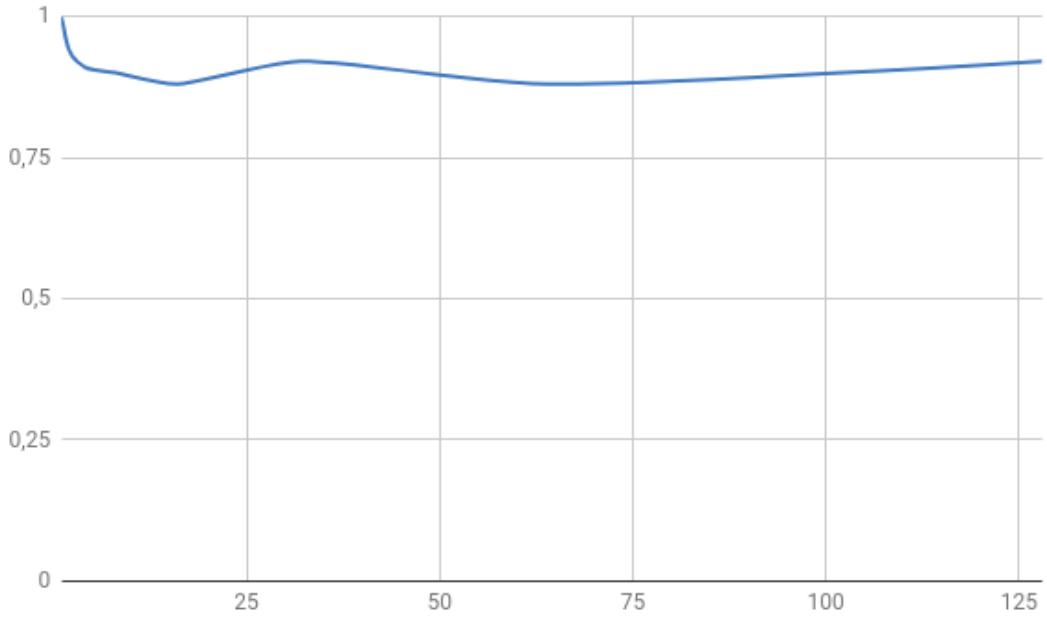


Fig. 3. The results of the fourth experiment for the first algorithm

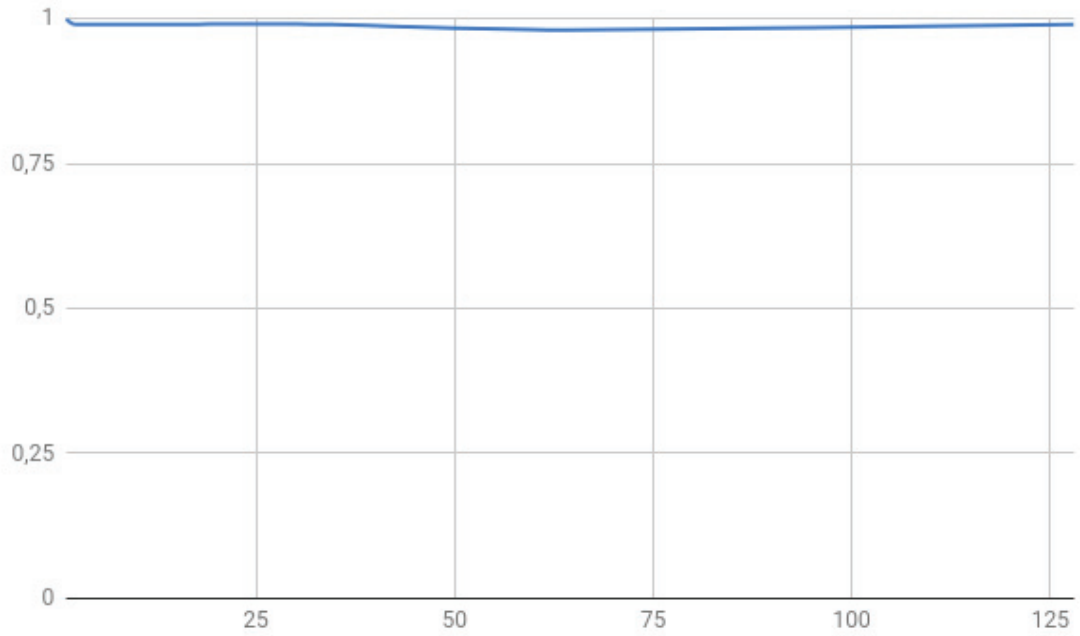


Fig. 4. The results of the fourth experiment for the second algorithm

In the fifth experiment, the dependence of efficiency on the number of tasks in the system was compared. The system has 10 nuclei, the time range for one task is from 1 to 10. The results can be seen in Fig. 5 and rice 6. If the number of tasks is less than the number of kernels in a cluster, then the efficiency is independent of the

algorithm and has a rather low value. When increasing the number of tasks, the efficiency increases for both algorithms.

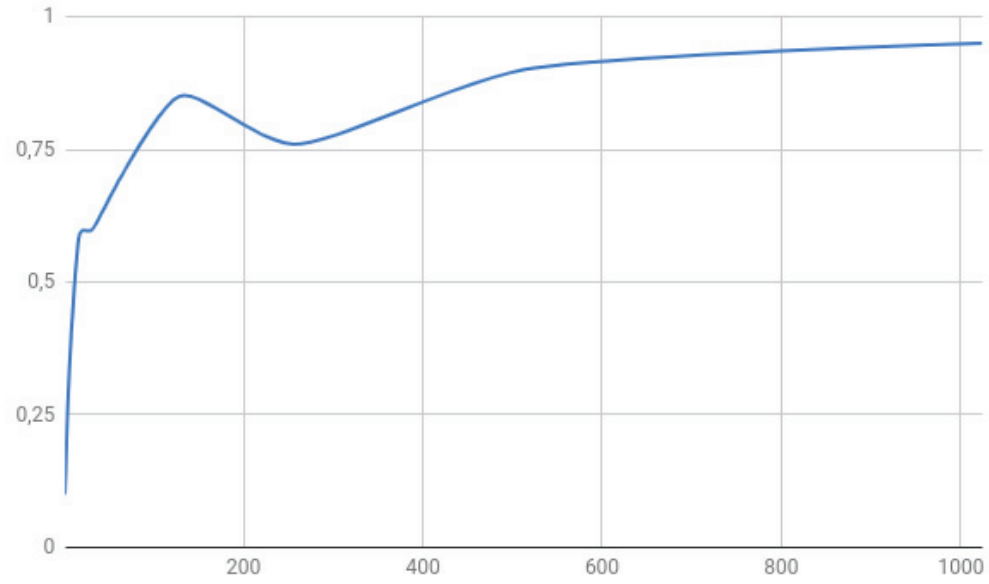


Fig. 5. Results of the fifth experiment for the first algorithm

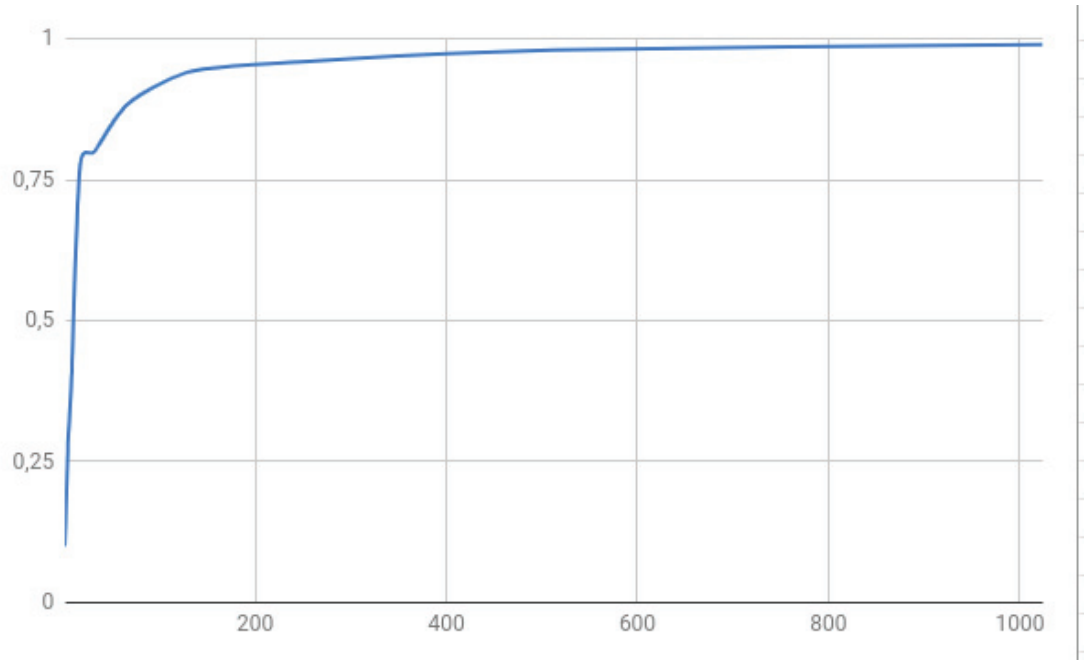


Fig. 6. The results of the fifth experiment for the first algorithm

Conclusions. Nowadays, multithreading calculations, and therefore algorithms for distributing tasks between them, are used more and more often, which suggests that this topic is very relevant. Two algorithms of task planning were analyzed in the work. The proposed dynamic planning algorithm allows you to significantly increase efficiency, so it can be used in modern systems. The average efficiency of the static algorithm was 0.8 - 0.9, and the dynamic - 0.9 - 1, that is, you can get an increase to 20%.

Список використаних джерел

1. Computer cluster [електронний ресурс]. — Режим доступу: https://en.wikipedia.org/wiki/Computer_cluster

2. Distributed computing [електронний ресурс]. — Режим доступу: https://en.wikipedia.org/wiki/Distributed_computing

3. Introduction to Parallel Computing [електронний ресурс]. — Режим доступу: https://computing.llnl.gov/tutorials/parallel_comp/

4. Maui Cluster Scheduler [електронний ресурс]. — Режим доступу: http://www.adaptivecomputing.com/products/maui/

5. Moab HPC Suite – Basic Edition [електронний ресурс]. — Режим доступу: http://www.adaptivecomputing.com/support/download-center/moab-hpc-suite-basic-edition-download/

ДОВІДКА ПРО АВТОРІВ

Сімоненко Валерій Павлович — професор, доктор технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

Simonenko Valery Pavlovich — Professor, Doctor of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine «Kiev Polytechnic Institute. Igor Sikorsky»

E-mail: svp@comsys.kpi.ua

Жижко Дмитро Сергійович — студент 6 курсу кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут ім. Ігоря Сікорського»

Dmytro Zhyzhko Serhiiovych — 6th student of the Department of Computer Engineering, National Technical University of Ukraine «Kiev Polytechnic Institute. Igor Sikorsky»

E-mail: dmytro110101@gmail.com

РОЗШИРЕНА АНОТАЦІЯ

Жижко Дмитро,
Сімоненко Валерій

АЛГОРИТМ ДИНАМІЧНОГО РОЗПОДІЛУ ЗАДАЧ ДЛЯ КЛАСТЕРНОЇ СИСТЕМИ

Актуальність теми дослідження. У наш час комп'ютерні технології є майже у всіх сферах людської діяльності. Це, у свою чергу, змушує фахівців в області інформаційних та комп'ютерних технологій створювати нові інструменти, які могли б швидко і якісно виконувати завдання. Щороку програмне забезпечення стає більш гнучким і досконалим. Такі програми споживають більше апаратних ресурсів. Спочатку таке завдання вирішувалося збільшенням тактової частоти процесора, але тепер у більшості випадків це вже неможливо зробити, тому що досягнута певна межа, або це дуже дорого. Альтернативою є використання декількох ідентичних процесорів, які можуть виконувати різні частини однієї задачі паралельно. Такий підхід дає можливість значно скоротити час виконання завдання. Системи, які мають багато обчислювальних модулів, називаються кластерами або сітковими системами. Вони мають багато переваг перед некомерційними комп'ютерами. Основна перевага полягає в тому, що вони дозволяють здійснювати паралельне виконання і багатозадачність. Сучасні центри обробки даних мають тисячі і мільйони процесорів у своїй структурі, що дозволяє їм обслуговувати велику кількість користувачів. З іншого боку, кінцевий користувач може використовувати ці ресурси і не витрачати гроші на придбання особистих машин.

Цілі дослідження. Збільшення кількості обчислювальних одиниць і навантаження на них, з одного боку, і ціни компонентів, з іншого боку, змушують розробників системного програмного забезпечення створювати планувальники завдань, що дозволить їм досягти значної ефективності. Найважливішим компонентом менеджера завдань є алгоритм планування. З кількох причин, метою цієї статті є розробка такого алгоритму планування, який дозволить забезпечити достатній рівень ефективності кластерної системи. Існуючий алгоритм, заснований на статичному плануванні, дає середні показники продуктивності 70% -90%, величина яких може бути значно зменшена при значній деталізації обчислювального завдання. Запропонований алгоритм призначений для вирішення цієї проблеми.

Аналіз актуальних наукових досліджень та питань. Оскільки в останні роки спостерігається тенденція до збільшення кількості ядер у процесорах, а також зростаючої популярності технологій, що є основою для наукових досліджень на тему багатопотокових обчислень. Створено ряд алгоритмів і

стратегій, що дозволяють ефективно розподіляти ресурси кластерної системи. Наприклад, можна назвати стратегію ДРФ. Це дає можливість розподіляти ресурси між користувачами залежно від їх квот. Але цей алгоритм працює на більш високому рівні абстракції, що не дозволяє впливати на розподіл завдань між ядрами процесорів.

Невивчені частини загальних питань визначають. Основним недоліком існуючих алгоритмів є те, що вони мають статичне планування з використанням черг для кожного обчислювального блоку. З великою деталізацією завдань це призводить до значних втрат ефективності. Крім того, використання додаткових черг завдань вимагає більшої кількості системних ресурсів.

Мета дослідження. Метою дослідження є розробка нового алгоритму планування, який би підвищив ефективність роботи кластера для різних типів завдань.

Виклад основних матеріалів. Проведено порівняння алгоритмів статичного та динамічного планування. Розроблено модель для тестування. Проведено моделювання для різних типів завдань. Динамічний алгоритм показав хороші показники.

Висновки. Сьогодні все частіше використовуються багатопоточні розрахунки і, отже, алгоритми розподілу задач між ними, що говорить про те, що ця тема дуже актуальна. У роботі проаналізовано два алгоритми планування завдань. Запропонований алгоритм динамічного планування дозволяє істотно підвищити ефективність, тому його можна використовувати в сучасних системах.

Ключові слова: кластер, розподіл завдань, алгоритми планування.

UDC 004.72

Oleksandr Dolynnyi,
Alla Kohan

METHOD OF SDN CLUSTERING USING CONNECTIONS DENSITY DISTRIBUTION

Олександр Долинний,
Алла Коган

СПОСІБ КЛАСТЕРИЗАЦІЇ МЕРЕЖІ SDN З УРАХУВАННЯМ РОЗПОДІЛУ ЩІЛЬНОСТІ ЗВ'ЯЗКІВ

The paper proposes method of SDN clustering using connections density distribution that solves the problem of controller load balancing.

Key words: SDN, network clustering, controller placement problem.
Fig.: 2. Tabl.: 1. Bibl.: 3.

У роботі запропоновано спосіб кластеризації мережі SDN з урахуванням розподілу щільності зв’язків, що вирішує проблему рівномірного навантаження контролерів.

Ключові слова: SDN, кластеризація мережі, задача розташування контролера.
Рис.: 2. Табл.: 1. Бібл.: 3.

Relevance of research topic. The efficient SDN management is a pending problem because of the wide spread of such networks and their increase in scale during recent years. To solve this problem we need to define network structure efficiently, that is, to split the network in to a number of subnetworks and to place a controller at the optimal position in each of them.

Formulation of the problem. As SDN separates control and data planes, the problem of controller placement arises, that is, how many controllers should be in a network and how exactly they should be placed. In this paper we propose a method of controller placement based on connections density distribution.

Analysis of recent research and publications. Some works have proposed heuristic algorithms to solve the controller placement problem and defined this as multi-objective combinatory optimization task. However, for a large-scale network such algorithms demonstrate an unreasonably high solution time [1].

On the other hand, there exists an approach in which controller placement is determined by certain defined metrics.

One such method is density based controller placement (DBCP), which uses nodes

clustering algorithm based on density distribution. These nodes have higher quantity of connectivity inside the cluster and lower quantity of connectivity with nodes from other networks, thus there is one and only one controller in each subnetwork [2].

Selection of unexplored parts of the general problem. In a real life network when performing clustering the task of controller load balancing arises. This task has not been solved in the original DBCP algorithm.

Target setting. We proposed a modification of DBCP algorightm so as to ensure the balanced controller load.

The statement of basic materials.

The density based clustering algorithm three main steps: density analysis, density based clustering and controller placement. The metrics that is introduced in the modified algorithm for the choice of controller's position is the index of proximity to the cluster's margin.

In general, the clustering network algorithm based on density distribution consists of three main steps:

1) analyze the distribution local density distribution throughout network of routers;

2) according to the found values of density and distance from the current node to the node with a higher density, divide the network routers in clusters;

3) solve the task of selecting a placement for one controller in each subnetwork according to the given criterion.

During the first stage, the topology of the network is analyzed for the presence of higher connected groups of routers. For each router, the density of the local connections d and the distance r to the router with the higher value of the local density is calculated.

The metric that is additionally introduced in the modified algorithm is the index of proximity to the boundary of the cluster s_{zp} to distinguish vertices with similar local densities of nodes. Let the router V have a local density d. For each router V_i we define the set of neighboring routers $N(Vi)$, each of them has a greater or equal local density. Then, according to the formulas of the information entropy theory [3], the index of proximity to the cluster boundary can be calculated by the following formula:

$$s_{\text{гр}\,i} = \sum_{j}^{N(Vi)} \frac{d\,j}{D} \log \frac{d\,j}{D}, \text{ де } D = \sum_{j}^{N(Vi)} d\,j$$

Then, during the second stage of clustering, the cluster assignment procedure for the current node needs to be corrected. As the firstly allocated cluster was defined as closest to the current vertex, then we will sort the vertices by increasing the proximity index to the cluster boundary in modified algorithm.

For all vertices included in the $N(Vi)$ set, we sort by the local density value in descending order, and we try to attach a vertex to a cluster that contains a router with a higher local density value and contains the ability to expand.

Figure 1 examines the example of the network, while the simulation has set the limit of 6 vertices to the size of the cluster. The simulation results are reproduced in Figure 2 and in Table 1.

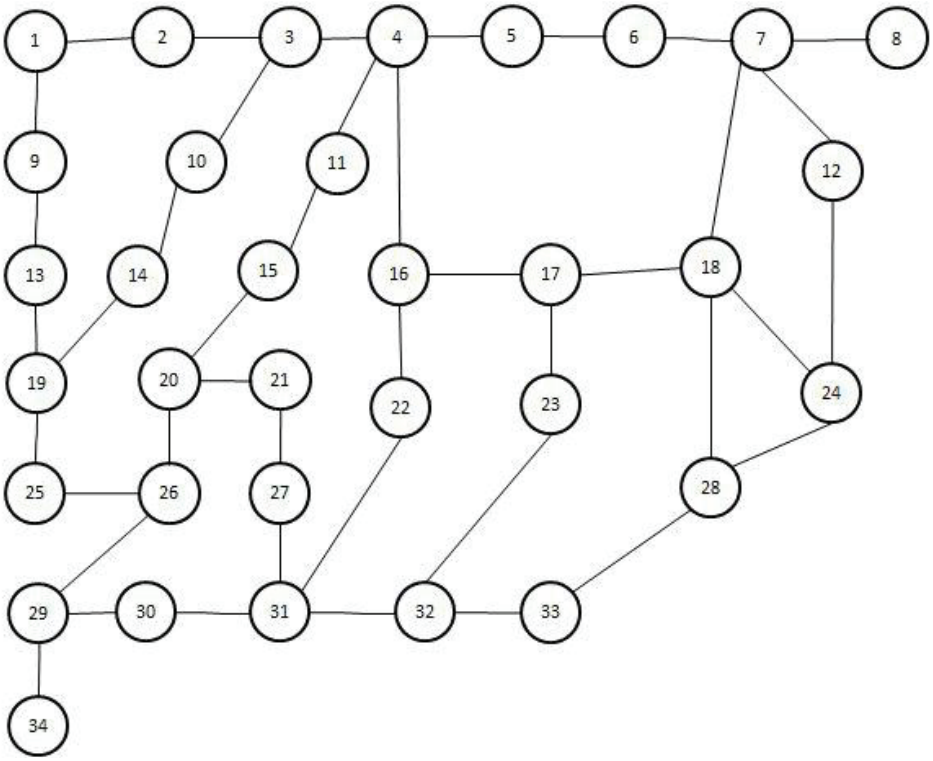


Fig. 1. Initial network graph

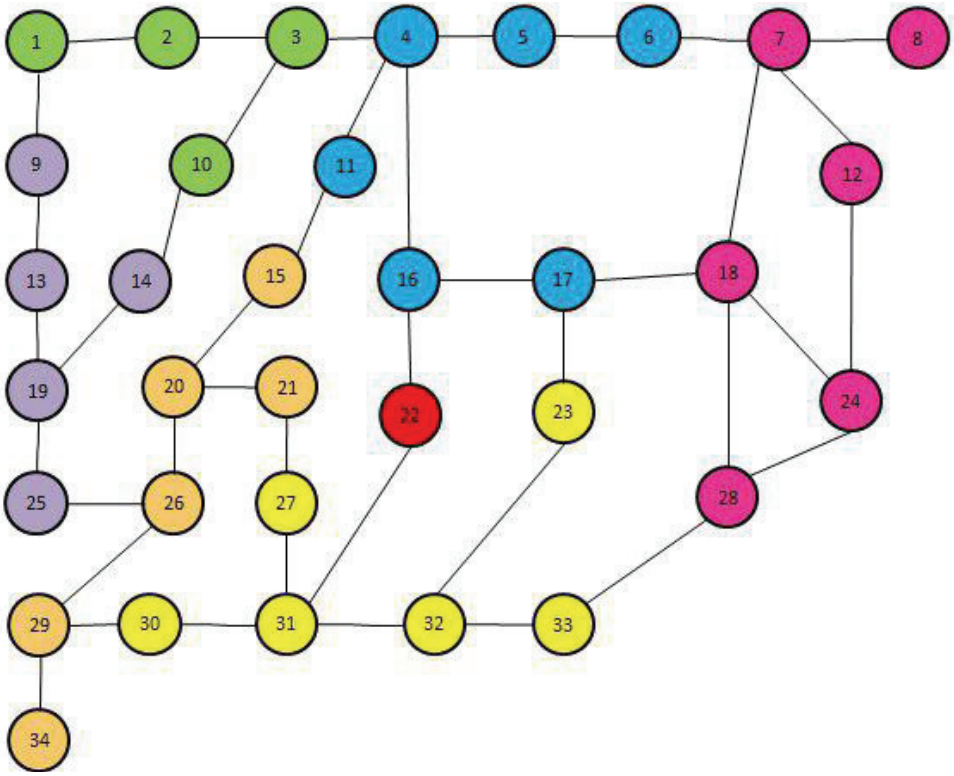


Fig. 2. Clustering algorithms' results considering maximum load limitation

Table 1

Clustering algorithms' results considering maximum load limitation

Node Id	***d***	***r***	s_{zp}	***Node Id***	***d***	***r***	s_{zp}
1	4	1	0.91	18	9	3	0.0
2	8	1	0.0	19	6	2	0.0
3	8	1	0.99	20	7	1	0.0
4	10	3	0.0	21	5	1	0.99
5	6	1	0.95	22	7	1	1.0
6	6	1	0.98	23	6	1	1.0
7	8	1	0.0	24	6	1	0.98
8	4	1	0.0	25	6	1	0.98
9	4	1	0.99	26	8	3	0.0
10	5	1	0.96	27	6	1	0.0
11	6	1	0.0	28	7	1	0.0
12	6	1	0.98	29	6	1	0.99
13	5	1	0.0	30	7	1	0.0
14	5	1	0.99	31	9	3	0.0
15	5	1	0.99	32	8	1	0.0
16	9	1	0.0	33	6	1	0.99
17	8	1	1.0	34	3	1	0.0

Conclusion. In our paper we have proposed an algorithm of network clustering using connections density distribution. An example of the algorithm's work has been demonstrated. The use of the proposed algorithm will allow to increase network clustering efficiency taking into account controller load.

References

1. Jianxin Liao, Haifeng Sun, Jingyu Wang, Qi Qi, Kai Li, Tonghong Li. (2017). *Density cluster based approach for controller placement problem in large-scale software defined networkings*. In Computer Networks (Vol. 112 – pp. 24–35).

2. A. Sallahi , M. St-Hilaire. (2015). *Optimal model for the controller placement problem in software defined networks*. In IEEE Commun. Lett. (Vol. 19 (1) – pp. 30–33).

3. A. Clauset, M.E. Newman, C. Moore. (2004). *Finding community structure in very large networks. In Physical Revolution*. (Vol 70 (6) – pp. 264–277).

Autors

Dolynnyi Oleksandr – student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Долинний Олександр Валерійович – студент кафедри обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

Коган Алла Вікторівна – старший викладач, кафедра автоматизованих систем обробки інформації і управління, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Kohan Alla – Senior lecturer, Department of Computer-Aided Management And Data Processing Systems, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

E-mail: a.v.kohan433@gmail.com

EXTENDED ABSTRACT

Oleksandr Dolynnyi, Alla Kohan

METHOD OF SDN CLUSTERING USING CONNECTIONS DENSITY DISTRIBUTION

Relevance of research topic. Актуальність теми дослідження. The efficient SDN management is a pending problem because of the wide spread of such networks and their increase in scale during recent years. To solve this problem we need to define network structure efficiently, that is, to split the network in to a number of subnetworks and to place a controller at the optimal position in each of them.

Formulation of the problem. As SDN separates control and data planes, the problem of controller placement arises, that is, how many controllers should be in a network and how exactly they should be placed. In this paper we propose a method of controller placement based on connections density distribution.

Analysis of recent research and publications. Some works have proposed heuristic algorithms to solve the controller placement problem. However, for a large-scale network such algorithms demonstrate an unreasonably high solution time. On the other hand, there exists an approach in which controller placement is determined by certain defined metrics. One such method is density based controller placement (DBCP).

Selection of unexplored parts of the general problem. In a real life network when performing clustering the task of controller load balancing arises. This task has not been solved in the original DBCP algorithm.

Target setting. We proposed a modification of DBCP algorightm so as to ensure the balanced controller load.

The statement of basic materials. The density based clustering algorithm three main steps: density analysis, density based clustering and controller placement. The metrics that is introduced in the modified algorithm for the choice of controller's position is the index of proximity to the cluster's margin.

Conclusion. In our paper we have proposed an algorithm of network clustering using connections density distribution. An example of the algorithm's work has been demonstrated. The use of the proposed algorithm will allow to increase network clustering efficiency taking into account controller load.

UDC 004.724.4

Iryna Larina, Alla Kohan

TRAFFIC ENGINEERING METHOD FOR SOFTWARE-DEFINED NETWORKS

Ларіна Ірина Сергіївна,
Коган Алла Вікторівна

СПОСІБ КОНСТРУЮВАННЯ ТРАФІКА В ПРОГРАМНО-КОНФІГУРОВАНИХ МЕРЕЖАХ

The paper considers traffic engineering (TE) in software-defined networks (SDN). A brief overview of the features of the SDN that allows improving of traffic engineering efficiency is presented. Given the features of the SDN technology, we propose a traffic engineering method that reduces packet loss and forwarding delay through dynamic re-routing. An example of the traffic engineering process in the conditions of an estimated link load changes is given.

Keywords: traffic engineering, dynamic re-routing, software-defined networks.
Fig. 2. Bible: 8.

У роботі розглянуті питання конструювання трафіка (TE) в програмно-конфігурованих мережах (SDN). Наведений короткий огляд особливостей SDN, які позволяють підвищити ефективність конструювання трафіка. З урахуванням особливостей технології SDN запропонований спосіб конструювання трафіка, який дозволяє знизити втрату пакетів та затримку пересилання даних за рахунок динамічної ремаршрутизації. Наведено приклад моделювання процесу конструювання трафіка в умовах прогнозованої зміни завантаження каналів зв'язку.

Ключові слова: конструювання трафіка, динамічна ремаршрутизація, програмно-конфігуровані мережі.
Рис.:2. Бібл.: 8.

Relevance of research topic. Modern trends in the rapid growth of cloud computing, Big Data, the Internet of Things and other technologies require greater flexibility and speed from computer networks. Also, modern networks are characterized by their large size and heterogeneity of equipment from different vendors. This complicates the task of traffic engineering (TE), which aims to optimize the network. One of the main tasks of traffic engineering for large networks is the task of load balancing of the network links.

Target setting. In order to solve the problem of traffic engineering, the use of technology of software-defined networks (SDN) is becoming promising, as it allows easier traffic management and more effective utilization of network resources [1]. Unlike traditional computer networks, SDN has a global view of the network state and can control its processes in a flexible way. This is achieved with the help of a centralized controller, which is responsible for the configuration and management of the network at the software level.

Analysis of recent research and publications. In [2] an overview of different traffic engineering methods for SDN is presented. The advantages of using SDN technology to solve the problem of load balancing are given.

In order to increase the efficiency of traffic management in networks, multipath routing methods are widely used. A centralized generation of multiple paths in SDN based on multipath routing can reduce traffic engineering time and improve the quality of service (QoS) [2]. Equal-Cost-Multi-Path (ECMP) algorithm is a popular solution for multipath routing in SDN, however it does not consider the load of network resources [3].

In [4] and [5] authors propose to determine big traffic flows that are routed on the least loaded paths. In [6] authors suggest distributing traffic based on pre-calculated optimal and suboptimal routes. These methods, however, lack sufficient flexibility to be able to dynamically respond to changes in network load. This is taken into account in [7], whose authors propose a traffic engineering method with the ability to perform dynamic re-routing.

Defining the unexplored parts of the general problem. The dynamic nature of the networks requires a quick response to changes in the state of network links. Therefore, the task is to develop a method for traffic engineering with the ability to perform dynamic re-routing to avoid links overloading.

The research objective. The task is to develop a method for traffic engineering in software-defined networks with the ability to dynamically reroute packets for predicted download of communication lines.

Statement of the main material. Statement of the main material. In this paper, we propose a method for traffic engineering that allows performing load balancing for links in the network

The calculation of the paths takes place centrally in the SDN controller. With the help of the modified wave algorithm [8] a set of all paths is formed between different nodes of the network. The advantage of this algorithm is that when paths are formed between two nodes of the network, paths between nodes arranged between them are formed simultaneously. The path information is stored in the routing tables on the network switches that are updated by the SDN controller. Since the SDN controller has a global view of the entire network, it can quickly update route information on switches.

The task that is to be solved during traffic engineering depends on the choice of the metric which will be used by controller to determine the best routes. This can be the length of the path, the delay time, the load capacity of the channel, and so on. In this paper, as a metric, the link utilization at the given time, as well as the predicted link utilization is used. The link utilization for link d_i is calculated by formula:

$$d_i = \frac{V_{tr}}{B_i}, \quad (1)$$

where *Vtr* is traffic volume, and *Bi* is link's bandwidth.

Forecasted link utilization is defined as maximum value of the forecasted link utilization change over a given time period Δt. We consider that SDN has the forecasted statistic of link utilization change over some time, for example, over a day.

When traffic enters the network, it is routed on the best path. The best path is defined as a path that has a minimum current utilization, and the forecasted utilization for which is not the maximum among other alternative paths. The SDN controller updates the route metric with some specified frequency. If the current link utilization value or its forecasted value reaches the maximum, then there is a need for re-routing to avoid overloading the channel, otherwise network performance will be degraded.

If the traffic has been assigned to a route in which the maximum utilization is observed or expected, another optimal path will be selected for the transfer of the next packet. Consider the re-routing process as an example. Let us consider a traffic flow that needs to be forwarded from node 1 to 5(Fig.1).

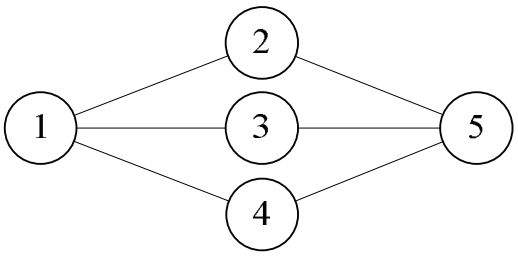


Fig.1. A graph for a part of network topology

Consider that the traffic flow was assigned to the route 1-2-5, for which a full utilization of link 2-5 is forecasted. A new route needs to be determined to re-route the traffic in order to avoid packet loss. For this, the SDN controller, knowing the required packet transfer time *t*, compares the forecasted utilization of alternatives routes over *t* (Fig.2).

Route which has the minimum link utilization over time t among other routes will be chosen as an optimal route. In the example above, this is route 1-4-5. Although the current utilization for this route is greater than the one for route 1-3-5, if traffic is sent there, there is a possibility of overloading the route, which will lead to degrading of network performance.

This allows us to distribute traffic more evenly and to avoid future need for additional re-routing because of the maximum link utilization.

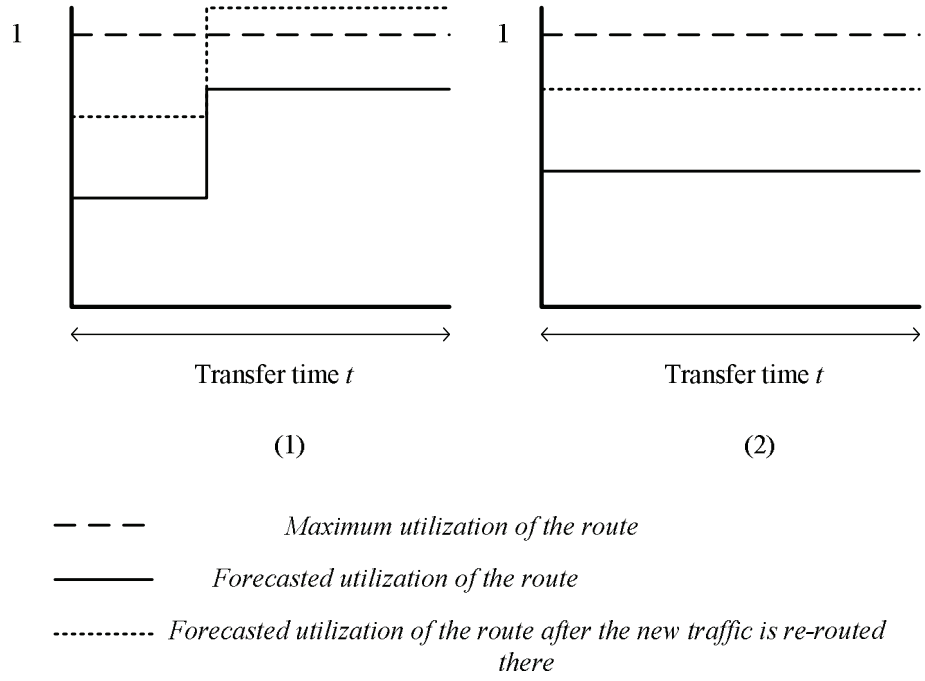


Fig.2. Forecasted change of utilization for routes 1-3-5 (1) and 1-4-5 (2). Maximum link utilization is 1

Conclusions. The problem of traffic engineering in modern networks was considered. And overview of existing solutions was made and the task of developing an effective way of traffic engineering in software-defined networks was defined. A method for traffic engineering with the ability of dynamic re-routing, which reconfigures traffic based on information about the forecasted link utilization change, was proposed. The proposed method allows increasing the efficiency of traffic engineering and ensuring a more even loading of communication links in the network. As a further study, the task of forecasting the link utilization can be considered.

References

1. He J. (2015). Achieving Near-Optimal Traffic Engineering in Hybrid Software Defined Networks / Jun He, Wei Song // IFIP Networking Conference (pp. 1–9).

2. Xia, W. (2015). A Survey on Software-Defined Networking / W. Xia, Y. Wen, C. Heng Foh, D. Niyato & H. Xie // IEEE COMMUNICATION SURVEYS & TUTORIALS. (Vol.1, No.1, pp. 27–51).

3. Xuan Liu, Sudhir Mohanraj, Michał Pioro & Deep Medhi (2014). Multipath Routing From a Traffic Engineering Perspective: How Beneficial is It? // 22nd IEEE International Conference on Network Protocols (ICNP) (pp.143-154).

4. M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, A. Vahdat, “Hedera: dynamic flow scheduling for data center networks”.: Proceedings of Networked Systems Design and Implementation Symposium, NSDI’10, vol. 10, April 2010, pp. 19–19.

5. A. R. Curtis, W. Kim, and P. Yalagandula. “Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection”, 30th IEEE Int. Conf. Comp. Commun. IEEE INFOCOM 2011, Shanghai, China, 10-15 April 2011, pp. 1629–163.

6. Jain, S.; Kumar, A.; Mandal, S.; Ong, J.; Poutievski, L.; Singh, A.; Venkata, S.; Wanderer, J.; Zhou, J.; Zhu, M.; et al. (2013) B4: Experience with a globally-deployed software defined WAN // ACM SIGCOMM Computer Communication Review (Vol. 43, pp. 3–14).

7. Kulakov, Yurii, Kohan, Alla & Kopychko, Sergii. (2020). Traffic Orchestration in Data Center Network Based on Software-Defined Networking Technology // Advances in Computer Science for Engineering and Education II (pp.228-237).

8. As’ad Mahmoud As’ad. (2018). A Method of Multipath Routing in SDN Networks // Advances in Computer Science and Engineering (Vol.17, No. 1, pp. 11-17).

Autors

Ларіна Ірина Сергіївна – студентка, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Irina Larina – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: larina.isb@gmail.com

Коган Алла Вікторівна – старший викладач, кафедра автоматизованих систем обробки інформації і управляння, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Kohan Alla - Senior lecturer, Department of Computer-Aided Management And Data Processing Systems, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: a.v.kohan433@gmail.com

РОЗШИРЕНА АНОТАЦІЯ

Ларіна Ірина Сергіївна, Коган Алла Вікторівна

СПОСІБ КОНСТРУЮВАННЯ ТРАФІКА В ПРОГРАМНО-КОНФІГУРОВАНИХ МЕРЕЖАХ

Актуальність теми дослідження. Сучасні тенденції стрімкого зростання хмарних обчислень, Big Data, Інтернету речей та інших технологій вимагають від комп'ютерних мереж більшої гнучкості та швидкодії. Також сучасним мережам характерні великий розмір та неоднорідність обладнання від різних виробників. Це ускладнює задачу конструювання трафіка (ТЕ), яка має за мету оптимізацію роботи мережі.

Постановка проблеми. Для вирішення задачі конструювання трафіка все більш перспективним стає використання технології програмно-конфігурованих мереж (SDN). Технологія представляє інтерес завдяки архітектурній особливості, яка полягає в наявності централізованого контролеру, що відповідає за конфігурацію та керування мережею.

Аналіз останніх досліджень і публікацій. Протягом останніх кількох років зростає кількість робіт, присвячених проблемі конструювання трафіку в мережах SDN. Багатошляхові методи маршрутизації широко використовуються для підвищення ефективності конструювання трафіку в мережах, що скорочує час конструювання трафіку і покращує якість обслуговування (QoS). Розроблені способи конструювання трафіку, що враховують об’єм потоків трафіку і поточний стан каналів для оптимального розподілу трафіку в мережі.

Виділення недосліджених частин загальної проблеми. Динамічний характер мереж потребує швидкої реакції на зміни стану ліній зв’язку. Тому поставлене завдання розробити спосіб конструювання трафіка з можливістю своєчасної динамічної ремаршрутизації пакетів для уникнення перенавантаження каналів.

Постановка завдання. Завданням є розробити спосіб конструювання трафіка в програмно-конфігурованих мережах з можливістю динамічної ремаршрутизації пакетів за прогнозованим завантаженням ліній зв’язку.

Викладення основного матеріалу. У даній роботі запропонований спосіб конструювання трафіка, який дозволяє забезпечити більш рівномірне завантаження каналів в мережі. Запропонований спосіб включає в себе можливість динамічно ремаршрутизації, якщо контролер спостерігає або очікує максимальне завантаження каналів на шляху трафіку.

Висновки. Був запропонований спосіб конструювання трафіка з можливістю динамічної ремаршрутизації, який проводить реконфігурацію трафіка на основі інформацію про прогнозовану зміну навантаження каналів.

Ключові слова: конструювання трафіка, динамічна ремаршрутизація, програмно-конфігуровані мережі.

UDC 004.451.44

Demchyk Valerii, Tyzun Vitalii,
Rusanova Olga, Korochkin Aleksandr

THE ORGANIZATION OF PARALLEL COMPUTATIONS IN HETEROGENEOUS COMPUTING SYSTEMS

Демчик Валерій, Тизунь Віталій,
Русанова Ольга, Корочкін Олександр

ОРГАНІЗАЦІЯ ПАРАЛЕЛЬНИХ ОБЧИСЛЕНЬ В ГЕТЕРОГЕННИХ КОМП'ЮТЕРНИХ СИСТЕМАХ

The article deals with the method of computing in heterogeneous multicore CPU+GPU systems. The results of an investigation of the effectiveness of methods for the task of text recognition using the technology of neural networks.

Key words: CPU, GPU, core, thread, parallelism, granularity, fork-join.
Fig.: 4. Bibl.: 13.

В статті розглядається спосіб організації обчислень в багатоядерних гетерогенних CPU+GPU системах. Наводяться результати дослідження ефективності методики для задач розпізнавання тексту з використанням технологій нейронних мереж.

Ключові слова: CPU, GPU, ядро, потік, паралелізм, зернистість, fork-join.
Рис.: 4. Бібл.: 13.

Relevance of research topic. Heterogeneous computer systems (HCS) are computer systems that contain several heterogeneous computing elements. HCS with CPU / GPU architecture it's a type of HCS systems equipped with a graphic processor. In a simplified format, the GPU model can be described as a set of a large number of simple processing elements of the same type. Each of these elements is much cheaper than its CPU analog. This contributed to the significant development of the GPU in terms of increasing the number of cores in it, and at present, the average number of cores in the GPU reaches a half a thousand. However, as practice has shown, if a heterogeneous computer system, which includes both the CPU and the GPU, is not busy working with complex graphics, then the computing power, that is presented in the current GPU, is superfluous, calculations are conducted so fast that most of the time the GPU cores idle in waiting for the next task.

Using this powerful and cheap computing power is a clear and logical solution of the problem of increasing computing productivity. With the integration into the

GPU of programmable shader blocks, it became possible to program normal user computing on this device. This technique was called GPGPU (General-purpose Computing on Graphics Processing Units) and became an important breakthrough in the development of modern computer technology. The first Nvidia graphics processors that support CUDA technology (Compute Unified Device Architecture - the implementation of GPGPU technology from this company) were not cheap and affordable. However, other GPGPU implementations appeared quite quickly, including the OpenCL framework, which allows you to program for all types of GPU, not only for Nvidia, but, for example, AMD. Also, apart from the fact, that OpenCL supports more GPUs than CUDA, these GPUs are usually cheaper and simpler, this technology is usable even with regular graphic cards.

Formulation of the problem. When programming for the CPU+GPU HCS, the most important part is to understand in which cases it is advisable to connect the GPU calculations, and in which cases better performance can be gained only by the use of the CPU. It's important to understand, when the increase in GPU performance reduces delays due to the data transfers between it and the CPU. Also, since both components of such a heterogeneous system are capable of computing, a possible option for concurrent computing that a CPU may not wait all the GPU computing time, but process some of the calculations on its own computing power. Therefore, it is important to find a balance, an optimal division of the input data between tasks, which will provide minimizing of the idle time during heterogeneous computing.

Analysis of recent research and publications. In recent years, more and more scientific articles and diploma papers about a calculation using graphic processors have appeared [4-12]. However, an overwhelming majority of them consider this question only from the point of view of choosing the better device for calculations between CPU and GPU. And there is not so much works about using both of devices for calculations at the same time. Moreover, in most of them the problem is considered in the context of solving classical problems of linear algebra, cryptography, and implementations of hyperparallel test algorithms [5-8]. Features of the deployment of high-speed neural networks are considered [9] only on the examples of some typical problems for neural networks, among which there is no one example of the most popular tasks - recognition and classification of images. The solution to this problem within a heterogeneous computer system is presented in one work [10], but only for mobile systems, which are obviously less developed and productive, than classical stationary systems.

The question of finding an optimal distribution of computational load between the CPU and the GPU was also considered only in the context of solving typical mathematical problems and implementation of test parallel algorithms [11].

Identification of unexplored parts of the general problem. Research of the efficiency of parallel computing in heterogeneous computer systems is usually carried

out on the classical problems for such studies - vector-matrix operations and implementations of various test parallel algorithms [5-8]. However, the main area in which today such systems are actually used are systems of artificial intelligence and neural networks. This is explained by the fact that by itself the neural network involves the simultaneous execution of a large number of elementary tasks, its structure is similar to the structure of the GPU. It is necessary to show the described problem on one of the classical tasks of this sphere.

It is also important to consider, that the heterogeneity of the system, and, accordingly, the need for the exchange of data between its components can lead to a situation, where the time, spent on the preparation and transmission of data and the collection of results, can slow down the acceleration of distributed computing. Therefore, it is necessary to consider in more detail the search for an optimal distribution of the data between the components of the system in order to provide a minimal idle time of one computing processor relatively to the other.

Also, in this work an own method for increasing the efficiency of parallel computing in heterogeneous computer systems is proposed, which is based on the application of combined parallelism [1] [2].

Problem statement. The task is to develop a program for recognizing text on images based on a neural network and focused on parallel work in a heterogeneous computer system. The program has to implement combined parallelism, as well as the ability to divide the percentage of processing data between the CPU and the GPU. The input data for a task is an image with letters, numbers, characters in the PNG format, as well as the percentage of GPU loading. Output data is the text string and the time it took to receive result.

Developed program has to be tested in various heterogeneous computer systems and show conclusions about the effectiveness of the proposed approach.

Combined parallelism. The simplified version of the CPU / GPU interaction scheme looks like this: from the CPU through the interface (for the external GPU it is PCI - Peripheral component interconnect) a set of instructions that must be performed on each core of the GPU is sent; Through the GPU controller, each core is configured for these instructions; The CPU then sends a set of data with which the GPU has to work; The GPU controller distributes this data between available cores, and after completing computing, it collects the results and sends it to the CPU. The instruction sets for relatively simple GPUs are small in size (several rows of program code) and simple in their structure (usually elementary mathematical and logical operations). We can say that the cores of the graphics processor itself implements the fine-grained parallelism calculations. However, with the advent of PCI interfaces of 2.1 version and above in heterogeneous computer systems, the possibility of parallel data transfer between CPU and GPU has appeared, which allows you to organize combined parallelism in the system.

This approach is based on the simultaneous use in the program of two types of parallelism: medium-grained and fine-grained. At the same time, the parallel program includes a set of traditional threads along the number of cores of the CPU (parallelism of medium-grain size). Each of these threads implements fine-grained parallelism by creating sub-threads using appropriate Fork-Join tools [7]. These small threads are used only for calculations. Additionally, as noted above, each of the medium-grain threads can interact with the GPU without delay, sending the necessary data to the graphics processor and taking the results of its work. In the GPU, the calculations are transferred to a large number of small threads, each on a separate core.

Previous studies [1] [2] proved the effectiveness of combined parallelism in the organization of overloaded parallel computing.

Analysis of the means of implementation. The main part of the program is written using the C# language. This multiparadigm language allows you to write program code for necessary task quickly and conveniently. In addition, in previous studies [1] [2], the means of C# language demonstrated their high efficiency in the implementation of all types of parallelism, including the combined one.

Calculations on the graphics processor are organized by the OpenCL framework, because it integrates seamlessly with the C# language and supports a large lineup of conventional GPUs. In addition, existing research shows its high performance, at the level of Nvidia CUDA [12].

Test results. Testing of programs was carried out on two different heterogeneous computer systems, which had the following parameters:

1. CPU: Intel i5-7200u, 2 cores, 4 threads, maximum frequency 3.1 GHz. GPU: AMD Radeon R5 M420, 1030 MHz, 320 processors, 2 GB of memory;

2. CPU: Intel Core i7-7700HQ, 4 cores, 8 threads, maximum frequency 3.8 GHz. GPU: NVIDIA GeForce GTX 1050 Ti, 1030 MHz, 768 processors, 4 GB of memory;

Software: Windows 10, .NET Framework 4.7, OpenCL 2.2.

Graphs below showing the dependence of the program's running time on the number of threads involved in it, as well as the data distribution between the CPU and the GPU. The graphs are presented separately for the situation, when the program processed a large amount of data (text recognition of 1000 characters), and for the situation, where the program worked with a small amount of data (200-character text recognition). Graphics are shown for both systems (1-2) in which the testing was conducted.

The following two graphs show the dependence of the program's performance on the amount of text that is submitted for recognition for cases where all data processing is carried out only on one of the elements of the system. The point of intersection on these graphs shows the turning point of the dependence. That is, if you submit text that is larger than this volume, then the efficiency of the use of recognition calculations on the CPU is reduced, and the efficiency of the use of the graphics processor is increased.

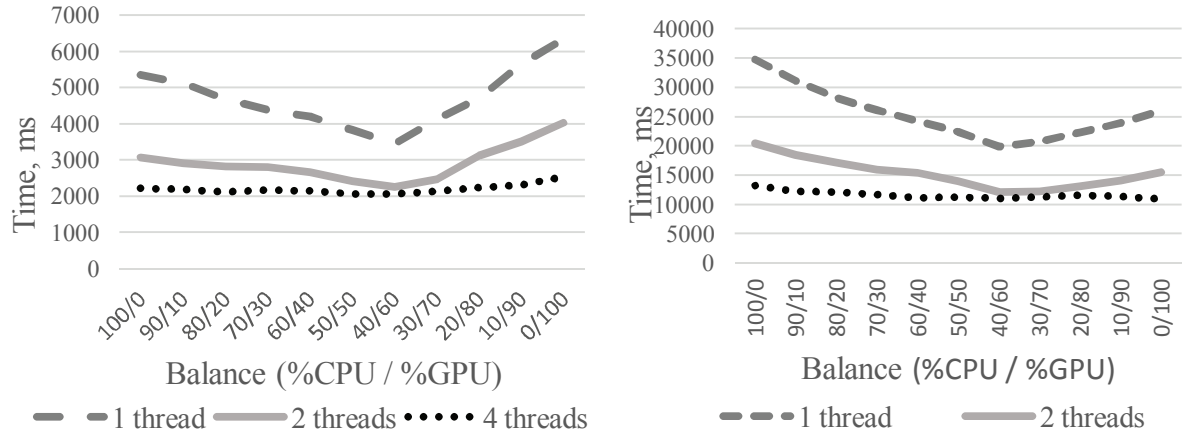


Fig. 1. Graphic of calculation speed depending on data distribution.
200-character text (left) and 1000 characters (right). System 1

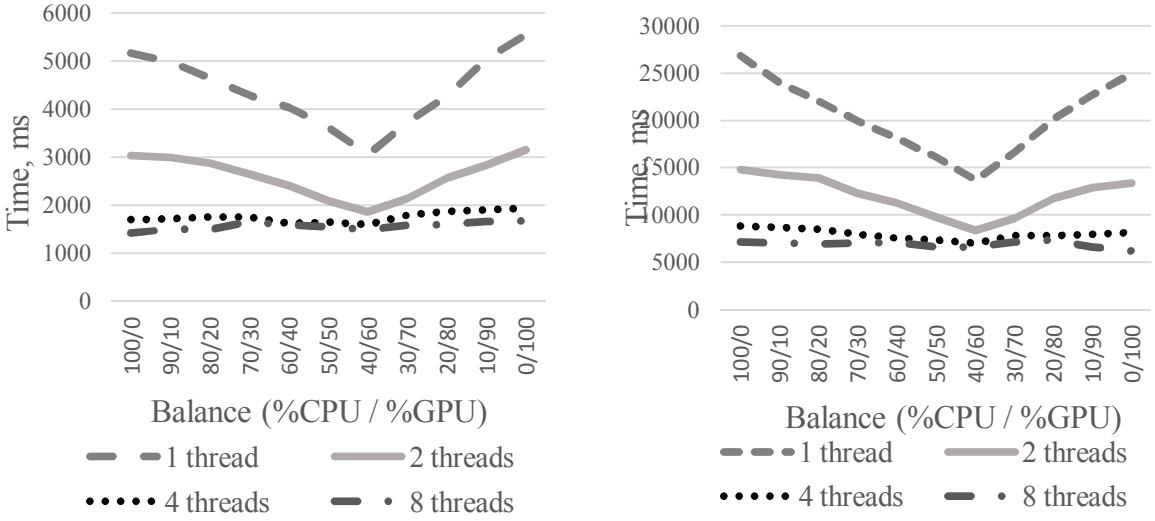


Fig. 2. Graphic of calculation speed depending on data distribution.
200-character text (left) and 1000 characters (right). System 2

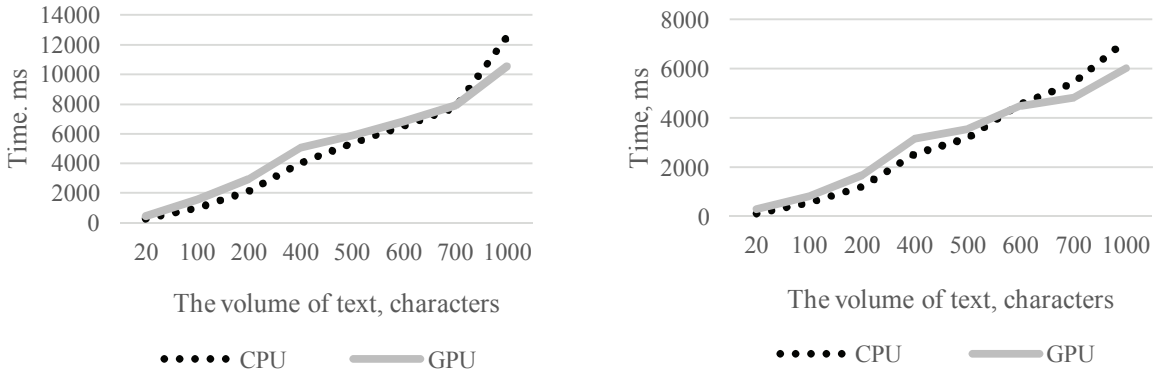


Fig. 3. Graph of the calculation time dependence of the program on the volume of text when recognizing on individual elements. Systems 1 (left) and 2 (right)

The following two graphs show the dependence of the program's performance on the amount of text that is submitted for recognition. The data in this case is distributed as follows: 40% for the CPU and 60 % for the GPU, since, as seen from the previous graphs, this kind of distribution achieves maximum program performance in all cases.

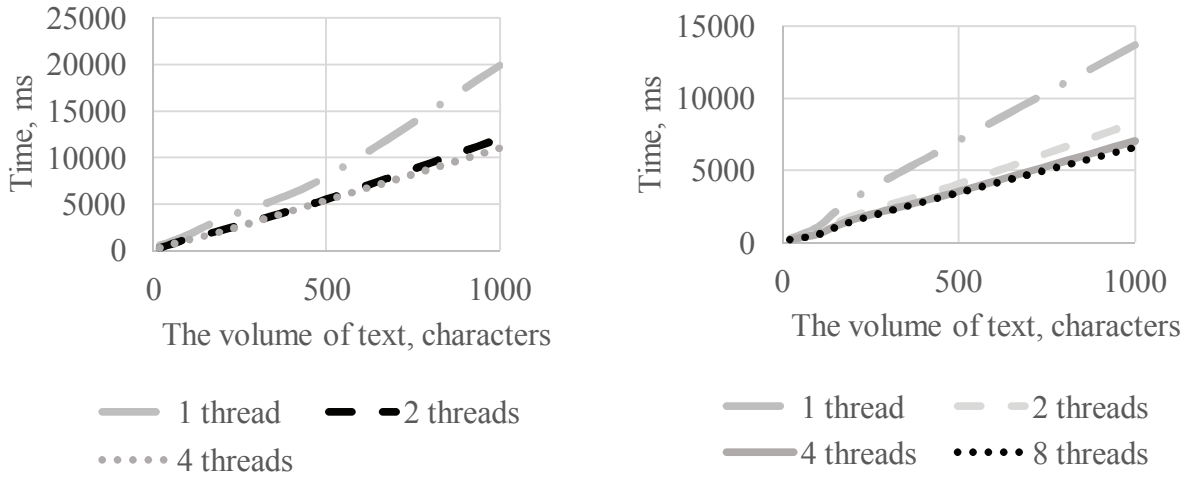


Fig. 4. Graph of calculation time according to the volume of text and number of threads. Balance point is 40/60. Systems 1 (left) and 2 (right)

Conclusions. The results of the test showed the effectiveness of heterogeneous computing systems in implementing the solution of the problem by means of C# and OpenCL. Applying the calculations to the graphics processor, along with the calculations at the CPU, allowed to significantly reduce the program execution time. In the process of research, in practice, the hypothesis of the existence of some of the most effective balance of data split between the CPU and the GPU has been confirmed.

Regarding the direct speed of the program work, the turning point can be considered as a text recognition of 600 characters. From the results it is clear that in the process of recognizing the text of less than 600 characters, the loss of time due to the exchange of data with the GPU is quite critical. It is because of sending a large percentage of data to the GPU nullifies the entire increase in performance from its use. It is desirable to send relatively a small amount of data to the graphics processor, so that it is accepted before the completion of computing on the central processor. In the case of text recognition over 600-700 characters, the situation is diametrically opposed. In this case, it is advisable to use the central processor only the role of administration and a small amount of computations, and the main part of data should be send to the graphics processor.

For all target systems considered, regardless of the amount of data processed by the program, the balance of 60/40 was the most optimal, that is, if the program

organizes the distribution of data, so that 60% of it is sent to the GPU, and the remaining 40% was left to process on the CPU, then it will provide the minimal idle, which allows you to get the most possible performance. Further offset of the balance in the direction of the GPU (for processing more than 600 characters) or CPU (for processing less than 600 characters) led to insignificant increases of performance, compared with the previous increases.

Additionally, it should be noted that the optimal balance obtained for the pattern recognition and classification problem is somewhat different from the optimal balance for linear algebra problems [11] in the direction of more GPU calculations.

The use of combined parallelism has also reduced the calculation time. With each subsequent added thread, you can observe a proportional decrease in the program's running time. It is also shows that the balance of calculations on the central and graphics processor remained unchanged with each subsequent added flow, since, on the one hand, the computing speed increased on the CPU, and on the other hand, the number of threads of interaction with the GPU increased, which turned into reduced the idle of GPU cores and idle due to data transfer between the CPU and the GPU.

So, in all cases, the most effective is the maximum possible use of cores and threads on the central processor, no matter how much we use the parallel calculations ob graphics processor.

Based on the foregoing, it can be admitted that the best approach to implement a text recognition system is to conduct preliminary testing on a target heterogeneous computer system, which will take extra time, but will ensure that the most effective proportion of the data distribution on this system is found.

References

1. Демчик В. В. Дослідження ефективності дрібнозернистого паралелізму в багатоядерних комп'ютерних системах / В. В. Демчик, О. В. Корочкін, О. В. Русанова // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка : зб. наук. праць. – К. : Век+, 2018. – № 66. – С. 56 – 61.

2. Демчик В. В. Застосування дрібнозернистого паралелізму для підвищення ефективності паралельних та розподілених обчислень / В. В. Демчик, О. В. Корочкін // Безпека. Відмовостійкість. Інтелект. Збірник праць міжнародної науково-практичної конференції ICSFTI2018. Київ, Україна, 10-12 травня 2018 р. / КПІ ім. Ігоря Сікорського –К. : КПІ ім. Ігоря Сікорського, Вид–во «Політехніка», 2018. – С. 362 – 368.

3. Жуков І.А., Корочкін О.В. Паралельні та розподілені обчислення: Навч. посібник [Текст]. – К.: Корнійчук, 2005. – 226 с. – ISBN 996-7599-36-1.

4. Hyesoon Kim, Richard Vuduc, Sara Baghsorkhi. Performance Analysis and Tuning for General Purpose Graphics Processing Units (GPGPU). — Morgan & Claypool Publishers, 2012. — 96 p.

5. Lin Cheng. Intelligent scheduling for simultaneous CPU-GPU applications by thesis // Graduate College of the University of Illinois at Urbana-Champaign, 2017 Urbana, Illinois [Електронний ресурс]. Режим доступу: http://rsim.cs.uiuc.edu/Pubs/Lin_thesis.pdf

6. Victor W Lee, Changkyu Kim, Jatin Chhugani, Michael Deisher, Daehyun Kim, Anthony D. Nguyen, Nadathur Satish, Mikhail Smelyanskiy, Srinivas Chennupaty, Per Hammarlund, Ronak Singhal and Pradeep Dubey. Debunking the 100X GPU vs. CPU Myth: An Evaluation of Throughput Computing on CPU and GPU // Throughput Computing Lab, Intel Corporation Intel Architecture Group, Intel Corporation [Електронний ресурс]. Режим доступу: https://www.academia.edu/36236172/Debunking_the_100X_GPU_vs._CPU_Myth_An_Evaluation_of_Throughput_Computing_on_CPU_and_GPU

7. T. Brandes, A. Arnold, T. Soddemann, D. Reith. CPU vs. GPU - Performance comparison for the Gram-Schmidt algorithm // Eur. Phys. J. Special Topics 210 – K.: EDP Sciences, Springer-Verlag, 2012. – № 210. – C. 73–88.

8. Haneesha H. K., Chandrashekhar B. N., Lakshmi H., Sunil. Performance Evaluation of CPU-GPU with CUDA Architecture Hybrid Computing, R&D // Nitte Meenakshi Institute of Technology, Bangalore-64.

9. Amr M. Kayid, Yasmeen Khaled, Mohamed Elmahdy. Performance of CPUs/GPUs for Deep Learning workloads // The German University in Cairo [Електронний ресурс]. Режим доступу: https://www.researchgate.net/publication/325023664_Performance_of_CPUsGPUs_for_Deep_Learning_workloads

10. Sipi Seppälä. Performance of Neural Network Image Classifcation on Mobile CPU and GPU // Aalto University MASTER'S THESIS 2018 [Електронний ресурс]. Режим доступу: https://pdfs.semanticscholar.org/946d/3f843ea93f22cc9c7e30af42a682139ad1e6.pdf

11. Ana Lucia Varbanescu. Heterogeneous CPU+GPU computing // University of Amsterdam. [Електронний ресурс]. Режим доступу: http://www.es.ele.tue.nl/~heco/courses/ASCI-schools/ASCI_springschool_2017/ASCI_HetCompCPU-GPU_part1.pdf

12. Kamran Karimi, Neil G. Dickson, Firas Hamze. A Performance Comparison of CUDA and OpenCL. - D-Wave Systems Inc. [Електронний ресурс]. Режим доступу: https://arxiv.org/ftp/arxiv/papers/1005/1005.2581.pdf

13. Lea, Doug. A Java Fork/Join Framework, In Proceedings of ACM Java Grande 2000 Conference (San Francisco, California, June 3-5, 2000)

Authors

Demchyk Valerii – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: kirintor830@gmail.com

Демчик Валерій Валентинович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Tyzun Vitalii – student, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: vitaliy.tyzun@gmail.com

Тизунь Віталій Юрійович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Rusanova Olga – docent, candidate of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: olga.rusanova.v@gmail.com

Русанова Ольга Веніамінівна – доцент, кандидат технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Korochkin Aleksandr – docent, candidate of Technical Sciences, Department of Computer Engineering, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

E-mail: avcora@gmail.com

Корочкін Олександр Володимирович – доцент, кандидат технічних наук, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

EXTENDED ABSTRACT

Demchyk Valerii, Tyzun Vitalii,
Rusanova Olga, Korochkin Aleksandr

THE ORGANIZATION OF PARALLEL COMPUTATIONS IN HETEROGENEOUS COMPUTING SYSTEMS

Relevance of the research. The method of computing in multi-core heterogeneous CPU+GPU systems using the technology of «combined parallelism» is considered. The results of research on the effectiveness of the method for text recognition task using the technology of neural networks are presented.

Target setting. The heterogeneity of the computational elements of a heterogeneous computer system leads to the problem of a long idle time, when one element of the system is waiting another to complete its task, which leads to total delay in the entire system.

Actual scientific researches and issues analysis. Over the past few years, there are more articles on this topic, but they simply choose between a CPU or a GPU, and only classical problems of linear algebra and hyperparallel test algorithms are considered.

Uninvestigated parts of general matters defining. The lack of research on the optimal distribution of computations between different elements of the system. Lack of research on solving real problems, such as text recognition with neural network technologies.

The research objective. The objective is to develop a program for text recognition based on a neural network and focused on parallel work in a heterogeneous computer system. The program has to implement combined parallelism, as well as the ability to adjust the percentage of processing on the CPU and GPU. It is necessary to test the developed program in various heterogeneous computer systems.

The statement of basic materials. A description of the main ideas and approaches that were implemented during the research. The extensive testing of the developed program in several different real heterogeneous computer systems has been carried out.

Conclusions. The proposed method for organizing calculations in heterogeneous computer systems has shown its effectiveness. Also, the hypothesis of the existence of an effective load distribution for the considered task between the elements of the target system was confirmed: 40% on the CPU and 60% on the GPU.

Key words: CPU, GPU, core, thread, parallelism, granularity, fork-join.

Наукове видання

Безпека. Відмовостійкість. Інтелект

Збірник праць міжнародної науково-практичної конференції
ICSFTI2019
14–15 травня 2019 р.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Свідоцтво про державну реєстрацію: ДК № 5354 від 25.05.2017 р.,
просп. Перемоги, 37
м. Київ, 03056

Підп. до друку 08.07.2019. Формат 60×84 1/16. Папір офс. Гарнітура Times.
Спосіб друку – ризографічний. Ум. друк. арк. 5,11. Обл.-вид. арк. 8,50. Наклад 80 пр.
Зам. № 19-093

Видавництво «Політехніка» КПІ ім. Ігоря Сікорського
вул. Політехнічна, 14, корп. 15
Київ, 03056
тел. (44) 204-81-78