http://doi.org/10.35784/iapgos.3505

received: 17.02.2023 | revised: 29.04.2023 | accepted: 22.09.2023 | available online: 30.09.2023

OVERLOAD AND TRAFFIC MANAGEMENT OF MESSAGE SOURCES WITH DIFFERENT PRIORITY OF SERVICE

Valerii Kozlovskyi[1], Valerii Kozlovskyi[2], Andrii Toroshanko[2], Oleksandr Toroshanko[3], Natalia Yakymchuk[4]

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Institute of Special Communications and Information Protection, Kyiv, Ukraine, [2]National Aviation University, Faculty of Cyber Security and Software Engineering, Department of Information Protection System, Kyiv, Ukraine, [3]Taras Shevchenko National University of Kyiv, Department of Cyber Security and Information Protection, Kyiv, Ukraine, [4]Lutsk National Technical University, Faculty of Computer and Information Technologies, Department of Electronics and Telecommunications, Lutsk, Ukraine

Abstract. *The scheme of dynamic management of traffic and activity of message sources with different priority of service is considered. The scheme is built on the basis of the neuroprognostic analysis model and the gradient descent method. For prediction and early detection of overload, the apparatus of the general theory of sensitivity with indirect feedback and control of activity of message sources is used. The control algorithm is started at the bottleneck of the network node. It uses a recursive prediction approach where the neural network output is referred to as many steps as defined by a given prediction horizon. Traffic with a higher priority is served without delay using the entire available bandwidth. Low-priority traffic will use the remaining bandwidth not used by higher-priority traffic. An algorithm for estimating the maximum available bandwidth of a communication node for traffic with a low service priority has been developed. This approach makes it possible to improve the efficiency of channel use without affecting the quality of service for high-priority traffic.*

Keywords: telecommunication network, overload prediction, sensitivity function, neural network, gradient descent method, service priority

PRZECIĄŻENIE I ZARZĄDZANIE RUCHEM ŹRÓDEŁ WIADOMOŚCI O RÓŻNYCH PRIORYTETACH USŁUG

Streszczenie. *Rozważono schemat dynamicznego zarządzania ruchem i aktywnością źródeł komunikatów o różnym priorytecie obsługi. Schemat zbudowany jest w oparciu o model analizy neuroprognostycznej oraz metodę gradientu. Do prognozowania i wczesnego wykrywania przeciążenia wykorzystuje się aparaturę ogólnej teorii wrażliwości z pośrednim sprzężeniem zwrotnym i kontrolą aktywności źródeł komunikatów. Algorytm sterowania jest uruchamiany w wąskim gardle węzła sieci. Wykorzystuje metodę predykcji rekurencyjnej, w której dane wyjściowe sieci neuronowej są odnoszone do tylu kroków, ile określono w danym horyzoncie predykcji. Ruch o wyższym priorytecie jest obsługiwany bez opóźnień z wykorzystaniem całej dostępnej przepustowości. Ruch o niskim priorytecie będzie wykorzystywał pozostałą przepustowość niewykorzystaną przez ruch o wyższym priorytecie. Opracowano algorytm szacowania maksymalnej dostępnej przepustowości węzła komunikacyjnego dla ruchu o niskim priorytecie usługi. Takie podejście umożliwia poprawę efektywności wykorzystania kanałów bez wpływu na jakość obsługi ruchu o wysokim priorytecie.*

Słowa kluczowe: sieć telekomunikacyjna, predykcja przeciążenia, funkcja czułości, sieć neuronowa, metoda opadania gradientu, priorytet usługi

Introduction

The quality of service provision in the telecommunications network is largely determined by routing algorithms, traffic management and operation in overload conditions.

In the theory of telecommunications, overload is defined as the loss of information during transmission caused by an increase in network load [3, 6, 12]. The task of overload management is to develop appropriate algorithms to prevent or reduce such loss and, first of all, each user should be provided with mechanisms for determining and receiving data from the network. For example, if a user wishes to obtain low-latency mass service, the system must provide a mechanism to achieve the goal. If the network is unable to prevent the loss of user data, then it is necessary to try to limit the loss as much as possible, and, subsequently, to try to be fair to all affected users.

The effectiveness of the traffic and overload management system largely depends on the routing methods used and the speed of the computing facilities, which should ensure minimal data delay in the network, as well as avoiding or minimizing the probability of overload [4, 13]. The accuracy of overload prediction and feedback control of traffic and data flows in high-speed computer networks is highly dependent on data transmission delays between network communication nodes.

As a result, the responses of the control commands will take effect within the network after some delay, and the control information received at the data sources or network access points may turn out to be outdated [1, 18].

There is a distinction between active and reactive overload management, which coexist in most networks. In a strict active scheme, the overload management mechanism is the reservation of network resources. In a reactive scheme, data sources need to monitor and respond to changes in the network state to prevent overload. Both management methods have their strengths and weaknesses. In systems with active management, users can be guaranteed lossless data delivery, but the number of active users must be limited. Reactive management allows much greater flexibility in the allocation of resources, but the probability of overloading increases.

The key performance indicators of the telecommunications network and the quality of user service largely depend on the traffic and overload management mechanisms used. The above conditions the relevance and necessity of research in this direction.

1. Literature review

In [18], a control method based on the sensitivity function of the performance of the telecommunications network is considered for traffic management. The sign of the performance sensitivity function provides the optimal direction for adjusting the data source speed.

In [7] proposed a traffic and overload management algorithm for mass service systems with uniform time distribution (Least Favorable Distribution, LFD). In modern telecommunication networks, flow distributions have self-similar properties, so the obtained asymptotic estimates will lead to unreasonably optimistic conclusions.

The results obtained in [7, 19] are of a general nature, their use requires new non-traditional approaches to solving the problem as a whole. Realistic estimates can be obtained by applying neural network models that must adapt to load spikes and variations in the probability distributions of application flows [18].

Increasing the memory volume of input buffers to prevent overload causes the problem of buffer bloat [19]. The number of unprocessed packets in the buffers increases, which leads to delays in their processing and can cause an avalanche process of buffer overflow and packet loss.

Works [5, 8] present a detailed classification of modern methods of overload management, identified advantages and disadvantages of their application in certain conditions of network operation. In particular, the RED (Random Early Detection) method, which is used in the TCP protocol to detect and prevent overload, is considered. However, the mentioned works are, to a certain extent, review in nature. Analytical expressions and quantitative comparative evaluations of the RED method and other methods, in particular, Tail Drop, WRED (Weighted Random Early Detection), etc., are not provided.

In works [6, 9], the TCP Veno algorithm, which is quite common in overload management systems, was investigated. The use of this algorithm is quite effective in wireless networks with a high rate of lost packets. It tries to isolate non-overload losses so as not to include overload control algorithms where it is not needed. Meanwhile, the very method of recognizing the nature of losses according to the TCP Veno algorithm is, in fact, quite trivial (linear classifier). But it requires a detailed analysis of statistics of a sufficiently large volume, which limits its use.

2. Formulation of the problem

In works [13, 18], a model of dynamic neuroprognostic analysis using the sensitivity function is proposed for optimal traffic management and overloading of the telecommunications network [15, 17]. The algorithm is based on traffic management and the activity of message sources with different service priorities. Traffic with a higher priority is served without delay using the available part of the bandwidth. Low-priority traffic will use the remaining bandwidth not used by higher-priority traffic.

Fig. 1. shows a model of an overload control system with joint processing of traffic with different priorities [16].

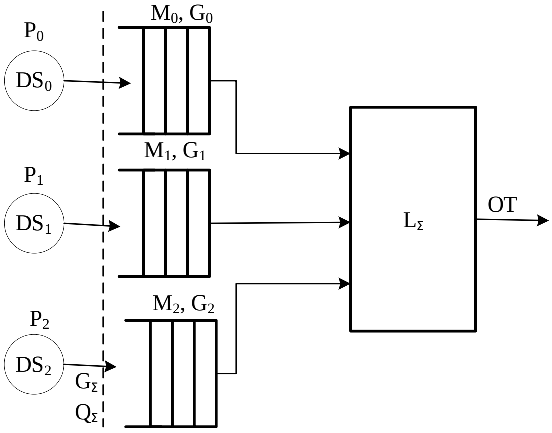

Fig. 1. Traffic management with joint processing in systems with different priorities

The vertical dashed line represents the joint control of the packet arrival rate taking into account the generalized threshold value Q_{Σ} for traffic with priority classes P_0, P_1 and P_2. For such a model, the total outgoing traffic L_{ε} is formed from the condition of analysis and taking into account the total capacity G_{Σ} of the output link of the node.

Fig. 2 shows a model of an overload control system with isolated traffic processing L_2 with a lower priority class [16].

A corresponding queue is organized for each type of traffic on the network node. Class P_0 has the highest priority, P_2 – the lowest.

In Fig. 1 and Fig. 2 the following designations are accepted (indexes indicate the corresponding priority):

- DS_0, DS_1, DS_2 – sources of input data;
- M0, M1, M2 – blocks of buffer memory for incoming packets;
- M_0, M_1, M_2 – the number of incoming packets with the corresponding priority, expected to be issued at time t;
- G_{Σ} – the total number of packets of all priorities expected to be issued at time t;
- Q_{Σ}, Q_2 – the threshold (limit) value of the queue length of total traffic and traffic with priority P_2,
- OT – outgoing traffic;
- HPOT – higher priority outgoing traffic;
- LPOT – lower priority outgoing traffic.

This approach allows for improvement in the efficiency of using the channel without affecting the quality of service for high-priority traffic. When the residual bandwidth changes over time, it is necessary to constantly adjust the traffic speed only of low-priority sources. This simplifies the construction of the incoming traffic control scheme. At the same time, the task of estimating the maximum available bandwidth for low-priority traffic arises.

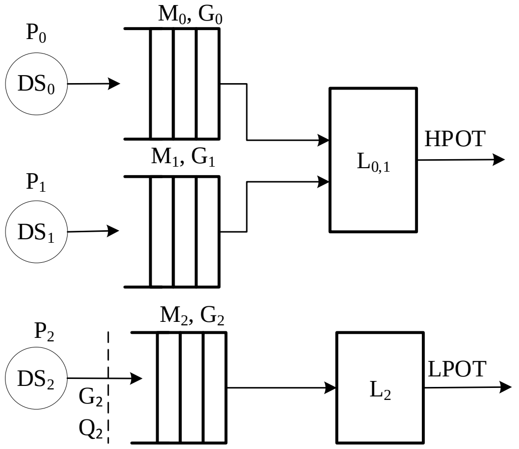

Fig. 2. Traffic management with separate processing in systems with different priority

3. Estimate available bandwidth for low-priority traffic

The structural diagram of the traffic management, forecasting and overload detection system is shown in Fig. 3. The scheme is built on the basis of a neural network using the sensitivity function of the key parameter of the network's efficiency – productivity [15, 16]. On the IR (Input Regulator) device the input traffic is regulated by the feedback control signal U.

The output queue length G is controlled by the OR (Output Regulator) device, taking into account the current G and the predicted $G^{\wedge}$ output queue length.

The deviation E of the current value of the queue size from its threshold value $Q(t)$ is carried out by the CQ (Control Queue) device. As a rule, the parameter Q is a constant value, i.e. $Q(t) = Q$.

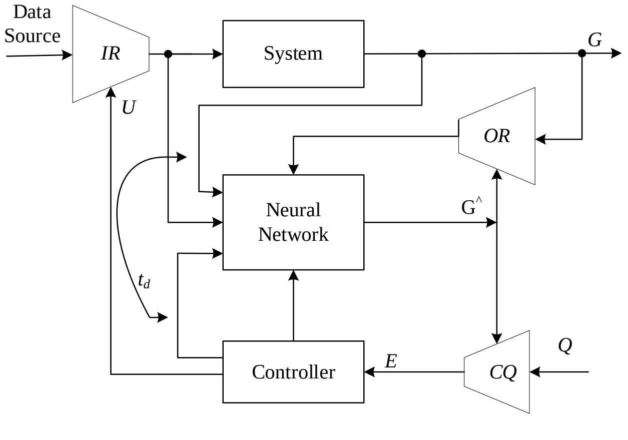

Fig. 3. Traffic control scheme

In [16], a mathematical model was developed for a transmission system to a network buffer with one data source and a fixed circular delay t_d.

We will evaluate the available bandwidth for low-priority traffic.

Let's mark $Y(t)$ as the available bandwidth at time t for low-priority traffic P_2. Available bandwidth $Y(t)$ is a function of node usage as well as traffic $L_{0,1}$ with higher priority P_0 i P_1 (Fig. 2). The expected available rate $Y^{\wedge}(t)$ will be equal to the bandwidth not used by the high-priority traffic plus the bandwidth required to fill the buffer capacity to a given threshold value Q at a given time t.

A nonlinear network system can be described by the following discrete expression [10, 11]:

$$Y(t) = F\begin{bmatrix} y(t-1), \dots, y(t-1), u(t-t_d), \\ \dots, u(t-t_d-m) \end{bmatrix} \tag{1}$$

where $Y(t)$ – scalar output (i.e., service delay, queue length, etc.); $u(t)$ – scalar input (feedback control signal U); $F[*]$ is an unknown nonlinear function generated by a neural network; $1 \dots m$ are orders of functions $y(t)$ and $u(t)$; $t_d \geq 1$ – circular delay.

The task of the control algorithm is to generate the control signal $u(t)$ so that the output of the system $Y(t)$ is as close as possible to the threshold value Q. The neural model for the unknown system (1) can be expressed as

$$Y^{\wedge}(t+1) = F^{\wedge}\begin{bmatrix} y(t), \dots, y(t-B+1), u(t-t_d+1), \\ \dots, u(t-t_d-m+1) \end{bmatrix} \tag{2}$$

where $Y^{\wedge}(t+1)$ – the predicted one-step (one-step) value of the output of the neural network $F^{\wedge}$; $B > d$ – prediction horizon.

If the neural network is properly configured, then the squared error $(Y(t+1) - Y^{\wedge}(t+1))^2 = \varepsilon$ is insignificant, so the value of $Y^{\wedge}(t+1)$ is close to the predicted output systems (1). As a result, the control signal $u(t-d_t+1)$ can be chosen so that the value of $Y^{\wedge}(t+1)$ at the next step $t+1$ is as close as possible to the value of Q.

Since the neural model is an asymptotic control system, it can be used to predict the next R values of the system output at the observation interval T:

$$Y^{\wedge} = [y(t+d_t), y(t\ +d_t+1), \dots, y(t+B)]^T \tag{3}$$

and the error vector can be obtained as:

$$E = [e(t+d_t), e(t+d_t+1), \dots, e(t+L)]^T$$

$$e(t+i) = r(t+i) - y^{\wedge}(t+1), d_t \leq i \leq L$$

where

$$R = [r(t+d_t), r(t+d_t+1), r(t+L)]^T$$

is a further predicted values of the system output at the observation interval T.

Given the above objective function for calculating the remaining bandwidth for traffic with a lower priority can be determined through the deviation index J from the predicted value of the controlled parameter as follows:

$$J = \tfrac{1}{2}\textstyle\sum_{i=d_t}^{L}\ [r(t+i) - y^{\wedge}(t+1)]^2.$$

Let us define feedback control signals U as

$$U = [u(t), u(t+1), \dots, u(t+B-1]^T.$$

Then the task of finding the optimal value of the remaining bandwidth for traffic with a lower priority is to find the values of the feedback control signals U such that the deviation index J is minimal:

$$U^{i+1} = u^i - \frac{\partial J}{\partial U^k}; \frac{\partial J}{\partial U^k} = -\frac{\partial Y^{\wedge k}}{\partial U^k}E^k$$

where $U^k(t+i)$ denotes the *k-th* iteration of control signal determination $u(t+i)$.

To solve this problem, you can use the following rule of gradient descent [16, 17]:

$$u^{k+1}(t+1) = u^k(t+i) + \frac{\partial Y^{\wedge k}}{\partial U^k}E^k$$

$$u^0(t+i) = u(t-1) + i;\ i = 0,1, \dots, L-1$$

k – iteration index (1, 2, …); $u^k\ (t+i)$ – k-th iteration of the control signal formation $u(t+i)$.

Control sequence

$$u^{k-1}(t-i), i = 0,1, \dots, B-1$$

is used to determine the predicted sequence

$$y^{\wedge k}(t+j), j = d_t, \dots, B$$

For the first iteration ($k = 1$), the control sequence starts with the value $u^0(n+i)$, which was determined in the last period. Then the deviation index J is minimized iteratively until its extremum is found.

According to the principle of receding control horizon [2, 14], only the first control signal is used to determine the control sequence U:

$$u(t+1) = u(t) +\ \textstyle\sum_{i=d_t}^{B} e(t+i)\frac{\partial y^{\wedge}(t+i)}{\partial u(t)}$$

One iteration is performed in each period.

Thus, the algorithm for calculating the maximum available bandwidth of a communication node for traffic with a low service priority involves the following steps:

1. Determination of the prediction horizon using the developed analytical expression and based on the specified performance requirements and other key parameters of the network.
2. Determination of the threshold values of the main network parameters to ensure the specified quality of service. The described method can be used both for systems with a fixed threshold and, if necessary, for dynamic systems with time-varying thresholds.
3. Selection of feedback control functions U that minimize the difference between the expected and actual value of the controlled output of the system. Minimization is performed by a neural network based on the considered gradient descent rule.
4. Use the first element of the control function U as the next control input and repeat the entire process in the next cycle.

4. Conclusions

Dynamic neural models of traffic management in a telecommunications network with different service priorities are considered. The proposed algorithm for forecasting and prevention of node overload in a network bottleneck. It is shown that an attempt to combat overload by simply increasing the buffer capacity does not lead to a solution to the problem, but on the contrary, leads to bufferbloat and an unacceptable increase in service delays.

An algorithm and analytical expressions are developed for calculating the maximum available bandwidth of a communication node for traffic with a low service priority. The algorithm involves the following 4 steps: definition of the horizon based on the requirements for ensuring the key parameters of network efficiency; determination of threshold values of the main parameters of the network to ensure the given quality of service; the selection of control functions of the feedback U, which ensure the minimum difference between the expected and the actual value of the controlled output of the system; using the first element of the control function U as the next control input, repeat the entire process in the next cycle.

The principle of traffic separation and processing according to the established priority classes makes it possible to use the bandwidth of the network in the most optimal way without losing the quality of user service. Traffic with a higher priority is served without delay using the entire available bandwidth. Low-priority traffic will use the remaining bandwidth not used by higher-priority traffic.

The algorithm in question uses a recursive prediction approach, where the output of the neural network refers to itself to go as far as the prediction horizon requires. Another approach is a single model in which all predictions are made at the same time at different nodes in the output layer of the neural network.

For prediction and early detection of overload, the apparatus of the general theory of sensitivity with indirect feedback and control of the activity of message sources is used. The resulting solutions allow you to significantly save channel and computing resources of the network.

References

[1] Bonaventure O.: Computer Networking: Principles, Protocols and Practices. Release. 2018.

[2] Golmohammadi A.: Prioritizing Service Quality Dimensions: A Neural Network Approach. World Academy of Science, Engineering & Technology 42, 2010, 602–605.

[3] Göransson P. et al.: Software Defined Networks: A Comprehensive Approach, 2nd ed. Morgan Kaufmann, 2017.

[4] Klymash M. M., Strykhaliuk B. M., Kaidan M. V.: Teoreticheskiye osnovy telekommunikatsionnykh setyei. LAP LAMBERT Academic Publishing, Saarbrücken 2014.

[5] Korolkova A. V., Kulyabov D. S., Tchernoivanov A. I.: On the Classification of RED Algorithms. Bulletin of the Russian Peoples' Friendship University 3, 2009, 34–46.

[6] Kurose J. F., Keith W. R.: Computer Networking: A Top-Down Approach, 7th Ed. Pearson Education, Inc., 2017.

[7] Lu Z. et al.: Overload Control for Signaling Congestion of Machine Type Communications in 3GPP Networks. PLOS ONE, 2016. [http://doi.org/10.1371/journal.pone.0167380].

[8] Maximov V. V., Chmykhun S. O.: Classification of algorithms of controlling networks congestions. Scientific proceeding of Ukrainian Research Institute of Communication 5(33), 2014, 73–79.

[9] Maxymov V. V., Chmykhun S. O.: Research of the algorithm of controlling congestion TCP Veno. Telecommunication and Information Technologies 4, 2015, 30–36.

[10] Shooman M. L.: Reliability of Computer Systems and Networks – Fault Tolerance, Analysis, Design. JohnWiley&Sons, Inc., NewYork 2002.

[11] Snarskyy A. A., Lande D. V.: Modelyrovanye slozhnыkh setey. Kyiv 2015.

[12] Stallings W.: Foundations of Modern Networking: SDN, NFV, QoE, IoT, and Cloud. Pearson Education, Inc., Old Tappan, New Jersey 2016.

[13] Tanenbaum A. S., Wetherall D. J.: Computer Networks. Prentice Hall, Cloth, 2011.

[14] Tasad R., Ruggieri M.: Technology Trends in Wireless Communications. Artech House, Boston – London 2003.

[15] Tkachuk A. et al.: Basic Stations Work Optimization in Cellular Communication Network. D. Cagánová et al. (eds.), Advances in Industrial Internet of Things, Engineering and Management, EAI. Springer Innovations in Communication and Computing, 2021, 1–19.

[16] Toroshanko O. S.: Multi-step model for prognostication and detection of telecommunication network overload. Telecommunication and Information Technologies 2(63), 2019, 35–43.

[17] Toroshanko Ya. I.: Sensitivity analysis of systems of mass service on the base of model of adaptation and regulation of foreign traffic. Herald of Khmelnytskyi national university 6(243), 2016, 171–175.

[18] Vinogradov N. et al.: Development of the Method to Control Telecommunication Network Congestion Based on a Neural Model. Eastern-European Journal of Enterprise Technologies 2(9), 2019, 67–73.

[19] Vynohradov N. A., Drovovozov V. Y., Lesnaya N. N., Zembytskaya A. S.: Analyz nahruzky na sety peredachy dannыkh v systemakh krytychnoho prymenenyya. Zvyazok 1(61), 2006, 9–12.

D.Sc. Valerii Kozlovskyi
e-mail: valerey@ukr.net

Research interests: Microwave devices. Heterogeneous transmission lines.
Author of nearly 200 publications.

http://orcid.org/0000-0003-0234-415X

D.Sc. Valerii Kozlovskyi
e-mail: vv_k@nau.edu.ua

Research interests: cyber security, neural networks, traffic control.
Author of nearly 150 publications

http://orcid.org/0000-0002-8301-5501

Andrii Toroshanko
e-mail: atoroshanko@gmail.com

Research interests: heterogeneous networks, self-similar traffic. network congestion.
Author of 7 publications.

http://orcid.org/0000-0002-0816-657X

Ph.D. Oleksandr Toroshanko
e-mail: toroshanko@gmail.com

Research interests: wireless sensor networks, traffic control, cyber security, neural networks.
Author of nearly 30 publications

http://orcid.org/0000-0002-2354-0187

Ph.D. Natalia Yakymchuk
e-mail: n.yakymchuk@lntu.edu.ua

Research interests: diagnostics and control of the telecommunication networks state, end-to-end diagnostics, congestion management.

http://orcid.org/0000-0002-8173-449X