Investigation of preemptive priority single-server queueing systems with peaked traffic flows

Seferin Mirtchev

In this article, a model of single-server preemptive priority queueing systems with a peaked arrival process, generally distributed service time and infinite waiting position is described by using the Polya distribution to define the peaked traffic flow. The model is obtained by using the generalized Pollaczek-Khinchin formula. This article is a summary of the previous research by the author. In the article, new results of the dependence of the mean waiting time in the systems with four priority classes and preemptive priority from the offered traffic with different values of the peakedness coefficient of the arrival process and different values of the variation coefficient of service process is presented. It is shown that the performance of such single-server preemptive priority queueing systems varies vastly depending on the peakedness of the arrival and service processes.

Keywords – non-preemptive priority, single-server queue, peaked arrival process.

Изследване на едноканални системи с чакане и абсолютен приоритет при неравномерен входящ поток (Сеферин Мирчев). *В статията е описан модел на едноканални телетрафични системи с чакане и абсолютен приоритет при неравномерен процес на постъпване на заявките, произволно разпределение на времето на тяхното обслужване и безкрайна опашка за чакане, като се използва разпределението на Пойа, за да се опише неравномерният трафичен поток. Моделът е получен чрез използването на обобщената формула на Полачек -Хинчин. Тази статия е обобщение на предишните изследвания на автора. В статията са представени нови резултати на зависимостта на средното време за чакане в система с четири приоритетни класа с абсолютен приоритет от постъпващия трафик при различен коефициент на неравномерност на постъпващия процес и различен коефициент на вариации на процеса на обслужване. Показано е, че характеристиките на тази едноканална телетрафична система с чакане и абсолютен приоритет се променят значително в зависимост от неравномерностите на процесите на постъпване и на обслужване.*

1. Introduction

The queueing systems represent an interesting area that is widely exploited in a number of real-life situations. In order to provide different service levels to different users, the queueing systems often utilize priority mechanisms. Using priorities can easily provide the distinction between these different levels of service [1]. For example in telecommunications networks, priority classes can be employed. The priority is usually identified in an appropriate header field of the transmitted protocol units, e.g. in the DiffServ Code Point (DSCP) field in IP packets, or the first three bits in the Ethernet frame's Q-tag field, or the Cell Loss Priority (CLP) flag in the ATM cells, the channel priority mechanisms in the IEEE 802.15.4 standard, and etc. Priority management is also widely used in some production processes, transport control, healthcare, and population protection. The new telecommunications technologies, such as Bluetooth, ZigBee, etc., used in the Internet of Things (IoT) area, allow the interconnection of a large number of devices that are seen as uneven sources of traffic. The variety of applications using IoT requires servicing with different priorities of the individual tasks in a single system and can be determined by the Service Level Agreement (SLA) between users and service providers [2].

Priority service issues arise in many practical situations in telecommunications networks. Practically in packet switched networks, it is important to define the strategy of sharing the communications resources provided to a large number of different traffic flows with different service requirements. Priority mechanisms can be either static or dynamic. The serving devices can handle tasks using either preemptive or non-preemptive priority disciplines. Regarding the use of priority services in data transmission networks, in-depth research is being carried out to improve the average delay for short

messages at the expense of the long ones or to meet the stringent requirements of traffic flows that are sensitive to delays or losses [3].

One of the fundamental dependencies in the queueing theory defines the average queue length and the average waiting time in a single-server queueing system M/G/1 with a Poisson arrival process, generally distributed service time and an infinite queue is given by the Pollaczek-Khinchin formula (PKF) [4]. Many authors offer a different generalization of PKF, for instance, a M/G/1 queue with an occasional service failure [5], throughput analysis of input-buffered switches [6], two-phase Markov modulated processes [7], point-to-point communications networks [8], etc.

The Polya/G/1 teletraffic system is a generalization of the M/G/1 queue, with a Polya arrival flow. This generalization leads to a significant increase in the complexity of the analysis.

Various models based on the M/G/1 teletraffic system have been suggested, such as a state dependent one [9], processor sharing disciplines [10], rest periods [11], etc. In [12], a new teletraffic model of a multichannel waiting system with a peaked arrival flow, described by Polya distribution, and constant service time is offered. All these models make it possible to take into account the influence of the peaked arrival flows in IP networks more accurately.

In [13], a single-channel Polya/D/1 teletraffic system is proposed, and all its interesting features and parameters are evaluated. The idea is based on an analytical continuation to the classical single channel M/D/1 system, using a Polya distributed arrival process.

In [14], a M/G/1 teletraffic system is investigated, where all sources are the same, but they are assigned a randomly selected priority level before requests are entering the system. A transformation of Laplace-Stieltjes is used to determine the parameters of the assigned priorities. It is shown that the model determines the average waiting time, limited by the mean time using of the two service mechanisms - FIFO and LIFO. Lastly, it is pointed out that this new approach can increase the efficiency of the server when a new source appears.

In [15], an optimal strategy behavior of high priority users for a M/G/1 queue with two priority classes is presented. When a highly priority request is received, it can be decided whether it should be served with preemptive priority or wait for the completion of a low-priority request that is currently served. Optimal strategies and numerical results are given.

A priority M/G/1 model with accumulation is analyzed in [16]. It allows controlling the waiting time for each class of requests. The authors present an in-depth analysis of a dynamic queue behavior with a preemptive priority for two disciplines serving low-priority requests – repeating the same request and repeating the next one.

This article is a summary of the previous research by the author. The goal is to present new results through the developed model of a single-channel teletraffic system with a priority service, a peaked traffic flow, generally distributed service time and an infinite queue [17], [18]. The model is developed based on a generalization of the classical M/G/1 model and of the PKF formula [19]. In this article the evaluation of this preemptive priority single-server queueing system, based on the new numerical results, is presented.

2. Arrival process with Polya distribution

The peaked arrival processes is described by Polya distribution with two parameters – intensity λ and peakedness β [20]. The probability $P_i(t)$ for the arrival of i requests with in the time interval t is determined by the following formula:

$$(1) \quad P_i(t) = \left(\frac{\lambda t}{1+\beta\lambda t} \right)^i \frac{1(1+\beta) \ \ldots \ [1+(i-1)\beta]}{i!} P_o(t) \ ,$$

where $P_o(t) = (1+\beta\lambda t)^{-\frac{1}{\beta}}$.

The mean value $M(t)$ and the variance $V(t)$ of the number of arrivals for the time interval t are respectively:

$$(2) \quad M(t) = \lambda t; \quad V(t) = \lambda t(1+\beta\lambda t) \ .$$

The coefficient of peakedness z of the number of arrivals is:

$$(3) \quad z = \frac{V(t)}{M(t)} = 1+\beta\lambda t > 1 \ .$$

3. Pollaczek-Khinchin's generalized formula for Polya/G/1 system

The M/G/1 teletraffic system model is one of the most frequently studied models in the telecommunications and computer networks. The model of a teletraffic Polya/G/1 system is a generalization of the above model. The Polya/G/1 system has a peaked input process, described by the Polya distribution, with an arrival intensity – λ and a coefficient of peakedness – z, generally distributed service time (independent of the input process) with a mean value – τ and a coefficient of variation – C_t. The

offered traffic $A = \lambda\tau$ must be less than 1 Erl in order for the teletraffic system to be stationary.

The PKF formula for a Polya/G/1 teletraffic system is obtained using the Kendal recursion [19]. The mean waiting time for a single-channel system with a peaked traffic flow (i.e. the time a request has to wait in the queue for a service) is:

$$W_q = \frac{\tau(A+z-1)\left(C_t^2+1\right)}{2(1-A)}. \quad (4)$$

As shown in [1], the mean value of the residual time t_R for servicing a request at a random point of time during its service (i.e. the time left to finish servicing the current request) is:

$$E(t_R) = \frac{\tau\left(C_t^2+1\right)}{2}. \quad (5)$$

The average residual service time R of a server at a random point of represents the average release time of the server, if it is busy at the moment. The probability of the server having a request currently being served is equal to the offered traffic A (as in a single-channel system with an unlimited queue the probability of the system not being occupied is $1 - A$). Therefore, the fraction of the average residual service time R taken by the mean residual time t_R of a request is determined by the offered traffic. Then the average residual service time R of the server, which may be busy or free at a random point of time, becomes:

$$R = \frac{A\tau\left(C_t^2+1\right)}{2}. \quad (6)$$

The average waiting time W_q for any request can be divided into two parts:

1. The average residual service time R of the server;
2. The average service time of the previous requests that have already arrived and are waiting in the queue:

$$W_q = R + \tau L'_q, \quad (7)$$

where L'_q is the average number of requests waiting in the queue (i.e. the average queue length) when new request arrives.

With conversion and substitution, one can get:

$$L'_q = \frac{(A^2+z-1)\left(C_t^2+1\right)}{2(1-A)}. \quad (8)$$

Using (4), the average queue length at the arrival of a new request can be expressed by the average waiting time, i.e.:

$$L'_q = \frac{(A^2+z-1)}{\tau(A+z-1)} W_q. \quad (9)$$

From (7) and (9) one can get the following:

$$W_q = R + k W_q; \quad (10)$$

$$W_q = \frac{R}{1-k}, \quad (11)$$

where: $k = (A^2+z-1)/(A+z-1)$.

4. Teletraffic Polya/G/1 system with preemptive priority

In communications networks, the subscribers are often served in several priority classes. Usually subscribers of first class have the highest priority.

In a Polya/G/1 single-channel waiting system with a preemptive priority [17], the requests of class i arrive with an arrival intensity λ_i, a coefficient of peakedness z_i, and an average service time τ_i. The offered traffic is $A_i = \lambda_i\tau_i$. The coefficient of variation of the service time is C_{ti}. In the present study, the First Input – First Output (FIFO) queueing discipline is used for serving the requests within each of the priority classes (Fig. 1).

Instead of working with the individual arrival and departure processes, one could consider the joined process of two or more priority classes. It is possible to describe the joined arrival process of the p priority classes by means of a Polya arrival process with the following intensity:

$$\lambda_p = \sum_{i=1}^{p} \lambda_i. \quad (12)$$

The resulting number of arrivals peakedness from the first to the p^{th} priority classes then becomes a weighted sum of the individual classes peakedness:

$$z_p = \sum_{i=1}^{p} \lambda_i z_i / \lambda_p. \quad (13)$$

Then the average service time of the joined departure process of the p priority classes is:

$$\tau_p = \sum_{i=1}^{p} \lambda_i \tau_i / \lambda_p. \quad (14)$$

The coefficient of variation of the departure process of the joined departure process of the p priority classes is:

$$C_{tp} = \sum_{i=1}^{p} \lambda_i C_{ti} / \lambda_p. \quad (15)$$

The offered traffic for the customers with priority p and higher becomes:

$$A_p = \sum_{i=1}^{p} \lambda_i\tau_i = \lambda_p\tau_p. \quad (16)$$

The total offered traffic is:

$$A = \lambda\tau = \sum_{i=1}^{N} A_i = \sum_{i=1}^{N} \lambda_i\tau_i. \quad (17)$$

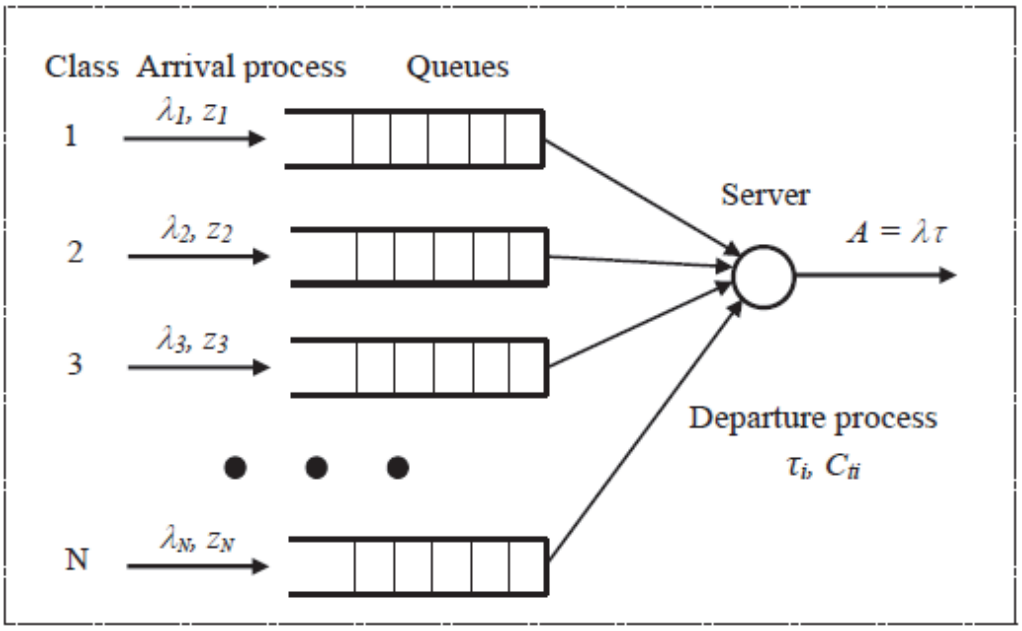


Fig.1. The priority model of the Polya/G/1 single-server teletraffic system.

In the case of having a Polya/G/1 queue with preemptive resume priority, the arrival of the new request with higher priority interrupt immediately the service of a customer with lower priority. The service of the former will be resumed when the server becomes free from where it was interrupted. In other words, for the higher priority requests, the lower priority customers do not exist. Preemptive priorities are mostly used for performance analysis of operating systems, emergency and civil prevention systems.

The mean residual service time at a random point of time for the customers with priority p and higher becomes:

$$R_p = A_p \tau_p \left(C_{tp}^2 + 1\right)/2 . \quad (18)$$

The average waiting time for a class p customer is the two components sum:

1. The average waiting time due to higher or same priority requests already waiting in the queue. This time is equivalent to average waiting time for the queueing system without priority with only the first p classes requests:

$$R_p/(1-k_p), \quad (19)$$

$$\text{where } k_p = \frac{A_p^2 + z_p \lambda_p - 1}{A_p + z_p \lambda_p - 1} . \quad (20)$$

2. The average waiting time for servicing the requests with higher priority arriving during the waiting and service time of class-p customer. These arriving higher priority requests interrupt class-p customer:

$$\left(W_{qp} + \tau_p\right) \sum_{i=1}^{p-1} \lambda_i \tau_i . \quad (21)$$

By summing these two components is obtained:

$$W_{qp} = R_p/(1-k_p) + \left(W_{qp} + \tau_p\right) A_{p-1}, \quad (22)$$

which results in this formula:

$$W_{qp} = \frac{R_p}{(1-k_p)(1-A_{p-1})} + \frac{\tau_p A_{p-1}}{1-A_{p-1}} . \quad (23)$$

For customers with the highest priority one can get the generalized PKF formula. The first class customers servicing is not interrupted by the arrival of lower-priority requests:

$$W_{q1} = \frac{R_1}{1-k_1} , \quad (24)$$

where k_1 is given by (20)

The average waiting time in the queue for customers of class 2 is:

$$W_{q2} = \frac{R_2}{(1-k_2)(1-A_1)} + \frac{\tau_2 A_1}{1-A_1} , \quad (25)$$

where k_2 is given by (20).

And the mean waiting time in the queue for the third priority is:

$$W_{q3} = \frac{R_3}{(1-k_3)(1-A_1-A_2)} + \frac{\tau_3 (A_1 + A_2)}{1-A_1-A_2} , \quad (26)$$

where k_3 is given by (20).

The average waiting time in the queue for customers of class 2 is:

(27) $$W_{q4} = \frac{R_4}{(1-k_4)(1-\sum_{i=1}^{3} A_i)} + \frac{\tau_4 \sum_{i=1}^{3} A_i}{1-\sum_{i=1}^{3} A_i},$$

where k_4 is given by (20).

5. Numeric results

Using a computer program and the formulas from the previous sections, results for the mean waiting time have been obtained for a given value of the coefficient of peakedness – z, the coefficient of variation of the service time – C_t and the offered traffic – A is obtained.

In the figure 2 the mean waiting time W_{qi} in the Polya/G/1 teletraffic model with four priority classes and a preemptive priority as a function of the offered traffic A is shown. The results have been obtained for different values of the coefficient of peakedness z (i.e. 1.0, 1.2 and 1.4) and zero coefficient of the variation of the service time, i.e. C_t=0.

The offered traffic to each priority class has the same value and the service time in each class equals the others, i.e. τ_i=*0.001 s*. For comparison, the figure also shows the mean waiting time W_q for the same teletraffic system but without priorities. It can be seen that the increase in the number of arrival requests leads to a significant increase in the mean waiting time of the low-priority requests. In addition, the mean waiting time for class 4 is several times higher than that in the non-priority system.

In the figure 3 the mean waiting time W_{qi} for the same values of the coefficient of peakedness z as in the figure 2 but with a coefficient of variation C_t equal to 2 is depicted. It can be seen that the peakedness of the service process significantly increases the mean waiting time for both low- and high-priority classes. The results show that the mean waiting time for low-priority classes increases significantly when the offered traffic approaches 1 Erl and has a much higher value in the case of greater peakedness of the incoming and servicing processes.

6. Conclusion

In this article a model for analyzing a single-channel teletraffic system with a peaked arrival traffic flow, a generally distributed service time, an infinity queue, and a preemptive priority, based on the generalized Pollaczek-Khinchin's formula, a peaked arrival traffic flow described by the Polya distribution, and the classical queueing M/G/1 system with priorities is described.

It is shown that the peakedness of arrival requests and the variation in service time lead to a significant increasing of delays in the single-channel preemptive priority system and, consequently, to a significant increasing of the queue length for low-priority traffic classes.

The analysis of preemptive priority teletraffic systems with peaked arrival and service processes provides guidelines for traffic flow analysis in the modern telecommunication networks and systems, which is important in their design.

The presented new results make it possible to assess the characteristics of preemptive priority queueing systems in fixed and mobile networks with traffic classification, in specific applications in cloud technologies and in point-to-point communications. Other possible applications include e-health, disaster and emergency protection, emergency calls, highly reliable robots, etc.

REFERENCES

[1] Iversen, V. Teletraffic engineering and network planning. DTU Fotonik, 2015, http://orbit.dtu.dk/files/118473571/Teletraffic_34342_V_B_Iversen_2015.pdf

[2] Goleva, R., Garcia, N., Mavromoustakis, C.X., Dobre, C., Mastorakis, G., Stainov, R., Chorbev, I., Trajkovik, V., AAL and ELE Platform Architecture, in Dobre, C., Mavromoustakis, C.X., Garcia, N., Goleva, R., Mastorakis, G. (Editors), Ambient Assisted Living and Enhanced Living Environments: Principles, Technologies and Control, 1st Edition, Elsevier, Butterworth-Heinemann, 2016, pages 544, Biomedical engineering book series, pp. 171-210.

[3] Sztrik J. Basic Queueing Theory. GlobeEdit, 2016.

[4] Zukerman M. Introduction to Queueing Theory and Stochastic Teletraffic Models, City University of Hong Kong, 2016, https://arxiv.org/pdf/1307.2968.pdf

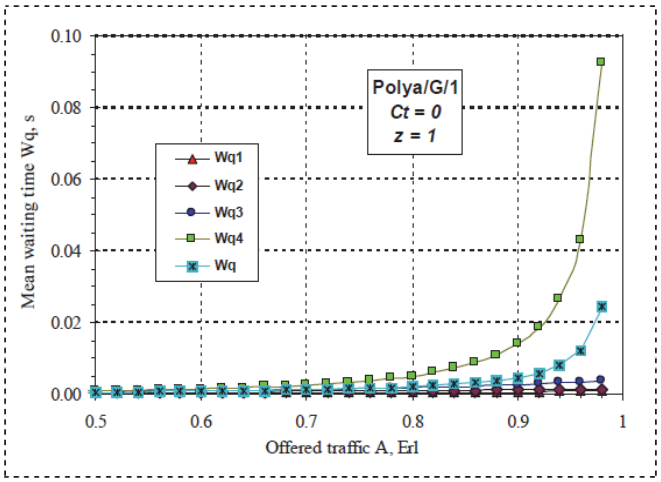
[5] Jain G., K. Sigman. A Pollaczek-Khintchine Formula for M/G/1 Queues with Disasters, Journal of Applied Probability, Vol. 33, No 4, 1996, pp. 1191-1200, doi: 10.2307/3214996

[6] Chang C., D. Lee, C. Yu. Generalization of the Pollaczek-Khinchin formula for throughput analysis of input-buffered switches, Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, Vol. 2, 2005, pp. 960-970, doi: 10.1109/INFCOM.2005.1498325

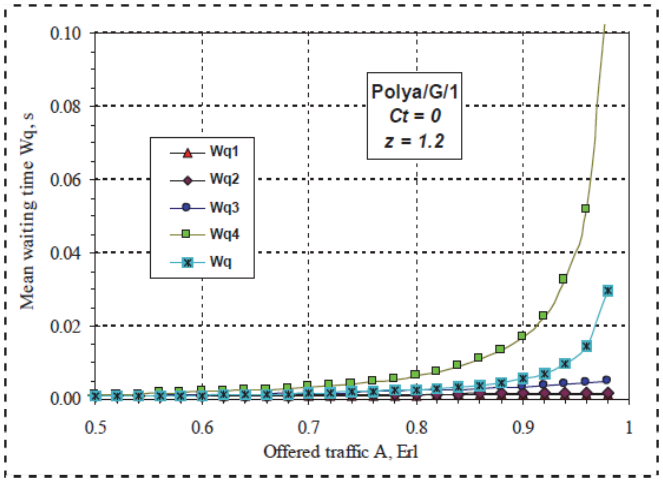
[7] Huang L., T. Lee. Generalized Pollaczek-Khinchin formula for Markov channels, IEEE Transactions on Communications, Vol. 61, No 8, 2013, pp. 3530–3540, doi: 10.1109/TCOMM.2013.061913.120712

[8] Zhang J., T. Lee, T. Ye, W. Hu. On Pollaczek-Khinchine Formula for Peer-to-Peer Networks, Eprint arXiv:1605.08146, 2016, bib. code:2016arXiv160508146Z

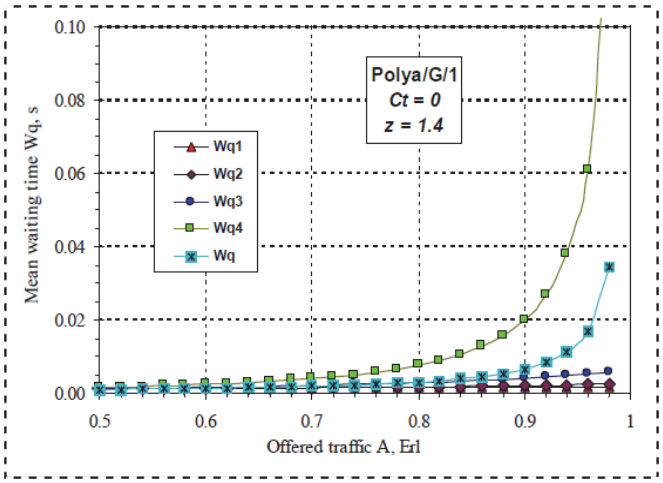
[9] Abouee-Mehrizi H., O. Baron. State-dependent M/G/1 queueing systems, Queueing System, Vol. 82, No 1, 2016, pp. 121-148, doi: 10.1007/s11134-015-9461-y



a). For coefficient of peakedness z=1.0 and coefficient of variation C_t=0

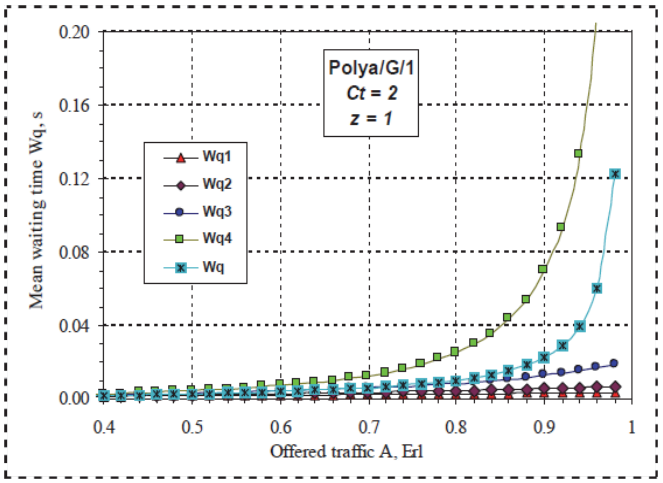


b). For coefficient of peakedness z=1.2 and coefficient of variation C_t=0

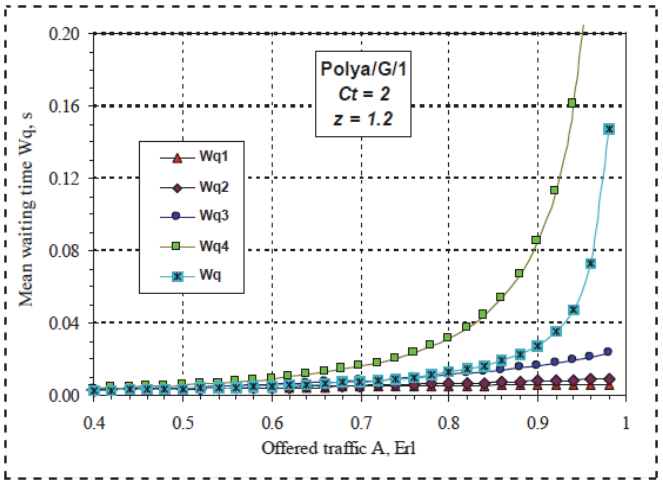


c). For coefficient of peakedness z=1.4 and coefficient of variation C_t=0

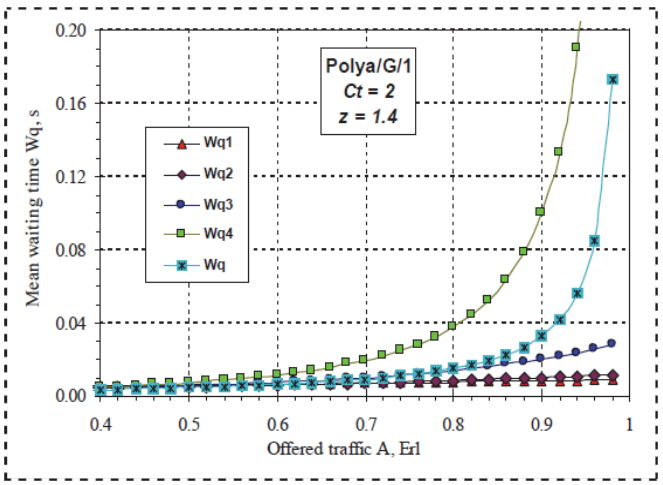
Fig.2. The mean waiting time in a Polya/G/1 queue with four preemptive priority classes as a function of the total offered traffic, for defined values of the coefficient of peakedness of the number of arrivals and fixed service time.



a). For coefficient of peakedness z=1.0 and coefficient of variation C_t=2



b). For coefficient of peakedness z=1.2 and coefficient of variation C_t=2



c). For coefficient of peakedness z=1.4 and coefficient of variation C_t=2

Fig.3. The mean waiting time in a Polya/G/1 queue with four preemptive priority classes as a function of the total offered traffic, for defined values of the coefficient of peakedness of the number of arrivals and peaked service process.

[10] Sigman K. Using the M/G/1 queue under processor sharing for exact simulation of queues, Annals of Operations Research, Vol. 241, No 1-2, 2016, pp. 23-34, doi: 10.1007/s10479-013-1408-2

[11] Giambene G., S. Puzovic. Non-Saturated Performance Analysis for WiMAX Broadcast Polling Access, IEEE Transactions on Vehicular Technology, Vol. 62, No 1, 2013, pp. 306-325, doi: 10.1109/TVT.2012.2216296

[12] Mirtchev S., R. Goleva. New constant service time Polya/D/n traffic model with peaked input stream, Simulation Modelling Practice and Theory, Vol. 34, No 11-12, 2013, pp. 200-207, doi: 10.1016/j.simpat. 2012.08.004

[13] Mirtchev S., R. Goleva, V. Alexiev. Evaluation of Single Server Queueing System with Polya Arrival Process and Constant Service Time, Proceedings of the International Conference on Information Technologies (InfoTech-2010) Varna, Bulgaria, 2010, pp. 203-212.

[14] Haviv M. The performance of a single-server queue with preemptive random priorities. Performance Evaluation, No103, 2016, pp. 60–68.

[15] Xu B., X. Xu, X. Wang. Optimal balking strategies for high-priority customers in M/G/1 queues with 2 classes of customers. J. Appl. Math. Comput., No 51, 2016, pp. 623–642.

[16] Fajardo V. A Generalization of M/G/1 Priority Models via Accumulating Priority. Thesis for the degree of Doctoral of Philosophy in Statistics. Waterloo, Ontario, Canada, 2015.

[17] Mirtchev, S., Goleva, R., Atamian, D., Ganchev, I. Investigation of priority queue with peaked traffic flows. Proceedings of the ACM Symposium on Applied Computing, 2018, pp. 1017-1019.

[18] Mirtchev S. Study of Preemptive Priority Single-server Queue with Peaked Arrival Flow, Proc. X National Conference with International Participation "Electronica 2019", 2019, pp. 59-62, DOI: 10.1109/ ELECTRONICA.2019.8825636

[19] Mirtchev S., I. Ganchev. A Generalized Pollaczek-Khinchin formula for the Polya/G/1 queue. Electronics Letters, Vol. 53, No 1, 2017, pp. 27-29.

[20] Ramos H., D. Almorza, J. Garcia-Ramos. On Characterizing the Polya Distribution, ESAIM: Probability and Statistics, No 6, 2002, (6), pp. 105-112, doi: 10.1051/ps:2002005

Prof. DSc Eng. Seferin T. Mirtchev *has graduated telecommunications at Technical University of Sofia (TUS) in 1981. He is with Department of Communication Networks TUS, vice president of the Union of Electronics, Electrical Engineering and Telecommunications (CEEC), member of IEEE and has research interest in teletraffic engineering, switching systems, quality of service, cloud computing.*

tel.: +359 2 965 2254 *e-mail: stm@tu-sofia.bg*

Received on: 29.02.2020