

Model of optimal distribution of network resources with constraints on quality of service indicators

Amirsaidov Ulugbek, Qodirov Azamat

Department of Data Communication Networks and Systems, Faculty of Telecommunication Technologies, Tashkent University of Information Technologies, Tashkent, Uzbekistan

Article Info

Article history:

Received Sep 6, 2022

Revised Oct 6, 2022

Accepted Nov 19, 2022

Keywords:

Buffer resource

Heterogeneous flow

Optimal distribution

Quality of service

Queuing model

ABSTRACT

Existing algorithms and mathematical models of queuing at the nodes of a telecommunications network are considered in this paper. The necessity of coordinated solutions to problems of distribution of channel and buffer resources of the network is shown. A model for the optimal distribution of channel and buffer resources on network nodes has been developed. The optimization (minimization) criterion is the total average packet delay with constraints on the quality of service (QoS) indicators of heterogeneous flows. The optimization problem is presented as a constrained nonlinear programming problem and solved using the “fmincon” program of the optimization toolbox MATLAB package.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Qodirov Azamat

Department of Data Communication Networks and Systems, Faculty of Telecommunication Technologies

Tashkent University of Information Technologies

108 Amir Temur Avenue, Tashkent 100200, Uzbekistan

Email: azamattuit2013@gmail.com

1. INTRODUCTION

In modern switching equipment, many queue management mechanisms are implemented in terms of their formation and provisioning [1]–[4]. In the algorithm without priority service first in first out (FIFO), all incoming packets are put into one queue and served in the order of arrival. Therefore, the FIFO algorithm does not provide a differentiated traffic quality of service (QoS). The priority queuing (PQ) algorithm provides for the division of all network traffic into a small number of classes with the assignment of a certain numerical attribute to each class. In the beginning, the queue with the highest priority packets is served. After servicing all packets of a given queue, the service receives the queue with the lower priority, and so on. From the principle of operation of the PQ, it is obvious that the constant presence of high-priority traffic in the queue will lead to significant service delays and loss of low-priority traffic. In fair queuing (FQ) algorithms, a channel resource is allocated that is proportional to the weights assigned to them. Allocation of a constant channel resource in the absence of packets in the queue leads to inefficient use of the channel bandwidth. In a hybrid service algorithm low latency queuing (LLQ) for traffic sensitive to delays, one queue is allocated for servicing which a certain bandwidth of the channel is reserved. The remaining queues are serviced by the FQ algorithm. Therefore, the LLQ algorithms also have the disadvantage of FQ algorithms. A characteristic feature of existing solutions for managing buffer and channel resources is a statistical strategy for their distribution. The allocation of buffer space and channel capacity is still primarily an administrative matter. In this regard the tasks of the dynamic management of channel and buffer resources, depending on the characteristics of the traffic arriving at the node, the value of the available resource and the required indicators of the QoS are relevant.

2. ANALYSIS OF EXISTING QUEUING MODELS

To date, there are quite a few approaches to the mathematical description of queue management processes that differ in the problem statement, the type of mathematical apparatus used, the expected accuracy of calculations, and the level of complexity of calculating the desired values. An important role in choosing one or another approach is also played by the scope of models and methods for solving the problems of analyzing possible options and strategies for queue management or synthesis problems related to the prospect of further practical implementation in new network mechanisms. The most adequate approach for describing queue management processes is the approach based on the use of the apparatus of differential-difference or integral equations of the interface state. In this case, the state of the interface can be understood as the average queue length, utilization factor, and average packet transmission rate. One example of such solutions is the use of nonlinear differential equations to describe the dynamics of changes in the utilization of interface queues, obtained on the basis of various approximations [4], [5]. According to Abualhaj *et al.* [6], the possibility of using neural networks and fuzzy logic methods for organizing queue management is substantiated. A compromise option from the point of view of the level of adequacy of the mathematical description of queue management processes and the expected computational complexity of the resulting solutions can be the approach presented in [7], [8]. It is based on the optimization problem statement of queue management in terms of balanced use of the available resource and queue utilization.

In multiprotocol label switching (MPLS) technology, the paradigm of ensuring a balanced load of network resources is called traffic engineering (TE), an important place that is occupied by solutions for balancing queues traffic engineering queues (TEQ) [9]. The ideas of TEQ were developed in [10]–[14]. Queue balancing is based on the introduction of appropriate metrics, and the applied network problem is reduced to an optimization problem of linear programming [10], [11]. These models in [5], [6] do not take into account the utilization of the router interface queues by their length. A model of dynamic guaranteed queuing is considered [14]–[19]. Furthermore, packet delay and buffer size are evaluated. According to Jithender and Mehar [20], a queue balancing model was proposed and studied, which takes into account the utilization of the queues in terms of the bandwidth allocated for serving each queue. The advantages of the model include full compliance of the solutions obtained in it with the concept of TEQ. The balancing process is based on the processing of individual flows of aggregated traffic. The disadvantages include the need to statically assign each bandwidth queue. According to Jaganathan and Vadivel [21] the mechanisms of vehicle queues at controlled crossroads are studied. The proposed methods can be used on telecommunication network nodes. According to AL-Allaf and Jabbar [22], the intelligent fish swarm inspired protocol (IFSIP) routing protocol and the queuing model (M/G/1) on network nodes are proposed to minimize the global packet delay. According to Lemeshko *et al.* [23] a reconfigurable nonlinear gentle random early dropping (RNLGRED) packet queuing scheme was proposed, designed to stabilize the queue length around a certain target value.

In flow models, the following problem of traffic distribution is considered in [22]–[24]. There are M classes of traffic distinguished by the type of priority N of serviced queues. The traffic intensity of the i -th class is $d_i (i = \overline{1, M})$. The part of the throughput of the outgoing channel assigned to the j -th queue is equal to $c_j (j = \overline{1, N})$. The following conditions must be met:

$$\sum_{j=1}^N c_j \leq c \quad (1)$$

$$\sum_{i=1}^M d_i \leq \sum_{j=1}^N c_j \quad (2)$$

The dynamic nature of the distribution of traffic packets in the queue of the network node is carried out by introducing a variable of x_{ij} , which means a part of the i -th traffic that will be sent for service to the j -th queue. To prevent the overloading of queues, the following conditions are introduced:

$$\sum_{i=1}^M \sum_{j=1}^N x_{ij} < c_j \quad (3)$$

$$\sum_{i=1}^M d_i x_{ij} < c_j \quad (4)$$

Since a thread of one class can only be served within one queue, the desired variable of x_{ij} takes only two values: 0 or 1, $x_{ij} = \{0,1\}$. The problem of determining the values of variables of x_{ij} is formulated as an optimization problem with different objective functions. According to Lebedenko *et al.* [24], the following objective function is minimized:

$$F(x) = \sum_{i=1}^M \sum_{j=1}^N f_{ij} x_{ij} \quad (5)$$

where f_{ij} characterizes the relative cost of using the resources of the j -th queue by packets of the i -th traffic. According to Lebedenko *et al.* [24], a vector is chosen as the desired variable:

$$\vec{x} = \begin{bmatrix} x_{ij} \\ c_j \end{bmatrix}, (i = \overline{1, M}, j = \overline{1, N}) \quad (6)$$

The following objective function is minimized:

$$F(x) = \vec{s}^t \vec{x} \quad (7)$$

where coordinates of the vector of the weight coefficient:

$$\vec{s} = \begin{bmatrix} s_{ij} \\ s_j \end{bmatrix} \quad (8)$$

Characterize the conditional cost (s_{ij}) of using the resources of the j -th queue by packets of the i -th traffic, as well as the cost (s_j) of allocating the j -th queue of one or another amount of bandwidth of the outgoing data transmission channel. According to Abualhaj *et al.* [6], for the purpose of balanced utilization of buffer resources, the following condition is additionally introduced:

$$f(p_j)Q_j \leq \alpha \quad (9)$$

Where α is the upper dynamically controlled utilization limit of the host queues. The problem of minimizing the following objective function is solved:

$$\min \alpha \quad (10)$$

Where the controlled variables are x_{ij} , c_j and α .

In the work of [21], the objective function of minimizing the sum of average queue lengths is used:

$$F = \sum_{j=1}^N f(p_j)Q_j \quad (11)$$

Where is some function of the characteristics of packets of the j -th queue with the priority of p_j . The value of the function of $f(p_j)$ must be greater, the higher the priority. In this case, a flow with a higher priority is served better than a flow with a lower priority.

According to Lemeshko *et al.* [25], conditions (3) and (4) are supplemented with conditions for preventing queue overloading along their length:

$$Q_j \leq Q_j^{\max} \quad (12)$$

Where Q_j^{\max} is the maximum capacity of the queue. The optimization problem associated with minimizing the objective function of the form is solved as (13):

$$F(x) = \sum_{i=1}^M \sum_{j=1}^N f_{ij} x_{ij} + \alpha \quad (13)$$

The objective function (13) includes functions (5) and (10). Optimization problems of (5) and (7) are linear programming problems, and problems (10), (11) and (13) are non-linear programming problems. In problems (5) and (7), the optimization results essentially depend on the metrics of f_{ij} and \vec{s}^t . The smaller the metrics of f_{ij} and \vec{s}^t , the more this queue will be utilized. Minimization of the objective functions (10), (11) and (13) ensures a balanced utilization of the queues of the network node [24], [25]. The general disadvantages of the considered flow queuing methods are: i) the amount of buffer resources is set and is not distributed among the queues; ii) in practice, each traffic is allocated its own queue and there is no need to distribute traffic between queues; and iii) the problem of optimal distribution of channel and buffer resources has not been considered, taking into account the limited QoS indicators. As a result of the analysis, in order

to ensure the best QoS indicators, this paper proposes a model for the optimal distribution of channel and buffer resources, taking into account constraints on delays and packet losses of heterogeneous flows.

3. THE PROPOSED MODEL FOR THE DISTRIBUTION OF CHANNEL AND BUFFER RESOURCES

Each flow of packets of the i -th class is assigned the i -th queue, $i = \overline{1, N}$. The total buffer resource with capacity L is distributed among the queues. To do this, a controlled variable y_i is introduced, which characterizes the share of buffer memory allocated for the i -th queue. In this case, the following conditions must be met:

$$\sum_{i=1}^N y_i = 1 \quad (14)$$

$$\sum_{i=1}^N y_i L = L \quad (15)$$

In order to distribute the total bandwidth of the communication channel (C), a controlled variable x_i is introduced, which characterizes the share of the bandwidth of the channel allocated for servicing the i -th queue. In this case, the following conditions must be met:

$$\sum_{i=1}^N x_i = 1 \quad (16)$$

$$\sum_{i=1}^N x_i C = C \quad (17)$$

The condition for preventing queue congestion is:

$$d_i < x_i C, \quad i = \overline{1, N} \quad (18)$$

The problem of optimal distribution of channel and buffer resources is solved by minimizing the sum of average delays (t) of packets of heterogeneous flows:

$$\min \sum_{i=1}^N t_i \quad (19)$$

With taking into account the following constraints:

$$t_i \leq t_{ip}, \quad i = \overline{1, N} \quad (20)$$

$$P_i \leq P_{ip}, \quad i = \overline{1, N} \quad (21)$$

Where t_{ip} and P_{ip} are the allowable values of the average delay and packet loss of the i -th flow in the network node. If each queue is an exponential queuing system with a limited queue of the M/M/1/L type [5], [6], then the probability of packet loss due to buffer overflow is defined by the following expression:

$$p_i = \frac{1-\rho_i}{1-\rho_i^{l_i+2}} \rho_i^{l_i+1} \quad (22)$$

Where ρ_i is the utilization of the i -th queue $i = \overline{1, N}$: $\rho_i = \frac{d_i}{x_i C}$; l_i is the capacity of buffer memory of the i -th queue: $l_i = y_i L$. The average queue length is determined by (23) [6], [7]:

$$Q_i = \frac{\rho_i^2}{(1-\rho_i)^2} p_{0i} \{1 - \rho_i^{l_i} [l_i(1 - \rho_i) + 1]\} \quad (23)$$

Where p_{0i} is the probability of no packets in the i -th queue:

$$p_{0i} = \frac{1-\rho_i}{1-\rho_i^{l_i+2}} \quad (24)$$

Average waiting time (W_i) and packet delay (T_i) are determined as (25) and (26):

$$W_i = \frac{Q_i}{d_i}, i = \overline{1, N} \tag{25}$$

$$T_i = W_i + 1/x_i c, i = \overline{1, N} \tag{26}$$

The problem statement of optimal distribution of channel and buffer resources taking into account constraints is a non-linear programming problem. To solve the problem of nonlinear programming, we will use the capabilities of the MATLAB system [15], represented by the optimization toolbox package and the fmincon program:

$$[z, fval] = fmincon(myfun', z0, A, B, Aeq, Beq, lb, ub, 'confun') \tag{27}$$

Where myfun is a non-linear optimization criterion (19). A and B are linear inequality constraints (18), Aeq and Beq are linear equality constraints (14-17), confun is non-linear constraints (20-21), $z_0 = x_0 \cup y_0$ —sets the initial values of the required variables x and y, taking values from 0 (lb) to 1 (ub).

4. ANALYSIS OF NUMERICAL RESULTS

In order to analyze the solution to the problem (19), taking into account constraints (20) and (21), we consider an example in which the following were used as initial data: i) number of threads (queues)-8; ii) the total bandwidth of the communication channel-230 packets/s; iii) total buffer memory-80 packets; iv) intensity of packets arrival—from 1 packets/s to 11 packets/s; and v) admissible values of average delay and packet loss probabilities of eight flows on a network node:

$$t_p = [0.02 \ 0.04 \ 0.05 \ 0.07 \ 0.09 \ 0.1 \ 0.15 \ 0.2];$$

$$P_p = [0.1 \ 0.05 \ 0.01 \ 0.005 \ 0.001 \ 0.0005 \ 0.0001 \ 0.00005].$$

Low-numbered packet flows have more stringent delay time requirements, while high-numbered packet flows have a higher packet loss rate. The results of the distribution of the total bandwidth (x) and the total capacity of buffer memory (y) between the queues at different packet arrival intensities for each flow are shown in Figures 1 and 2. Flows with more stringent requirements for packet delay are allocated the most bandwidth of the communication channel, and flows with more stringent requirements for the probability of packet loss are allocated a larger capacity of buffer memory. The results of calculating the average delay (t) and packet loss probability (P) for each packet flow are shown in Figures 3 and 4. It can be seen from Figures 3 and 4 that the channel and buffer resource allocations carried out satisfy the specified requirements for the delay time and packet loss probability of each flow.

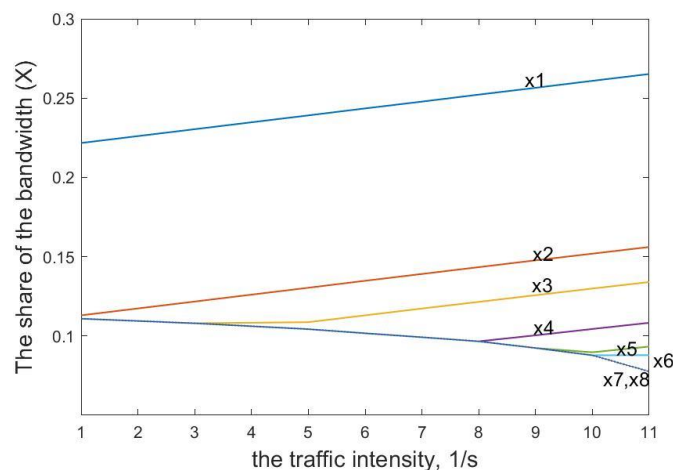


Figure 1. The dependence of the share of the bandwidth of the communication channel allocated for queues on the traffic intensity of each flow

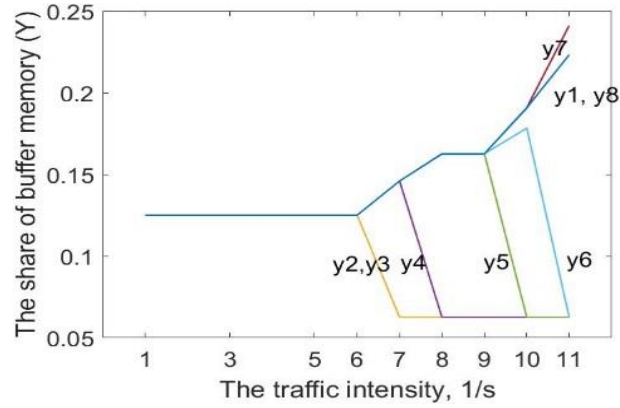


Figure 2. The dependence of the share of buffer memory allocated for queues on the traffic intensity of each flow

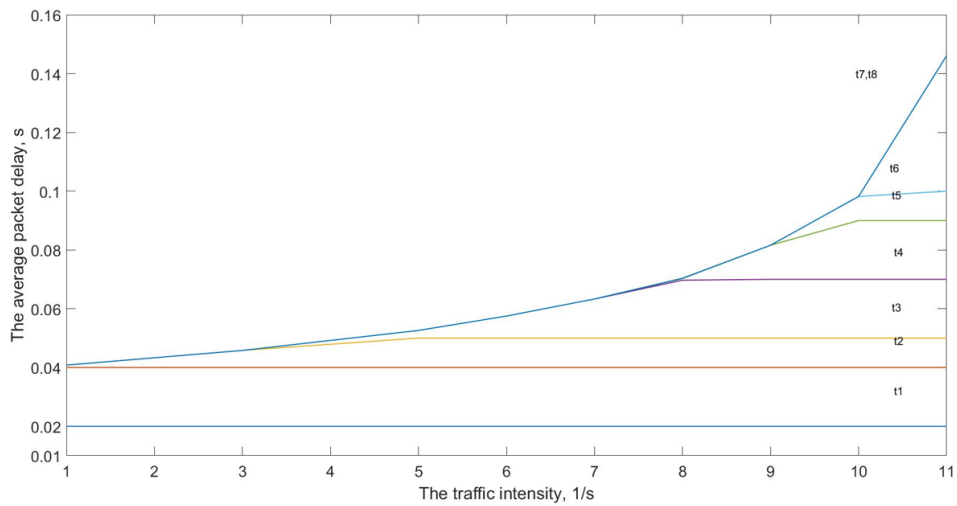


Figure 3. The dependence of the average packet delay on the traffic intensity of each flow

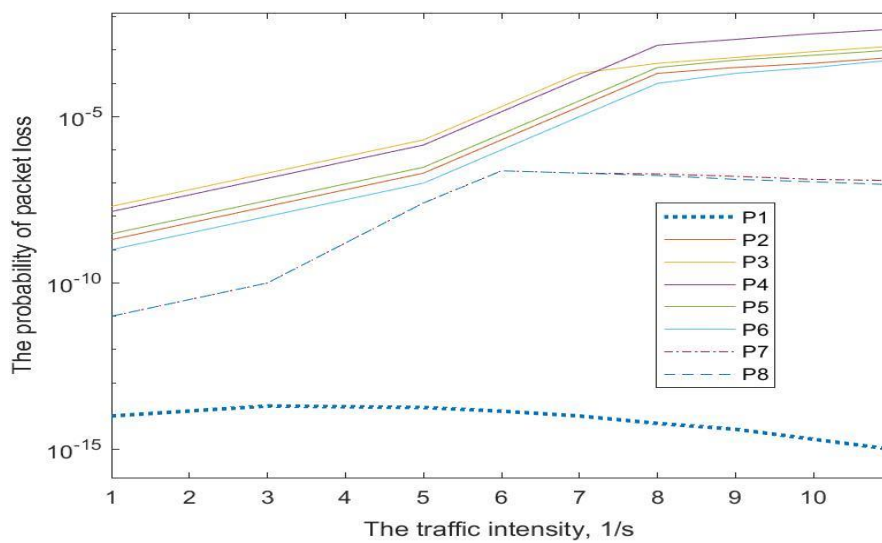


Figure 4. The dependence of the average packet delay on the traffic intensity of each flow

5. CONCLUSION

Increasing the QoS indicators is inextricably linked with the improvement of queue management mechanisms. One of the main areas of QoS is the development of flow models of queuing systems that will allow you to fully realize the benefits of dynamic traffic control based on the current utilization of channel and buffer resources. A model of dynamic queue servicing is proposed, which differs from the known ones in that the mechanisms of dynamic allocation of channel and buffer resources are implemented, and allow differentiated guaranteed servicing of flows of various classes, taking into account constraints on the QoS indicators. The development of the proposed approach is seen as an increase in consistency in solving queue management problems with other traffic management problems, for example, access control, flow routing, and resource reservation.




REFERENCES

- [1] J. Sérgio and B. Martins, "Quality of Service in IP Networks," in *Managing IP Networks*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2004, pp. 57–142. doi: 10.1002/0471722987.ch3.
- [2] M. E. G. Mustafa and S. A. Talab, "The Effect of Queuing Mechanisms First in First out (FIFO), Priority Queuing (PQ) and Weighted Fair Queuing (WFQ) on Network's Routers and Applications," *Wireless Sensor Network*, vol. 08, no. 05, pp. 77–84, 2016, doi: 10.4236/wsn.2016.85008.
- [3] K. A. Alnowibet, "Nonstationary erlang loss queues and networks," Ph.D. dissertation, North Carolina State Univ., Raleigh, NC, USA, 2004.
- [4] T.-Y. Tsai, Y.-L. Chung, and Z. Tsai, "Introduction to Packet Scheduling Algorithms for Communication Networks," in *Communications and Networking*, Sciyo, 2010. doi: 10.5772/10167.
- [5] S. H. Hosseini, M. Shabani, and B. N. Araabi, "A Neuro-Fuzzy Control for TCP Network Congestion," 2009, pp. 93–101. doi: 10.1007/978-3-540-89619-7_10.
- [6] M. M. Abualhaj, M. M. Al-Tahravi, A. H. Hussein, and S. N. Al-Khatib, "Fuzzy-Logic Based Active Queue Management Using Performance Metrics Mapping into Multi-Congestion Indicators," *Cybernetics and Information Technologies*, vol. 21, no. 2, pp. 29–44, Jun. 2021, doi: 10.2478/cait-2021-0017.
- [7] O. Lemeshko, T. Lebedenko, and A. Al-Dulaimi, "Improvement of Method of Balanced Queue Management on Routers Interfaces of Telecommunication Networks," in *2019 3rd International Conference on Advanced Information and Communications Technologies (AICT)*, Jul. 2019, pp. 170–175. doi: 10.1109/AIACT.2019.8847749.
- [8] Y. Li, S. Panwar, and T. Laboratories, "Performance Analysis of MPLS TE Queues for QoS Routing," *Simulation series*, vol. 36, no. 3, pp. 170–174, 2004.
- [9] O. Lemeshko, A. S. Ali, and O. Simonenko, "A queue management model on router of active network," in *The Experience of Designing and Application of CAD Systems in Microelectronics*, Feb. 2015, pp. 419–421. doi: 10.1109/CADSM.2015.7230891.
- [10] A. V. Lemeshko, A. Salem Ali, and Z. A. Sabeeh, "Researching and Designing of the Dynamic Flow-Based Queue Balancing Models on Telecommunication Network Routers," *International Journal of Wisdom Based Computing*, vol. 2, no. 1, pp. 47–51, 2012.
- [11] O. Lemeshko, O. Yeremenko, and M. Yevdokymenko, "MPLS Traffic Engineering Solution of Multipath Fast ReRoute with Local and Bandwidth Protection," in *International Conference on Computer Science, Engineering and Education Applications*, 2020, pp. 113–125. doi: 10.1007/978-3-030-16621-2_11.
- [12] S. Haryadi, "Telecommunication Traffic Unit and Traffic Mathematical Model," 2018. osf.io/preprints/inarxiv/jf5ry
- [13] A. A. Qodirov, "Modeling the Gilbert Model for Communication Channels Based on Artificial Intelligence," in *2021 International Conference on Information Science and Communications Technologies (ICISCT)*, Nov. 2021, pp. 1–3. doi: 10.1109/ICISCT52966.2021.9670347.
- [14] S. O. Hassan *et al.*, "Random early detection-quadratic linear: an enhanced active queue management algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 4, pp. 2262–2272, Aug. 2022, doi: 10.11591/eei.v11i4.3875.
- [15] F. Wahida Binti Zulkefli, P. Ehkan, M. N. M. Warip, and N. Y. Phing, "A efficacy of different buffer size on latency of network on chip (NoC)," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 2, pp. 438–442, Jun. 2019, doi: 10.11591/eei.v8i2.1422.
- [16] A. U.B. and Q. A.A., "Implementation of the Reinforcement Learning Mechanism in the Random Access Channel Procedure," in *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, Oct. 2020, pp. 1–5. doi: 10.1109/AICT50176.2020.9368724.
- [17] U. Amirsaidov and A. Qodirov, "A packet delay assessment model in the data-link layer the LTE," *JOIV : International Journal on Informatics Visualization*, vol. 5, no. 4, pp. 402–408, Dec. 2021, doi: 10.30630/joiv.5.4.601.
- [18] J. Prados-Garzon, A. Laghrissi, M. Bagaa, and T. Taleb, "A Queuing Based Dynamic Auto Scaling Algorithm for the LTE EPC Control Plane," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–7. doi: 10.1109/GLOCOM.2018.8648023.
- [19] L. Mas, J. Vilaplana, J. Mateo, and F. Solsona, "A queuing theory model for fog computing," *The Journal of Supercomputing*, vol. 78, no. 8, pp. 11138–11155, May 2022, doi: 10.1007/s11227-022-04328-3.
- [20] J. Jithender and A. Mehar, "Estimation Of Cycle By Cycle Queue Length At Approaches Of Signalized Intersection Under Mixed Traffic Conditions," *Suranaree Journal of Science and Technology*, vol. 28, no. 2, pp. 1–9, 2021.
- [21] R. Jaganathan and R. Vadivel, "Intelligent Fish Swarm Inspired Protocol (IFSIP) For Dynamic Ideal Routing in Cognitive Radio Ad-Hoc Networks," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 1063–1074, Nov. 2021, doi: 10.12785/ijcds/100196.
- [22] A. AL-Allaf and A. I. A. Jabbar, "Reconfigurable Nonlinear GRED Algorithm," *International Journal of Computing and Digital Systems*, vol. 9, no. 5, pp. 1009–1022, Sep. 2020, doi: 10.12785/ijcds/090521.
- [23] O. Lemeshko, T. Lebedenko, and M. Holoveshko, "Development and Research of Active Queue Management Method on Interfaces of Telecommunication Networks Routers," 2021, pp. 1–20. doi: 10.1007/978-3-030-71892-3_1.
- [24] T. N. Lebedenko, A. V. Simonenko, and F. A. R. Arif, "A queue management model on the network routers using optimal flows aggregation," in *2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET)*, Feb. 2016, pp. 605–608. doi: 10.1109/TCSET.2016.7452129.




- [25] L. Oleksandr, N. Olena, and V. Tetiana, "Hierarchical coordination method of inter-area routing in telecommunication network," in *2016 International Conference Radio Electronics & Info Communications (UkrMiCo)*, Sep. 2016, pp. 1–4. doi: 10.1109/UkrMiCo.2016.7739626.

BIOGRAPHIES OF AUTHORS



Amirsaidov Ulugbek    graduated from Tashkent University of Information Technology (TUIT) in 1980 with a specialty Engineer of Electrical Communication with an honors diploma. He actively participates in scientific, technical, and scientific-methodical conferences at the republican and international levels. He is an Associate Professor and a Doctor of Science in the Data Communication Networks and Systems Department at the Telecommunications technologies faculty of the TUIT. His professional interests include network research and development, next-generation network architecture, and network performance issues, including quality of service. He has more than 50 scientific articles and conference materials, 2 textbooks on technical subjects, 2 monographs, and 4 certificates for a software product. He can be contacted at email: amirsaidov.ulugbek@gmail.com.



Qodirov Azamat    graduated from Tashkent University of Information Technology (TUIT) in 2015 with a specialty in Telecommunication Engineering with an honors diploma. He actively participates in scientific, technical, and scientific-methodical conferences at the republican and international levels. He is an Assistant Professor and a Ph.D. student at the Data Communication Networks and Systems Department at Telecommunications technologies faculty of TUIT. His professional interests include network research and development, next-generation network architecture, machine learning, reinforcement learning, and network performance issues, including quality of service. He has more than 17 scientific articles and conference materials. He can be contacted at email: azamattuit2013@gmail.com.