

## ASYMPTOTIC ANALYSIS FOR WAITING TIME IN A RETRIAL QUEUE WITH MULTIPLE INPUT STREAMS

YANG SONG<sup>✉</sup> AND QI LIU\*<sup>✉</sup>

**Abstract.** Under the classical retrial policy, we consider a single-server  $M/G/1$  queue with multiple input streams and orbits. Different types of customers have corresponding arrival rates, general distributions of service time and retrial rates. Assume that the retrial rates for different types of customers linearly converge to zero. We firstly derive the first-order asymptotics of the orbit queue lengths. Subsequently, we find that the joint asymptotic distribution of the number of retrials follows a multidimensional geometric distribution. Finally, we obtain the joint asymptotic distribution of waiting times, which follows a multidimensional exponential distribution. This result indicates that the waiting times for different types of customers are independent of each other.

**Mathematics Subject Classification.** 60K25, 68M20.

Received November 5, 2024. Accepted January 2, 2026.

### 1. INTRODUCTION

In a retrial queueing system, when arriving customers find the server occupied, they choose to join an orbit and retry for service after a random interval. Retrial queues are generally used to model the problems in telecommunication systems, computer networks, call centers and so on. Readers may refer to [1, 2] for a comprehensive overview of the fundamental methodologies.

In many scenarios, customers are often classified by their service requirements or attributes. So the retrial queueing models with multi-class customers are of great significance in practice. Extensive research exists on such queueing systems (simply called multi-class retrial queues) with different retrial policies. As to the constant retrial policy, by using regenerative approach, Avrachenkov *et al.* [3] deduced necessary stability conditions for  $GI/G/c$  multi-class retrial queue; then in [4], they obtained the sufficient stability conditions for  $M/G/c$  queues with finite waiting space. For a multi-class  $M/G/1$  retrial queue with balking, Morozov *et al.* [5] established the necessary and sufficient stability condition, also obtained performance measures under exponential service times assumption. With classical retrial policy, stability conditions for the multi-class  $GI/G/c$  and  $M/G/1$  retrial queue were established in [6, 7], respectively. In these aforementioned studies, exponential retrial intervals are considered. However, in [8], the stability condition was studied for a multi-class  $M/G/1$  model with general retrials. Apart from concerns on the stability conditions, the approximation of probability distribution of orbit queue length was obtained for a multi-class  $M/M/1$  retrial queue with sufficiently small retrial rate [9]. Also,

---

*Keywords.* Retrial queue, waiting time, asymptotic analysis, number of retrials, orbit queue length.

School of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P.R. China.

\*Corresponding author: [iuliuqi@nuaa.edu.cn](mailto:iuliuqi@nuaa.edu.cn)

for a multi-class  $M/M/c$  retrial queue, as the retrial rate going to infinity, Shin and Moon [10] provided the joint stationary distribution for the number of customers in service and orbit.

As stated in [2], it is well known that multi-class retrial queues with parallel orbits are difficult to analyze, since the joint queue length process is a random walk on the multidimensional integer lattice. Specific to the retrial queueing models with double orbits, researchers adopted boundary value problem or kernel method to analyze the stationary distribution of the joint orbit queue length, see [11–13]. Recently, for an  $M/M/1$  retrial model with double orbits, the parametric non-linear programming approach was used to solve for the average queue length and waiting time [14]. Both queue length and waiting time are important performance measures for evaluating a queueing model, while little work attends to the joint distribution of customers' waiting times in different orbits. Comparing to the queue length, waiting time is more difficult to handle with. One reason is that, under the classical retrial policy, customers joining the orbit later may be served before those joining earlier. So it is impossible to study a specific customer's waiting time. While for any tagged customer in each orbit, studying the joint stationary distribution of customers' waiting times remains intractable.

Among the research about waiting time for retrial queueing systems with single-class customers, much of them focused on the Laplace transform, moment, numerical approximation, and tail asymptotics [15–19]. Besides, we notice that for a single-server retrial queue with finite sources, by allowing the number of customer sources to approach infinity, the authors employed asymptotic analysis and derived the asymptotic distribution of customer's waiting time [20–22]. In these work, it is not necessary to derive the exact expressions for the generating functions of number of retrials, nor for the characteristic functions of waiting times. Instead, on the basis of limit condition, what they need are asymptotic solutions for those equations of generating functions or characteristic functions.

Inspired by these, we extend the study to a single-server retrial queue with multiple types of customers and orbits. Each type of customer has its own arrival rate, retrial rate, and general service time distribution. When a type  $j$  ( $j = 1, \dots, N$ ) customer's retrial rate linearly approaches zero, we investigate the joint asymptotic distribution of waiting times for the tagged customers in different orbits. It is worth mentioning that, under the classical retrial policy, although each customer's retrial rate goes to zero, the total retrial rate in the orbits might not be negligible.

The proposed retrial queueing model effectively characterizes the behaviors of one kind of computer virus. When spreading to computers *via* the Internet, the virus may lurk unseen in computer's memory and wait for another attack later. See each potential attack as a retrial. Due to the protection of antivirus software, the potential attack may fail. After a random amount of time, the virus may launch another attack. Suppose computers process the incoming data as soon as possible. Compared with the service time, the latency period of virus is much longer. Our proposed model just describes the inactive state for computer virus. It provides a realistic and analytical framework to understand the impact of computer virus. In this work, the main contributions are theoretical results obtained under the condition that retrial rates converge to zero. Numerical experiments demonstrate that the asymptotic results agree with the simulated distributions, when retrial rates are relatively small. This effectively verifies the correctness of the theoretical results. Furthermore, taking advantage of the asymptotic results of waiting time or number of retrials, we can explore the protection strategies to mitigate the impact of computer virus.

The remaining of the paper is organized as follows. Section 2 introduces the model discussed in this paper. In Section 3, we analyze the first-order asymptotics of the orbit queue length. In Section 4 and 5, for the customers joining the orbits, we obtain the asymptotic distributions of number of retrials and waiting times, respectively. In Section 6, we conduct several numerical experiments to illustrate the correctness of our theoretical results. Finally, conclusions are presented in Section 7.

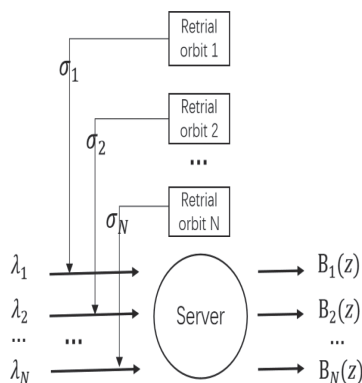


FIGURE 1. Queueing system.

## 2. MODEL DESCRIPTION

We consider a single-server retrial queueing system with multiple input streams and retrial orbits, as shown in Figure 1. Let  $j = 1, \dots, N$ . The customers arrive at the system following  $N$  independent Poisson streams with rate  $\lambda_j$ , and the server holds at most one customer each time. Denote total arrival rate  $\lambda = \lambda_1 + \dots + \lambda_N$ . For a type  $j$  customer, its service time  $T_j$  follows a general distribution with probability distribution function  $B_j(z)$ . If the server is available, an arriving customer of type  $j$  receives service immediately; otherwise, the customer joins orbit  $j$  and seeks service again after a random amount of time. The intervals between two successive retrials of a type  $j$  customer are independent and exponentially distributed random variables with parameter  $\sigma_j$ . The corresponding Laplace transform is  $\varphi_j(\alpha) = \frac{\sigma_j}{\alpha + \sigma_j}$ .

For a tagged customer of type  $j$ , define two random variables: the residual waiting time  $W_{\text{res}}^{(j)}(t)$  represents the interval from time  $t$  till the service beginning of the tagged customer; the residual number of retrials  $R_{\text{res}}^{(j)}(t)$  represents the number of retrials within  $W_{\text{res}}^{(j)}(t)$ . At time  $t$ , let  $Q_j(t)$  denote the number of customers in orbit  $j$  and  $C(t)$  the state of the server. When the server is idle,  $C(t) = 0$ ; when the server is busy serving a type  $j$  customer,  $C(t) = j$ . For the customer under service at time  $t$ , introduce an elapsed service time  $Z(t)$  as a supplementary variable, then we construct a continuous-time Markov chain  $\{(C(t), Q_1(t), \dots, Q_N(t), Z(t)) : t \in [0, \infty)\}$  on the state space  $\{0, 1, \dots, N\} \times \{0, 1, \dots\}^N \times [0, \infty)$ . Assume  $\sum_{j=1}^N \lambda_j E[T_j] < 1$ , the system is stable. Tagging any customer in orbit  $j$ , we denote its waiting time as  $W_j$  and the number of retrials before receiving service as  $R_j$ . Let  $W_{\text{res}}^{(j)} = \lim_{t \rightarrow \infty} W_{\text{res}}^{(j)}(t)$ ,  $R_{\text{res}}^{(j)} = \lim_{t \rightarrow \infty} R_{\text{res}}^{(j)}(t)$ ,  $C = \lim_{t \rightarrow \infty} C(t)$ ,  $Q_j = \lim_{t \rightarrow \infty} Q_j(t)$ ,  $Z = \lim_{t \rightarrow \infty} Z(t)$ . Set  $D_0 = P(C = 0)$ ,  $D_j(z) dz = P(C = j, z \in (z, z + dz))$ ,  $D_j = \int_0^\infty D_j(z) dz$ . Then define the stationary distributions and probability density functions as

$$\begin{aligned}
 P_0(m_1, \dots, m_N) &= P(C = 0, Q_1 = m_1, \dots, Q_N = m_N), \\
 P_j(m_1, \dots, m_N, z) &= P(C = j, Q_1 = m_1, \dots, Q_N = m_N, Z < z), \\
 p_j(m_1, \dots, m_N, z) dz &= dP_j(m_1, \dots, m_N, z) \\
 &= P(C = j, Q_1 = m_1, \dots, Q_N = m_N, Z \in (z, z + dz)), \\
 P_j(m_1, \dots, m_N) &= P(C = j, Q_1 = m_1, \dots, Q_N = m_N) \\
 &= \int_0^\infty p_j(m_1, \dots, m_N, z) dz.
 \end{aligned}$$

For  $j = 1, \dots, N$ , they satisfy the following balanced equations

$$\sum_{j=1}^N (\lambda_j + m_j \sigma_j) P_0(m_1, \dots, m_N) = \sum_{j=1}^N \int_0^\infty \mu_j(z) p_j(m_1, \dots, m_N, z) dz, \tag{2.1}$$

$$\sum_{l=1}^N \lambda_l p_j(m_1, \dots, m_l - 1, \dots, m_N, z) = \frac{dp_j(m_1, \dots, m_N, z)}{dz} + \left[ \sum_{l=1}^N \lambda_l + \mu_j(z) \right] p_j(m_1, \dots, m_N, z),$$

with boundary conditions

$$\lambda_j P_0(m_1, \dots, m_N) + (m_j + 1) \sigma_j P_0(m_1, \dots, m_{j-1}, m_j + 1, m_{j+1}, \dots, m_N) = p_j(m_1, \dots, m_N, 0), \tag{2.2}$$

where  $\mu_j(z) = B'_j(z)/(1 - B_j(z))$ . It is worth noting that, if  $m_j = -1$  for any  $j = 1, \dots, N$ , then  $p_j(m_1, \dots, m_j, \dots, m_N, z) = 0$ .

Next, for  $j = 1, \dots, N$ , we introduce the following steady-state characteristic functions

$$H_0(u_1, \dots, u_N) = \sum_{m_1=0}^\infty \dots \sum_{m_N=0}^\infty e^{i(u_1 m_1 + \dots + u_N m_N)} P_0(m_1, \dots, m_N),$$

$$H_j(u_1, \dots, u_N, z) = \sum_{m_1=0}^\infty \dots \sum_{m_N=0}^\infty e^{i(u_1 m_1 + \dots + u_N m_N)} p_j(m_1, \dots, m_N, z),$$

$$H_j(u_1, \dots, u_N) = \int_0^\infty H_j(u_1, \dots, u_N, z) dz,$$

where  $i = \sqrt{-1}$  is the imaginary unit. Notice that

$$\sum_{m_1=0}^\infty \dots \sum_{m_N=0}^\infty e^{i(u_1 m_1 + \dots + u_N m_N)} m_j P_0(m_1, \dots, m_N) = -i \frac{\partial H_0(u_1, \dots, u_N)}{\partial u_j}, \quad (j = 1, \dots, N).$$

Multiplying both sides of (2.1) and (2.2) by  $e^{i(u_1 m_1 + \dots + u_N m_N)}$ , then summing  $m_j$  from 0 to  $\infty$  for  $j = 1, \dots, N$ , we get

$$\sum_{j=1}^N \lambda_j H_0(u_1, \dots, u_N) - \sum_{j=1}^N \sigma_j i \frac{\partial H_0(u_1, \dots, u_N)}{\partial u_j}$$

$$- \sum_{j=1}^N \int_0^\infty \mu_j(z) H_j(u_1, \dots, u_N, z) dz = 0,$$

$$\frac{\partial H_j(u_1, \dots, u_N, z)}{\partial z} + \left[ \sum_{l=1}^N \lambda_l + \mu_j(z) \right] H_j(u_1, \dots, u_N, z)$$

$$- \sum_{l=1}^N \lambda_l e^{i u_l} H_j(u_1, \dots, u_N, z) = 0, \quad (j = 1, \dots, N),$$

$$\lambda_j H_0(u_1, \dots, u_N) - i \sigma_j e^{-i u_j} \frac{\partial H_0(u_1, \dots, u_N)}{\partial u_j} = H_j(u_1, \dots, u_N, 0), \quad (j = 1, \dots, N). \tag{2.3}$$

By now, we have established the equations (2.3) for characteristic functions of the orbit queue lengths.

## 3. ASYMPTOTICS FOR THE ORBIT QUEUE LENGTH

Due to the difficulty in obtaining the explicit solution to equations (2.3), in this section, we propose the important assumption for our model, that is  $\sigma_j \rightarrow 0$  ( $j = 1, \dots, N$ ). Under this condition, suppose  $\sigma_j = \gamma_j \sigma$ , where  $\gamma_j$  ( $j = 1, \dots, N$ ) could be different constants and  $\sigma \rightarrow 0$ , we employ the asymptotic analysis method to investigate the asymptotic solution to equations (2.3), and the result about first-order asymptotics of orbit queue lengths is shown in Theorem 3.1.

**Theorem 3.1.**

$$\lim_{\sigma \rightarrow 0} E[e^{iu_1 Q_1 + \dots + iu_N Q_N}] = e^{iu_1 \frac{\kappa_1}{\sigma} + \dots + iu_N \frac{\kappa_N}{\sigma}}, \quad (3.1)$$

where  $\kappa_j$  ( $j = 1, 2, \dots, N$ ) satisfies

$$\gamma_j \kappa_j = \lambda_j \sum_{l=1}^N (\lambda_l + \gamma_l \kappa_l) \int_0^\infty [1 - B_l(z)] dz. \quad (3.2)$$

In addition, the stationary probability distributions of server's state are given by

$$D_0 = \left\{ 1 + \sum_{j=1}^N (\lambda_j + \gamma_j \kappa_j) \int_0^\infty [1 - B_j(z)] dz \right\}^{-1}, \quad (3.3)$$

$$D_j = (\lambda_j + \gamma_j \kappa_j) D_0 \int_0^\infty [1 - B_j(z)] dz.$$

*Proof.* For  $j = 1, \dots, N$  and  $k = 0, 1, \dots, N$ , denote  $\epsilon = \sigma$ ,  $u_j = \epsilon w_j$  and make the following notations

$$F_0(w_1, \dots, w_N, \epsilon) = H_0(u_1, \dots, u_N),$$

$$F_j(w_1, \dots, w_N, z, \epsilon) = H_j(u_1, \dots, u_N, z),$$

$$F_j(w_1, \dots, w_N, \epsilon) = \int_0^\infty F_j(w_1, \dots, w_N, z, \epsilon) dz,$$

$$F_j(w_1, \dots, w_N, z) = \lim_{\epsilon \rightarrow 0} F_j(w_1, \dots, w_N, z, \epsilon),$$

$$F_k(w_1, \dots, w_N) = \lim_{\epsilon \rightarrow 0} F_k(w_1, \dots, w_N, \epsilon).$$

According to equations (2.3), we easily obtain

$$\begin{aligned} & \sum_{j=1}^N \lambda_j F_0(w_1, \dots, w_N, \epsilon) - \sum_{j=1}^N \gamma_j i \frac{\partial F_0(w_1, \dots, w_N, \epsilon)}{\partial w_j} \\ & - \sum_{j=1}^N \int_0^\infty \mu_j(z) F_j(w_1, \dots, w_N, z, \epsilon) dz = 0, \\ & \frac{\partial F_j(w_1, \dots, w_N, z, \epsilon)}{\partial z} + \left[ \sum_{l=1}^N \lambda_l + \mu_j(z) \right] F_j(w_1, \dots, w_N, z, \epsilon) \\ & - \sum_{l=1}^N \lambda_l e^{i\epsilon w_l} F_j(w_1, \dots, w_N, z, \epsilon) = 0, \quad (j = 1, \dots, N), \\ & \lambda_j F_0(w_1, \dots, w_N, \epsilon) - i\gamma_j e^{-i\epsilon w_j} \frac{\partial F_0(w_1, \dots, w_N, \epsilon)}{\partial w_j} = F_j(w_1, \dots, w_N, 0, \epsilon), \quad (j = 1, \dots, N). \end{aligned} \quad (3.4)$$

Following two steps, we investigate the equations (3.4) and show the main result in Theorem 3.1.

**Step 1.** Under  $\epsilon \rightarrow 0$ , consider the limits in equations (3.4). So,

$$\begin{aligned} \sum_{j=1}^N \lambda_j F_0(w_1, \dots, w_N) - \sum_{j=1}^N \gamma_j i \frac{\partial F_0(w_1, \dots, w_N)}{\partial w_j} \\ - \sum_{j=1}^N \int_0^\infty \mu_j(z) F_j(w_1, \dots, w_N, z) dz = 0, \\ \frac{\partial F_j(w_1, \dots, w_N, z)}{\partial z} + \mu_j(z) F_j(w_1, \dots, w_N, z) = 0, \quad (j = 1, \dots, N), \\ \lambda_j F_0(w_1, \dots, w_N) - i\gamma_j \frac{\partial F_0(w_1, \dots, w_N)}{\partial w_j} = F_j(w_1, \dots, w_N, 0), \quad (j = 1, \dots, N). \end{aligned} \tag{3.5}$$

For (3.5), we confirm its solution with form

$$\begin{aligned} F_0(w_1, \dots, w_N) &= D_0 \Phi(w_1, \dots, w_N), \\ F_j(w_1, \dots, w_N, z) &= D_j(z) \Phi(w_1, \dots, w_N), \quad (j = 1, \dots, N). \end{aligned} \tag{3.6}$$

Substituting (3.6) into (3.5), we get

$$\begin{aligned} D_0 \sum_{j=1}^N \lambda_j - iD_0 \sum_{j=1}^N \gamma_j \frac{\partial \Phi(w_1, \dots, w_N) / \partial w_j}{\Phi(w_1, \dots, w_N)} - \sum_{j=1}^N \int_0^\infty \mu_j(z) D_j(z) dz = 0, \\ \frac{dD_j(z)}{dz} + \mu_j(z) D_j(z) = 0, \quad (j = 1, \dots, N), \\ \lambda_j D_0 - i\gamma_j D_0 \frac{\partial \Phi(w_1, \dots, w_N) / \partial w_j}{\Phi(w_1, \dots, w_N)} = D_j(0), \quad (j = 1, \dots, N). \end{aligned} \tag{3.7}$$

Note that the solution to (3.7) has the following form

$$\Phi(w_1, \dots, w_N) = e^{i(w_1 \kappa_1 + \dots + w_N \kappa_N)}. \tag{3.8}$$

Then rewrite (3.7) as

$$\begin{aligned} \sum_{j=1}^N (\lambda_j + \gamma_j \kappa_j) D_0 - \sum_{j=1}^N \int_0^\infty \mu_j(z) D_j(z) dz = 0, \\ \frac{dD_j(z)}{dz} + \mu_j(z) D_j(z) = 0, \quad (j = 1, \dots, N), \\ (\lambda_j + \gamma_j \kappa_j) D_0 = D_j(0), \quad (j = 1, \dots, N). \end{aligned} \tag{3.9}$$

In (3.9), the last two equations can yield the first equation, so we solve it to get

$$\begin{aligned} D_j(z) &= (\lambda_j + \gamma_j \kappa_j) D_0 e^{-\int_0^z \mu_j(v) dv} \\ &= (\lambda_j + \gamma_j \kappa_j) D_0 [1 - B_j(z)]. \end{aligned}$$

Furthermore,

$$D_j = \int_0^\infty D_j(z) dz = (\lambda_j + \gamma_j \kappa_j) D_0 \int_0^\infty [1 - B_j(z)] dz.$$

Using the normalization condition  $D_0 + D_1 + \dots + D_N = 1$ , we obtain

$$D_0 = \left\{ 1 + \sum_{j=1}^N (\lambda_j + \gamma_j \kappa_j) \int_0^\infty [1 - B_j(z)] dz \right\}^{-1}.$$

**Step 2.** In (3.4), for  $j = 1, \dots, N$ , integrate the second equation with respect to  $z$  from 0 to  $\infty$ , then add it to the third equation. Subsequently, sum  $j$  from 1 to  $\infty$  for the obtained equation, and add it to the first equation in (3.4), then we get

$$\sum_{j=1}^N i\gamma_j (e^{-i\epsilon w_j} - 1) \frac{\partial F_0(w_1, \dots, w_N, \epsilon)}{\partial w_j} + \sum_{j=1}^N \left[ \sum_{l=1}^N \lambda_l (1 - e^{i\epsilon w_l}) \right] F_j(w_1, \dots, w_N, \epsilon) = 0. \quad (3.10)$$

Considering  $e^{\pm i\epsilon w_j} = 1 \pm i\epsilon w_j + o(\epsilon)$  as  $\epsilon \rightarrow 0$ , and substituting (3.6) and (3.8) into (3.10), we obtain

$$\sum_{j=1}^N \left( \gamma_j \kappa_j D_0 - \lambda_j \sum_{l=1}^N D_l \right) w_j = 0,$$

then

$$\gamma_j \kappa_j D_0 = \lambda_j \sum_{l=1}^N D_l, \quad (j = 1, \dots, N). \quad (3.11)$$

Substituting (3.3) into (3.11), we obtain that  $\kappa_j (j = 1, \dots, N)$  are the positive solutions of (3.2).

Let

$$H(u_1, \dots, u_N) = \sum_{m_1=0}^{\infty} \dots \sum_{m_N=0}^{\infty} e^{i(u_1 m_1 + \dots + u_N m_N)} P(Q_1 = m_1, \dots, Q_N = m_N),$$

then

$$H(u_1, \dots, u_N) = H_0(u_1, \dots, u_N) + H_1(u_1, \dots, u_N) + \dots + H_N(u_1, \dots, u_N).$$

Since

$$\lim_{\epsilon \rightarrow 0} H_k(u_1, \dots, u_N) = \lim_{\epsilon \rightarrow 0} F_k(w_1, \dots, w_N, \epsilon) = F_k(w_1, \dots, w_N), \quad (k = 0, 1, \dots, N),$$

and

$$\begin{aligned} F_0(w_1, \dots, w_N) + F_1(w_1, \dots, w_N) + \dots + F_N(w_1, \dots, w_N) \\ = (D_0 + D_1 + \dots + D_N) e^{i(w_1 \kappa_1 + \dots + w_N \kappa_N)} = e^{i(w_1 \kappa_1 + \dots + w_N \kappa_N)}, \end{aligned}$$

so,

$$\lim_{\epsilon \rightarrow 0} H(u_1, \dots, u_N) = \lim_{\epsilon \rightarrow 0} E[e^{i\epsilon(w_1 Q_1 + \dots + w_N Q_N)}] = e^{i(w_1 \kappa_1 + \dots + w_N \kappa_N)},$$

which yields (3.1). The proof is completed. □

**Remark 3.2.** For  $j = 1, \dots, N$ , the result shown in (3.1) implies  $E[|Q_j \sigma - \kappa_j|] \rightarrow 0$  as  $\sigma \rightarrow 0$ , which means  $Q_j \sigma$  converges to  $\kappa_j$  in distribution as  $\sigma \rightarrow 0$ . Each orbit queue length weakly converges to a specific value, which is the foundation for deriving the asymptotic distribution of the residual number of retrials.

#### 4. NUMBER OF RETRIALS

In this section, we analyze the joint asymptotic distribution of the number of retrials for all types of tagged customers. Initially, in Section 4.1, under the condition  $\sigma \rightarrow 0$ , the asymptotic analysis method is utilized to solve the equations for the conditional generating functions of the residual number of retrials, then we derive the asymptotic distribution for the residual number of retrials. Subsequently, in Section 4.2, the joint asymptotic distribution for the number of retrials is derived.

**4.1. Asymptotic distribution for the residual number of retrials**

For  $j = 1, \dots, N$ ,  $r_j \in \{1, 2, 3, \dots\}$ , denote

$$\begin{aligned} & \Pi_0(r_1, \dots, r_N, m_1, \dots, m_N) \\ &= P\{R_{\text{res}}^{(1)} = r_1, \dots, R_{\text{res}}^{(N)} = r_N | C = 0, Q_1 = m_1, \dots, Q_N = m_N\}, \\ & \Pi_j(r_1, \dots, r_N, m_1, \dots, m_N, z) \\ &= P\left\{R_{\text{res}}^{(1)} = r_1, \dots, R_{\text{res}}^{(N)} = r_N | C = j, Q_1 = m_1, \dots, Q_N = m_N, Z \in (z, z + dz)\right\} \end{aligned}$$

as the conditional distribution of the residual number of retrials. By conditioning on the next event occurring in the queueing system, we derive the following balanced equations based on the Markovian property, that is,

$$\begin{aligned} & \sum_{j=1}^N (\lambda_j + m_j \sigma_j) \Pi_0(r_1, \dots, r_N, m_1, \dots, m_N) \\ &= \sum_{j=1}^N \lambda_j \Pi_j(r_1, \dots, r_N, m_1, \dots, m_N, 0) \\ & \quad + \sum_{j=1}^N (m_j - 1) \sigma_j \Pi_j(r_1, \dots, r_N, m_1, \dots, m_{j-1}, m_j - 1, m_{j+1}, \dots, m_N, 0) \\ & \quad + \sum_{j=1}^N \sigma_j \Pi_j(r_1, \dots, r_{j-1}, r_j - 1, r_{j+1}, \dots, r_N, m_1, \dots, m_{j-1}, m_j - 1, m_{j+1}, \dots, m_N, 0)^*, \\ & \quad - \partial \Pi_j(r_1, \dots, r_N, m_1, \dots, m_N, z) / \partial z \tag{4.1} \\ &= - \left[ \sum_{l=1}^N (\lambda_l + \sigma_l) + \mu_j(z) \right] \Pi_j(r_1, \dots, r_N, m_1, \dots, m_N, z) \\ & \quad + \sum_{l=1}^N \lambda_l \Pi_j(r_1, \dots, r_N, m_1, \dots, m_l, m_l + 1, m_{l+1}, \dots, m_N, z) \\ & \quad + \sum_{l=1}^N \sigma_l \Pi_j(r_1, \dots, r_{l-1}, r_l - 1, r_{l+1}, \dots, r_N, m_1, \dots, m_N, z) \\ & \quad + \mu_j(z) \Pi_0(r_1, \dots, r_N, m_1, \dots, m_N), \quad (j = 1, \dots, N). \end{aligned}$$

For the first equation in (4.1), the last term with superscript \* only appears when  $r_j = 1$ , ( $j = 1, \dots, N$ ).

Define the steady-state generating functions

$$\begin{aligned} & G_0(z_1, \dots, z_N, m_1, \dots, m_N) \\ &= E\left\{z_1^{R_{\text{res}}^{(1)}} \dots z_N^{R_{\text{res}}^{(N)}} | C = 0, Q_1 = m_1, \dots, Q_N = m_N\right\} \\ &= \sum_{r_1=1}^{\infty} \dots \sum_{r_N=1}^{\infty} z_1^{r_1} \dots z_N^{r_N} \Pi_0(r_1, \dots, r_N, m_1, \dots, m_N), \\ & G_j(z_1, \dots, z_N, m_1, \dots, m_N, z) \\ &= E\left\{z_1^{R_{\text{res}}^{(1)}} \dots z_N^{R_{\text{res}}^{(N)}} | C = j, Q_1 = m_1, \dots, Q_N = m_N, Z \in (z, z + dz)\right\} \\ &= \sum_{r_1=1}^{\infty} \dots \sum_{r_N=1}^{\infty} z_1^{r_1} \dots z_N^{r_N} \Pi_j(r_1, \dots, r_N, m_1, \dots, m_N, z), \quad (j = 1, \dots, N). \end{aligned}$$

TABLE 1. Equivalence between the notations.

Notations	Represent for
$G_0(z_1, \dots, z_N, m_1, \dots, m_N)$	Generating function of $(R_{\text{res}}^{(1)}, \dots, R_{\text{res}}^{(N)})$ when $C = 0$
$G_j(z_1, \dots, z_N, m_1, \dots, m_N, z)$	Generating function of $(R_{\text{res}}^{(1)}, \dots, R_{\text{res}}^{(N)})$ when $C = j$
$\epsilon$	$\sigma = \sigma_j / \gamma_j$ ( $\gamma_j$ is a constant)
$x_j$	$m_j \epsilon$
$S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon)$	$G_0(z_1, \dots, z_N, x_1 / \epsilon, \dots, x_N / \epsilon)$
$S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon)$	$G_j(z_1, \dots, z_N, x_1 / \epsilon, \dots, x_N / \epsilon, z)$
$K_0(z_1, \dots, z_N, x_1, \dots, x_N)$	$\lim_{\epsilon \rightarrow 0} S_0(z_1, \dots, z_N, m_1, \dots, m_N, \epsilon)$
$K_j(z_1, \dots, z_N, x_1, \dots, x_N, z)$	$\lim_{\epsilon \rightarrow 0} S_j(z_1, \dots, z_N, m_1, \dots, m_N, z, \epsilon)$
$K(z_1, \dots, z_N, x_1, \dots, x_N)$	$K_0(z_1, \dots, z_N, x_1, \dots, x_N), K_j(z_1, \dots, z_N, x_1, \dots, x_N, z)$
$H(z_1, \dots, z_N)$	$K(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N)$

Multiply both sides of equations in (4.1) by  $z_1^{r_1} \cdots z_N^{r_N}$  and sum all  $r_j$  ( $j = 1, \dots, N$ ) from 1 to  $\infty$ , then it yields

$$\begin{aligned}
& - \sum_{j=1}^N (\lambda_j + m_j \sigma_j) G_0(z_1, \dots, z_N, m_1, \dots, m_N) + \sum_{j=1}^N \lambda_j G_j(z_1, \dots, z_N, m_1, \dots, m_N, 0) \\
& + \sum_{j=1}^N (m_j - 1) \sigma_j G_j(z_1, \dots, z_N, m_1, \dots, m_{j-1}, m_j - 1, m_{j+1}, \dots, m_N, 0) \\
& + \sum_{j=1}^N \sigma_j z_j G_j(z_1, \dots, z_{j-1}, 1, z_{j+1}, \dots, z_N, m_1, \dots, m_{j-1}, m_j - 1, m_{j+1}, \dots, m_N, 0) = 0, \\
& \partial G_j(z_1, \dots, z_N, m_1, \dots, m_N, z) / \partial z - \left[ \sum_{l=1}^N (\lambda_l + \sigma_l) + \mu_j(z) \right] G_j(z_1, \dots, z_N, m_1, \dots, m_N, z) \\
& + \sum_{l=1}^N \lambda_l G_j(z_1, \dots, z_N, m_1, \dots, m_{l-1}, m_l + 1, m_{l+1}, \dots, m_N, z) \\
& + \left( \sum_{l=1}^N \sigma_l z_l \right) G_j(z_1, \dots, z_N, m_1, \dots, m_N, z) \\
& + \mu_j(z) G_0(z_1, \dots, z_N, m_1, \dots, m_N) = 0, \quad (j = 1, \dots, N).
\end{aligned} \tag{4.2}$$

Now we have established the equations (4.2) satisfied by the generating functions of the residual number of retrials, and we aim at solving  $G_0(z_1, \dots, z_N, m_1, \dots, m_N)$  and  $G_j(z_1, \dots, z_N, m_1, \dots, m_N, z)$  ( $j = 1, \dots, N$ ) from it. By similar techniques used in the proof of Theorem 3.1, we make some transformations for  $\sigma$  and  $m_j$  ( $j = 1, \dots, N$ ) in (4.2). To enhance readability of the following analysis process, Table 1 provides the equivalence between the notations in both columns in advance.

Let

$$\begin{aligned}
S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon) &= G_0(z_1, \dots, z_N, m_1, \dots, m_N), \\
S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) &= G_j(z_1, \dots, z_N, m_1, \dots, m_N, z), \quad (j = 1, \dots, N).
\end{aligned}$$

Rewrite (4.2) as

$$\begin{aligned}
 & - \sum_{j=1}^N (\lambda_j + x_j \gamma_j) S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon) + \sum_{j=1}^N \lambda_j S_j(z_1, \dots, z_N, x_1, \dots, x_N, 0, \epsilon) \\
 & + \sum_{j=1}^N (x_j - \epsilon) \gamma_j S_j(z_1, \dots, z_N, x_1, \dots, x_{j-1}, x_j - \epsilon, x_{j+1}, \dots, x_N, 0, \epsilon) \\
 & + \sum_{j=1}^N \epsilon \gamma_j z_j S_j(z_1, \dots, z_{j-1}, 1, z_{j+1}, \dots, z_N, x_1, \dots, x_{j-1}, x_j - \epsilon, x_{j+1}, \dots, x_N, 0, \epsilon) \\
 & = 0, \\
 & \partial S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) / \partial z \\
 & - \left[ \sum_{l=1}^N (\lambda_l + \epsilon \gamma_l) + \mu_j(z) \right] S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) \\
 & + \sum_{l=1}^N \lambda_l S_j(z_1, \dots, z_N, x_1, \dots, x_{l-1}, x_l + \epsilon, x_{l+1}, \dots, x_N, z, \epsilon) \\
 & + \left( \sum_{l=1}^N \epsilon \gamma_l z_l \right) S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) \\
 & + \mu_j(z) S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon) \\
 & = 0, \tag{4.3} \\
 & \hspace{15em} (j = 1, \dots, N).
 \end{aligned}$$

Recalling Theorem 3.1 and Remark 3.2, equations (4.3) still hold after replacing  $x_j$  with  $\kappa_j$ , ( $j = 1, \dots, N$ ). To deal with (4.3) with  $\sigma \rightarrow 0$ , we conduct analysis in two steps. At first, we derive the relationship shown in (4.8). It means, when the server is in different states, the conditional generating functions of the remaining number of retrials are same. Then using this result, Taylor expansion, as well as the first-order asymptotics of  $S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon)$  and  $S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon)$  ( $j = 1, \dots, N$ ), we get equations (4.11). According to Remark 3.2, after some calculations, we obtain the expressions for  $G_0(z_1, \dots, z_N, m_1, \dots, m_N)$  and  $G_j(z_1, \dots, z_N, m_1, \dots, m_N, z)$  ( $j = 1, \dots, N$ ).

More details are presented now.

**Step 1.** Let

$$\begin{aligned}
 \lim_{\epsilon \rightarrow 0} S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon) &= K_0(z_1, \dots, z_N, x_1, \dots, x_N), \\
 \lim_{\epsilon \rightarrow 0} S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) &= K_j(z_1, \dots, z_N, x_1, \dots, x_N, z), \quad (j = 1, \dots, N),
 \end{aligned}$$

then from equations (4.3), we get

$$\begin{aligned}
 & - \sum_{j=1}^N (\lambda_j + x_j \gamma_j) K_0(z_1, \dots, z_N, x_1, \dots, x_N) + \sum_{j=1}^N \lambda_j K_j(z_1, \dots, z_N, x_1, \dots, x_N, 0) \\
 & + \sum_{j=1}^N x_j \gamma_j K_j(z_1, \dots, z_N, x_1, \dots, x_N, 0) = 0, \tag{4.4} \\
 & \frac{\partial K_j(z_1, \dots, z_N, x_1, \dots, x_N, z)}{\partial z} - \mu_j(z) K_j(z_1, \dots, z_N, x_1, \dots, x_N, z) \\
 & + \mu_j(z) K_0(z_1, \dots, z_N, x_1, \dots, x_N) = 0, \quad (j = 1, \dots, N).
 \end{aligned}$$

For  $j = 1, \dots, N$ , solving the second equation in (4.4), we get

$$\begin{aligned} K_j(z_1, \dots, z_N, x_1, \dots, x_N, z) e^{-\int_0^z \mu_j(v) dv} \\ = K_j(z_1, \dots, z_N, x_1, \dots, x_N, 0) - K_0(z_1, \dots, z_N, x_1, \dots, x_N) \int_0^z e^{-\int_0^v \mu_j(y) dy} \mu_j(v) dv. \end{aligned} \quad (4.5)$$

If  $z \rightarrow \infty$ , then  $\exp(\int_0^z \mu_j(v) dv) \rightarrow \infty$ , which yields the component on the right-hand side of (4.5) converges to zero, *i.e.*,

$$\begin{aligned} K_j(z_1, \dots, z_N, x_1, \dots, x_N, 0) &= K_0(z_1, \dots, z_N, x_1, \dots, x_N) \int_0^\infty e^{-\int_0^v \mu_j(y) dy} \mu_j(v) dv \\ &= K_0(z_1, \dots, z_N, x_1, \dots, x_N), \quad (j = 1, \dots, N). \end{aligned} \quad (4.6)$$

Combining (4.5) with (4.6), we obtain

$$K_j(z_1, \dots, z_N, x_1, \dots, x_N, z) = K_0(z_1, \dots, z_N, x_1, \dots, x_N), \quad (j = 1, \dots, N). \quad (4.7)$$

For convenience, we denote

$$\begin{aligned} K(z_1, \dots, z_N, x_1, \dots, x_N) &= K_0(z_1, \dots, z_N, x_1, \dots, x_N), \\ &= K_j(z_1, \dots, z_N, x_1, \dots, x_N, z), \quad (j = 1, \dots, N). \end{aligned} \quad (4.8)$$

**Remark 4.1.** The relationship in (4.8) illustrates that the generating function of the joint distribution of the residual number of retrials is neither relevant to the server's states, nor to the elapsed service time of the customer being served.

**Step 2.** Applying Taylor formula, we have

$$\begin{aligned} S_j(z_1, \dots, z_N, x_1, \dots, x_{j-1}, x_j \pm \epsilon, x_{j+1}, \dots, x_N, z, \epsilon) \\ = S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) \pm \epsilon \frac{\partial S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon)}{\partial x_j} + o(\epsilon). \end{aligned} \quad (4.9)$$

On the basis of (4.8), we provide the solution to (4.3) in the form of first-order asymptotics, that is,

$$\begin{aligned} S_0(z_1, \dots, z_N, x_1, \dots, x_N, \epsilon) &= K(z_1, \dots, z_N, x_1, \dots, x_N) + \epsilon s_0(z_1, \dots, z_N, x_1, \dots, x_N) + o(\epsilon), \\ S_j(z_1, \dots, z_N, x_1, \dots, x_N, z, \epsilon) &= K(z_1, \dots, z_N, x_1, \dots, x_N) + \epsilon s_j(z_1, \dots, z_N, x_1, \dots, x_N, z) + o(\epsilon), \end{aligned} \quad (4.10)$$

in which the lowercase  $s_0(\cdot)$  and  $s_j(\cdot)$  ( $j = 1, \dots, N$ ) on the right-hand side denote the first-order partial derivatives, respectively. Plug (4.9) and (4.10) into (4.3), then we obtain

$$\begin{aligned} - \sum_{j=1}^N (\lambda_j + x_j \gamma_j) s_0(z_1, \dots, z_N, x_1, \dots, x_N) + \sum_{j=1}^N (\lambda_j + x_j \gamma_j) s_j(z_1, \dots, z_N, x_1, \dots, x_N, 0) \\ - \sum_{j=1}^N x_j \gamma_j \frac{\partial K(z_1, \dots, z_N, x_1, \dots, x_N)}{\partial x_j} - \sum_{j=1}^N \gamma_j K(z_1, \dots, z_N, x_1, \dots, x_N) \\ + \sum_{j=1}^N \gamma_j z_j K(z_1, \dots, z_{j-1}, 1, z_{j+1}, \dots, z_N, x_1, \dots, x_N) = 0, \end{aligned} \quad (4.11)$$

$$\begin{aligned} \mu_j(z) s_0(z_1, \dots, z_N, x_1, \dots, x_N) - \mu_j(z) s_j(z_1, \dots, z_N, x_1, \dots, x_N, z) \\ + \frac{\partial s_j(z_1, \dots, z_N, x_1, \dots, x_N, z)}{\partial z} + \sum_{l=1}^N \lambda_l \frac{\partial K(z_1, \dots, z_N, x_1, \dots, x_N)}{\partial x_l} \\ + \sum_{l=1}^N \gamma_l (z_l - 1) K(z_1, \dots, z_N, x_1, \dots, x_N) = 0, \quad (j = 1, \dots, N). \end{aligned}$$

In (4.11), multiply both sides of the first equation by  $D_0$  and the other equations by  $D_j(z)$ , respectively. Then integrate these equations for  $z$  over the interval from 0 to  $\infty$ , and sum  $j$  from 1 to  $N$ . Finally, add up the resulting equations together. We have

$$\begin{aligned}
 & \left[ \sum_{j=1}^N \int_0^\infty \mu_j(z) D_j(z) dz - \sum_{j=1}^N (\lambda_j + x_j \gamma_j) D_0 \right] s_0(z_1, \dots, z_N, x_1, \dots, x_N) \\
 & + \sum_{j=1}^N (\lambda_j + x_j \gamma_j) D_0 s_j(z_1, \dots, z_N, x_1, \dots, x_N, 0) \\
 & - \sum_{j=1}^N \int_0^\infty \mu_j(z) D_j(z) s_j(z_1, \dots, z_N, x_1, \dots, x_N, z) dz \\
 & + \sum_{j=1}^N \int_0^\infty D_j(z) \frac{\partial s_j(z_1, \dots, z_N, x_1, \dots, x_N, z)}{\partial z} dz \\
 & + \left\{ \sum_{j=1}^N \left[ \lambda_j \sum_{l=1}^N \int_0^\infty D_l(z) dz - x_j \gamma_j D_0 \right] \right\} \frac{\partial K(z_1, \dots, z_N, x_1, \dots, x_N)}{\partial x_j} \\
 & + \left\{ \left[ \sum_{i=1}^N \gamma_i (z_i - 1) \right] \sum_{j=1}^N \int_0^\infty D_j(z) dz - D_0 \sum_{j=1}^N \gamma_j \right\} K(z_1, \dots, z_N, x_1, \dots, x_N) \\
 & + \sum_{j=1}^N \gamma_j z_j D_0 K(z_1, \dots, z_{j-1}, 1, z_{j+1}, \dots, z_N, x_1, \dots, x_N) = 0.
 \end{aligned} \tag{4.12}$$

**Remark 4.2.** According to Remark 3.2,  $x_j$  can be substituted by  $\kappa_j$  in (4.3) for  $j = 1, \dots, N$ . Since equation (4.12) is derived from (4.3), it also works to substitute  $x_j$  with  $\kappa_j$  in (4.12).

Following Remark 4.2, from the first equation in (3.9) as well as the equation (3.11), we find that the coefficients of  $s_0(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N)$  and that of  $\partial K(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N) / \partial x_j$  are both zero. Furthermore, applying integration by parts and combining the rest equations in (3.9), we have

$$\begin{aligned}
 & \left[ \sum_{j=1}^N (\lambda_j + x_j \gamma_j) D_0 s_j(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N, 0) \right. \\
 & \quad \left. - \sum_{j=1}^N \int_0^\infty \mu_j(z) D_j(z) s_j(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N, z) dz \right] \\
 & + \sum_{j=1}^N \int_0^\infty D_j(z) \frac{\partial s_j(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N, z)}{\partial z} dz = 0.
 \end{aligned}$$

Denote  $H(z_1, \dots, z_N) = K(z_1, \dots, z_N, \kappa_1, \dots, \kappa_N)$  and consider  $D_0 + \sum_{j=1}^N D_j = 1$ . Equation (4.12) is equivalent to

$$\begin{aligned}
 & \left\{ (1 - D_0) \sum_{i=1}^N \gamma_i (z_i - 1) - D_0 \sum_{i=1}^N \gamma_i \right\} H(z_1, \dots, z_N) \\
 & + D_0 \sum_{i=1}^N \gamma_i z_i H(z_1, \dots, z_{i-1}, 1, z_{i+1}, \dots, z_N) = 0.
 \end{aligned} \tag{4.13}$$

For any  $j_1 \in \{1, \dots, N\}$ , from (4.13) we derive that

$$H(1, \dots, 1, z_{j_1}, 1, \dots, 1) = \frac{z_{j_1} D_0}{1 - z_{j_1} (1 - D_0)}.$$

In addition, for any  $j_1, j_2 \in \{1, \dots, N\}$  and  $j_1 \neq j_2$ , we obtain

$$H(1, \dots, 1, z_{j_1}, 1, \dots, 1, z_{j_2}, 1, \dots, 1) = \frac{z_{j_1} D_0}{1 - z_{j_1} (1 - D_0)} \frac{z_{j_2} D_0}{1 - z_{j_2} (1 - D_0)}.$$

From above process, with condition  $\sigma \rightarrow 0$ , we have the generating function of the residual number of retrials for single-class tagged customer, as well as the joint generating function for two-class tagged customers. In this way, we can successively obtain the joint generating function of the residual number of retrials for  $j$  ( $j = 3, \dots, N$ ) classes of tagged customers. We omit repetitive steps and show the general expression as

$$H(z_1, \dots, z_N) = \frac{z_1 D_0}{1 - z_1 (1 - D_0)} \frac{z_2 D_0}{1 - z_2 (1 - D_0)} \cdots \frac{z_N D_0}{1 - z_N (1 - D_0)}, \quad (4.14)$$

from which, we get the following theorem.

**Theorem 4.3.** *For  $N$  tagged customers of different types, the joint asymptotic distribution of the residual number of retrials is*

$$\lim_{\sigma \rightarrow 0} P\left(R_{\text{res}}^{(1)} = r_1, R_{\text{res}}^{(2)} = r_2, \dots, R_{\text{res}}^{(N)} = r_N\right) = D_0^N (1 - D_0)^{\sum_{j=1}^N r_j - N}, \quad (4.15)$$

where  $j = 1, \dots, N$  and  $r_j = 1, 2, 3, \dots$ .

## 4.2. Asymptotic distribution for the number of retrials

In a stable system, if a primary type  $j$  ( $j = 1, \dots, N$ ) arrival finds the server idle with probability  $D_0$ , it accepts service immediately. In this case, the number of retrials is zero. However, if the primary arrival finds the server busy with probability  $1 - D_0$ , the arrival customer joins the orbit  $j$  immediately. For this case, the number of retrials  $R_j$  is equal to the residual number of retrials  $R_{\text{res}}^{(j)}$ . Therefore, from (4.15), we obtain the following theorem.

**Theorem 4.4.** *The joint asymptotic distribution of the number of retrials for  $N$  types of customers joining the orbit is*

$$\lim_{\sigma \rightarrow 0} P(R_1 = r_1, R_2 = r_2, \dots, R_N = r_N) = D_0^N (1 - D_0)^{\sum_{j=1}^N r_j - N},$$

where  $j = 1, \dots, N$  and  $r_j = 1, 2, 3, \dots$ .

**Remark 4.5.** Theorem 4.4 indicates that as  $\sigma$  goes to zero, the joint asymptotic distribution for the number of retrials of  $N$  types of customers joining the orbit follows an  $N$ -dimensional geometric distribution. When  $\sigma$  is sufficiently small, the number of retrials of different types of customers are independent of each other.

## 5. ASYMPTOTIC DISTRIBUTION FOR THE WAITING TIME

In previous section, the asymptotic distribution of the residual number of retrials is obtained. In this section, we utilize the relationships among residual waiting time  $W_{\text{res}}^{(j)}$ , residual number of retrials  $R_{\text{res}}^{(j)}$  and retrial interval to derive the asymptotic distribution of the residual waiting time, then further determine the asymptotic distribution of the waiting time.

Recall that the interval between two successive retrials follows an exponential distribution. Moreover, Theorem 4.4 shows that the number of retrials  $(R_1, \dots, R_N)$  follows a multivariate geometric distribution when  $\sigma$  approaches zero. Obviously,

$$W_{\text{res}}^{(j)}(t) = \tau_1^{(j)} + \tau_2^{(j)} + \dots + \tau_{R_{\text{res}}^{(j)}(t)}^{(j)}, \tag{5.1}$$

where  $\tau_k^{(j)}$  ( $k = 1, \dots, R_{\text{res}}^{(j)}(t)$ ) denotes the  $k$ th retrial interval of the tagged customer with type  $j$ . Therefore, due to the memoryless property of exponential distribution and geometric distribution,  $(W_{\text{res}}^{(1)} \leq t_1, W_{\text{res}}^{(2)} \leq t_2, \dots, W_{\text{res}}^{(N)} \leq t_N | R_{\text{res}}^{(1)} = r_1, R_{\text{res}}^{(2)} = r_2, \dots, R_{\text{res}}^{(N)} = r_N)$  follows an Erlang distribution. Furthermore, adopting the Law of Total Expectation, (4.15) and (5.1), we can get the following theorem.

**Theorem 5.1.** *When  $\sigma$  is small enough, the Laplace transform of the joint stationary distribution of the residual waiting time is given by*

$$E\left\{e^{-\alpha_1 W_{\text{res}}^{(1)} - \alpha_2 W_{\text{res}}^{(2)} - \dots - \alpha_N W_{\text{res}}^{(N)}}\right\} \approx \frac{\prod_{l=1}^N D_0 \sigma_l}{\prod_{l=1}^N (\alpha_l + D_0 \sigma_l)}.$$

**Remark 5.2.** This expression reveals that, if  $\sigma$  is small enough, the joint stationary distribution of  $(W_{\text{res}}^{(1)}, \dots, W_{\text{res}}^{(N)})$  can be approximated by an  $N$ -dimensional exponential distribution with parameter  $D_0 \sigma_l$  ( $l = 1, 2, \dots, N$ ).

Notice that

$$W_j = \begin{cases} 0, & R_j = 0, \\ \tau_1^{(j)} + \tau_2^{(j)} + \dots + \tau_{R_j}^{(j)}, & R_j > 0. \end{cases}$$

This means when the number of retrials is zero, the arrival customer receives service immediately and does not need to join the orbit. Conversely, when the number of retrials is positive, the customer must be waiting in the orbit. Following the analysis in Section 4.2, when the system is stable, the distribution of the number of retrials is identical to that of the residual number of retrials. Thus, the waiting time of customer who joins the orbit also has the same distribution as the residual waiting time.

Overall, we finally have the main result of this work.

**Theorem 5.3.** *When  $\sigma$  is small enough, for any  $t_j > 0$  ( $j = 1, 2, \dots, N$ ), we have*

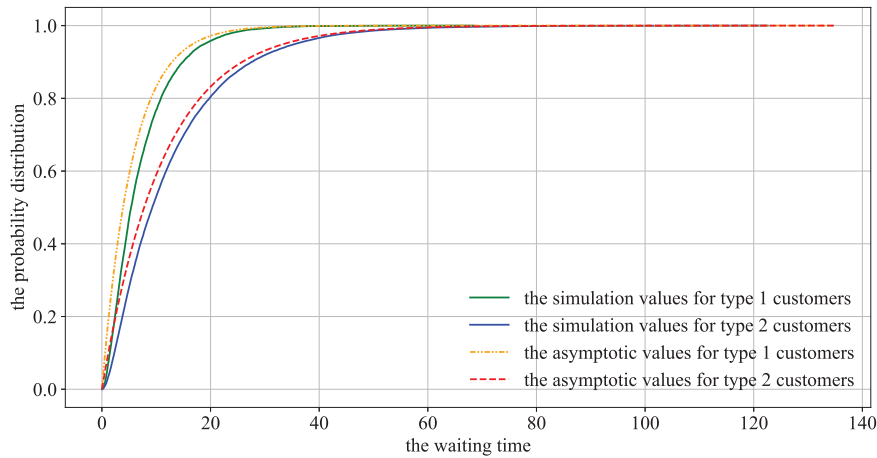
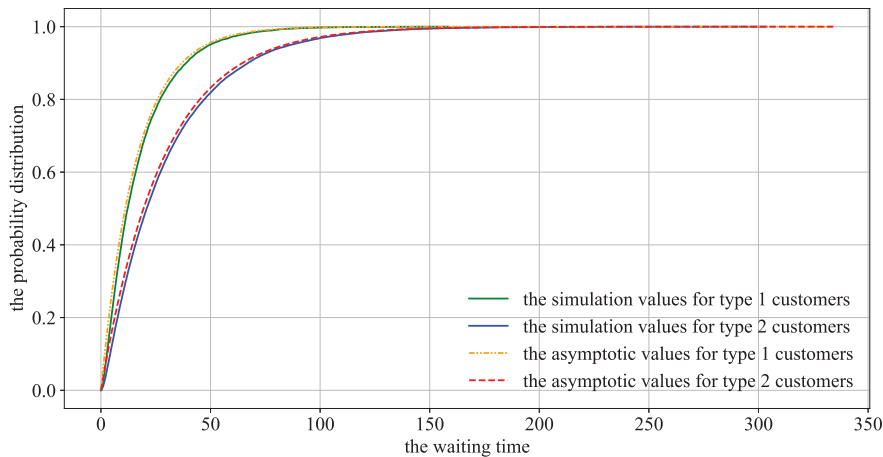
$$P(W_1 \leq t_1, W_2 \leq t_2, \dots, W_N \leq t_N) \approx \prod_{l=1}^N \left[1 - e^{-(1 - \sum_{j=1}^N D_j) \sigma_l t_l}\right].$$

**Remark 5.4.** The above theorem also implies if  $\sigma$  is small enough, the waiting times of different types of customers are independent of each other.

## 6. EXPERIMENTS

In this section, we present numerical experiments to verify the correctness of our asymptotic result for the waiting time distribution. Given the parameters of the model, by numerical simulation, we obtain the stationary distributions of waiting time for various types of customers. These simulation values are used to compare with the asymptotic results provided by Theorem 5.3. The comparison shows they match well. Moreover, we study how parameters (such as arrival rate and service rate) affect the asymptotic approximations.

Consider a special case where  $N = 2$ . For  $j = 1, 2$ , suppose the service time of type  $j$  customers follows an exponential distribution  $F_j(t) = 1 - e^{-\mu_j t}$ . Let  $(\lambda_1, \lambda_2) = (0.02, 0.06)$  and  $(\mu_1, \mu_2) = (0.9, 0.7)$ . We simulate

FIGURE 2. Waiting time distribution for  $(\sigma_1, \sigma_2) = (0.2, 0.1)$ .FIGURE 3. Waiting time distribution for  $(\sigma_1, \sigma_2) = (0.07, 0.04)$ .

the process of customers going through the system, including arrival, service, retrial and departure. To ensure the accuracy, the simulation duration is set to 5 000 000 units of time. Since the queueing system is stable, we record the waiting times of all customers in each orbit and plot the stationary distribution curves.

If  $(\sigma_1, \sigma_2)$  takes different values, for example,  $(\sigma_1, \sigma_2) = (0.2, 0.1)$ ,  $(\sigma_1, \sigma_2) = (0.07, 0.04)$ ,  $(\sigma_1, \sigma_2) = (0.007, 0.004)$ , the curves of stationary distribution for waiting time, including both simulation and asymptotic curves, are presented in Figures 2–4. Obviously, for  $j = 1$  or  $j = 2$ , the smaller the retrial rate is, the closer the two curves are.

For clarity, let  $\xi_j$  ( $j = 1, 2$ ) represent the maximum absolute difference between the simulated values and the asymptotic results. Table 2 presents the values of  $\xi_j$  ( $j = 1, 2$ ) for different retrial rate pairs. Assume an acceptable error is 0.05, then the asymptotic results are credible when the retrial rate is no more than 0.04.

To verify the correctness of asymptotic results for large values of  $N$ , simulation experiments are conducted for  $N \in \{3, 5, 7, 9\}$ . Given a vector  $x = (x_1, \dots, x_n)$ , where  $n \geq N$ , let  $x_{(N)} = (x_1, \dots, x_N)$  be a truncated vector of  $x$ . For  $n = 9$ , denote  $\lambda = (\lambda_1, \dots, \lambda_9) = (0.02, 0.03, 0.01, 0.04, 0.02, 0.03, 0.01, 0.02, 0.05)$ ,  $\mu = (\mu_1, \dots, \mu_9) =$

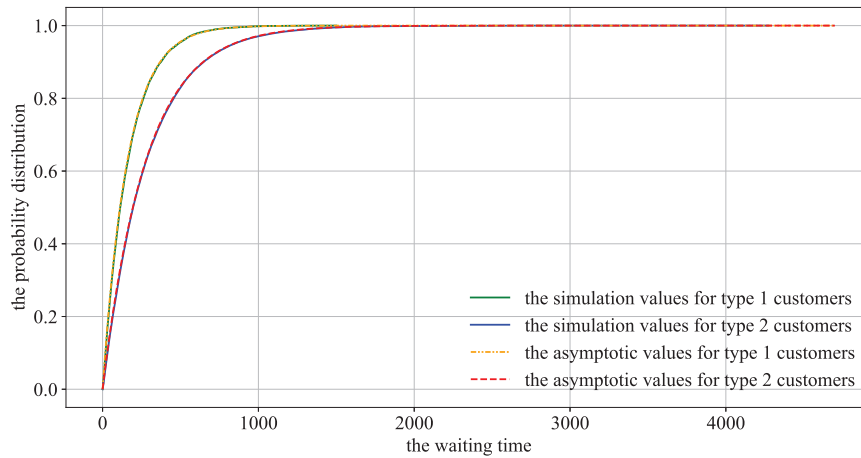


FIGURE 4. Waiting time distribution for  $(\sigma_1, \sigma_2) = (0.007, 0.004)$ .

TABLE 2. The maximum error between simulation and asymptotics.

$(\sigma_1, \sigma_2)$	(0.2, 0.1)	(0.1, 0.08)	(0.08, 0.06)	(0.06, 0.04)	(0.04, 0.02)	(0.02, 0.008)
$\xi_1$	0.1503	0.0929	0.0797	0.0573	0.0369	0.0232
$\xi_2$	0.0880	0.0737	0.0585	0.0407	0.0243	0.0103

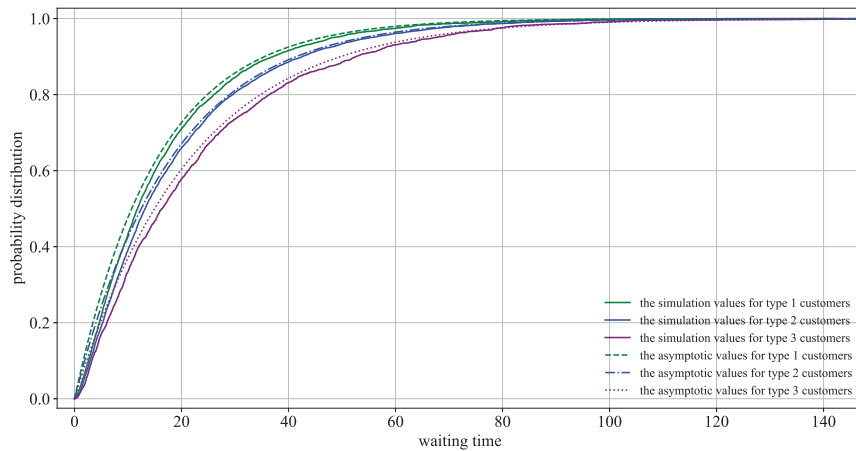
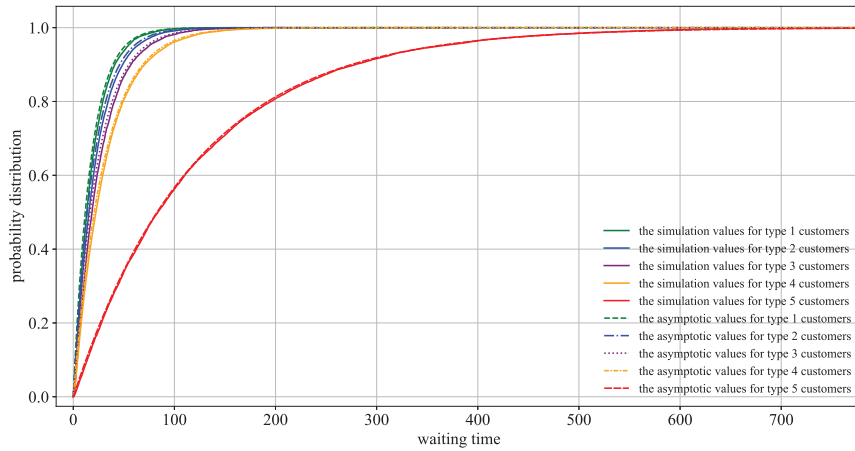
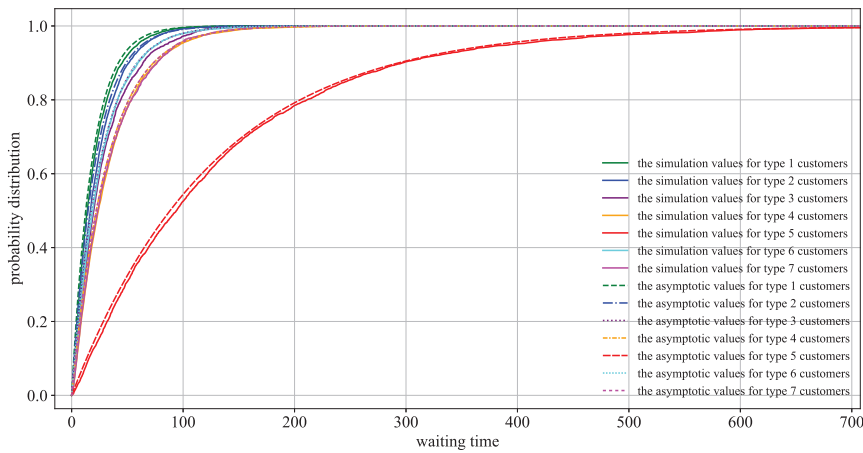


FIGURE 5. Waiting time distribution for  $N = 3$ .

$(0.9, 0.8, 0.7, 0.6, 0.9, 0.8, 0.7, 0.4, 0.6)$ , and  $\sigma = (\sigma_1, \dots, \sigma_9) = (0.07, 0.06, 0.05, 0.04, 0.01, 0.05, 0.04, 0.05, 0.03)$ . When  $N$  takes values in  $\{3, 5, 7, 9\}$ , let arrival rate, service rate and retrial rate be  $\lambda_{(N)}$ ,  $\mu_{(N)}$ , and  $\sigma_{(N)}$ , respectively. All simulation results and asymptotic results are shown in Figures 5–8.

Next, we take different service time distributions into consideration. For  $N = 3$ , if the service time follows Erlang or hyperexponential distribution, we verify the asymptotic results in our work remain reliable. Both experiments use the same arrival rate  $(\lambda_1, \lambda_2, \lambda_3) = (0.02, 0.06, 0.03)$  and retrial rate  $(\sigma_1, \sigma_2, \sigma_3) = (0.04, 0.01, 0.06)$ . For a type  $j$  ( $j = 1, 2, 3$ ) customer, if its service time follows Erlang distribution  $G_j(x) =$

FIGURE 6. Waiting time distribution for  $N = 5$ .FIGURE 7. Waiting time distribution for  $N = 7$ .

$1 - e^{-\mu_j x} \sum_{i=0}^{k_j-1} \frac{(\mu_j x)^i}{i!}$ , where  $x \geq 0$ , set  $k_1 = k_2 = k_3 = 2$ ,  $\mu_1 = 0.5$ ,  $\mu_2 = 0.7$ ,  $\mu_3 = 0.8$ . If the type  $j$  ( $j = 1, 2, 3$ ) customer has service time following hyperexponential distribution  $H_j(x) = 1 - \sum_{i=1}^{k_j} p_{ji} e^{-\mu_{ji} x}$ , where  $x \geq 0$ , set  $k_1 = k_2 = k_3 = 2$ ,  $\mu_{11} = 0.7$ ,  $\mu_{12} = 0.8$ ,  $\mu_{21} = 0.6$ ,  $\mu_{22} = 0.7$ ,  $\mu_{31} = 0.5$ ,  $\mu_{32} = 0.6$ . The corresponding results are presented in Figures 9 and 10.

Now, we study how the arrival rate and service rate affect the asymptotic approximation. Consider  $N = 3$  and take the fixed parameters  $(\sigma_1, \sigma_2, \sigma_3) = (0.08, 0.03, 0.06)$ . For simplicity, we suppose  $\lambda_1 = \lambda_2 = \lambda_3$  and  $\mu_1 = \mu_2 = \mu_3 = 0.7$  for different type  $j$  ( $j = 1, 2, 3$ ) customer. With the arrival rate increasing, the average errors  $\bar{\xi}_j$  between the asymptotic solutions and simulation results are calculated and shown in Table 3. Also, suppose  $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$  and  $\mu_1 = \mu_2 = \mu_3$ , with the service rate increasing, the values of  $\bar{\xi}_j$  are shown in Table 4. It can be observed that the average error decreases with the increases of the arrival rate or service rate. The explanation is that, for fixed retrial rate, when arrival rate or service rate increases gradually, the retrial rate gradually becomes smaller relative to them. This leads to a decreasing error between the asymptotics and the simulation.

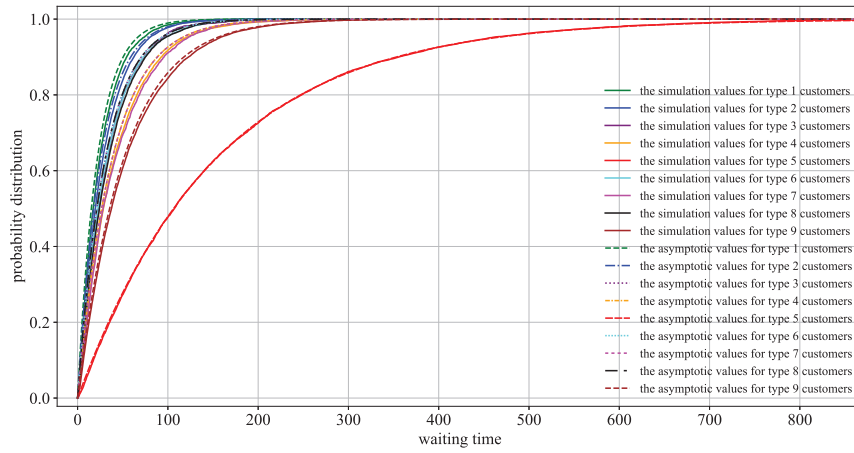


FIGURE 8. Waiting time distribution for  $N = 9$ .

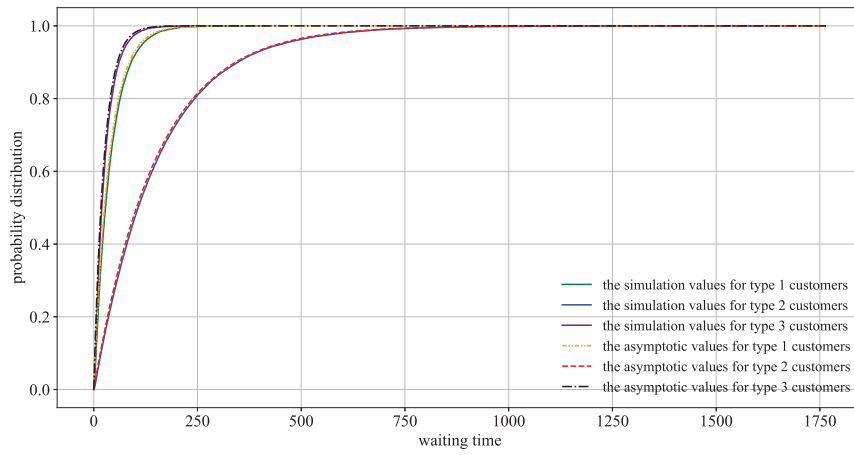


FIGURE 9. Service time follows Erlang distribution.

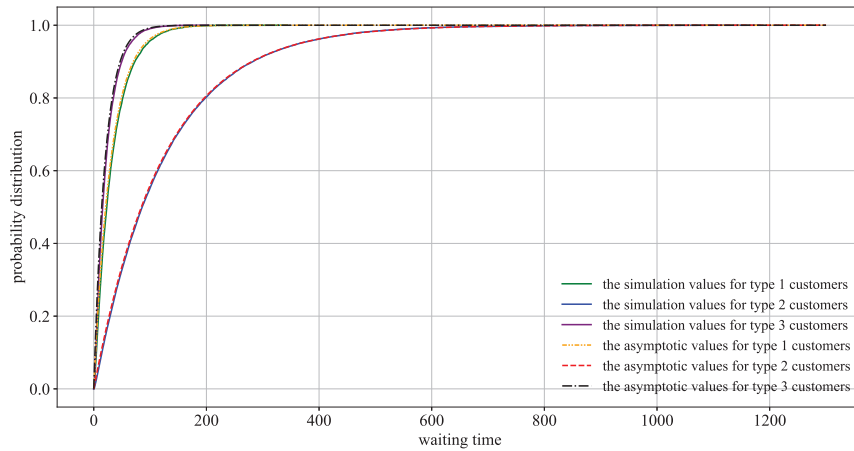


FIGURE 10. Service time follows hyperexponential distribution.

TABLE 3. The average errors with different arrival rates.

$\lambda_j$ ( $j = 1, 2, 3$ )	0.03	0.06	0.09	0.12	0.15	0.18
$\bar{\xi}_1$	0.0444	0.0401	0.0327	0.0279	0.0229	0.0176
$\bar{\xi}_2$	0.0202	0.0173	0.0163	0.0149	0.0130	0.0122
$\bar{\xi}_3$	0.0360	0.0322	0.0270	0.0233	0.0186	0.0155

TABLE 4. The average errors with different service rates.

$\mu_j$ ( $j = 1, 2, 3$ )	0.5	0.8	1.1	1.4	1.7	2
$\bar{\xi}_1$	0.0319	0.0306	0.0252	0.0213	0.0204	0.0189
$\bar{\xi}_2$	0.0186	0.0148	0.0120	0.0096	0.0078	0.006
$\bar{\xi}_3$	0.0277	0.0234	0.0218	0.0186	0.0162	0.0137

## 7. CONCLUSION

In this work, we analyze a single-server queueing system with  $N$  types of customers and  $N$  retrial orbits, with the assumption that the retrial rate of each type of customers linearly converges to zero. We obtain the following asymptotic results in turn: (a) first-order asymptotic property of the orbit queue length, *i.e.*, Theorem 3.1; (b) joint asymptotic distribution of the residual number of retrials, *i.e.*, Theorem 4.3; (c) joint asymptotic distribution of the total number of retrials, *i.e.*, Theorem 4.4; (d) joint asymptotic distribution of the residual waiting time, *i.e.*, Theorem 5.1; (e) joint asymptotic distribution of the total waiting time, *i.e.*, Theorem 5.3.

Among these results, the first-order asymptotic property of orbit queue length indicates each queue length weakly converges to a specific value, which is the foundation of the sequence analysis. Also, it is worth mentioning that in steady state, the residual number of retrials and the total number of retrials are identically distributed. Just for this reason, the asymptotic distribution of the total number of retrials can be directly derived from that of the residual number of retrials. Similarly, the residual waiting time and the total waiting time of any customer in the orbit are identically distributed. So, taking advantage of the relationships among the residual waiting time, the residual number of retrials, and the retrial interval, we have the asymptotic distribution of the residual waiting time. At the same time, the asymptotic distribution of the total waiting time is finally obtained.

At the end of the work, we conduct several numerical experiments to demonstrate the correctness of our conclusions, no matter for large values of  $N$ , or for different service time distributions.

### FUNDING

This work was supported by the National Natural Science Foundation of China (No.12201296).

### DATA AVAILABILITY STATEMENT

The research data associated with this article are included in the article.

### REFERENCES

- [1] G. Falin and J. Templeton, *Retrial Queues*. Chapman, Hall, London (1997).
- [2] J.R. Artalejo and A. Gómez-Corral, *Retrial Queueing Systems: A Computational Approach*. Springer, Berlin (2008).
- [3] K. Avrachenkov, E. Morozov, R. Nekrasova and B. Steyaert, Stability analysis and simulation of N-class retrial system with constant retrial rates and Poisson inputs. *Asia-Pac. J. Oper. Res.* **31** (2014) 1440002.

- [4] K. Avrachenkov, E. Morozov and B. Steyaert, Sufficient stability conditions for multi-class constant retrial rate systems. *Queueing Syst.* **82** (2016) 149–171.
- [5] E. Morozov, A. Rumyantsev, S. Dey and T.G. Deepak, Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking. *Perform. Eval.* **134** (2019) 102005.
- [6] E. Morozov and T. Phung-Duc, Stability analysis of a multiclass retrial system with classical retrial policy. *Perform. Eval.* **112** (2017) 15–26.
- [7] E. Morozov and S. Rogozin, Stability condition of multiclass classical retrials: a revised regenerative proof. *Ann. Math. Inf.* **56** (2022) 71–83.
- [8] R. Nekrasova, Stability analysis of a multi-class retrial queue with general retrials and classical retrial policy, in 2021 28th Conference of Open Innovations Association (FRUCT). IEEE (2021) 328–333.
- [9] A. Nazarov, T. Phung-Duc and Y. Izmailova, Multidimensional central limit theorem of the multiclass  $M/M/1/1$  retrial queue, in Distributed Computer and Communication Networks, edited by V.M. Vishnevskiy, K.E. Samouylov and D.V. Kozyrev. Lecture Notes in Computer Science (LNCCN). Vol. 12563. Springer International Publishing, Cham (2020) 298–310.
- [10] Y.W. Shin and D.H. Moon,  $M/M/c$  retrial queue with multiclass of customers. *Methodol. Comput. Appl. Probab.* **16** (2014) 931–949.
- [11] K. Avrachenkov, P. Nain and U. Yechiali, A retrial system with two input streams and two orbit queues. *Queueing Syst.* **77** (2014) 1–31.
- [12] Y. Song, Z. Liu and Y.Q. Zhao, Exact tail asymptotics: revisit of a retrial queue with two input streams and two orbits. *Ann. Oper. Res.* **247** (2016) 97–120.
- [13] I. Dimitriou, A two-class retrial system with coupled orbit queues. *Probab. Eng. Inf. Sci.* **31** (2017) 139–179.
- [14] S.S. Sanga and M. Jain, Fuzzy modeling of single server double orbit retrial queue. *J. Ambient Intell. Human. Comput.* **13** (2022) 4223–4234.
- [15] G. Falin and C. Fricke, On the virtual waiting time in an  $M/G/1$  retrial queue. *J. Appl. Probab.* **28** (1991) 446–460.
- [16] G.I. Falin, On the waiting time in a single-channel queueing system with secondary calls. *Moscow Univ. Comput. Math. Cybern.* **4** (1977) 83–87.
- [17] J.R. Artalejo, G.I. Falin and M.J. Lopez-Herrero, A second order analysis of the waiting time in the  $M/G/1$  retrial queue. *Asia-Pac. J. Oper. Res.* **19** (2002) 131–148.
- [18] R.D. Nobel and H.C. Tijms, Waiting-time probabilities in the  $M/G/1$  retrial queue. *Statistica Neerlandica* **60** (2006) 73–78.
- [19] J. Kim, J. Kim and B. Kim, Regularly varying tail of the waiting time distribution in  $M/G/1$  retrial queue. *Queue. Syst.* **65** (2010) 365–383.
- [20] E. Sudyko, A.A. Nazarov and J. Sztrik, Asymptotic waiting time analysis of a finite-source  $M/M/1$  retrial queueing system. *Probab. Eng. Inf. Sci.* **33** (2018) 387–403.
- [21] A. Nazarov, J. Sztrik, A. Kvach and A. Tóth, Asymptotic sojourn time analysis of finite-source  $M/M/1$  retrial queueing system with collisions and server subject to breakdowns and repairs. *Ann. Oper. Res.* **288** (2020) 417–434.
- [22] A.A. Nazarov, J. Sztrik, Jr. and A. Kvach, Asymptotic waiting time analysis of finite source  $M/GI/1$  retrial queueing systems with conflicts and unreliable server. *Publicationes Mathematicae Debrecen* **101** (2022) 397–419.

**Please help to maintain this journal in open access!**



This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting [subscribers@edpsciences.org](mailto:subscribers@edpsciences.org).

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.