STATISTICAL ANALYSIS OF EMPIRICAL NETWORK TRAFFIC DATA FROM PROGRAM MONITORING

Radi Romansky *

Department of Electronics and Electro-energy, College of Energy and Electronics,
Technical University of Sofia
Bulgaria

* Corresponding Author, e-mail: rrom@tu-sofia.bg

Abstract: Information servicing of remote client requests in a network environment is highly dependent on the regularity of the traffic and its parameters. With the development of modern digital technologies, the peculiarities of organizing processes to achieve high functionality and good performance with low loss of time and resources are increasing. This necessitates conducting adequate experiments for reliable traffic investigation and optimization of distributed service. Different approaches are presented in the literature, mainly aimed at model research and measurement (monitoring) of real data for observed parameters. The paper proposes a combined approach to combine program monitoring with statistical analysis of the measured data to ensure better accuracy of the obtained estimates for selected traffic parameters. A justification of the chosen approach is made on the basis of a brief summary of traffic features and a discussion of recommendations for efficient distributed service. Program monitoring was carried out, and the accumulated empirical data from it are a subject of statistical analysis and the obtained evaluations are summarized.

Key words: networking, distributed information servicing, investigation, program monitoring, statistical assessments.

1. INTRODUCTION

The importance of information and its processing in the contemporary digital society is common knowledge, which is confirmed by the research in [1] with the conclusions drawn about the growing interest in content, information search and its processing. In this direction, the investigation of distributed systems and the information and communication processes taking place in them is an important point in the optimization of the distributed information service [2]. Two main alternatives are usually discussed when organizing research experiments:

✓ Application of modelling methods, such as mathematical formalization, analytical (deterministic or stochastic) modelling [3], simulation [4], statistical (empirical) modelling [5] etc.; ✓ Measurement (monitoring) using software, hardware or combined means [6]. The choice of an approach to conduct the research is mainly determined by the set goal and the architectural features of the research object in order to increase the efficiency of routing [7]. In this respect, when researching information services in distributed environments, the characteristics of network processes, the expected estimates of network traffic and system performance, throughput, etc., are essential.

One possibility to conduct an effective investigation of the network characteristics of distributed service traffic is to apply combined methods to use the individual advantages of each method while at the same time reducing the influence of the inevitable disadvantages. One such possibility is presented in this paper, where network traffic data obtained from an empirical measurement is further analysed using statistical tools. The goal is to efficiently process a formed sample of real data using statistical methods to obtain estimates for selected parameters of network processes. To present the research, the following section summarizes the basic characteristics of network traffic and some related works are discussed. The specific experimental results are presented in sections 3 and 4.

2. NERWORK TRAFFIC AND DISTRIBUTED SERVICE INVESTIGATION

Routing in computer networks is generally dynamic in nature and mainly depends on traffic organization and communication parameters. For example, in wireless sensor networks (WSN), one of the solved problems when developing routing protocols is to reduce energy consumption, and as noted in [8], in such cases, the lifetime of the network will be extended. To confirm this, the article makes a comparative study of the possibilities for developing energy-saving routing protocols and makes a concrete proposal.

As pointed out in [9], various techniques are applied for network traffic classification and new ones are even developed, but the existing literature still lacks a thorough study for a comprehensive summary and detailed analysis. In this respect, article [9] proposes five categories for classification of the traffic based on preliminary defined characteristics and criteria, which are discussed and several basic questions are raised for future resolution. One point of view on the subject is briefly presented below.

Network traffic is very diverse and stochastic in nature, and can be defined in two basic categories: functional (transactional) traffic and background traffic. Two main approaches are applied to its observation.

1. Observation of the details of each, or at least of a set of frames, passed through the network environment – allows viewing of each frame passed through the network or frames filtered by a special criterion. This method is particularly

suitable when dealing with network problems, since the content of the frames is extremely important in this process, as well as when distributing the flow when processing large data sets in clusters. In [2], it is stated that public cloud providers offer a variety of resources for use in clusters, presenting a cooperative system for optimizing cluster configurations. The main goal of the system is to predict the processing time of the data using specialized regression models, and the authors confirm the effectiveness by achieving average values of 3% for the absolute errors.

2. Observation of statistically processed results for the load of network segments and for the distribution of network protocols and services by segments. The detailed content of each frame is not available and in many cases is not buffered during the network analyser operation. This method is suitable for long-term monitoring of the parameters of a network segment in order to perform optimization. A study of object monitoring when storing rare information in different remote servers is presented in [5], with the goal of minimizing the total search time. By applying a statistical approach, it has been confirmed that initial grouping of servers and subsequent search in selected groups allows to significantly reduce data search time.

Some features of network traffic related to distributed service investigation and evaluation can be summarized as follows.

♦ It is possible certain heavily loaded network segments to have high volumes of network traffic, leading to investigation problems (e.g., access delays, real-world lag, real-world frame loss, etc.). A major task in developing network solutions is to increase performance by reducing end-to-end delay, an analysis of which is presented in [11]. A framework for connecting end machines and a mathematical analytical method for calculating the expected average end-to-end delay using a network simulator are proposed in this article.

♦ Conducting network traffic research, including monitoring, should be based on a predefined goal and strategy to ensure the necessary effectiveness of research tools. To achieve efficiency in the research, it is good to analyse the belonging of a given property or a certain characteristic to the set goal. To implement this, a sequential semantic-based emulation technique is proposed in [12], by developing a specialized automated procedure accepting as input a formal specification and the structural semantics of the distributed system. A program to emulate the possible evolutions of the considered system is generated as a result of the execution.

♦ Due to the existence of different types of traffic (some of which are less important for the overall performance of the information service), it is necessary to focus the specific investigation on appropriate types to ensure the adequacy of the obtained estimates and conclusions. One example is the study of traffic and information flows in IP-based networks using monitoring tools and mathematical distributions presented in [6]. Protocols and exchanged packets are traced with estimates of transmission delays, forming mathematical distributions for network functionality analysis.

♦ In order to optimize the experiments and present the estimates, it is necessary to choose the appropriate research method for the specific task, which will reduce the possible "distortions" of the estimates. In this respect, the means used to display the accumulated experimental results (a sample of registrations) and the assessments formed on the sample are also of great importance. For example, a set of procedures for multivariate analysis of spatial-temporal information and interaction of geographically connected systems is presented in [10]. The goal is intellectualization of management decision-making systems. In this respect, algorithms for intelligent support of decision-making processes are proposed and an analysis of the achieved efficiency in geographically distributed systems is carried out.

3. PROGRAM MONITORING

Empirical measurement was carried out using software monitors "Iris" and "Distinct Network Monitor". Subject of research are dependencies related to average length of packages, total number of packages, number of packages for different network protocols, number of packages for different IP protocols, incoming and outgoing traffic, etc. Summary of selected empirical monitoring data distributed by protocols is presented in Table 1. As can be seen from the graphical interpretation of Figure 1, MAC and IP traffic are significant compared to the levels of ARP and DNS traffic.

Table 1. Summarized empirical data from conducted monitoring

Total number of packages	*Distribution by network protocols*		*Distribution of IP packages by protocols ICMP, TCP u UDP*			*Input and output packages*	
	ARP	IP	ICMP	TCP	UDP	IN	OUT
2044	4	2040	3	2004	33	1206	838
5113	4	5109	0	5003	106	2650	2463
8025	2	8023	0	8005	18	4205	3820
12176	8	12168	0	12138	30	6366	5810

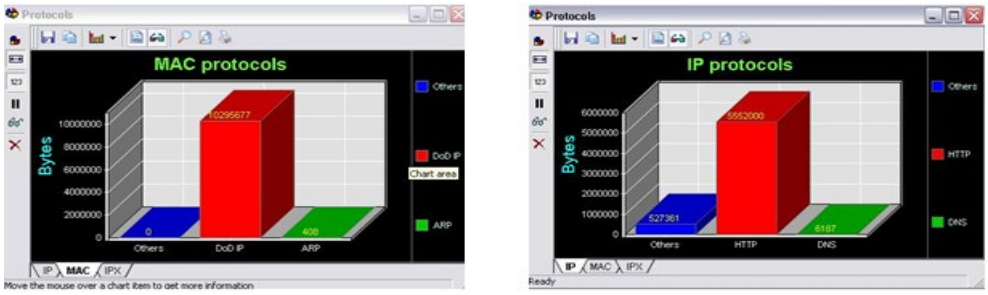


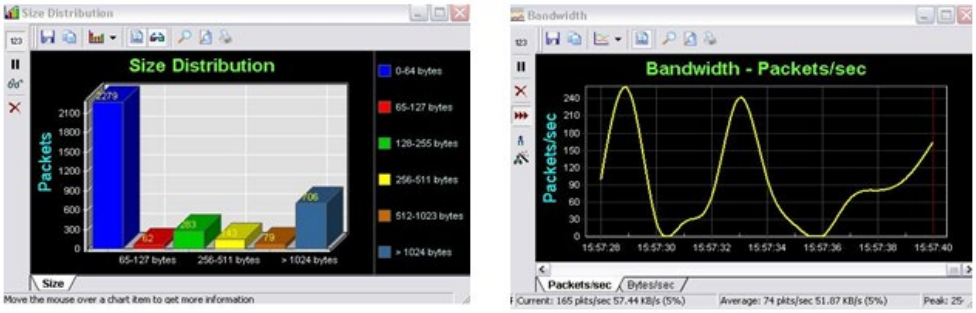
Figure 1. Graphical interpretation of monitored data

Regarding the distribution of packages by size, both monitors are applied, and the results are similar – figure 2(a, b). The formed estimates show that the main traffic in a traditional information service in a distributed environment is formed by mostly short packets with a length of up to 64 B.

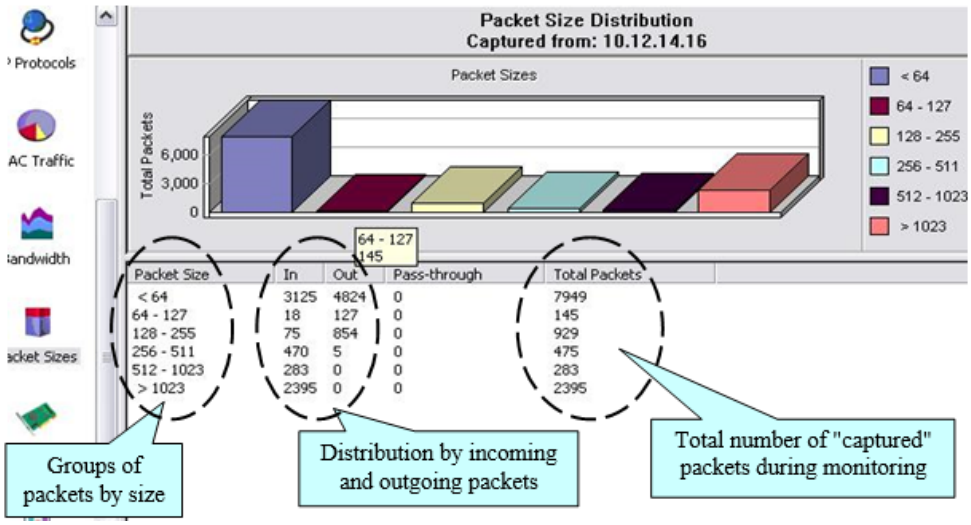
Through statistical analysis of the sample (accumulated empirical data) for the ratio of IP traffic to total traffic, maximum values (relative to the other components) were calculated, with the correlation coefficient tending to 1. At the same time, for ARP traffic, a mean value of 4.5 and a variance of 6.33 were determined, with a correlation of about 0.5.

4. STATISTICAL ANALYSIS OF MONITORED DATA

The use of mathematical statistics allows combining the rigorous formalization of statistical models with the correctness of measured real data in a given time period of a real working network environment. In this case, the mathematical statistics would help to overcome the relatively private nature of a given measurement, which is dependent on a number of external factors at any given moment. The organization of the investigation is summarized below.



a) distribution base on size and throughput



b) distribution of packages by size

Figure 2. Network traffic investigation

(1) *Defining key observed factors:* ✓ basic number of users; ✓ average value of the selection times (average click time); ✓ waiting time for user selection by pressing a mouse button (click time); ✓ server bandwidth; ✓ bandwidth of user access; ✓ number of URLs; ✓ average request service time; ✓ supported network traffic, incl. incoming traffic and outgoing traffic; ✓ time for interpreting the URL domains using the DNS server of the client machine (time for DNS); ✓ time to establish a connection with the server (time to connect); ✓ time to receive the first byte of the response to a request to the server (TFB); ✓ average length of packages; number of packets for the different network protocols and number of packets for the different IP protocols.

(2) *Defining dependency groups for analysis:* ✓ study of the dependence between the number of users and basic parameters when loading distributed objects; ✓analysing system parameters with varying number of users for each test, making all user requests sent at a given time to the server identical; ✓ dependencies on the relationship between the number of URLs and basic system parameters; ✓ dependencies for analysing the relationship between packet count and packet length for different protocols; ✓ investigating the relationship between incoming, outgoing and total traffic.

(3) *Program measurement experiments and accumulated empirical data*. Registrations are realized when accessing a form, static and dynamic site (respectively "form", "stat" and "dyna" in Table 2).

Table 2. Empirical results of conducted monitoring for group 1

users	*server bandwidth (kb/s)*			*aver. click time (ms)*			*aver. request time (ms)*			*time to first byte (ms)*		
	form	stat	dyna	form	stat	dyna	form	stat	dyna	form	stat	dyna
2	5	20	65	12000	45	1000	45	45	1000	0	35	7500
4	15	37	12	3000	80	15500	80	80	15000	2500	65	1250
6	35	42	37	2500	100	3500	100	100	3500	750	85	1250
8	50	42	12	2000	150	6000	150	150	6000	1000	110	4000
10	25	42	37	7500	190	11000	185	185	11000	2500	170	1000

(4) *Statistical analysis of the formed samples in the following directions.*

- Investigation of the dependence of information service parameters on the number of users – a summary of part of the results is given in the Table 2. The graphical visualization of the measurement of "server bandwidth" (with simulated 8 users) is presented in figure 3. The performed correlation analysis for the parameters "server bandwidth" and "users" determines -0.238 for the correlation coefficient, with the values for "server bandwidth" remaining relatively constant in the range of 40-42kb/s.
- Investigation of the dependence of information traffic parameters on the number of URLs. Figure 4 shows the registrations for the variable "network traffic" at a fixed number of URLs, as from the carried out statistical analysis average value 54.75 and variance 1940 are determined, and regarding an existing

dependence between the two variables are calculated covariance 118.75, correlation coefficient +0.243 and regression line with parameters 39.051 (for the displacement) and 0.966 (for the regression coefficient).

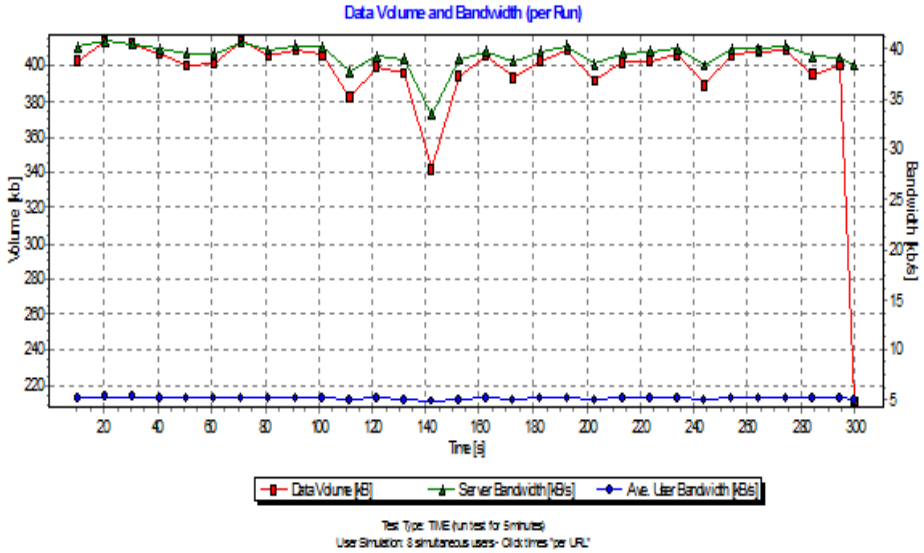


Figure 3. Experimental results for server bandwidth (users=8)

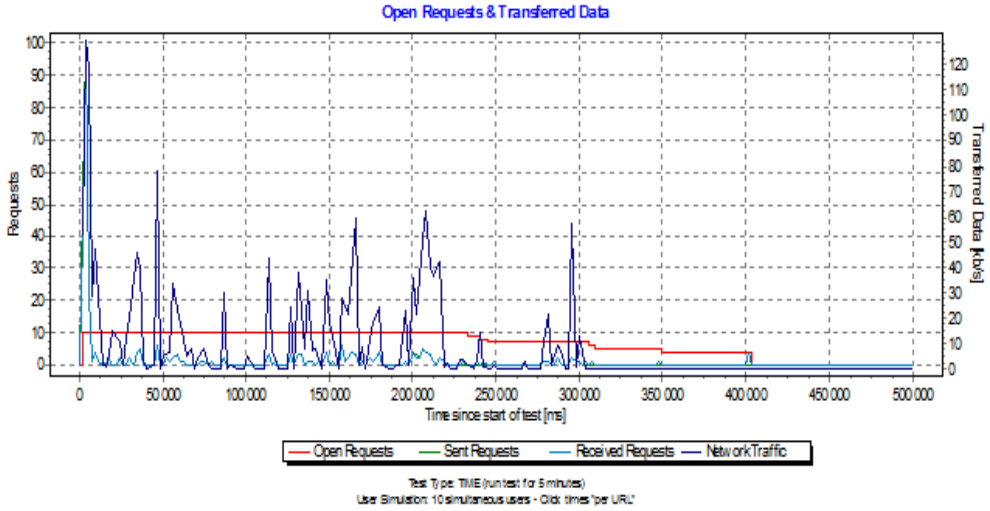


Figure 4. Registered values for network traffic (count URL = 30)

- Investigation of the "User Wait Time" for information service and its distribution in relation to active users (Users). Example estimations obtained by using Webserver Stress Tool are presented in Figure 5.

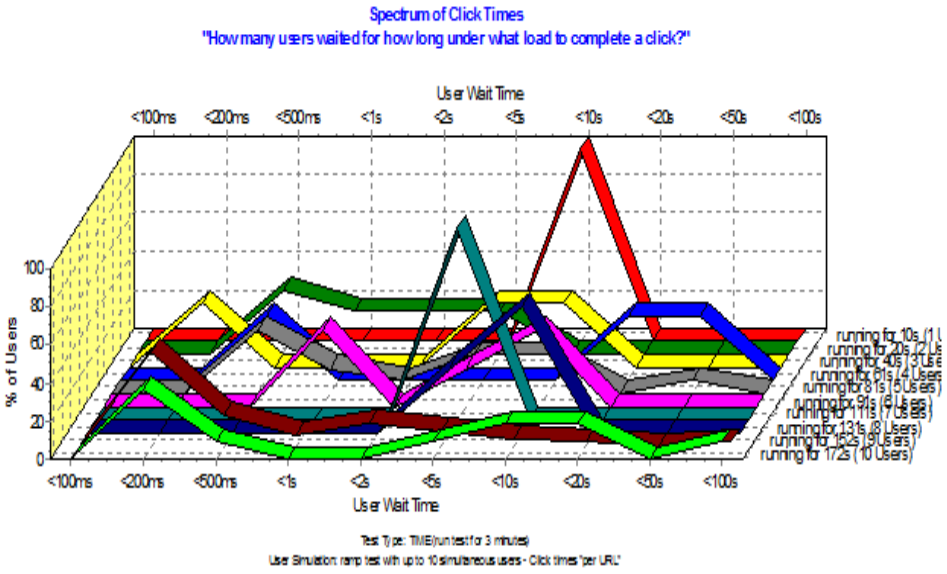


Figure 5. Distribution of "waiting time"/"users" in information service

5. CONCLUSION

The research presented in the article is focused on one specific part of the network processes and service of the requests and does not pretend to scale of the possible problems. The results above should be seen as an addition to other research carried out by the author, both in the field of empirical investigation of network processes and directed at different aspects of modelling. The diversity of today's digital society constantly poses topical problems to be solved and defines opportunities for future research in various aspects of distributed information services. One such area is the management of processes in traditional computer networks, which is usually done manually through the command line, which according to [13] leads to reduced efficiency and possible configuration deviations, as well as susceptibility to errors. To overcome these challenges, the paper proposes an approach to improve network management in traditional networks by using Ansible playbook to dynamically configure the multiple devices and push the archive files to Git for version control.

Another discussed direction is resource allocation and access management, as a similar problem oriented to spatially distributed networks with multiple inputs multiple outputs (MIMO), is considered in [14]. In this network type, user requests are served by subsets of nearby transmitters (distributed units) controlled by a centralized unit. In this respect, the article proposes a hybrid metric based on developed algorithms for optimizing the system resource through user planning and regulating the throughput of the distributed system.

The use of information technologies in various spheres of contemporary society inevitably leads to an increase in the amount of information maintained and exchanged. The transfer of these data (including personal data) through network channels is one of the sides of communications, reflecting on traffic parameters, but one other important side should also be taken into account – protection of the personal privacy of traffic participants [15, 16]. The attractive possibilities and conveniences of modern network technologies are also accompanied by increasing challenges for privacy, which led to the development and adoption of the General Data Protection Regulation (GDPR). In this reason, in [17] it is stated that "*a link should be established between policy statements and the program code, with the support of a formalized analysis*". In support of this statement, the article formalizes the notion of privacy policy in accordance with object-oriented distributed systems, defining specific instructions for securely protecting the available data. A language for specification of policies, as well as a high-level modelling language for distributed systems are proposed to implement the proposed approach.

REFERENCES

[1] K. Ann Renninger, Suzanne E. Hidi. Interest development, self-related information processing, and practice, *Theory Into Practice*, No. 1 (vol. 61), 2022, pp. 23-34. doi: 10.1080/00405841.2021.1932159

[2] J. Will, L. Thamsen, D. Scheinert, J. Bader, O. Kao. C3O: Collaborative cluster configuration optimization for distributed data processing in public clouds. *2021 IEEE International Conference on Cloud Engineering (IC2E)*, San Francisco, USA, Nov. 2021, pp. 43-52, doi: 10.1109/IC2E52221.2021.00018.

[3] M. Mehdi Hassani, R. Berangi. An analytical model to calculate blocking probability of secondary user in cognitive radio rensor retworks. *International Journal on Information Technologies and Security*, No. 2 (vol. 10), 2018, pp. 3-12.

[4] R. Romansky. Evaluation of experimental data from monitoring and simulation of network communication parameters. *International Journal on Information Technologies and Security*, No. 2 (vol. 14), 2022, pp. 75-86.

[5] I.V. Atlasov, V.E. Bolnokin, O.Ja.Kravets, D.I. Mutin, G.N. Nurutdinov. Statistical models for minimizing the number of search queries. *International Journal on Information Technologies and Security*, No. 3 (vol. 12), 2020, pp. 3-12.

[6] I. Nedyalkov, Al. Stefanov, G. Georgiev. Studying and characterization of the data flows in an IP-based network. *International Journal on Information Technologies and Security*, No. 1 (vol. 11), 2019, pp. 3-12.

[7] O.Ja. Kravets, I.V. Atlasov, I.A. Aksenov, A.S. Molchan, O.Yu. Frantsisko, P.A. Rahman. Increasing efficiency of routing in transient modes of computer network operation. *International Journal on Information Technologies and Security*, No. 2 (vol. 13), 2021, pp. 3-14.

[8] M. E. A. Ibrahim, A. E. S. Ahmed, H. Almujahed. A comparative study of energy saving routing protocols for Wireless Sensor Networks. *International Journal on Information Technologies and Security*, No. 2 (vol. 11), 2019, pp. 3-16.

[9] J. Zhao, X. Jing, Z. Yan, W. Pedrycz. Network traffic classification for data fusion: A survey. *Information Fusion*, vol. 72, Aug 2021, pp.22-47, https://doi.org/10.1016/j.inffus.2021.02.009.

[10] V.V.Goryachko, O,N. Choporov, A.P.Preobrazhenskiy, O.Ja.Kravets. The use of intellectualization management decision-making in the interaction of territorially connected systems. *International Journal on Information Technologies and Security*, No. 1 (vol. 12), 2020, pp. 87-97.

[11] A. Alsheikhy. Estimating end-to-end delay on a networking environment using developed framework. *International Journal on Information Technologies and Security*, No. 1 (vol. 14), 2022, pp. 3-16.

[12] L. Di Stefano, R. De Nicola, O. Inverso. Verification of distributed systems via sequential emulation, *ACM Transactions on Software Engineering and Methodology*, No. 3 (vol.31), July 2022, art. 37, pp. 1-41. (https://doi.org/10.1145/3490387).

[13] Blessing Bwalya, Aaron Zimba. An SDN approach to mitigating network management challenges in traditional networks. *International Journal on Information Technologies and Security*, No. 4 (vol. 13), 2021, pp. 3-14.

[14] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, K. V. Srinivas. Distributed resource allocation optimization for user-centric cell-free MIMO networks. *IEEE Transactions on Wireless Communications*, No. 5 (Vol. 21), May 2021, pp. 3099-3115, doi: 10.1109/TWC.2021.3118303.

[15] R. Romansky. *Digital Age and Personal Data Protection*, ISBN: 978-620-4-73564-1, LAP LAMBERT Academic Publishing, 14 January 2022 (124 p.)

[16] R. Romansky. Privacy and Data Protection in the Contemporary Digital Age. *International Journal on Information Technologies and Security*, No. 4 (vol. 13), 2021, pp. 99-110.

[17] Sh. Tokas, O. Owe, T. Ramezanifarkhani. Static checking of GDPR-related privacy compliance for object-oriented distributed systems. *Journal of Logical and Algebraic Methods in Programming*, ISSN 2352-2208, Vol. 125, Feb 2022, art. 100733, https://doi.org/10.1016/j.jlamp.2021.100733.

Information about the author:

Radi Romansky is a full professor at Technical University of Sofia, Doctor (Dr) in Computer Engineering and Doctor of Science (D.Sc.) in Informatics and Computer Science; Full member of European Network of Excellence on High Performance and Embedded Architectures and Compilation (HiPEAC). He has over 215 scientific publications and over 25 books. Areas of scientific interests: ICT, informatics, computer architectures, computer modelling, privacy and data protection, etc.

Manuscript received on 12 June 2022