DELAYS IN QUEUES OF QUEUING SYSTEMS WITH STATIONARY REQUESTS FLOWS

Boris Ya. Likhttsinder,
Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia,
lixt@psuti.ru

Yulia O. Bakai,
Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia,
ov.bakai@gmail.com

During all periods of the study of telecommunications systems traffic, the analysis was based on mass service theory. The subjects of the study here are request flows to be processed by some limited performance resources. Resource constraints and the random nature of requests' receipt lead to refusals in processing or queues. The first works devoted to the analysis of teletraffic belong to A. K. Erlang. Request flows represented flows of requests for connections in networks with channel switching. Since requests were received from a large number of independent users, the flows of such requests could be defined as stationary, ordinary with no effect, or as recurring requests, with an exponential distribution of time intervals between neighboring requests. Connection request flows to a telephone exchange node are a superposition of a large number of low-intensity flows from independent users. Therefore, the fixed Poisson flow model describes the real flows in telephone exchanges with channel switching quite well. Therefore, the stationary Poisson flow model describes real flows in telephone exchanges with channel switching rather well. The emergence of telecommunications networks with packet switching, especially multiservice networks, showed the impossibility of using Poisson flow models for their analysis. The article is devoted to the analysis of delays in queues of queuing systems with correlated stationary flows of general type requests. The traffic of packets in multiservice networks is typically characterized by a high degree of correlation. On the basis of interval methods of analysis, the relations generalizing the Khinchin-Pollaczek formula for the average value of waiting time in queuing systems with flows of the general kind of requests are obtained. The main parameters to be analyzed when outputting the above formulas are time intervals between neighboring requests. It is shown that the values of time delays in queues depend on the dispersion and dispersion index of a random value characterizing the degree of additional maintenance of processed requests.

Information about authors:

Boris Ya. Likhttsinder, Doctor of Technical Science, Povolzhskiy State University of Telecommunications and Informatics, Professor of Network and Communication Systems Department, Samara, Russia

Yulia O. Bakai, Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia

Introduction

Many different models with non-exponential laws of probability distribution between requests come into being [1]. However, all these intervals are considered mutually independent. Such models poorly reflect the properties of flows, because flows of packets in real networks are strongly correlated and have a batch nature [1,3]. The description of such flows can be carried out using self-similar process models [4, 5]. The main disadvantage of self-similar process models is their complexity. The overwhelming number of researches in the field of self-similar flows occurs in isolation from the processes of their processing in the queuing service system. Attempts to obtain analytic ratios that determine delays and queue sizes in queuing systems with self-similar flows are usually accompanied by significant difficulties and have not led to any significant practical results. Consequently, the analytical study not only of the flow itself but also of the queuing systems in which it is processed seems to be hardly realistic [1]. Indeed, queues in the queuing systems are formed as a result of the interaction of two flows, one of which is a flow of requests coming to the system and the other one is a flow of requests leaving the queue. The specified flows are strongly correlated. As a result, the total flow loses its long-term dependence typical for self-similar processes [3].

The insufficient efficiency of applying self-similar process models to the analysis of packet teletraffic resulted in a number of models controlled by the Markov chain. A special case of such models was called "models with hyper-exponential event flows" [6]. In a further development, the flows were named MC-flows (Markov Chain), MAP-flows (Markovian Arrival Process), and their generalization - group BMAP (Batch Markovian Arrival Process) input flows. The evolution of such models is well represented in the review [1]. It is also noted that it is important to take into account the correlation properties of teletraffic because the presence of positive correlation significantly increases the size of the queues and delays in the queuing systems. The influence of correlation dependencies in an incoming teletraffic flow has been considered several times in the papers of the author of this paper, where a direct dependence of the average queue size on the correlation coefficients of the input teletraffic flow has been established. [2, 3]

Additional Maintenance Time

All classical theory of the queuing systems is based on the assumption of mutual independence of incoming requests. Even the well-known Kendall classification [7] does not imply the possibility of considering the correlation properties of the investigated flows, but only considers the laws of distribution of the corresponding probabilities. There are a lot of works considering different laws of interval distribution of probabilities between neighboring requests, not Poisson's distributions of probability numbers of requests on some time interval and some others. However, they presuppose the mutual independence of incoming requests.

One of the promising, in our opinion, directions of studying packet traffic is the Interval method [3], which allows replacing the analysis of time intervals between neighboring requests and time intervals of processing requests with the analysis of one random value – the number of requests received during consecutive time intervals of processing requests. We have shown that the dispersion and correlation properties of the specified random

value, at a given loading, fully characterize the average queue size in the queuing systems [8, 9].

In this article, as a unified random value characterizing the input request flow and its processing in the queuing systems, we will consider another random value u_i representing the difference between the request maintenance time τ_i and the time interval ϑ_i between two neighboring requests on the i-th interval $u_i = \tau_i - \vartheta_i$. If the interval of processing τ_i turns out to be bigger than the interval ϑ_i, then the request maintenance on this interval is not completed in time. Therefore, the value is called the additional maintenance time. This value is included as the only external variable in the balance equation set by Lindley [10],

$$w_i = w_{i-1} + u_i$$
$$u_i = \tau_i - \vartheta_i \ \text{ if } \ w_{i-1} + \tau_i - \vartheta_i \geq 0 \tag{1}$$
$$u_i = -w_{i-1} \ \text{ if } \ w_{i-1} + \tau_i - \vartheta_i < 0 .$$

In all cases when the second inequality is executed, $w_i = 0$. Obviously,

$$w_{i-1} = \sum_{j=-\infty}^{i-1} u_j . \tag{2}$$

L. Kleinrock in his work [11] noted the determining influence of the specified value on the size of delays and queues in the queuing systems. Believing that the values of τ_i and ϑ_i are not correlated and mutually independent, he significantly limited the requirements for the flows. However, packet flows in multi-service telecommunications networks do not satisfy these requirements because they have a high degree of correlation.

Figure 1 shows the fraction of the influence of correlation properties of the video traffic stream on the average size of the queues (the upper chart) and the fraction obtained without considering the influence of correlation (the lower chart) with the same values of the load factor ρ. The difference is impressive and shows that correlated dependencies of packet traffic flows cannot be ignored.

Let us square both parts of equation (1). After averaging at a significant interval of time, we obtain:

$$\overline{u_i \cdot w_{i-1}} + \frac{1}{2}\overline{u^2}_i = 0 . \tag{3}$$

By substituting

$$w_{i-1} = \sum_{j=-\infty}^{i-1} u_j = w_0 + \sum_{j=1}^{i-1} u_j$$

We get:

$$\overline{u_i \cdot w_{i-1}} + \frac{1}{2}\overline{u^2}_i = \overline{w_0 u_i} + \overline{u_i \sum_{j=1}^{i-1} u_j}_i + \frac{1}{2}\overline{u^2}_i = .$$

$$\overline{w_0 u_i} + \overline{u_i(\sum_{j=1}^{i} u_j - u_i)} + \frac{1}{2}\overline{u^2}_i = 0 , \text{ or}$$

$$\overline{u_i \sum_{j=1}^{i} u_j} - \frac{1}{2}\overline{u^2}_i = -\overline{w_0 u_i} \tag{4}$$
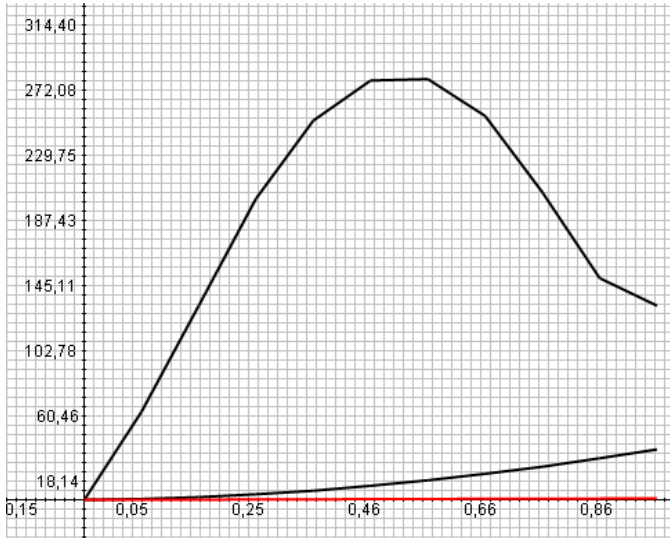
Fig. 1 Fractions of influence of video traffic stream correlation
properties on average queue size

Accepting the hypothesis of the attenuation of the correlation
bond, we can consider that $\overline{w_0 u_i} = \overline{w_0}\ \overline{u_i}$, hence,

$$ \overline{u_i \sum_{j=1}^{i} u_j} - \frac{1}{2}\overline{u^2}_i = -\overline{w_0}\ \overline{u_i} . \tag{5} $$

The first member of the specified difference represents the
average value of the product of the function by its integral (sum),
therefore

$$ \overline{u_i \cdot \sum_{j=1}^{i} u_j} = \sum_{k=0}^{\infty} R_u(k) = \sum_{k=1}^{\infty} R_u(k) + \overline{u_i}^2 . \tag{6} $$

Here $R_u(k)$ – the value of the correlation coefficient of the
variable u_i , when shifted by k intervals. By substituting in (5),
we get:

$$ \sum_{k=1}^{\infty} R_u(k) + \overline{u_i}^2 - \frac{1}{2}\overline{u^2}_i = \sum_{k=1}^{\infty} R_u(k) + \frac{1}{2}\overline{u^2}_i = -\overline{w_0}\ \overline{u_i} , $$

or

$$ \sum_{k=1}^{\infty} R_u(k) + \frac{1}{2}D_u + \frac{1}{2}(\overline{u_i})^2 = \overline{w_0}\ \overline{u_i} , $$

Thus,

$$ \overline{w} = \frac{D_u + 2\sum_{k=1}^{\infty} R_u(k) + \overline{(u_i)}^2}{2(-\overline{u_i})} . \tag{7} $$

Let us replace the variable $v_i = -u_i$, in this case, the numera-
tor value will not change. We will also introduce the dispersion
index

$$ J_v = 1 + 2\sum_{k=1}^{\infty} r_v(k) , \tag{8} $$

where $r_v(k) = \dfrac{R_v(k)}{D_v}$ is the normalized value of the correlation

coefficient. Then, we will obtain a fundamental ratio suitable for
the analysis of correlated input flows of requests that generalizes
the known Khinchin-Pollaczek formula:

$$ \overline{w} = \frac{D_v J_v}{2\overline{v_i}} + \frac{\overline{v_i}}{2} . \tag{9} $$

**Algorithm of forming an array of the random variable v_i,
characterizing the traffics**

Let us take a look at the sequence of source data processing.
Arrays of packet lengths l_i $[Bit]$ and packets arrival mo-
ments t_i $[us]$ are introduced. The limit value of the number of
packets M, the ordinal number of the incoming packet $i = 0;$,
the starting point $t_0 [\sec] = 0\ [us]$, the starting value of the vari-
able $v_0 = 0\ [us]$ are set. The total processing time of the packets
$T_0[c] = 0\ [us]$, the total waiting time $W_0 = 0$, the initial value
of the waiting time in the queue $w_0 = 0$, the bandwidth value of
the channel C_0 $[Mbps]$.

1. Switching to the next package $i := i + 1$.
2. The moment $t_{i+1} [\sec]$ of the next packet arrival time is en-
tered.
3. Interpackage interval is determined $\vartheta_i = t_{i+1} - t_i\ [us]$
4. The coefficient $k = \dfrac{8}{C_0}\ [\dfrac{us}{Bit}]$ is calculated.
4. The length of the next packet in bytes is entered – l_i $[Bit]$.
5. The packet processing time is determined in microseconds
$\tau_i = k\, l_i [\dfrac{us}{pack}]$.
6. The total processing time of packages in microseconds is
determined $T_i = T_{i-1} + \tau_i\ [us]$
7. Additional maintenance time is determined
$u_i := \tau_i - \vartheta_i, v_i := u_i\ [us]$.
8. Remaining time is determined $\Delta_i = w_{i-1} + u_i\ [us]$
9. If $\Delta_i < 0$, then $v_i := -w_{i-1}\ [us]$
10. Recording v_i into an array V.
11. Determining the current waiting time in the queue
$w_i = w_{i-1} + v_i\ [us]$
12. Determining the total processing time $T_i = T_{i-1} + \tau_i\ [us]$.
13. Total waiting time $W_i = W_{i-1} + w_i\ [us]$.
14. If $i < M - 1$, moving to p.1. $\rho_0 = \dfrac{T_M}{t_M}$.
15. Average delay value $\overline{w} = \dfrac{W_M}{M}\ [us]$.
16. Load factor

17. Average intensity $\lambda = \dfrac{M}{t_M}$.

18. Average packet processing time $\overline{\tau} = \dfrac{T_M}{M}\,[us]$.

19. Average interval between packets $\overline{\vartheta} = \dfrac{t_M}{M}$.

20. Average queue size $\overline{q} = \overline{w}\lambda\,[us]$.

21. Output of results and array $V\,[us]$.

22. Organizing 10 cycles for the values C_x, corresponding to

$\rho_x = 0,1;\ 0,2...0,9\ ;!,0$. $C_x = C_0\,\dfrac{\rho_0}{\rho_x}$, and drawing charts for all

the average values of the parameters and dependence C_x on the

coefficient ρ_x.

Conclusions

Thus, we obtained a fundamental ratio (9) directly linking the dispersion and dispersion index of a variable with the average maintenance delay value in a single-line queuing system.

Mathematical expectation, dispersion, and dispersion index of the specified random variable fully determine the average size of the queues in the queuing systems, at the specified value of its load factor, and the given algorithm shows the method of obtaining the values of this variable.

References

1. Vishnevskiy, V.M.; Dudin, A.N. (2017). Queueing systems with correlated arrival flows and their applications to modeling tele-communication networks. *Automatics and telemechanics*. Vol. 8. P. 3-59.
2. Likhttsinder B.Ya. (2015). Interval'nyj metod analiza mul'tiservisnogo trafika setej dostupa. *Elektrosvyaz'*. Vol. 12. P. 52-54.
3. Likhttsinder B.Ya. (2018). Traffic of multiservice access networks (interval analysis and design). Moscow: Hotline – Telecom. 290 p.
4. Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. (1994). On the self-similar nature of Ethernet traffic. *SIGCOMM Comput. Commun*. Vol. 23. No.4. P. 183-193. DOI:https://doi.org/10.1145/167954.166255
5. Tsybakov B. S. (1999). Teletraffic model based on the self-similar random process. *Radio engineering*. Vol. 5. P. 24-31.
6. Kuzmin I. V. (1981) Osnovy modelirovaniya slozhnyh sistem: Uchebnoe posobie dlya studentov vuzov. Kiev: Vyshcha Shkola. Head publishing house. 360 p.
7. Kleinrock L. Queuing systems. Vol. 2. New York, Wiley, 1974. 519 p.
8. Likhttsinder B.Ya. (2015). Korrelyacionnye svojstva dlin ocheredej v sistemah massovogo obsluzhivaniya s potokami obshchego vida. *ICT*. Vol. 13. No.3. P. 276-280.
9. Likhttsinder B.Ya. (2015). Korrelyacionnye svyazi v pachechnyh potokah sistem massovogo obsluzhivaniya. *Telecommunications*. Vol. 9. P. 8-12.
10. Lindley, D. (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 48. No.2. P. 277-289. doi:10.1017/S0305004100027638
11. Kleinrock L. (2013). Queuing systems: Theory. New York, Chester, 1975. 432 p.

ЗАДЕРЖКИ В ОЧЕРЕДЯХ СМО СО СТАЦИОНАРНЫМИ ПОТОКАМИ ЗАЯВОК

Лихтциндер Борис Яковлевич, *Поволжский государственный университет телекоммуникаций и информатики (ПГУТИ), Самара, Россия, lixt@psuti.ru*

Бакай Юлия, *Поволжский государственный университет телекоммуникаций и информатики, Самара (ПГУТИ), Россия, ov.bakai@gmail.com*

Аннотация

Статья посвящена анализу задержек в очередях систем массового обслуживания (СМО), с коррелированными стационарными потоками заявок общего вида. Для трафика пакетов в мультисервисных сетях характерна высокая степень коррелированности. На основании интервальных методов анализа, получены соотношения, обобщающие формулу Хинчина-Поллачека для среднего значения времени ожидания в системах массового обслуживания, с потоками заявок общего вида. Основными анализируемыми параметрами при выводе указанных формул являются временные интервалы между соседними заявками. Показано, что значения временных задержек в очередях, зависят от дисперсии и индекса дисперсии случайной величины, характеризующей степень дообслуживания обрабатываемых заявок.

Ключевые слова: *системы массового обслуживания, потоки заявок, временные задержки, очереди, ковариация, загрузка.*

Литература

1. *Вишневский В.М., Дудин А.Н.* Системы массового обслуживания с коррелированными входными потоками и их применение для моделирования телекоммуникационных сетей / Автоматика и телемеханика №8 , 2017. С. 3-59.

2. *Лихтциндер Б.Я.* Интервальный метод анализа мультисервисного трафика сетей доступа // Электросвязь. 2015. №12. С. 52-54.

3. *Лихтциндер Б.Я.* Трафик мультисервисных сетей доступа (интервальный анализ и проектирование). М.: Горячая линия – Телеком, 2018. 290 с.

4. *Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson.* On the self-similar nature of Ethernet traffic // SIGCOMM Comput. Commun. Vol. 23, №4, 1994. pp. 183-193.

5. *Цыбаков Б.С.* Модель телетрафика на основе самоподобного случайного процесса // Радиотехника, 1999, № 5. С. 24-31.

6. Основы моделирования сложных систем: Учебное пособие для студентов вузов / Под общ.ред. И.В. Кузьмина.- Киев: Вища школа. Головное изд-во, 1981. 360 с.

7. *Клейнрок Л.* Вычислительные системы с очередями. Т. 2. Пер. с англ. М.: Мир. 1979.

8. *Лихтциндер Б.Я.* Корреляционные свойства длин очередей в системах массового обслуживания с потоками общего вида // ИКТ. 2015. Т.13. №3. С. 276-280.

9. *Лихтциндер. Б.Я.* Корреляционные связи в пачечных потоках систем массового обслуживания // Телекоммуникации. 2015. № 9. С. 8-12.

10. *Lindley D.* The theory of queues with a single server // Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 48, №2, 1952. P. 277-289. doi:10.1017/S0305004100027638

11. *Клейнрок Л.* Теория массового обслуживания. М.: Книга по требованию 2013. 429 с.

Информация об авторах:

Лихтциндер Бория Яковлевич, *д.т.н., профессор. Поволжский государственный университет телекоммуникаций и информатики (ПГУТИ), Самара, Россия*

Бакай Юлия, *Поволжский государственный университет телекоммуникаций и информатики, (ПГУТИ), Самара, Россия*