Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 149 (2019) 177–184

Procedia Computer Science

www.elsevier.com/locate/procedia

ICTE in Transportation and Logistics 2018 (ICTE 2018)

Minimization method for average packet delay in data transmission networks

Sergiy Panchenko[a,]*, Karyna Trubchaninova[a], Ilja Korago[b]

[a] Ukranian State University of Railway Transport, Kharkov, Ukraine 61000
[b]Institute of Transport, Riga Technical University, Kalku 1, LV-1006, Latvia

Abstract

Increasing the speed of transmission of information stream in the railway logistics system of cargo delivery can be achieved by developing new methods of routing information stream, taking into account changes in the structure of the data transmission network and the characteristics of transmission channels and are based on researches of the values of the average data packet delay of these information streams. The mathematical model of rational distribution of data stream on the routes in a data transmission network was formulated on the basis of the use of adaptive routing algorithms to optimize the structure of the telecommunications network of railway transport as an important component of the logistics system for cargo delivery. The way of calculation of the maximum values of the intensity of data stream between separate knots of a telecommunication network is proposed. The found dependences of change of error probability in a data packet on its length allow to maximize the message-transmission rate on data transmission channels for a given probability of data transmission distortion taking into account memory limitations and minimizing processor system costs when processing a message.

Keywords: Railway transport; Logistics system; Freight transportation; Adaptive routing; Average data packet delay

1. Introduction

The operational model of the world's railways undergoes considerable transformation [1]. The railway industry introduces digitalization and automation, there is an increasing need for personalized mobility and new logistics

* Corresponding author. Tel.: +380-577301081.
E-mail address: karyna.trubchaninova@kart.edu.ua

1877-0509
10.1016/j.procs.2019.01.121

solutions that ensure the transparency of the transportation process. The railways try to improve their transportation services for the carriage of goods in accordance with the requirements set by customers of supply chains [2,3]. They are building complex logistics systems of cargo delivery that require the transfer of large volumes of information stream between a large number of participants in the transportation process and customers. Information act as a promoter of the logistics system, and inaccurate information or its absence can cause interruptions in the logistics system and the appearance of failures in the cargo delivery chains. For effective information support of cargo support in the logistics chain, it is important to optimize the structure of the telecommunications network of railway transport on the basis of the use of adaptive routing algorithms. The importance of telecommunication networks in transportation systems, in particular in High-speed railway (HSR), is emphasized in the research [4]. This work focuses on building radio resource management (RRM) systems for HSR wireless telecommunications. Research [5] is directed at applying wireless sensor networks to improve the quality of the transportation process on the railway transport, and also to support the use of mobile units (cars, locomotives) for better control and optimal use. Therefore, the issue of optimizing the structure of the telecommunications network of railway transport on the basis of the use of adaptive routing algorithms becomes relevant. In modern data transmission networks, there is a direct dependence of network capacity on the capacity of routers that handle internetwork traffic. Due to the high workload of routers, getting a really optimal data transfer route is quite difficult. Most data networks use the criteria for the maximum carrying capacity of the data transmission channel when choosing a transmission route. Domestic and foreign networks don't use routing algorithms, that additionally calculate information about possible partial changes of the network structure. At the same time, the choice of route, taking into account the dynamics of the network, can improve the carrying capacity of the channels, the average transmission delay and its variation.

2. Mathematical formulation for determining the average latency of a data packet in a network

Key parameters having an impact on the value of average data packet delay in a corporate data transmission network is length of routes and intensity of data flows that are delivered by them. As an efficiency index for distribution of information flows by routes let us choose objective function that is defined by the expression [6-8]:

$$F = \frac{1}{c_u} \cdot \sum_{j=1}^{h_r} \sum_{a=1}^{h_m} c_{m_a}^{j} \cdot l_{m_a}^{j}; \qquad l_{m_a}^{j} = \sum_{j=1}^{h_{w_a}^{j}} l_{w_b}, \tag{1}$$

where:

- c_u – total intensity of the up-diffused flows of data in DTS
- h_r – an amount of informative streams is between the great number of knots of network
- h_m – a number of routes is for the transmission of j -stream in distribution
- $c_{m_a}^{j}$ – intensity value (bit units per second) of j -flow by the route m_a^j
- l_{w_b} – length of transmission data channel w_b being a part of route m_a^j
- $h_{w_a}^{j}$ – length of route, conditioned by the number of channels of DT, that is included in a route

From expression (1) it turns out that the chosen objective function of the task of information flow distribution by routes in the network is defined by the total product of route length values and intensity values of data flows transmitted over them taking into account the value of total intensity of input flow in DTN. Let us define the problem of rational distribution of data flows by routes in DTN. For each pair of different nodes x and y of DTN the input process of packets that arrive is supposed to be stationary and have intensity u_{r_j} . Therefore, u_{r_j} is the intensity of input flow (measured in bit units per second) that comes in DTN in node x and addressed node y . The goal of data flow distribution lies in the fact that the flow with intensity u_{r_j} to be divided among several routes from the sender to the addressee in such a way so that the resulting general flow by routes in the network minimizes the objective function defined by the expression (1).

To make more accurate formulation of the task of rational data flow distribution let us make the following definitions:

- R – set of pairs of network nodes between which the information exchange takes place
- M^j – set of all routes connecting j-pair of network nodes
- $c^j_{m_a}$ – intensity value (bit units per second) of j-flow by the route m^j_a

Then, the totality of all flows by routes $\left(m^j_a \middle| r_j \in R \wedge a \in M^j\right)$ has to meet the following limitation:

$$\sum_{a \in M^j} m^j_a = u_{r_j} \ for\ all\ r_j \in R;\ m^j_a \geq 0\ for\ all\ a \in M^j, r_j \in R. \tag{2}$$

Therefore, the task of data flow distribution by routes with the aim of minimization of average data packet delay in DTN can be formulated as a task of identification of flows by routes that minimize the objective function (1) under above-mentioned limitations. This is the main task of adaptive routing that will be considered later [8].

3. Calculation of minimum values of data flow intensities between separate network nodes

For performing the data flow distribution in DTN it is necessary to make the previous calculation of intensity values of data flows circulating in it. The way to calculate maximum values of flow intensities between separate network nodes is a component part of the adaptive routing method. The data for transmission from network nodes come at random times t. The duration of transmission of every separate data flow is also a random value. The totality of network nodes gives rise to stochastic data flow with u intensity, the service of which is done by switching nodes (PSC) of DTN. The data transmission network having the totality of PSC is a complex multi-phase queuing system (QS) because one and the same data packet is serviced by a number of PSC. Let us research the behaviour of single-phase QS that describes the operation of a separate network node. Let us agree to examine the system in the stationary operation mode. In real DTN this condition over a long time period (24 hours) is not fulfilled. But in limited time intervals ($t_0 \leq 1$ hour) we can assume the stationary of the data flow. It is commonly thought that the intervals between data packet arrival for transmission generated by network nodes are independent and evenly distributed. Forming the model DTN, it is assumed that the data flow has properties of ordinariness and absence of aftereffect. The possibility of obtaining exactly h packets for time period t_0 under data packet intensity i equals to:

$$P_h(t_0) = \left(u \cdot t_0\right)^h \cdot exp\left(-u \cdot t_0\right) \big/ h!. \tag{3}$$

Mathematical expectation of data packet number that arrive at the time period t_0 equals to:

$$m = u \cdot t_0. \tag{4}$$

The dispersion of the Poisson distribution equals to its mathematical expectation:

$$\delta = u \cdot t_0. \tag{5}$$

It means that the number of data packets that arrive per time unit can vary rather widely that corresponds to the physical nature of the phenomenon. The data flow as any random process can be characterized by some constant value $u_{a,i}$. This characteristic takes into account not momentary data quantity that are delivered from the node y_a to the node y_i, characteristic for the transmission speed, but the nature of its change with respect to time. In practice we can find the value of data flow intensity between separate network nodes relying on the principle inherent in the interaction of two objects in DTN, according to which the data exchange intensity between nodes is directly proportional to the product of node capacities and carrying capacities of data transmission channels between them and inversely proportional to the distance between these nodes [6-7]. For this DTN with the help of unoriented weighted graph let us specify:

$$S = (Y, \varphi_y, W, l_w, p_w), \tag{6}$$

where:

- Y – set of points $y_i \in Y$ of graph S that are in isomorphism with network nodes, their number $h_y = |Y|$
- $\varphi_y : Y \rightarrow N_+$ – weight function that defines for every node y_i the performance of its processor φ_{y_i} (operations per second)
- W – set of edges of graph S (edge $w_{a,i} \in W$ is defined in graph S between points y_a and y_i if between correspondent nodes there is a data transmission channel)
- $l_w : W \rightarrow N_+$ – weight function that defines length $l_{w_{a,i}}$ to every data transmission channel $w_{a,i} \in W$
- $p_w : W \rightarrow N_+$ – weight function that defines carrying capacity $p_{w_{a,i}}$ to every data transmission channel $w_{a,i} \in W$

For the set of points Y the level graph B is constructed and the matrix is formed:

$$H_b = \left\| h_{b_{a,i}} \right\|, \tag{7}$$

where $h_{b_{a,i}}$ is a number of DTN hierarchy levels through which a data packet has to go under exchange between points y_a and y_i.

The level graph represents subordination of nodes of data transmission system (DTS). The intensity of data flows circulating between nodes in vertical direction, as rule, is higher than between nodes on the same level. With the help of Dantzig algorithm, the shortest ways between every two points y_a and y_i of the graph S are determined, and the matrix is formed:

$$L_m = \left\| l_{m_{a,i}} \right\|, \tag{8}$$

where $l_{m_{a,i}}$ – length of the shortest way between points y_a and y_i.

The carrying capacity of the way between the points y_a and y_i is defined with the formula:

$$p_{m_{a,i}} = \min_{w_j \in m_{a,i}} p_{w_j}, \tag{9}$$

where p_{w_j} – carrying capacity of the edge w_j of graph S being a component part of way $m_{a,i}$.

Let u_a denote maximum value of the total intensity of data flows of point y_a, which it exchanges with all points of the set Y, and let u_a denote maximum value of data flow intensity between points y_a and y_i. Then, the maximum value of total data exchange intensity of point y_a with all other points of the set Y is defined by the expression:

$$u_a = l_p \cdot \varphi_{y_a} \cdot p_{m_a} \cdot \sum_{i=1}^{h_y} l_{y_{a,i}} \cdot \sum_{i=1}^{h_y} h_{B_{a,i}} \Big/ h_y \cdot h_0 \cdot \sum_{i=1}^{h_y} p_{m_a} \cdot l_{y_{a,i}} \cdot h_{B_{a,i}}, \quad (10)$$

where:

- l_p – length of data packet
- p_{m_a} – average weighted carrying capacity of the way between y_a and other points of set Y
- h_0 – number of auxiliary operations of processor, between separate operations of input-output of data packet in data transmission channel
- $l_{y_{a_i}}$ – average weighted distance between y_a and other points of set Y
- $h_{B_{a_i}}$ – average weighted radius of graph B with the centre in point y_a

Average weighted carrying capacity of the way between y_a and other points of set Y:

$$p_{m_a} = \left(\sum_{i=1}^{h_y - 1} \frac{p_{m_{a,i}} \cdot l_{m_{a,i}} \cdot h_{B_{a,i}}}{\varphi_{y_{a,i}}} \right) \Big/ \left(\sum_{i=1}^{h_y - 1} \frac{l_{m_{a,i}} \cdot h_{B_{a,i}}}{\varphi_{y_{a,i}}} \right). \quad (11)$$

Average weighted distance between y_a and other points of set Y:

$$l_{y_a} = \left(\sum_{i=1}^{h_y - 1} \frac{p_{m_{a,i}} \cdot l_{m_{a,i}} \cdot \varphi_{y_{a,i}}}{h_{B_{a,i}}} \right) \Big/ \left(\sum_{i=1}^{h_y - 1} \frac{p_{m_{a,i}} \cdot \varphi_{y_{a,i}}}{h_{B_{a,i}}} \right). \quad (12)$$

Average weighted radius of graph B with the centre in point y_a:

$$h_{B_a} = \left(\sum_{i=1}^{h_y - 1} \frac{p_{m_{a,i}} \cdot h_{B_{a,i}} \cdot \varphi_{y_{a,i}}}{l_{m_{a,i}}} \right) \Big/ \left(\sum_{i=1}^{h_y - 1} \frac{p_{m_{a,i}} \cdot \varphi_{y_{a,i}}}{l_{m_{a,i}}} \right). \quad (13)$$

As u_a is the maximum value of the total data exchange intensity of the point y_a, the distribution of u_a between other points of the set Y plays an important role. The obtained value u_a is distributed between interacting points including the point y_i according to the expression:

$$u_{a,i} = \left(\frac{u_a \cdot \varphi_{y_a} \cdot p_{m_i}}{l_{m_{a,i}} \cdot h_{B_{a,i}}} \right) \Big/ \left(\sum_{i=1}^{h_y - 1} \frac{p_{m_{a,i}} \cdot \varphi_{y_i}}{l_{m_{a,i}} \cdot h_{B_{a,i}}} \right). \quad (14)$$

For every pair of points y_a and y_i using expression (14) the values $u_{a,i}$ and $u_{i,a}$,which are in the general case may be not equal to each other, are calculated. According to $u_{a,i}$ and $u_{i,a}$ the arithmetic mean of the maximum value of data flow intensity between y_a and y_i points is calculated:

$$\overline{u_{a,i}} = u_{a,i} + u_{i,a}/2. \quad (15)$$

Recalculation of the maximum values of data flow intensities between network nodes is done in case of its structure change. The offered way of the evaluation of data flows circulating in DTN makes it possible to define maximum values of data flow intensities between separate network nodes taking into account its disposition and distances between them. The obtained values of data flow intensities are used under their distribution in the process of adaptive routing of data flows [7-8].

4. Determination of rational data packet length

The offered way of determination for a rational data packet length is a component part of the adaptive routing method and makes it possible to maximize the speed of data transmission over DT channels depending on the value of probability of distortion of one bit of data transmission. For messages transmitted in DTN, the data packet length l_p is chosen to be constant. The data packet length l_p cannot be too short because, under the fixed length of auxiliary part (header) of the packet, the part of message being transmitted in one packet decreases. Besides, the temporary utilization of ECM for assembly and disassembly of messages and the memory capacity for storage of packet descriptors and headers go up. At long length l_p and given reliability of data transmission over DT channel, the probability of data packet transmission with an error goes up and, as consequence, the frequency of repeated packet transmission goes up, which decreases the efficiency of DTN operation, and also causes increase of memory loss of ECM due to not full occupation of space that is reserved for the last message packet by information (for every data packet in ECM memory the page of fixed length is reserved). As seen above and taking into account practical recommendations, the rational data packet l_p transmitted over the route can be defined with the help of the following expression:

$$l_p = min(l_{p_1}, l_{p_2}), \quad (16)$$

where:

- l_{p_1} – rational packet length in terms of memory saving and minimization of processor utilization under message processing
- l_{p_2} – data packet length providing maximum speed of message transmission over DT channel under the given probability of distortion of one bit of data transmission

The obtained value l_p is rounded off to the nearest value that equals to:

$$l_p = 2^{\mu+1}, \quad (17)$$

where μ – integral number.

Providing that the length of transmitted messaged in DTN is distributed according to the exponential law with the mathematical expectation equal to l_s bits, in terms of memory saving and taking into account system utilization of processor for assembly and disassembly of messages that go up with the decrease of the packet length and also taking into account the tendency for increase of transmitted messages length, the rational data packet length is defined by the expression:

$$l_{p_1} = k_1\left(l_b + \sqrt{k_2 \cdot l_b \cdot l_s}\right), \tag{18}$$

where:

- k_1 – coefficient of processor utilization for assembly and disassembly of message
- l_b – length of data packet header (bits)
- k_2 – coefficient of blockiness

The value of the packet length l_p at which the speed of message transmission over DT channel V_p takes maximum value corresponds to the packet length l_{p_2}. The speed of transmission for a part of message in one data packet over DT channel is defined by the expression:

$$V_p = \left(l_p - l_b - \frac{h_b}{l_s/l_p + 1}\right) \Big/ \left(t_y + \frac{l_p}{V_w}\left(1 + P_0 + P_0^2 + \ldots\right)\right) = \left(l_p - l_b - \frac{h_b \cdot \left(l_p - l_b\right)}{l_s + l_p - l_b}\right) \Big/ \left(t_y + \frac{l_p}{V_w} + \frac{P_0}{1 - P_0} \cdot \frac{l_p}{V_w}\right), \tag{19}$$

where:

- h_b – number of free bits in the last data packet
- t_y – period of switching for data packet in packet switching centre (PSC) (sec)
- P_0 – probability of error in data packet
- V_w – speed of data transmission in DT channel (bit/sec)

The number of free bits in the last data packet is defined from the expression:

$$h_b = l_p - \left(mod\left(l_s, \left(l_p - l_b\right)\right) + l_b\right), \tag{20}$$

where $mod\left(l_s, \left(l_p - l_b\right)\right)$ – remainder on dividing l_s by $\left(l_p - l_b\right)$.

The error probability in the data packet can be defined as:

$$P_0 = 1 - \left(1 - P_c\right)^{l_p}, \tag{21}$$

where P_c – probability of distortion of one bit of data transmission.

Dependence of error probability in data packet P_0 on values P_c is given in Fig. 1.

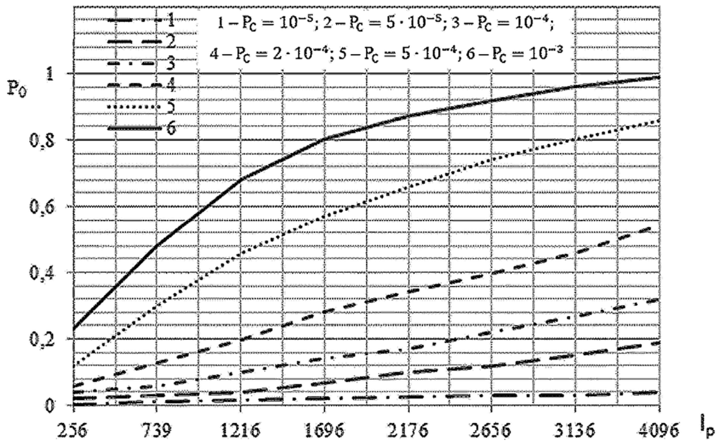

Fig. 1. Change of error probability in data packet depending on its length.

5. Conclusion

Therefore, the offered method of definition for the rational packet length makes it possible to maximize the speed of message transmission over DT channels and thus, in general, allow to improve the quality of the transport process in the railway transport, as well as support the use of moving units (wagons, locomotives) for better control and optimal use.

References

[1] Paula, Fraga-Lamas, Tiago M. Fernández-Caramés, and Luis, Castedo. (2017) "Towards the Internet of Smart Trains." *A Review on Industrial IoT-Connected Railways*. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5492363/. [Accessed: Jun 21, 2017].

[2] Julija Freimane, Mareks Mezitis, Baglan Togizbaeva Markov Models of Fail-safe Devices of Automation and Telecontrol Systems of Railroad Transport. TransComp 2016. Zakopane Polija, December 2016, https://www.infona.pl/resource/bwmeta1.element.baztech-7155bd9a-5102-4d9a-9621-2467174bc92b/tab/summary, ISSN : 1232-3829

[3] Mihailovs, F.Sansyzbajeva, Z. Mezitis, M. (2017) "Simulation of the Interaction of Railway Station and Harbor." *Procedia Computer Science 104:* 222-226

[4] Shengfeng Xu, Gang Zhu, Bo Ai, and Zhangdui Zhong. (2016) "A Survey on High-Speed Railway Communications." *A Radio Resource Management Perspective*. Available: https://arxiv.org/pdf/1603.05368.pdf. [Accessed: Mar 17, 2016].

[5] Maryam Taherian, Mehrdad Maeen, and Majid Haghparast. (2017) "Promoting the Quality Level of Signaling in Railway Transportation System Taking Advantage from Wireless Sensor Networks Technology." Available: https://res.mdpi.com/computers/computers-06-00026/article_deploy/computers-06-00026.pdf?filename=&attachment=1. [Accessed: July 28, 2017].

[4] H., Kuchuk, and A. Kovalenko. (2016) "Use of methods for change of solution space for traffic control optimization of multiservice networks." *Systemy obrobky informatsii* [Systems for information processing] **5 (142)**:128-132.

[5] H., Kuchuk, and A. Kovalenko. (2015) "Method for minimization of average packet delay in virtual connections of support network for cloud service." *Systemy upravlinnia, navihatsii ta zviazku* [Systems of control, navigation and communication]. **2**: 117-120.

[6] K., Trubchaninova, and N., Svynarova. (2012) "Research of data packet delay value of data flows in data transmission network." *Collected scientific works of Ukrainian State Academy of Railway Transport* **133**: 17-19.

[7] M., Terrence, Y. K., Peter, Cheung, Lam, Kai-Pui, and Luk, Wayne. (2011) "Adaptive Routing in Network-on-Chips Using a Dynamic Programming Network." *IEEE Transactions on industrial electronics* **58** (**8**): 3701-3716.

[8] P., Geoffray, and T., Hoefler. (2008) "Adaptive Routing Strategies for Modern High Performance Network." *Stanford, California*: 165-174.

[9] A., Korolov, H., Kuchuk, A., and Pashnev. (2003) "Adaptive routing in corporate networks." *KhMU*: 224 p.

[10] V., Vishnevskiy. (2003) "Theoretical foundations of computer networking." *Technosphere*: 512.

[11] E., Tanenbaum. (2006) "Computer networks." *Piter*: 992.

[12] I., Pozdniak. (2006) "NGN routing methods." *Problems of technology and telecommunications*: 148-149.

[13] S, Ermolaev. (2008) "Ant optimization algorithms." *Infocommunication technology* **6(1)**:23-29.

[14] V, Bachinskiy, and V, Giorgizova-Gay. (2011) "Choice of dynamic routing protocol in corporate IP-network." *Systematic information and information technology* **1**:99-110.

[15] A, Vishnu, M, Koop, A, Moody et al. (2007) "Hot-spot avoidance with multipathing over infiniband: An mpi perspective." In CCGRID '07: Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid, IEEE Computer Society:479–486.