Kuchuk N. H.
National Technical University «Kharkiv Polytechnic Institute», Kharkiv

COMPREHENSIVE PERFORMANCE CRITERION
FOR HYPER-CONVERGED INFRASTRUCTURE

The purpose of the article: find a quantitative relationship between structural network parameters, an indicator of the efficiency of the use of resources and the quality of the services provided by the network for a network with hyperconverged infrastructure. The optimal values of the degree of load of the channels are obtained. For each link in a hyperconverged infrastructure network, the required number of transmission channels and the required throughput are calculated. This makes it possible to calculate the throughput of communication channels during network synthesis.
Keywords: hyperconvergence, performance criteria, channel loading

Кучук Н. Г.
Національний технічний університет "Харківський політехнічний інститут", Харків

КОМПЛЕКСНИЙ КРИТЕРІЙ ЕФЕКТИВНОСТІ
ГІПЕРКОНВЕРГЕНТНОЇ ІНФРАСТРУКТУРИ

*Мета статті: знайти кількісний взаємозв'язок між структурно-мережевими параметрами, показником ефективності використання ресурсів та якістю послуг, що надаються мережею з гіперконвергентною інфраструктурою. **Результати.** Проведені дослідження дозволили сформулювати два завдання, які допускають рішення, що зручні для інтерпретації: мінімізація середнього часу затримки заявки в мережі при забезпеченні ймовірності відмови в обслуговуванні не більше допустимої; визначення мінімальної ймовірності відмови в обслуговуванні при забезпеченні середнього часу затримки пакету на заданому рівні. Проведено розрахунок величини середньої максимальної ймовірності відмови, яка визначається кількісними характеристиками величини трафіка, пропускної спроможності каналів, кількості каналів і ємності буферної пам'яті. Розрахована середня максимальна ймовірність втрати заявок. Проведений аналіз результатів досліджень показав, що отриманий результат дозволяє стверджувати про однозначне дотримання прийнятних оптимальних значень ступеня завантаження каналів мережі як мінімального середнього часу затримки пакетів в мережі при заданій допустимої ймовірності їх втрат, так і середньої максимальної ймовірності втрат пакетів в мережі. **Висновки.** Отримано оптимальні значення ступеня завантаження каналів. Для кожної ланки мережі з гіперконвергентною інфраструктурою розраховується необхідна кількість каналів передачі і необхідна пропускна здатність. Це дає можливість при синтезі мережі розрахувати пропускні спроможності каналів зв'язку та необхідний обсяг буферної пам'яті при відомих топології мережі і матриці тяжіння. Розрахунок проводиться за критерієм забезпечення середнього мінімального часу доставки повідомлень і максимальної ймовірності відмови в обслуговуванні пакетів мережі в допустимих межах.*
Ключові слова: гіперконвергентність, критерій ефективності, завантаження каналу

Кучук Н. Г.
Национальный технический университет «Харьковский политехнический институт», Харьков

КОМПЛЕКСНЫЙ КРИТЕРИЙ ЭФФЕКТИВНОСТИ
ГИПЕРКОНВЕРГЕНТНОЙ ИНФРАСТРУКТУРЫ

Цель статьи: найти количественную взаимосвязь между структурно-сетевыми параметрами, показателем эффективности использования ресурсов и качества предоставляемых

сетью услуг для сети с гиперконвергентной инфраструктурой. Получены оптимальные значения степени загрузки каналов. Для каждого звена сети с гиперконвергентной инфраструктурой рассчитывается необходимое количество каналов передачи и требуемую пропускную способность. Это дает возможность при синтезе сети рассчитать пропускные способности каналов связи и необходимый объем буферной памяти.

Ключевые слова: *гиперконвергентность, критерий эффективности, загрузка канала*

1. Introduction

What is hyperconverged infrastructure? Organizations currently use three-tier architectures. Which consist of a computing layer, a storage layer, and a network layer. It has been so for many years. These architectures have been effective in the past. But they require large creation costs. They are difficult to operate and scale. They cannot respond quickly to the needs of modern applications. Due to the poor adaptability of such architectures, organizations are considering using public clouds. They don't want to wait until the IT infrastructure can support their applications. But they are worried about problems: higher costs and lack of enterprise-class security. Modern equipment has made a three-tier architecture unnecessary. High-performance processors and memory in combination with modern storage technologies make it possible to integrate the entire infrastructure. You can work with it on the basis of servers x86 with internal SSD - disks. This infrastructure is called hyperconverged or HCI. Let's dwell on the advantages HCI- infrastructure. It saves money and improves adaptability. HCI - infrastructure significantly reduces capital expenditures. Because you only need servers x86. Specialized storage arrays, controllers, and fiber optic networks aren't needed. Operation is simplified. So to perform standard tasks does not require a lot of time and special skills. Besides, HCI- infrastructure helps to respond quickly to business needs. The equipment can be ready for use in a few hours. Operational settings can be ready in a few minutes. This eliminates the need for complex blocks such as logical volumes. Horizontal and vertical scaling is achieved by adding drives and servers x86. This occurs without interruption, lengthy reconfiguration or additional costs. With HCI, organizations can deploy cloud-like infrastructure in a local environment where costs are lower. But the level of control and security is higher than in a public cloud.

Hyperconverged infrastructure (HI) is characterized by simplicity of architecture and management, reduction of overhead costs, simplified interaction with vendors in environments with a high level of visualization [1, 2]. But these systems also have disadvantages. Therefore, before implementing HCI, it is necessary to evaluate its effectiveness.

Resource efficiency and user experience are interrelated. Therefore, resource efficiency is key to analysis and synthesis HCI. It is estimated by the ratio of load to module throughput. In the structure of a synthesized network with HCI (SNHCI) module throughput is taken equal to the load passing through it.

2. Analysis of the literature

High quality services require high bandwidth [3-14]. Therefore, the reduced use of network resources. For the SNHCI is right [4]

$$Q \cdot E = C, \tag{1}$$

where Q – assessment of a generalized indicator of service quality;

E – assessment of a generalized indicator of the effectiveness of the use of network resources;

C – constant value characterizing the network in terms of the quality of user service.

For rate Q we use the load factor of the i-th block SNHCI

$$\chi_i = F_i / (n_i V_i), \tag{2}$$

where F_i – flow in block;

V_i – bandwidth of one block channel;

n_i – number of channels in a given block.

Suppose $U_i = n_i \cdot V_i$ – channel capacity of the block, then

$$\chi_i \cdot U_i = F_i. \tag{3}$$

U_i – value may act as an estimate of E.

3. The purpose of the article

The purpose of the article: find a quantitative relationship between structural network parameters, an indicator of the efficiency of the use of resources and the quality of the services provided by the network for a network with hyperconverged infrastructure.

4. Research results

4.1. Task formalization

The conducted studies allowed us to formulate two tasks. These tasks allow solutions that are easy to interpret.

1. Problem 1. The task of minimizing the average delay time of the application in the network $\left(\overline{T}_{\text{spec}}\right)$ while ensuring the probability of a denial of service is not more than acceptable $\left(P_{\text{fail}}^{\text{accept}}\right)$:

$$\overline{T}_{\text{spec}}^{\min} = \min_{V,F} \overline{T}_{\text{spec}}; \qquad \overline{P}_{\text{fail}} \le P_{\text{fail}}^{\text{accept}}, \tag{4}$$

where F – the set of flows in SNHCI;

V – the set of bandwidth routes to SNHCI.

2. The task of determining the minimum probability of denial of service $\left(\overline{P}_{\text{fail}}\right)$ while ensuring the average packet delay time at a given level $\left(T_{\text{spec}}^{\text{accept}}\right)$:

$$\overline{P}_{\text{fail}}^{\min} = \min_{V,F} \overline{P}_{\text{fail}}; \qquad \overline{T}_{\text{spec}} \le T_{\text{spec}}^{\text{accept}}. \tag{5}$$

Both tasks can be considered as a dual nonlinear programming problem.

Each i-th link SNHCI modeled as queuing system (QS) of type M/M/n/m [6] with a limited queue. In the queuing system the Poisson stream of applications arrives with intensity λ_i, service intensity μ_i. The number of places in the queue for i-th link equally m_i. In the general case, queues are associated with the entrance to each link. The input is formed by a beam of n channels and a common buffer, containing m memory cells.

Problem 1 is solved by optimizing the average delay time by a complex indicator for each i-th network link [7]:

$$\chi_i = f(n_i, m_i), \tag{6}$$

where χ_i – the degree of loading of each of the n_i channels of the network link.

4.2. Initial data for solving the two problem

Consider the problem 2. Calculation of the average maximum probability of failure $\overline{P}_{fail}^{\max}$ determined by the quantitative characteristics of the parameters SNHCI: F_j, V_i, n, m (traffic size, channel capacity, number of channels and buffer memory capacity).

The average number of occupied channels for each link SNHCI defined as [8]:

$$\overline{z}_i = \rho_i (1 - \frac{\rho^{n_i+m_i}}{n_i^{m_i} \cdot n_i!} \cdot P_0), \tag{7}$$

where $\rho_i = \lambda_i / \mu_i$ – reduced flow rate;

$$\frac{\rho_i^{n_i+m_i}}{n_i^{m_i} \cdot n_i!} \cdot P_0 = P_{i\text{fail}} - \tag{8}$$

probability of denial of service to the next incoming to the communication center SNHCI packet due to lack of queue space;

P_0 – the probability that the channels are free, there is no queue and there are no service requests in QS:

$$P_0 = \left(\sum_{\alpha=0}^{n_i-1} \frac{(n_i \cdot \chi_i)^\alpha}{\alpha!} + \frac{(n_i \cdot \chi_i)^n}{n_i!} \cdot \frac{1-\chi_i^{m_i+1}}{1-\chi_i} \right)^{-1}. \tag{9}$$

The average number of applications in the queue can be found as:

$$\overline{r_i} = P_{i\text{fail}} \sum_{\alpha=1}^{m_i} \alpha \chi_i^{-(m_i-\alpha)}, \tag{10}$$

where $\chi_{,i} = \rho_i / n_i$ – degree of loading of the channel of the i-th link SNHCI.

Average number of applications per link QS determined from the expression:

$$\overline{W_i} = \overline{z_i} + \overline{r_i}. \tag{11}$$

Taking into account (7), (8), (10) expression (11) will take the form:

$$\overline{W_i} = n_i \chi_i (1 - P_{i\text{fail}}) + P_{i\text{fail}} \sum_{\alpha=1}^{m_i} \alpha \chi_i^{-(m_i-\alpha)}. \tag{12}$$

Ratio (12) true for any link SNHCI, wherein ρ independent of transmission direction:

$$\chi_i = \rho / n_i \approx \chi. \tag{13}$$

Loss probabilities in each i-th network link are comparable to each other, i.e.

$$P_{i\text{fail}} \approx \overline{P}_{\text{fail}}. \tag{14}$$

Using the Little Formula [9] and Kleinrock approximation [10] for k links

$$\gamma \overline{T}_{spec} = k \overline{W_i}. \tag{15}$$

Little formula allows you to go to the network level [9].

Taking into account (12), the average delay time of the application is determined:

$$\overline{T}_{spec} = \frac{1}{\gamma} \cdot k [\overline{P}_{\text{fail}} \sum_{\alpha=1}^{m_i} \alpha \chi^{-(m_i-\alpha)} + n \chi (1 - \overline{P}_{\text{fail}})]. \tag{16}$$

Let us set the acceptable packet delay time in the network, i.e.

$$\overline{T}_{spec} \leq T_{spec}^{\text{accept}}. \tag{17}$$

From (16) we determine the average maximum probability of loss of applications in SNHCI within the acceptable delay time:

$$\overline{P}_{\text{fail}}^{\max} = \frac{\gamma \overline{T}_{spec} - k n_i \chi}{k \sum_{\alpha=1}^{m_i} \alpha \chi^{-(m_i-\alpha)} - k n_i \chi}. \tag{18}$$

4.3. The solution to the optimization problem by a comprehensive indicator

Function (18) has an extremum (maximum), the search for which is the task of unconditional optimization. By calculating the partial derivative [11]

$$\partial \overline{P}_{\text{fail}} / \partial \chi = 0, \tag{19}$$

we get a local extremum. It is due to unimodality $\overline{P}_{\text{fail}}^{\max}$ is global.

Additivity of function (15) and conditions (11) determine the separability $\overline{P}_{\text{fail}}$. Therefore

$$\frac{\partial \overline{P}_{\text{fail}}}{\partial \chi} = \frac{d \overline{P}_{\text{fail}}}{d \chi} = 0. \tag{20}$$

To simplify the calculations, we denote

$$A = \frac{\gamma}{k} \cdot T_{\text{spec}}^{\text{accept}}.$$

Then (16) takes the form:

$$\overline{P}_{\text{fail}} = \frac{A - n_i \chi}{\sum_{\alpha=1}^{m_i} \alpha \chi^{-(m_i - \alpha)} - n_i \chi}. \tag{21}$$

On the other hand

$$A - n_i \chi = \frac{\sum_{\alpha=1}^{m_i} \alpha \chi^{-(m_i - \alpha)} - n_i \chi}{1 + \frac{1}{n_i} \sum_{\alpha=1}^{m_i} \alpha (m_i - \alpha) \chi^{-[m_i - (\alpha-1)]}}. \tag{22}$$

Substituting (22) into (21) we obtain an expression for determining the average maximum probability of failure in a time not exceeding T_{spec}^{accept}:

$$\overline{P}_{\text{fail}}^{\max} = \frac{1}{1 + \frac{1}{n_i} \sum_{\alpha=1}^{m_i} \alpha (m_i - \alpha) \chi^{-[m_i - (\alpha-1)]}}. \tag{23}$$

On the other hand, there are expressions (5), (7), which together, taking into account (11), can be represented as:

$$\overline{P}_{\text{fail}}^{\max} = \frac{(n_i \chi)^{n_i + m_i}}{n_i! \, n_i^{m_i}} \cdot \frac{1}{\sum_{\alpha=0}^{n_i} \frac{(n_i \chi)^{\alpha}}{\alpha!} + \frac{(n_i \chi)^{n_i}}{n_i!} \sum_{\alpha=1}^{m_i} \chi^{\alpha}}. \tag{24}$$

After transformation (21) we get:

$$\overline{P}_{\text{fail}}^{\max} = \frac{1}{\frac{n_i!}{\chi^{m_i}} \sum_{\alpha=0}^{n_i} \frac{(n_i \chi)^{-(n_i - \alpha)}}{\alpha!} + \sum_{\alpha=1}^{m_i} \chi^{-(m_i - \alpha)}}. \tag{25}$$

The left sides of expressions (20) and (25) are equal. Therefore, it is legitimate to assert:

$$\frac{n_i!}{\chi^{m_i}} \sum_{\alpha=0}^{n_i} \frac{(n_i \chi)^{-(n_i - \alpha)}}{\alpha!} + \sum_{\alpha=1}^{m_i} \chi^{-(m_i - \alpha)} = 1 + \frac{1}{n_i} \sum_{\alpha=1}^{m_i} \alpha (m_i - \alpha) \chi^{-[m_i - (\alpha-1)]}. \tag{26}$$

After simple transformations (26), we obtain the expression:

$$\sum_{\alpha=0}^{n_i} \frac{n_i!}{\alpha!}(n_i\chi)^{-(n_i-\alpha)} = \sum_{\alpha=1}^{m_i-1} (\frac{\alpha(m_i-\alpha)}{n_i\chi}-1)\chi^{\alpha}. \tag{27}$$

Define the roots of the equation (24). They vary within $0 \leq \chi < 1$. According to queuing theory [9] at $\chi \geq 1$ the average delay time of packets on the network and the probability of their loss increase significantly. Therefore, the roots of equation (27) are acceptable values of the degree of loading of channels χ_{accept}.

5. Analysis of research results

The obtained result (27) allows us to assert a unique correspondence between acceptable optimal values of the degree of load of network channels $\chi_{\text{accept}}^{opt}$, both the minimum average packet delay time in the network for a given acceptable probability of their loss, and the average maximum probability of packet loss in the network $P_{\text{fail}}^{\text{accept}}$. At a given allowable time, packet delays depend on the required probability of packet loss in the network $P_{\text{fail}}^{\text{accept}}$, and from the allowable delay time in the network T_{spec}^{accept}. We can say that packet delays are functions of discrete values of the number of channels n and the number of places in the queue m.

The equation of system (27) is a function of one variable χ. It makes it possible to independently determine an acceptable value of the channel load for each network link $\chi_{i\,\text{reason}}$.

But this equation is transcendental. Therefore, it is not possible to obtain an exact analytical solution (27). But they can be solved either numerically or graphically. Acceptable values $\chi_{i\,\text{reason}}$ determined by the coordinates of the intersection points of these curves with the axis $\chi_{i\,np}$. Graph analysis shows, that there are three solutions depending on the ratio of values n and m.

Define the values $\chi_{i\,\text{reason}}^{opt}$ and the value of the total flow at the input of the i-*th* link given by the gravitational matrix $\|F_{ij}\|$. The required number of transmission channels of the *i-th* link and their bandwidth will be determined based on the ratio:

$$\chi_{i\,\text{reason}}^{opt} = \frac{\lambda_i}{n_i \cdot \mu_i} = \frac{L \cdot \lambda_i}{n_i \cdot L \cdot \mu_i} = \frac{F_i}{V \cdot n_i};$$

$$\chi_{i\,\text{reason}}^{opt} = f(m_i, n_i), \tag{28}$$

where $F_i = L \cdot \lambda_i$ – total flow at the input of the *i-th* link;

$V_i = L \cdot \mu_i$ – bandwidth of each n_i channels.

Integrated metric optimization $\chi_{\text{reason}}^{opt}$ (28) allows you to vary the values V_i. Depending on the class of traffic F_i any combination of channels can be provided. At the same time, a virtual channel with variable bandwidth is formed. Such a channel does not depend on the required $T_{\text{spec}}^{\text{accept}}$. In this case, the probability of a denial of service package will remain within the acceptable range ($P_{\text{отк}}^{\text{accept}}$).

This allows you to provide a set of channels with a given bandwidth. An acceptable value remains $\chi_{i\,\text{reason}}^{opt}$. The average packet delay time in the network remains within the acceptable value $T_{\text{spec}}^{\text{accept}}$.

The numerical solution method is implemented using the program Mathcad. In fig.1 all

conditions are combined (7), (8), (16), (18), (26), (27). Analysis showed within the specified parameter values, you can exchange the channel capacity for the amount of buffer memory at the entrance to this channel.

At the same time, it is necessary to maintain the constancy of network quality indicators (delay time, probability of packet loss) permissible limits.
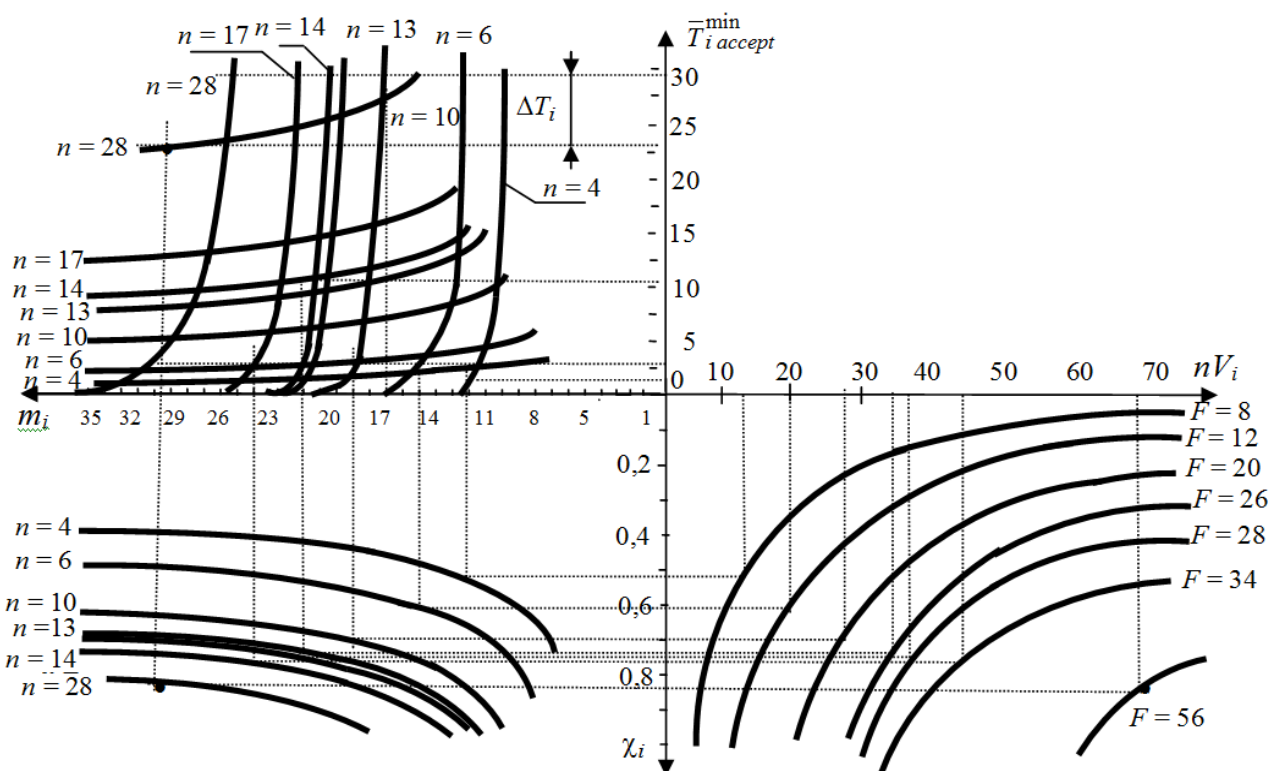

Fig. 1. Scheme of the relationship between structural network parameters

6. Conclusion

The optimal values of the degree of load of the channels are obtained. For each link in a hyperconverged infrastructure network, the required number of transmission channels and the required throughput are calculated. This makes it possible to calculate the throughput of communication channels during network synthesis. Also, the required amount of buffer memory with a known network topology and a given gravitational matrix. The calculation is based on the criterion of ensuring the average minimum message delivery time. And calculating the maximum probability of denial of service network packets within acceptable limits.

7. References

1.   Shmatkov S.I., Kuchuk, N.G. and Donets V.V. (2018). "Model of information structure of the hyperconvergent system of support of electronic computing resources of university e-learning". *Control systems, navigation and communication*. PNTU. Poltava. 2 (48). 97-100. Print.

2.   Merlac V., Smatkov S., Kuchuk N. and Nechausov A. (2018). "Resourses Distribution Method of University e-learning on the Hypercovergent platform", *Conf. Proc. of 2018 IEEE 9th International Conference on Dependable Systems, Service and Technologies. DESSERT'2018, Ukraine, Kyiv*. 136-140. Print. DOI: http://dx.doi.org/ 10.1109/DESSERT.2018.8409114

3.   Donets V., Kuchuk N., Shmatkov S. (2018). "Development of software of e-learning information system synthesis modeling process". *Modern information systems*. 2 (2). 117–121. Print. DOI: https://doi.org/10.20998/2522-9052.2018.2.20.

4.   Kuchuk N., Mozhaiev O., Mozhaiev M. and Kuchuk H. (2017). "Method for calculating of R-learning traffic peakedness", *4th International Scientific-Practical Conference Problems of*

Infocommunications Science and Technology, PIC S and T 2017. 359–362. Print. URL : http://dx.doi.org/10.1109/INFOCOMMST.2017.8246416

5.   Kovalenko A. and Kuchuk H. (2018). "Methods for synthesis of informational and technical structures of critical application object's control system". *Advanced Information Systems*. 2 (1). 22–27. Print. DOI: https://doi.org/10.20998/2522-9052.2018.1.04

6.   Sviridov A., Kovalenko A. and Kuchuk H. (2018), "The pass-through capacity redevelopment method of net critical section based on improvement ON/OFF models of traffic", *Advanced Information Systems*, 2 (2). 139–144. Print. DOI: https://doi.org/10.20998/2522-9052.2018.2.24

7.   Mozhaev O., Kuchuk H., Kuchuk N., Mozhaev M. and Lohvynenko M. (2017). "Multiservise network security metric". *IEEE Advanced information and communication technologies-2017. Proc. of the 2th Int. Conf.Lviv*. 133-136. Print.

8.   Kuchuk G., Kovalenko A., Komari I. E., Svyrydov A. and Kharchenko V. (2019). "Improving big data centers energy efficiency: Traffic based model and method". *Studies in Systems, Decision and Control, vol 171. Kharchenko, V., Kondratenko, Y., Kacprzyk, J. (Eds.)*. Springer Nature Switzerland AG. 161-183. Print. DOI: http://doi.org/10.1007/978-3-030-00253-4_8

9.   Svyrydov A., Kuchuk H. and Tsiapa O. (2018). "Improving efficienty of image recognition process: Approach and case study". *Proceedings of 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies, DESSERT 2018*. 593-597. Print. DOI: http://dx.doi.org/10.1109/DESSERT.2018.8409201

10.   Ruban I., Kuchuk H. and Kovalenko A. (2017). "Redistribution of base stations load in mobile communication networks". *Innovative technologies and scientific solutions for industries*. 1 (1). 75-81. Print. DOI : https://doi.org/10.30837/2522-9818.2017.1.075

11.   Amin Salih M. and Potrus M.Y. (2015). "A Method for Compensation of Tcp Throughput Degrading During Movement Of Mobile Node", *ZANCO Journal of Pure and Applied Sciences*, 27 (6). 59–68. Print.

12.   Mohammed, A. S. (2017). "Optimal Forecast Model for Erbil Traffic Road Data", *ZANCO Journal*. 29 (5). 137–145, DOI: https://doi.org/10.21271/ZJPAS.29.5.15

13.   Kuchuk G., Nechausov S. and Kharchenko V. (2015). "Two-stage optimization of resource allocation for hybrid cloud data store". *International Conference on Information and Digital Technologies. Zilina*, 266-271. Print. DOI: http://dx.doi.org/10.1109/DT.2015.7222982

14.   Sivaram M., Yuvaraj D., Amin Salih Mohammed, Porkodi V. and Manikandan V. (2018). "The Real Problem Through a Selection Making an Algorithm that Minimizes the Computational Complexity", *International Journal of Engineering and Advanced Technology*, 8 (2). 95-100. Print.

Author of the article

Кучук Ніна Георгіївна – к.п.н., доц., доцент кафедри обчислювальної техніки та програмування (Kuchuk Nina Heorhiivna – PhD in Pedagogics, Associate Professor, Associate Professor of the Department of Computer Science and Programming). Phone: +380 (99) 037 82 52. E-mail: nina_kuchuk@ukr.net.