Multi-server queue with batch arrivals

Liubov Korolkova[1,2,*], *Nematulla* Mashrabov[2], and *Alexandr* Murzin[1]

[1] South Ural State University (national research university), Aerospace Faculty, 454080 Chelyabinsk, Russian Federation
[2] South Ural State Agricultural University, Institute of Agricultural Engineers, 457100 Troitsk, Russian Federation

Abstract. A multi-server queueing system, that is loaded continuously in certain periods of time and which functions for a certain amount of time allocated for the functioning of the system, is considered. Based on the renewal theory, an expression is obtained for the distribution density of the number of arrivals served herewith the service time for each server can be different. In the numerical example, the distributions of the number of services for the systems consisting of one, two, five servers are obtained. The approach to optimization of the queue using the stochastic model of supply and demand is outlined. According to the model, the distributions of the number of services, the queue length as the number of unused arrivals, the number of idle servers as the number of unused services are calculated. Each of these values corresponds to the cost. Knowledge of the distribution functions of the model indicators makes it possible to calculate the cost parameters with dependence of unit costs on the number of servers. The optimal number of servers can be selected from the condition of the maximum of the total average cost.

1 Introduction

The queue theory is often used for functioning description of complicated systems. It is under consideration the multi-server queueing system which is contiguously downloaded in some periods of time, and can be considered as a particular case of a system with batch arrivals. Thus, in [1] the situation is considered when, under conditions of low loading, the service begins when a certain number of arrivals in the system is clustered and ends when the system is completely freed.

The articles examine queue with different service disciplines, a specific input process and/or service time; the possibility of a failure of the serving device [2, 3] is considered. In this case, at least one of these quantities has exponential distribution [4, 5]. In all articles, the characteristics of the system for the steady-state queue are studied. In [6, 7] steady state probability distributions were obtained. Some important performance measures such as the average number of arrivals in the system and the mean sojourn time have also been obtained in [7]. Arbitrary distributed times of arrivals and services are considered in [8], but only in the case of one server.

* Corresponding author: korolkovali@rambler.ru

In the works on optimizing queue, only Markov systems are considered. In [9], in the case of several criteria, the theory of decision-making was used. In [10, 11], optimization was carried out by queue simulating.

The paper studies the characteristics of a multi-server queue, at the input of which there is a batch arrivals. Service time is characterized by an arbitrary distribution and can be different on different servers. The service of arrivals is considered for horizontal time, which can be either deterministic t_H, or random, T_H. The last case has not been treated in the references.

An important characteristic of the general service process under consideration is the distribution of the number of services during the time T_H, t_H. Distribution is the basis for optimizing the queue by establishing the optimal number of servers and (or) determining the rational service time due to the modernization of servers.

2 Distribution density of the number of served arrivals

When servicing in the one server, the number of services completed within the horizontal time will be random with the distribution density (according to the renewal theory)

$$a_1(t_H;n) = a_1(n) = P\{N_1(t_H) = n\} = K_n(t_H) - K_{n+1}(t_H)\,, \tag{1}$$

where $K_n(t) = \int_0^t K_{n-1}(t-y)dE(y),\ n \geq 2,\ K_0(t) = \begin{cases} 0, t < 0, \\ 1, t \geq 0; \end{cases}$

$K_1(t) = E(t)$, $E(t)$ – service time distribution function.

In the case when the time T_H, during which the processes are studied is random, the integral of the "convolution type" $K_n(T_H)$ is computed as follows

$$K_n(T_H) = \int_0^\infty [1 - E_H(y)]dK_n(y)\,, \tag{2}$$

where $E_H(t) = P\{T_H \leq t\}$ – distribution function of the time T_H.

Simultaneous operation of m same servers during the time T_H will be characterized, respectively, by the total flow of services. In general, we assume that each server is characterized by its service time.

We obtain a formula for the distribution density of the number of arrivals served. The discrete analogue of (1) will have the following form

$$A_m^+(n) = \sum_{j=0}^{n} A_{m-1}^+(n-j) \cdot a_m(j)\,, \tag{3}$$

where $a_m(j)$ – distribution density of m-th service for T_H, t_H, $A_m^+(j)$ – distribution function of summary completion of m services. The "+" sign will further indicate that the indicator refers to the summary completion of the services.

The required distribution density

$$a_m^+(n) = \frac{A_m^+(n) - A_m^+(n-1)}{n-(n-1)} = A_m^+(n) - A_m^+(n-1)\,. \tag{4}$$

Further we obtain

$$A_m^+(n) - A_m^+(n-1) = \sum_{j=0}^{n} A_{m-1}^+(n-j)\cdot a_m(j) - \sum_{j=0}^{n-1} A_{m-1}^+(n-1-j)\cdot a_m(j) =$$

$$= \left[A_{m-1}^+(n-0)\cdot a_m(0) + A_{m-1}^+(n-1)\cdot a_m(1) + ... + A_{m-1}^+(n-n)\cdot a_m(n)\right] -$$

$$-\left[A_{m-1}^+(n-1-0)\cdot a_m(0) + A_{m-1}^+(n-1-1)\cdot a_m(1) + ... + A_{m-1}^+(n-1-(n-1))\cdot a_m(n-1)\right] =$$

$$= \left[A_{m-1}^+(n-0) - A_{m-1}^+(n-1-0)\right]\cdot a_m(0) + \left[A_{m-1}^+(n-1) - A_{m-1}^+(n-1-1)\right]\cdot a_m(1) + ...$$

$$+ \left[A_{m-1}^+(n-(n-1)) - A_{m-1}^+(n-1-(n-1))\right]\cdot a_m(n-1) + A_{m-1}^+(n-n)\cdot a_m(n).$$

Consider the multiplier for $a_m(0)$, which is according to (4)

$$A_{m-1}^+(n-0) - A_{m-1}^+(n-1-0) = a_{m-1}^+(n),$$

that is, the first product can be written as $a_{m-1}^+(n)\cdot a_m(0)$; the second composition

$$\left[A_{m-1}^+(n-1) - A_{m-1}^+(n-1-1)\right]\cdot a_m^+(1) = a_{m-1}^+(n-1)\cdot a_m(1),$$

etc. As a result, we obtain the following formula

$$a_m^+(n) = P\left\{\sum_{i=1}^{m} N_i(T_H, t_H) = n\right\} = \sum_{j=0}^{n} a_{m-1}^+(n-j)\cdot a_m(j). \quad (5)$$

With a random number of servers M with the distribution $b(m) = P\{M = m\}$ the number of completed services will have a distribution

$$a_M^+(n) = \sum_{m=0}^{\infty} a_m^+(n)\cdot b(m). \quad (6)$$

3 Numerical example

Consider a queue that is continuously loaded for 60 units of time. The service time by one server is distributed arbitrarily and does not exceed 10 units of time. The number of arrivals is not limited.

3.1 A single server queue

We assume that the service time has a Weibull distribution $E(t) = E_1(t) = 1 - \exp(-1.38\cdot 10^{-4}\cdot t^5)$; while the average service time is $\bar{t}$ = 5.58 units of time, coefficient of variation v = 0.23. On average, 10.57 arrivals will be served (calculation by formula $t_H / \bar{t}$) simulating. The distribution varies from 9 to 13.

If the service time is distributed according to $E(t) = E_2(t) = 1 - \exp(-0.138\cdot t^2)$ with average $\bar{t}$ = 2.54 units of time and v = 0.52, in the period of 60 units of time, 24.79 arrivals will be served. The distribution varies from 19 to 31.

Comparison of the calculation results shows how much the distribution of service time is: it affects both the average number and the dispersion of a possible number of services. In

this case, as indicated above, the maximum service time by the server does not exceed 10 units of time.

3.2 Two server queue

By the (5) we obtain the distributions of the number of services for three cases. If the number of arrivals at the entrance to the queue is not limited, then in the first case (E(t) = $E_1(t)$) for time t_H from 19 to 24 arrivals can be served; the average number of arrivals served is 21.15. In the second case ($E(t) = E_2(t)$) two servers can serve from 42 to 58 arrivals; the average number of arrivals served – 49.57.

Let us consider a more interesting case, when the service time by one server is distributed according to the distribution function $E_1(t)$, the second – according to the $E_2(t)$. The average number of services calculated using the distribution $a_2^+(n)$ is 35.36 units and can no longer be determined by a calculation from the simple relation $t_H / \bar{t}$.

The envelopes to the distribution densities for three cases are shown in Fig. 1.

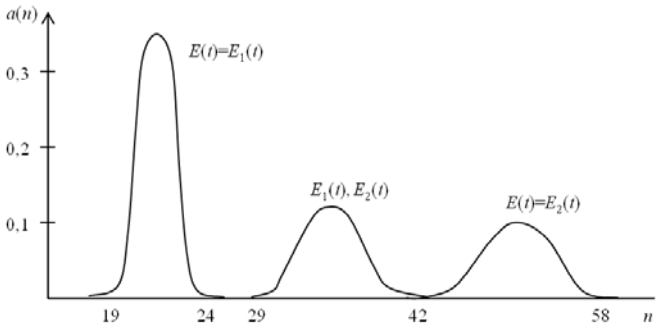

Fig. 1. Envelopes to the distribution densities of the served objects.

3.3 Five server queue

By the (5), we obtain similarly the distributions of the number of services. In the first case (E(t) = $E_1(t)$) five servers can serve from 49 to 58 arrivals; the average number of arrivals served is 52.86. In the second case (E(t) = $E_2(t)$) five servers can serve from 112 to 136 arrivals; average number of arrivals served – 123.94.

4 Approach to optimizing queue

Optimization of the queue can be carried out by applying to it a stochastic model of supply and demand, for example [12, 13].

Since the number of arrivals served by servers within the horizontal time is random, the total number of arrivals Z will be divided by the number Z_+ of arrivals served and the number Z_- of unserved arrivals by m servers during the time T_H, t_H. The value Z_- determines the queue length. In turn, the possible number of services will be divided by the number N_+ of used services and the number N_- of unused services. The random value N_- is the number of downtime servers. The distribution density $h_m(n)$ of the numbers N_+

$$h_m(n)=\begin{cases} a_m^+(n), & n<z, \\ 1-\sum_{n=0}^{z-1} a_m^+(n)=\sum_{n=z}^{\infty} a_m^+(n), & n=z. \end{cases} \quad (7)$$

The density $q_m(n)$ distribution of the number N_- of unused services and the density $g_m(n)$ distribution of the number Z_- of unserved arrivals, i.e. queue length distribution

$$q_m(n)=P\{N_-=n\}=\begin{cases} \sum_{j=0}^{z} a_m^+(j), & n=0, \\ a_m^+(z+n), & n\geq 1; \end{cases} \quad g_m(n)=P\{Z_-=n\}=\begin{cases} \sum_{j=z}^{\infty} a_m^+(j), & n=0, \\ a_m^+(z-n), & n\geq 1. \end{cases} \quad (8)$$

Consider two servers with service times $E_1(t)$, $E_2(t)$. According to the density distribution of services (Fig. 1), if batch arrivals will not exceed 29, it will be served fully during the time t_H = 60. On average over this period will be observed at 10 of the server downtime.

If the service received 35 arrivals, an amount that is approximately equal to the average number of served arrivals for t_H, then the average will be served 34.07 arrivals. The results of the calculations show that in 61.18% of the t_H periods the arrivals will be serviced in full, in 13.18% of the periods will not be served by one arrivals, in 10.57% of the periods – by two arrivals, etc. Similarly, in 51.2% periods there will be no server downtime, 13.64% of the periods will be idle for one server, etc.

If each of the values, $N_+=Z_+$, N_-, Z_- correspond to unit costs: $c^1_{Z_+}$ – service of one arrival, $c^1_{Z_-}$ – one non-service, $c^1_{N_-}$ – one downtime of the servers, then the average value of the costs $\bar{c}_{Z_+}(m)$, $\bar{c}_{N_-}(m)$, $\bar{c}_{Z_-}(m)$ is determined according to the formulas

$$\bar{c}_{Z_+}(m)=\bar{c}^1_{Z_+}\cdot\bar{z}_+(m), \quad \bar{c}_{N_-}(m)=\bar{c}^1_{N_-}\cdot\bar{n}_-(m), \ \bar{c}_{Z_-}(m)=\bar{c}^1_{Z_-}\cdot\bar{z}_-(m), \quad (9)$$

where $\bar{z}_+=\bar{n}_+$, $\bar{n}_-$, $\bar{z}_-$ – mathematical expectations.

Instead of unit costs, there may be value tables $c_{Z_+}(n)$, $c_{N_-}(n)$, $c_{Z_-}(n)$, disproportionate to numbers: serviced arrivals, servers downtimes, unused arrivals. In this case, the average values of costs $\bar{c}_{Z_+}(m)$, $\bar{c}_{N_-}(m)$, $\bar{c}_{Z_-}(m)$ are determined by the relations

$$\bar{c}_{Z_+}(m)=\sum_{n=0}^{\infty} nc_{Z_+}(n)h_m(n), \ \bar{c}_{N_-}(m)=\sum_{n=0}^{\infty} nc_{N_-}(n)q_m(n), \ \bar{c}_{Z_-}(m)=\sum_{n=0}^{\infty} nc_{Z_-}(n)g_m(n). \quad (10)$$

The optimal number of servers can be found by the maximum of the total average cost $\bar{c}_{z+}(m)-\bar{c}_{N_-}(m)-\bar{c}_{z_-}(m)$.

5 Conclusion

Based on the deduced dependence for calculating the distribution density of the number of served arrivals, the distributions of the number of services for a system consisting of one, two, five servers were obtained. A comparison of these indicators for differently distributed service durations not exceeding a certain value is made. One can highlight that the periods of service by each channel can be characterized by different distribution laws. No

restrictions are placed on T_H, t_H duration, hence it allows to study both stationary and non-stationary system.

Using the supply and demand system, the queue length and the number of unused services distributions are obtained, and also the method for determining the optimal number of servers is indicated.

Further development in research of support servers reliability account will be conducted.

The work was supported by Act 211 Government of the Russian Federation, contract № 02.A03.21.0011.

References

1. V.A. Lokhvitsky, A.V. Ulanov, *The numerical analyses of queuing system with hyperexponential distribution of cooling time*, Bull. of Tomsk State Univ., v. **4**, pp. 36-43 (2016)
2. C.S. Kim, V.I. Klimenok, D.S. Orlov, *Multi-line system for servicing a group-Markov flow and negative applications*, Autom. and Remote Control, v. **12**, pp. 106-122 (2006)
3. K. Kirupa, Dr. K. Udaya Chandrika, *Batch arrival retrial G-queue and an unreliable server with delayed repair*, Int. J. of Inn. Res. in Sc., Eng. & Techn., v. **3(5)**, pp. 12436-12444 (2014)
4. K. Kirupa, Dr. K. Udaya Chandrika, *Batch arrival retrial G-queue with additional multi-optional service and server breakdown*, J. of Eng. Comp. & App. Sc., v. **3(7)**, pp. 6-11 (2014)
5. S. Maragathasundari, S. Srinivasan, A. Ranjitham, *Batch arrival queueing system with two stages of service*, Int. J. of Math. Anal., v. **8(6)**, pp. 247-258 (2014)
6. R. Ramesh, S. Kumara, G. Ghuru, *A batch-arrival queue with multiple servers and fuzzy parameters: parametric programming approach*, Int. J. of Sc. & Res., v. **2(9)**, pp. 135-140 (2013)
7. G.I. Falin, *A single-server batch arrival queue with returning customers*, Eur. J. of Oper. Res. v. **201(3)**, pp. 786-790 (2010)
8. W. Kempa, *GI/G/1/∞ batch arrival queueing system with a single exponential vacation*, Math. Meth. of Op. Res. v. **69(1)**, pp. 81-97, (2009)
9. V.A. Chekmenev, M.P. Kalinina, *Optimization of the multilinear Markov queuing system with expectation, functioning in conditions of uncertainty*, Bull. KuzGSTU, v. **4**, pp. 3-6 (2003)
10. V.V. Afonin, V.V. Nikulin, *Optimization of Markov queuing systems with expectation in the MATLAB system*, Vestn. Astrakhan. State Tech. Univ., v. **2**, pp. 39-47, (2017)
11. I.N. Boyarshinova, T.R. Ismagilov, I.A. Potapova, *Simulation and optimization of queuing sistems*, Basic res., v. **9**, pp. 9-13 (2015)
12. I.V. Korolkov, L.I. Korolkova, *Calculation of system indicators with random integer demand and supply*, Int. scient.-pract. conf., pp. 338-341 (2003)
13. L.I. Korolkova , N. Mashrabov, *Analysis of the loading of equipment using the model of supply and demand*, Int. Techn. & Econ. J., v. **3**, pp. 105-108 (2015)