Multi-channel Queuing Systems with Markovian Impatience

Yury I. Ryzhikov(✉)

Institute for Informatics and Automation of the Russian Academy of Sciences,
39, 14-th Line VO, St. Petersburg 199178, Russian Federation
ryzhbox@yandex.ru

Abstract. The iterative Takahashi—Takami method is adjusted to calculate distribution of the number of requests in the multi-phase systems with H_2- service time and exponential distribution of the requests' "patience". The method of calculating the moments of waiting and sojourn time distributions for "successful" requests is also offered. The results are compared with the ones obtained from the simulation model. Application of the method is shown to calculate the successful request's sojourn time distribution in the queueing network.

Keywords: Queueing theory · Multi-phase systems · Iteration
Impatient requests

1 Introduction

Among many applications of the queuing theory, situations with *impatient customers* who have random restrictions on the request's sojourn time play a significant role. In telecommunications and military it could be some moving equipment with a limited time of staying in the zone of reach, in emergency situations people rescue, in court—lengthy legal procedures with deadlines, in medicine—critical patients whose conditions deteriorate rapidly in anticipation of emergency assistance, etc.

The simplest problem of this type—Markovian system with exponentially distributed service time [1–3]—has a very limited *practical* value. The attempt taken in [4] to generalize the approach for $M/H_2/n-H_2$ model proved ineffective due to the fast growth of the problem dimension.

A reasonable compromise would be a $M/H_2/n-M$ model. Specifics of implementation of an iterative method [5–7] are discussed below—with additional calculation of the system's sojourn time for "successful" requests, their ratio with respect to input flow, as well as calculating non-productive system losses due to incomplete service. This software model is in essence the only one that allows to generalize the problem on the *queueing networks*—thanks to the possibility to disregard the accumulated requests' patience due to its Markovian property.

A. Dudin et al. (Eds.): ITMM 2018/WRQ 2018, CCIS 912, pp. 83–92, 2018.
https://doi.org/10.1007/978-3-319-97595-5_7

2 Iterative Method for the Model $M/H_2/n - M$

The considered system receives a Poisson flow of requests of the intensity λ. The H_2- servicing can be presented as an exponentially distributed for requests of two types, selected with probabilities y_1 and y_2, with intensities μ_1 and μ_2 respectively. Any request's sojourn time in the system, regardless of its location (in the channel or in the queue), is limited by a random variable exponentially distributed with parameter γ.

Shown on Fig. 1 is a fragment of the diagram of transitions between microstates of the system $M/H_2/3 - M$ *by outgoing*, presented for 2-nd, 3-rd and 4-th layers.

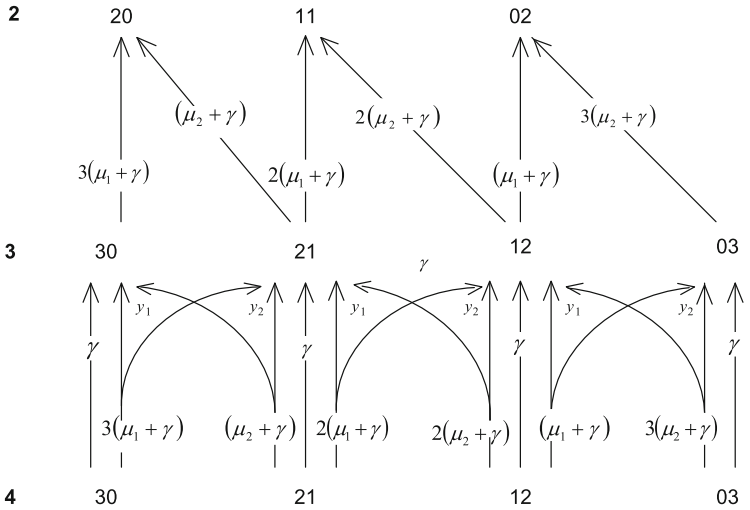

Fig. 1. Fragment of the withdrawal chart

Each layer corresponds to the number of system requests shown to the left (in our example—2, 3, 4). Code combinations like (2,1) indicate the distribution by type of requests being served; additional multipliers $\{y_i\}$ placed at the ends of arrows—the probabilities of selecting requests of the appropriate type from the queue. On layers $j > 3$ the difference will consist only in the intensity of additional *vertical* transitions $(\gamma, 2\gamma, 3\gamma, \dots)$, reflecting exit of the impatient requests from a queue of the current length.

Let S_j be a set of all possible system microstates in which exactly j of requests are being served, and let m_j be a number of elements in S_j. Further, in accordance with the diagram of transitions for the selected model, let's build the matrices of the intensities of infinitesimal transitions:

$A_j[m_j \times m_{j+1}]$—in S_{j+1} (on arrival requests),

$B_j[m_j \times m_{j-1}]$—in S_{j-1} (full completion of service request),

$D_j[m_j \times m_j]$—exit from the states of layer j
(the matrices sizes are indicated in the square brackets). Calculation of these matrices in case of H_2 approximation of the service time distribution is easily programmed.

Let's introduce vectors-strings $g_j = \{g_{j,1}, g_{j,2}, \ldots, g_{j,m_j}\}$ of probabilities for the system to be in the state (j, i), $j = 0, 1, \ldots$. Now it is possible to write down the vector-matrix equations of the transition balance

$$\begin{aligned} g_0 D_0 &= g_1 B_1, \\ g_j D_j &= g_{j-1} A_{j-1} + g_{j+1} B_{j+1}, \qquad j = 1, 2, \ldots \end{aligned} \tag{1}$$

Now we describe the general scheme of the iterative calculation of the stationary vectors of probabilities. Assume $t_j = \gamma_j / p_j$, where p_j is the cumulative probability of presence exactly j requests in the system, and define

$$x_j = p_{j+1}/p_j, \qquad z_j = p_{j-1}/p_j. \tag{2}$$

With the *bottom-up* passage of the layers in the iteration number m the system of equations (1) can be rewritten with respect to vectors of conditional probabilities of the microstates normalized to 1 within a layer:

$$\begin{aligned} t_0^{(m)} D_0 &= x_0 t_1^{(m)} B_1, \\ t_j^{(m)} D_j &= z_j t_{j-1}^{(m-1)} A_{j-1} + x_j t_{j+1}^{(m)} B_{j+1}, \qquad j = 1, 2, \ldots \end{aligned} \tag{3}$$

Using vectors-columns $\mathbf{1}_j = \{1, 1, \ldots, 1\}^T$ of size σ_j, the additional system conditions (3) for normalizing components to 1 can be written for all j

$$t_j \mathbf{1}_j = 1 \tag{4}$$

and the balance of the total intensities of transitions between adjacent layers

$$t_j^{(m)} A_j \mathbf{1}_{j+1} = x_j t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j. \tag{5}$$

In the case of the system with an unlimited queue, the calculation algorithm for the set of vectors $\{t_j\}$ and the numbers $\{x_j\}$ and $\{z_j\}$ satisfying ratios (3)–(5), relies on the existence of a limit vector of conditional probabilities $t = \lim_{j \to \infty} t_j$, which is a consequence of the stabilization of transition matrices at $j > n$. The algorithm is based on a sequential approximation to the desired characteristics for a bounded set of indices $j = \overline{0, N}$.

Let's rewrite the equations of the system (3) for $j \geq 1$ as

$$t_j^{(m)} = z_j \beta'_j + x_j \beta''_j, \tag{6}$$

where

$$\begin{aligned} \beta'_j &= t_{j-1}^{(m-1)} \; A_{j-1} D_j^{-1}, \\ \beta''_j &= t_{j+1}^{(m)} \quad B_{j+1} D_j^{-1}. \end{aligned} \tag{7}$$

In this and subsequent formulas the products of matrices can be calculated before the start of iterations. In particular, their products by $\mathbf{1}_j$ are equal to the row sums, and the products

$$t_{j-1}^{(m)} A_{j-1} \mathbf{1}_j = \lambda. \tag{8}$$

One of the central ideas of the Takahashi—Takami method is the assumption about stabilization of conditional probabilities vectors $\{t_j\}$ for $j \to \infty$ confirmed by calculations. It allows to close the calculation scheme by assuming that for a layer with sufficient large number $j = N$

$$\beta''_N = t_{N-1}^{(m-1)} B_{N+1} D_N^{-1}. \tag{9}$$

This assumption was a consequence of the transition matrices stabilization already at $j = n + 1$. In our case, $\{B_j\}$ and $\{D_j\}$ are stabilized only at $j \to \infty$, but due to increasing decline of "impatient" requests, the cumulate layer probabilities $\{p_j\}$, having other equal conditions, will decrease much faster, and hence the errors from the above assumption will play a lesser role. Therefore, we will still use the condition (9) for the boundary layer N. The acceptability of this assumption can be verified by repeating the calculation for an increased value of N.

What is left to specify is how to calculate $\{z_j\}$ and $\{x_j\}$. We rewrite (5) accounting (7):

$$(z_j \beta'_j + x_j \beta''_j) A_j \mathbf{1}_{j+1} = z_j t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j.$$

Hence we have the proportionality

$$z_j = c x_j$$

with a factor

$$c = \frac{t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j - \beta''_j A_j \mathbf{1}_{j+1}}{\beta'_j A_j \mathbf{1}_{j+1}}.$$

Since all the products $A_j \mathbf{1}_{j+1}$ (the row sums of the matrices of requests arrival intensities) in the considered case of Poissonian incoming flow are equal to λ, the last formula can be represented as

$$c = \frac{t_{j+1}^{(m)} B_{j+1} \mathbf{1}_j - \lambda \beta''_j \mathbf{1}_j}{\lambda \beta'_j \mathbf{1}_j}. \tag{10}$$

Substitution of (8) in (6) and multiplying both parts of the result by $\mathbf{1}_j$ give

$$1 = t_j^{(m)} \mathbf{1}_j = x_j^{(m)} (c \beta'_j + \beta''_j) \mathbf{1}_j.$$

So,

$$x_j^{(m)} = 1/[(c \beta'_j + \beta''_j) \mathbf{1}_j]. \tag{11}$$

A convenient criterion for terminating iterations is the condition

$$\max_j |x_j^{(m)} - x_j^{(m-1)}| \leq \varepsilon.$$

Due to rather obvious considerations, it is convenient to choose initial conditional distributions of the number of served requests of each type to be binomial with probabilities proportional to $\{y_i/\mu_i\}$.

After iterations termination, we can now move on to find cumulant probabilities. Assuming $p_0 = 1$, we sequentially calculate

$$p_{j+1} = p_j x_j, \qquad j = \overline{0, N-1}, \tag{12}$$

and then normalize them to 1. We recall that sufficiency of the chosen N is determined by the smallness of the last probabilities.

Let's compare the implementation of the described method and the corresponding simulation model for a three-channel system with H_2- service time distribution (coefficient of variation $v = 2$)—Table 1:

Table 1. Probabilities of system states

j	Simulation	Calculation	j	Simulation	Calculation
0	8.578e-2	8.556e-2	10	5.577e-3	5.640e-3
1	1.986e-1	1.981e-1	11	2.838e-3	2.812e-3
2	2.256e-1	2.252e-1	12	1.319e-3	1.330e-3
3	1.650e-1	1.654e-1	13	6.093e-4	5.973e-4
4	1.178e-1	1.181e-1	14	2.611e-4	2.553e-4
5	8.096e-2	8.101e-2	15	1.188e-4	1.040e-4
6	5.286e-2	5.295e-2	16	4.774e-5	4.042e-5
7	3.275e-2	3.284e-2	17	1.886e-5	1.502e-5
8	1.924e-2	1.929e-2	18	4.876e-6	1.822e-6
9	1.064e-2	1.072e-2	19	3.149e-6	5.966e-7

Simulation was carried out before acceptance for service of 2 million requests. Therefore, the agreement of the calculated and statistical probabilities of order 10^{-4} and more, for which the number of observations exceeded 100, should be considered very good.

3 Distribution of the Waiting Time

The biggest problem in the systems with impatient requests calculation is the definition of their temporaryś characteristics. Here, the known formula for the moments of waiting time distribution

$$w_j = q_{[j]}/\lambda^j, \qquad j = 1, 2, \ldots, \tag{13}$$

in which $\{q_{[j]}\}$ are the factorial moments of queue length distribution, as shown by simulation experiments, gives poor accuracy.

Let us calculate

- vectors-rows $g_k = p_k * t_k$ of the stationary probabilities of microstates, $k = 0, 1, \ldots,$
- diagonal matrices of total up-transition intensities with elements $\{\sigma_{k,i}\}$,
- diagonal matrices $\{U_k(s)\}$ of the Laplace-Stieltjes transformations (LST) of the distributions of up-transitions duration by the corresponding total intensities $\{\sigma_k\}$ with elements $\{\sigma_{k,i}/(\sigma_{k,i} + s)\}$,
- the product $\tilde{U}_n(s)$ of the matrix $U_n(s)$ by a unit vector-column,
- matrices $\{T_k\}$ with elements $\{b_{k,i,j}/\sigma_{k,i}\}$ of the probabilities of transitions on the overlying layer.

In addition, we replace the diagonal matrix $U_n(s)$ by the same name vector-column. It is not difficult to see that the LST of waiting time outputs directly on service from n-th layer is

$$\omega_n(s) = g_n \tilde{U}_n(s).$$

For the $(n+1)$-th layer we have

$$\omega_{n+1}(s) = g_{n+1}[U_{n+1}(s)T_{n+1}]\tilde{U}_n(s),$$

for the $(n+2)$—

$$\omega_{n+2}(s) = g_{n+2}[U_{n+2}(s)T_{n+2}U_{n+1}(s)T_{n+1}]\tilde{U}_n(s),$$

etc. Given the above rules for forming factors $F_k(s) = U_k(s)T_k$ included in these formulas, we can immediately define matrices $\{F_k\}$ as the sets of elements type $\{b_{k,i,j}/(\sigma_{k,i} + s)\}$. Summing up the results for all possible starting layers, we get the final formula for the LST of waiting time distribution:

$$\omega(s) = \left[\sum_{k=0}^{\infty} g_{n+k} \prod_{i=0}^{k} F_{n+i}(s)\right]\tilde{U}_n(s). \tag{14}$$

In this formula, the *inverted* product symbol is used to specify the inverse order of cofactors (be reminded that the multiplication of matrices in general is non-commutative). The initial value is $F_n(s) = I$.

Having calculated a table of LST values in a neighbourhood of zero, we can construct its approximation by the Newton interpolation polynomial, and obtain the moments of waiting time distribution by multiple differentiation of the latter.

All is left is to consider the possibility of "impatience" for the same labeled request. Because for each request, the probability to endure the time u is equal to $e^{-\gamma u}$, the distribution function of a successful waiting

$$W^+(t) = \int_0^t e^{-\gamma u} w(u)\, du.$$

Accordingly, the LST of this distribution

$$\omega^+(s) = \int_0^\infty e^{-st} d\left[\int_0^t e^{-\gamma u} w(u)\, du \right] dt.$$

The derivative of integral by the parameter in this case is $e^{-\gamma t} w(t)$. Hence,

$$\omega^+(s) = \int_0^\infty e^{-st} e^{-\gamma t} w(t)\, dt = \omega(s+\gamma).$$

Thus, LST of the *successful* waiting should be calculated according to (14) with replacing the argument s by $s + \gamma$.

The moments of successful waiting, obtained in this way, should be divided by the probability of a successful wait, including zero, that is, by

$$\pi_w = \sum_{k=0}^{n-1} p_k + \omega^+(\gamma).$$

We compare the numerical results obtained by this technique and by means of simulation (2 million of served requests). For a three-channel system with the intensity $\lambda = 1.5$ of the incoming Poissonian flow, average service time $b_1 = 4.0$, the service variation factor $v_b = 2.0$ and impatience intensity $\gamma = 0.2$ the results are summarized in Table 2.

Table 2. Moments of the distribution of successful waiting

Method	w_1^+	w_2^+	w_3^+
Imitation	0.483	1.048	3.319
Calculation	0.479	1.039	3.287

4 Distribution of a Successful Request Sojourn Time

When the request from a queue has been extracted, the assumption of the permissible patience having Markovian distribution allows to count down its patience *anew*. After all, we are only interested in "successful" requests which received complete servicing. Suppose the distribution of the latter be two-phase hyperexponential with parameters $\{y_m, \mu_m\}$. Then the j-th moment of the time of successful servicing

$$b_j^+ = \int_0^\infty \left[\int_0^\theta t^j b(t)\, dt \right] \gamma e^{-\gamma\theta}\, d\theta = \int_0^\infty \left[\int_0^\theta t^j \sum_{m=1}^{2} y_m \mu_m e^{-\mu_m t}\, dt \right]. \qquad (15)$$

It can be shown that

$$b_j^+ = j! \sum_{m=1}^{2} \frac{y_m}{\mu_m^j} - j! \sum_{m=1}^{2} \frac{y_m}{\mu_m^j} \cdot \sum_{i=0}^{j} \frac{\gamma}{\mu_m + \gamma} \left(\frac{\mu_m}{\mu_m + \gamma} \right)^i$$
$$= j! \sum_{m=1}^{2} \frac{y_m}{\mu_m^j} \left[1 - \frac{\gamma}{\mu_m + \gamma} \sum_{i=0}^{j} \left(\frac{\mu_m}{\mu_m + \gamma} \right)^i \right]. \tag{16}$$

In accordance with (15), the zero moment of the successful servicing can be considered as a probability π_s of the such. Substituting $j = 0$ in (15), we get

$$\pi_s = \sum_{m=1}^{2} u_m \left[1 - \frac{\gamma}{\mu_m + \gamma} \right] = \sum_{m=1}^{2} \frac{y_m \mu_m}{\mu_m + \gamma}. \tag{17}$$

It is of interest to estimate the volume of wasted service. The average service time of the interrupted request

$$\bar{\tau} = \int_0^\infty \theta \bar{B}(\theta) \gamma e^{-\gamma\theta} \, d\theta,$$

where $\bar{B}(\theta)$ is the complementary distribution function of the full service duration. In our problem

$$\bar{\tau} = \int_0^\infty \theta \left[\sum_{m=1}^{2} y_m e^{-\mu_m \theta} \right] \gamma e^{-\gamma\theta} \, d\theta \tag{18}$$
$$= \gamma \sum_{m=1}^{2} y_m \int_0^\infty \theta e^{-(\mu_m + \gamma)\theta} \, d\theta = \gamma \sum_{m=1}^{2} \frac{y_m}{(\mu_m + \gamma)^2}.$$

Total losses per unit of time will be

$$g = \lambda \pi_w (1 - \pi_s) \bar{\tau}.$$

The moments of $\{v_j^+\}$ of a successful stay in the system are calculated via convolution of $\{w_j^+\}$ and $\{b_j^+\}$, and the probability of a successful stay in the system

$$\pi_v = \pi_w \pi_s.$$

5 Calculation of a Network with Impatient Requests

The assumption that permissible patience has the Markovian distribution allows us to apply the results of previous section to the calculation of networks with hyper-exponential servicing, using their flow-equivalent decomposition. Since in our case the intensity of the exiting successful flow differs from the intensity of the incoming one, it is necessary to make the following changes in the usual scheme of open network calculation:

1. There should be no cyclic routes in the network.
2. The numbering and, respectively, the order of the nodes calculation must be determined on the basis of a preceding relationship—for example, using the well-known Floyd algorithm.
3. The nodes with impatience must be calculated using the above method. The intensities of the output flow for such nodes must be calculated via multiplying the incoming intensities by the corresponding probability π_v.

The network sojourn time calculation deserves a special consideration. When talking about most critical applications, it isn't enough to know the average sojourn time—such cases usually raise the question of the highest moments and/or calculation of the distribution function. Appropriate technique [8] is based on the construction of the LST for the network sojourn time distribution via the "nodal" LST, a routing matrix, and its' subsequent numerical differentiation at zero.

6 Conclusion

The main results of this work are as follows:

1. A diagram of transitions between microstates of the model $M/H_2/n - M$ is proposed taking into account exponentially distributed patience of all requests located in the system. On its basis, the rules are corrected to calculate the matrices $\{B_j\}$ and $\{D_j\}$ of the transition intensities.
2. Permissibility of using the formula (8) was justified, which allows to limit the number of accounted layers.
3. The stationary probabilities of system states, moments of successful waiting and sojourn time were obtained and compared with their analogs received by simulation.
4. The formulas for calculating average losses from interrupted service per unit of time and the intensity of the flow of successful requests were proposed
5. Application of these results to the calculation of the *queueing networks* service with impatient requests was demonstrated.

Acknowledgments. The work described in the paper was supported by state project 0073-2018-0003.

References

1. Takagi, H.: Waiting time in the M/M/m/(m+c) queue with impatient customers. Int. J. Pure Appl. Math. **90**(4), 519–559 (2014)
2. Aktekin, T., Soyer, R.: Bayesian analysis of queues with impatient customers: applications to call centers. Nav. Res. Logist. **59**, 441–456 (2012)
3. Boot, N.K., Tijm, H.: A multiserver queueing system with impatient customers. Manag. Sci. **45**(3), 444–448 (1999)

4. Ryzhikov, Yu.I., Ulanov, A.V.: Calculation of the hyperexponential queueing system $M/H_2/n - H_2$ with requests impatient in the queue. Bull. Tomsk State Univ.: Manag. Comput. Technol. Comput. Sci. **2**(27), 47–53 (2014). (in Russian)
5. Ryzhikov, Yu.I.: Iterative method for calculating multi-channel queueing systems - the basics, modifications and limiting opportunities. In: Proceedings of the 9th Russian Multiconference on Control Problems. Information Technologies in Management. SPb.: "Concern Elektropribor", pp. 224–233 (2016). (in Russian)
6. Seelen, L.P.: An algorithm for Ph/Ph/c Queues. Eur. J. Oper. Res. **23**, 118–127 (1986)
7. Takahashi, Y., Takami, Y.: A numerical method for the steady-state probabilities of a GI/G/c queuing system in a general class. J. Oper. Res. Soc. Jpn. **19**(2), 147–157 (1976)
8. Ryzhikov, Yu.I.: An algorithmic approach to queueing problems: monograph. A.F. Mojaysky VKA (Military Space Academy), St.-Petersburg (2013). (in Russian)