Investigation of Priority Queue with Peaked Traffic Flows

Seferin Mirtchev
Technical University of Sofia
8 Kl. Ohridski Blvd, 1000 Sofia
Bulgaria
stm@tu-sofia.bg

Rossitza Goleva
Technical University of Sofia
8 Kl. Ohridski Blvd, 1000 Sofia
Bulgaria
rgoleva@gmail.com

Dimitar Atamian
Technical University of Sofia
8 Kl. Ohridski Blvd, 1000 Sofia
Bulgaria
dka@tu-sofia.bg

Ivan Ganchev
University of Limerick, Ireland
& University of Plovdiv, Bulgaria
ivan.ganchev@ul.ie

ABSTRACT

In this paper[1], a new single-server priority queueing system with a peaked arrival process and generally distributed service time is analysed by using the Polya distribution to describe the peaked traffic flows. The mean waiting time in the case of infinite number of waiting places is obtained using a generalized Pollaczek-Khinchin formula. It is shown that the performance of such delay systems varies vastly depending on the peakedness of the input flow. To the best of our knowledge, such a priority queueing system with a peaked arrival process is analysed for the first time.

CCS CONCEPTS

• Networks → Network performance evaluation → Network performance analysis

KEYWORDS

Polya arrival process; non-preemptive priority; generalized Pollaczek-Khinchin formula; Polya/G/1 queue; mean waiting time

ACM Reference format:

S. Mirtchev, R. Goleva, D. Atamian and I. Ganchev. Investigation of Priority Queue with Peaked Traffic Flows. 2018. In *SAC 2018: SAC 2018: Symposium on Applied Computing, April 9–13, 2018, Pau, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/ 10.1145/3167132.3167407

1 INTRODUCTION

In order to offer different service levels for different groups of users, the queueing systems are often controlled by priority mechanisms. The priority queueing can easily provide service differentiation [1]. To achieve this, in the telecommunications networks for instance, priority classes can be used, which are supported by a corresponding field in the header of the Protocol Data Units (PDUs), e.g. the DiffServ Code Point (DSCP) in Internet Protocol (IPv4) packets, the first 3 bits in the VLAN protocol identifier of the 802.1Q (Q-tagged) Ethernet frames, the Cell Loss Priority (CLP) bit in Asynchronous Transfer Mode (ATM) cells, the Flow Label and Traffic Class in IPv6 packets, the priority channel access in IEEE 802.15.4, etc. The priority control is also widely used in production practice, transportation management, health insurance, civil protection, prevention services, etc. New communication technologies, like Bluetooth, ZigBee and others used for Internet of Things (IoT) services nowadays, allow interconnection of a big number of devices that are considered irregular traffic sources. The diversity of the applications built on the top of IoT requires schedules with priorities of tasks in the systems.

The priority queueing systems are used also in many mobile networks. The prioritization scheme for handover calls in cellular networks is discussed in multiple articles. For instance, a Dynamic Multilevel Queue Scheduling algorithm is proposed in [2]. In [3], a mathematical model to estimate the priority processing of handoff calls in cellular wireless networks is proposed.

The Pollaczek-Khinchin formula is considered a fundamental equation in the queueing theory [4]. Different generalizations of it were implemented in M/G/1 queues by: Markov modulated service processes with two-states [5], application-layer protocols and scheduling in peer-to-peer networks [6], etc.

Kleinrock analysed priority queues under the condition that user classes have defined priorities [6]. In [7], the Kleinrock's analysis of a time-dependent priority queue is extended to accumulating priority queue.

In [8], a priority queue-scheduling algorithm for resource allocation inside a data centre, running various kinds of application workloads, is proposed.

In the current paper, a new priority Polya/G/1 queue with a bursty arrivals and general distribution service times is analysed. Performance analysis is done in comparison to the standard M/G/1 model by using a generalized Pollaczek-Khinchin formula.

SAC 2018, April 9–13, 2018, Pau, France

ACM ISBN 978-1-4503-5191-1/18/04. . . $15.00
https://doi.org/10.1145/3167132.3167407

2 POLYA ARRIVAL PROCESS

A pure birth process, called a Polya arrival process, has average arrival rate λ and peakedness β of the input flow [9]. The probability $P_i(t)$ of having i arrivals within a time interval t is:

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

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

The coefficient of peakedness (z) of the Polya flow is the ratio of the variance $V(t)$ and the mean value $M(t)$ of the number of arrivals within a time interval t:

$$z=[V(t)]/[M(t)]=1+\beta\lambda t>1\,. \qquad (2)$$

3 GENERALIZED POLLACZEK-KHINCHIN FORMULA FOR POLYA/G/1 QUEUE

We used here a generalized version of M/G/1, called Polya/G/1, with Polya distributed arrivals described by two parameters – the rate λ and the coefficient of peakedness z. The general identically distributed service times are considered independent of the arrival process and have mean value τ and coefficient of variation C_t. It is assumed that the offered traffic $A=\lambda.\tau$ is less than *1 Erl* as to ensure system stability.

The generalized Pollaczek-Khinchin formula applied to the Polya/G/1 system is obtained using the Kendall's Recursion [10]. The mean waiting time for this queueing system is:

$$W_q=\frac{\tau(A+z-1)\left(C_t^2+1\right)}{2(1-A)}\,. \qquad (3)$$

The mean residual lifetime t_R at a random point of time is [1]:

$$E(t_R)=\left[\tau\left(C_t^2+1\right)\right]/2\,. \qquad (4)$$

As the offered traffic A is equal to the probability of finding a user being served, the mean residual service time is:

$$R=\left[A\tau\left(C_t^2+1\right)\right]/2\,. \qquad (5)$$

The mean waiting time (W_q) for an arbitrary user can be presented in two parts: 1) the mean residual service time; and 2) the waiting time experienced by the users that have already arrived but are waiting in the queue:

$$W_q=R+\tau L'_q\,, \qquad (6)$$

where L'_q is the mean number of waiting users at the instant of arrival.

By conversion and substitution, one can get the following:

$$L'_q=\frac{(A^2+z-1)\left(C_t^2+1\right)}{2(1-A)} \qquad (7)$$

Then from (3), (6) and (7), the following formula could be obtained:

$$W_q=R/(1-k)\,, \qquad (8)$$

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

4 POLYA/G/1 QUEUE WITH NON-PREEMPTIVE PRIORITY

In communication and computer networks, users are usually classified into N classes with different priorities. It is assumed also that a user of class p has higher priority than a user of class $p+1$. In a single-server non-preemptive priority queueing system, a new user waits until a server becomes idle even if it is serving a user with lower priority [9]. It also waits until all the users with higher priority and users arriving earlier with the same priority have been served.

In a Polya/G/1 single-server queue, the users of class i arrive in accordance to a Polya process with arrival intensity λ_i, coefficient of peakedness z_i of the number of arriving users, and mean service time τ_i. The offered traffic is $A_i=\lambda_i\tau_i$. C_{ti} denotes the service time distribution's coefficient of variation. A FCFS queueing discipline is assumed for each priority class.

We assume that the overall arrival process is a Polya arrival process, similarly to the individual arrival processes, with intensity λ, coefficient of peakedness z, mean service time τ, and coefficient of variation C_t as per the following formulas:

$$\lambda=\sum_{i=1}^{N}\lambda_i\,,\ \ z=\sum_{i=1}^{N}\frac{\lambda_i}{\lambda}z_i\,,\ \ \tau=\sum_{i=1}^{N}\frac{\lambda_i}{\lambda}\tau_i\,,\ C_t=\sum_{i=1}^{N}\frac{\lambda_i}{\lambda}C_{ti}\,. \qquad (9)$$

The total offered traffic becomes:

$$A=\lambda\tau=\sum_{i=1}^{N}A_i=\sum_{i=1}^{N}\lambda_i\tau_i\,. \qquad (10)$$

At a random moment of time, one can obtain the mean residual service time as:

$$R=\sum_{i=1}^{N}R_i=\sum_{i=1}^{N}\frac{A_i\tau_i\left(C_{ti}^2+1\right)}{2}\,. \qquad (11)$$

The mean time to wait (W_{q1}) for a highest-priority user is equal to the sum of the mean residual service time R and the mean waiting time due to other highest-priority users already arrived and waiting in the queue $\tau_1 L'_{q1}$, i.e.:

$$W_{q1}=R+\tau_1 L'_{q1}=R+k_1 W_{q1}\,;\ \Rightarrow W_{q1}=R/(1-k_1)\,, \qquad (12)$$

where $k_1=(A_1^2+z_1-1)/(A_1+z_1-1)$.

The mean waiting time W_{qp} for a class-p user consists of two components – the waiting time due to having other active users of classes 1 to p already waiting in the queueing system, and the mean time it takes to serve all newly arriving users of classes 1 to $p-1$ while this class-p user is still waiting in the queue.

By putting these two components together, one can get the following formula:

$$W_{qp}=\frac{R}{(1-k'_p)(1-\sum_{i=1}^{p-1}A_i)}\,, \qquad (13)$$

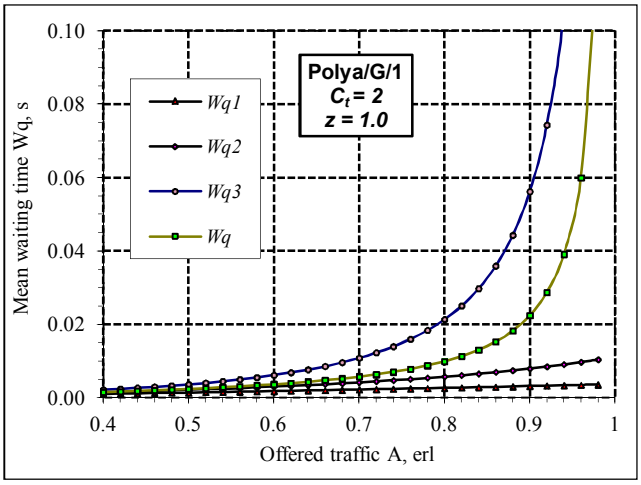
where

$$k'_p=\frac{\left(\sum_{i=1}^{p}A_i\right)^2+\sum_{i=1}^{p}\frac{z_i\lambda_i}{\sum_{j=1}^{p}\lambda_j}-1}{\sum_{i=1}^{p}A_i+\sum_{i=1}^{p}\frac{z_i\lambda_i}{\sum_{j=1}^{p}\lambda_j}-1}\,. \qquad (14)$$

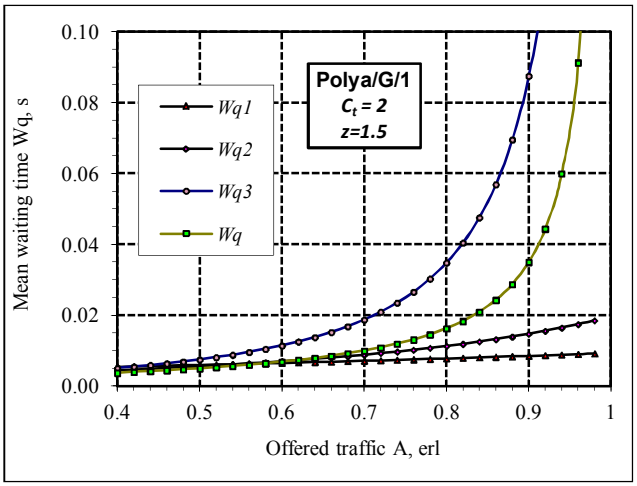
5 NUMERICAL RESULTS

The mean waiting times for a given value of the coefficient of peakedness (z) of the number of arrivals, the service time distribution's coefficient of variation (C_t), and the offered traffic (A) were calculated by means of a computer program based on the formulas presented in the previous sections.

Figure 1 demonstrates the mean waiting times (W_{qi}) in a Polya/G/1 system with three non-preemptive priority classes as a function of the total offered traffic A, for two values of the coefficient of peakedness (z) when $C_t = 2$. The offered traffic in each of the three classes is of the same volume and the mean service times are also equal, i.e. $\tau_1 = \tau_2 = \tau_3 = 0.001s$. The mean waiting time W_q in a system without priorities at all is also shown on the figure.



a) Non-preemptive priority queue with peakedness $z = 1.0$



b) Non-preemptive priority queue with peakedness $z = 1.5$

Figure 1: The mean waiting time in a Polya/G/1 queue with three non-preemptive priority classes as a function of the total offered traffic.

One could see that increasing the value of the coefficient of peakedness (z) of the number of arrivals leads to an increase of the mean time to wait for users in each service class; however, relatively, this increase is higher for low-priority classes. In addition, the mean waiting time for users in class 3 is even higher than that in a system without priorities at all, for all instances.

6 CONCLUSIONS

The presented methods for determining the mean time to wait in a single-server Polya/G/1 queue with non-preemptive priority allow accurate planning of telecommunications networks and overall improving of the quality of service in all kind of applications that will need priorities at different levels of service.

The impact of the peakedness of the number of arrivals and the service time variance on increasing the service delay and the queue length of the low-priority classes in the system has been demonstrated.

The proposed methods could be used to evaluate the priority queueing performance in fix and mobile communication networks with service differentiation, special applications in cloud computing, and peer-to-peer processing with pre-emption. E-health, civil protection, emergency calls, and reliable robotics are few of the possible applications of this kind.

ACKNOWLEDGMENTS

Our thanks go to ICT COST Action IC1303: Algorithms, Architectures and Platforms for Enhanced Living Environments (AAPELE), ICT COST Action IC1406: High-Performance Modelling and Simulation for Big Data Applications (cHiPSet), TD COST Action TD1405: European Network for the Joint Evaluation of Connected Health Technologies (ENJECT), and H2020 project on Advanced systems for prevention & early detection of forest fires 2016/PREV/03 (ASPIRES).

REFERENCES

[1] Iversen, V. Teletraffic engineering and network planning. DTU Fotonik, 2015

[2] Kumar R., K. Varshini. Multilevel priority packet scheduling scheme for wireless networks, International Journal of Distributed and Parallel Systems (IJDPS) Vol.5, No.1/2/3, 2014, pp. 69-76.

[3] Samanta R., P. Bhattacharjee, G. Sanyal. Performance Analysis of Cellular Wireless Network by Queuing Priority Handoff calls, International Journal of Electrical and Electronics Engineering, Vol. 3, No 8, 2009, pp. 472-477.

[4] 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

[5] 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.

[6] Kleinrock L. Queueing Systems, vol. II: Computer Applications, Wiley, New York, 1976.

[7] Stanford D., P. Taylor, I. Ziedins. Waiting time distributions in the accumulating priority queue, Queueing System, No 77, 2014, pp.297–330.

[8] Madhumathi R., R. Radhakrishnan. Priority Queue Scheduling Approach for Resource Allocation in Cloud, Asian Journal of Information Technology, No 15, 2016, pp. 472-480.

[9] 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

[10] 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.