mathematics

MDPI

Article

Diffusion Limit of Multi-Server Retrial Queue with Setup Time

Anatoly Nazarov [1], Alexander Moiseev [1], Tuan Phung-Duc [2,3,*] and Svetlana Paul [1]

[1] Institute of Applied Mathematics and Computer Science, Tomsk State University, 36 Lenin Ave., 634050 Tomsk, Russia; nazarov.tsu@gmail.com (A.N.); moiseev.tsu@gmail.com (A.M.); paulsv82@mail.ru (S.P.)

[2] Department of Policy and Planning Sciences, Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

[3] Public Policy Program, VNU Vietnam Japan University, My Dinh Campus, Nam Tu Liem, Hanoi, Vietnam

* Correspondence: tuan@sk.tsukuba.ac.jp

Received: 1 October 2020; Accepted: 14 December 2020; Published: 16 December 2020

Abstract: In the paper, we consider a multi-server retrial queueing system with setup time which is motivated by applications in power-saving data centers with the ON-OFF policy, where an idle server is immediately turned off and an off server is set up upon arrival of a customer. Customers that find all the servers busy join the orbit and retry for service after an exponentially distributed time. For this model, we derive the stability condition which depends on the setup time and turns out to be more strict than that of the corresponding model with an infinite buffer which is independent of the setup time. We propose asymptotic methods to analyze the system under the condition that the delay in the orbit is extremely long. We show that the scaled-number of customers in the orbit converges to a diffusion process. Using this diffusion limit, we obtain approximations for the steady-state probability distribution of the number of busy servers and that of the number of customers in the orbit. We verify the accuracy of the approximations by simulations and numerical analysis. Numerical results show that the retrial system under the limiting condition consumes more energy than that with an infinite buffer in front of the servers.

Keywords: multi-server retrial queue; setup time; two-phased service; asymptotic analysis; long delay condition

1. Introduction

Presently, cloud computing is used by many companies and individuals. In our everyday business, we use many kinds of sharing services such as Dropbox, Slack, Overleaf, etc. which are based on cloud computing. Cloud computing is even more important under the current situation where social distancing is encouraged to combat COVID-19, leading to remote work for a large portion of our society. Cloud computing is supported by data centers in which a huge number of servers are available. These servers consume a huge amount of energy and thus saving-energy is crucial in management of data centers and cloud computing. Since user traffic has peak-on and peak-off nature, it is desired that more server resource is allocated in peak-on period and less resource is allocated in peak-off period. Furthermore, because user traffic stochastically varies, we need control policies that add more resources when the workload is large and release resources when the workload is small. The same mechanism is also observed in 5G networks with network functions virtualization (NFV) [1,2]. Motivated by these situations, many power-saving mechanisms were proposed [3–5].

One of the most natural mechanisms is the ON-OFF policy which turns OFF a server once it has no job to handle and turns on an OFF server again once a job arrives. However, an OFF server

cannot be active immediately to serve the waiting job but needs a setup time, during which the server cannot process the job but consumes energy. Queues with setup time are appropriate models for these situations. A server is turned off when it has no job to process while it is turned on again when waiting jobs are available. These models are challenging because the underlying Markov chain has a non-homogeneous structure, i.e., the number of setup servers may depend on the number of jobs in the system. Gandhi et al. [4] study this problem and propose analysing the model using a renewal reward approach. Phung-Duc [6] analyzes the same model using a generating function approach and a matrix analytic method. Furthermore, structure of the optimal policy was studied by Maccio and Down [5,7].

Furthermore, from the user point of view, the request might be blocked if a resource is not allocated, i.e., all the servers are busy upon arrival. In practice, in case the request is blocked, a protocol will reconnect after some time. This time is random and depends on the number of retries. In some application, the retrial interval is doubled for two consecutive blockings.

Motivated by the above applications, we consider a multiserver retrial queues with setup time. In this model, the server is switched off immediately once it finishes serving a job and no new job is available. An off server will be switched on upon arrival of a job and the server changes to the setup state. After some setup time, the server becomes active so that it can serves the job. If all servers are occupied (serving a job or in setup), the arriving job is blocked and makes a retrial (joins the orbit) after an exponentially distributed time. A retrial job behaves the same as a fresh one. The model was first proposed and numerically analyzed in [8], where a non-trivial sufficient stability condition was derived. Some analytical results for single server case are obtained in [9]. In this paper, we study the model in depth, in which our focus is not a numerical solution but to obtain analytical solution under a specific regime. We consider the situation where the retrial time interval is relatively long. In such a regime, the number of retrial jobs explodes. However, using a proper scaling, we can obtain the explicit distribution for the number of jobs in the orbit. The explicit solution is then used to approximate the distribution of the number of customers in the orbit and that of the states of the servers.

It should be noted that multiserver retrial queues without setup time whose underlying Markov chain is two-dimensional are already very difficult to obtain an analytical solution [10,11]. This paper combines two challenging features, i.e., setup time and retrial. As a result, our model in this paper is formulated by a three-dimensional Markov chain and thus is even more challenging to obtain analytic results. Thus, we consider the system under the slow retrial asymptotic regime.

The slow retrial asymptotic regime was studied by Cohen [12], where the first order asymptotic (a law of large number) for multiserver retrial queue without setup time was obtained in the stationary regime. This result is extended in our paper to the model with setup time in non-stationary regime. Our result states that the scaled number of jobs in the orbit converges to a deterministic process which is characterized by an ordinary differential equation (ODE). Furthermore, we study the second order asymptotics, where the scaled number of jobs in the orbit weakly converges to a diffusion process. The limiting results are then used to obtain the approximation of performance measures. As related works, the second order asymptotic result for multiserver retrial queue without setup was derived in [13] (pp. 55–59). Some recent development of asymptotic analysis of retrial queues can be found in [14–16]. Furthermore, we refer to the books [13,17] for basic results of retrial queueing systems. Difussion limits for queueing models have been extensively studied [18–24]. The methodological difference is that we are based on characteristic function while [18–24] are based on sample path equations or other techniques.

The rest of our paper is organized as follows. Mathematical description of the problem is presented in Section 2. In Section 3, we obtain main equations for the problem solution. The first stage of the asymptotic analysis is made in Section 4, where we derive function $a(x)$ which will be used for further diffusion analysis. Based on the results of the first stage, we obtain the condition for the existence of the steady-state regime for our model in Section 5. The second stage of the asymptotic analysis is made in Section 6, where we derive function $b(x)$ which is used together with $a(x)$ for the asymptotic

diffusion analysis in Section 7. Using obtained continuous distributions, we build approximations for discrete distributions under study in Section 8. In Section 9, we analyze applicability areas of obtained approximations using numerical methods and simulation. Concluding remarks are presented in Section 10.

2. Mathematical Model

We consider a multi-server retrial queue with Poisson arrivals with intensity λ, infinite-capacity orbit and N servers. On the arrival of a customer, if all servers are busy, the customer goes to the orbit where he or she stays for a random time exponentially distributed with parameter σ (mean $1/\sigma$) and then tries to get a service again. If there is a free server, the customer occupies it. Each server serves a customer in two phases: the first one is a setup time which is exponentially distributed with parameter μ_1 and the second one is a real service which is exponentially distributed with parameter μ_2. After the service completed at the second phase, the customer leaves the system.

The service in the system has a specific feature: when one customer completes its service at the second phase and make its server free, and if we have another customer which is served at the first phase at another server, this customer (which was served at the first phase) immediately goes to the server that just becomes free and starts its service at the second phase. Therefore, its total time of servicing becomes less.

Let us denote:

- $n_1(t)$ is the number of servers that are working at the first phase at the instant t;
- $n_2(t)$ is the number of servers that are working at the second phase at the instant t;
- $i(t)$ is the number of customers in the orbit at the instant t;
- $P(n_1, n_2, i, t) = \mathrm{P}\{n_1(t) = n_1, n_2(t) = n_2, i(t) = i\}$ is the joint probability distribution of the stochastic process $\{n_1(t), n_2(t), i(t)\}$.

The goal of the study is to obtain the probability distribution $P(n_1, n_2, i, t) \equiv P(n_1, n_2, i)$ for the process $\{n_1(t), n_2(t), i(t)\}$ in the steady-state regime. The problem is solved using the method of asymptotic diffusion analysis [16,25] under an asymptotic condition of long delay in the orbit: $\sigma \to 0$ (mean retrial interval $1/\sigma \to \infty$).

3. Main Equations

Because the process $\{n_1(t), n_2(t), i(t)\}$ is a three-dimensional Markov chain, we can write the following equalities. For $n_1 + n_2 = 0$:

$$P(0,0,i,t+\Delta t) = P(0,0,i,t)(1-\lambda\Delta t)(1-i\sigma\Delta t) + P(0,1,i,t)\mu_2\Delta t + o(\Delta t).$$

For $1 \le n_1 + n_2 \le N-1$:

$$\begin{aligned} P(n_1,n_2,i,t+\Delta t) &= P(n_1,n_2,i,t)(1-\lambda\Delta t)(1-n_1\mu_1\Delta t)(1-n_2\mu_2\Delta t)(1-i\sigma\Delta t) \\ &+P(n_1-1,n_2,i,t)\lambda\Delta t + P(n_1+1,n_2-1,i,t)(n_1+1)\mu_1\Delta t \\ &+P(n_1+1,n_2,i,t)n_2\mu_2\Delta t + P(n_1-1,n_2,i+1,t)(i+1)\sigma\Delta t + o(\Delta t). \end{aligned}$$

For $n_1 + n_2 = N$:

$$\begin{aligned} P(n_1,n_2,i,t+\Delta t) &= P(n_1,n_2,i,t)(1-\lambda\Delta t)(1-n_1\mu_1\Delta t)(1-n_2\mu_2\Delta t) \\ &+P(n_1-1,n_2,i,t)\lambda\Delta t + P(n_1,n_2,i-1,t)\lambda\Delta t + P(n_1+1,n_2-1,i,t)(n_1+1)\mu_1\Delta t \\ &+P(n_1-1,n_2,i+1,t)(i+1)\sigma\Delta t + o(\Delta t). \end{aligned}$$

We derive the following system of differential balance equations from here. For $n_1 + n_2 = 0$:

$$\frac{\partial P(0,0,i,t)}{\partial t} = -(\lambda + i\sigma)P(0,0,i,t) + \mu_2 P(0,1,i,t).$$

For $1 \leq n_1 + n_2 \leq N-1$:

$$\frac{\partial P(n_1,n_2,i,t)}{\partial t} = -(\lambda + n_1\mu_1 + n_2\mu_2 + i\sigma)P(n_1,n_2,i,t) + \lambda P(n_1-1,n_2,i,t)$$
$$+(n_1+1)\mu_1 P(n_1+1,n_2-1,i,t) + n_2\mu_2 P(n_1+1,n_2,i,t) + (i+1)\sigma P(n_1-1,n_2,i+1,t).$$

For $n_1 + n_2 = N$:

$$\frac{\partial P(n_1,n_2,i,t)}{\partial t} = -(\lambda + n_1\mu_1 + n_2\mu_2)P(n_1,n_2,i,t) + \lambda P(n_1-1,n_2,i,t)$$
$$+\lambda P(n_1,n_2,i-1,t) + (n_1+1)\mu_1 P(n_1+1,n_2-1,i,t) + (i+1)\sigma P(n_1-1,n_2,i+1,t).$$

Denoting $j = \sqrt{-1}$ and using partial characteristic functions

$$H(n_1,n_2,u,t) = \sum_{i=0}^{\infty} e^{jui} P(n_1,n_2,i,t),$$

we obtain the following system of differential equations.

For $n_1 + n_2 = 0$:

$$\frac{\partial H(0,0,u,t)}{\partial t} = -\lambda H(0,0,u,t) + \mu_2 H(0,1,u,t) + j\sigma \frac{\partial H(0,0,u,t)}{\partial u}. \tag{1}$$

For $1 \leq n_1 + n_2 \leq N-1$:

$$\frac{\partial H(n_1,n_2,u,t)}{\partial t} = -(\lambda + n_1\mu_1 + n_2\mu_2)H(n_1,n_2,u,t) + \lambda H(n_1-1,n_2,u,t)$$
$$+(n_1+1)\mu_1 H(n_1+1,n_2-1,u,t) + n_2\mu_2 H(n_1+1,n_2,u,t) + j\sigma \frac{\partial H(n_1,n_2,u,t)}{\partial u} \tag{2}$$
$$-e^{-ju} j\sigma \frac{\partial H(n_1-1,n_2,u,t)}{\partial u}.$$

For $n_1 + n_2 = N$:

$$\frac{\partial H(n_1,n_2,u,t)}{\partial t} = -(\lambda + n_1\mu_1 + n_2\mu_2)H(n_1,n_2,u,t) + e^{ju}\lambda H(n_1,n_2,u,t)$$
$$+\lambda H(n_1-1,n_2,u,t) + (n_1+1)\mu_1 H(n_1+1,n_2-1,u,t) \tag{3}$$
$$-e^{-ju} j\sigma \frac{\partial H(n_1-1,n_2,u,t)}{\partial u}.$$

Let us use linear finite difference operators $\mathbf{A}$, $\mathbf{B}$, $\mathbf{I}_0$, $\mathbf{I}_1$ to rewrite the system (1)–(3) in the following compact form:

$$\frac{\partial \mathbf{H}(u,t)}{\partial t} = \left(\mathbf{A} + e^{ju}\lambda\mathbf{B}\right)\mathbf{H}(u,t) + \left(\mathbf{I}_0 - e^{-ju}\mathbf{I}_1\right) j\sigma \frac{\partial \mathbf{H}(u,t)}{\partial u}. \tag{4}$$

Here $\mathbf{H}(u,t)$ is $(N+1)\times(N+1)$ top-left triangle matrix with entries equal to $H(n_1,n_2,u,t)$ for indices $n_1 \geq 0$, $n_2 \geq 0$, $n_1 + n_2 \leq N$ and equal to zero for other numbers n_1, n_2. Operators in (4) are defined as follows:

$$[\mathbf{AH}(u,t)]_{n_1,n_2} = \begin{cases} -\lambda H(0,0,u,t) + \mu_2 H(0,1,u,t) & \text{for } n_1+n_2 = 0, \\ -(\lambda + n_1\mu_1 + n_2\mu_2)H(n_1,n_2,u,t) + \lambda H(n_1-1,n_2,u,t) & \\ \quad +(n_1+1)\mu_1 H(n_1+1,n_2-1,u,t) + n_2\mu_2 H(n_1+1,n_2,u,t) & \\ \qquad \text{for } 1 \leq n_1+n_2 \leq N-1, & \\ -(\lambda + n_1\mu_1 + n_2\mu_2)H(n_1,n_2,u,t) + \lambda H(n_1-1,n_2,u,t) & \\ \quad +(n_1+1)\mu_1 H(n_1+1,n_2-1,u,t) & \text{for } n_1+n_2 = N, \end{cases} \tag{5}$$

$$[\mathbf{BH}(u,t)]_{n_1,n_2} = \begin{cases} H(n_1,n_2,u,t) & \text{for } n_1+n_2=N, \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

$$[\mathbf{I}_0\mathbf{H}(u,t)]_{n_1,n_2} = \begin{cases} H(n_1,n_2,u,t) & \text{for } n_1+n_2\le N-1, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

$$[\mathbf{I}_1\mathbf{H}(u,t)]_{n_1,n_2} = H(n_1-1,n_2,u,t). \tag{8}$$

Summing up all equations (1)–(3), we derive

$$\begin{gathered} \frac{\partial}{\partial t}\left\{\sum_{n_1+n_2\le N} H(n_1,n_2,u,t)\right\} = \\ \left(e^{ju}-1\right)\left\{\lambda \sum_{n_1+n_2=N} H(n_1,n_2,u,t) + e^{-ju}\sum_{n_1+n_2\le N-1} j\sigma\frac{\partial H(n_1,n_2,u,t)}{\partial u}\right\}. \end{gathered} \tag{9}$$

Let us denote the summing operator for indices $n_1+n_2=N$ by $\mathbf{E}_1$, the summing operator for indices $n_1+n_2\le N-1$ by $\mathbf{E}_2$, and the total summing operator by $\mathbf{E}=\mathbf{E}_1+\mathbf{E}_2$. Then equation (9) together with (4) may be rewritten in the form

$$\begin{gathered} \frac{\partial \mathbf{H}(u,t)}{\partial t} = \left(\mathbf{A}+e^{ju}\lambda\mathbf{B}\right)\mathbf{H}(u,t) + \left(\mathbf{I}_0 - e^{-ju}\mathbf{I}_1\right) j\sigma\frac{\partial \mathbf{H}(u,t)}{\partial u}, \\ \frac{\partial}{\partial t}\left[\mathbf{E}\mathbf{H}(u,t)\right] = \left(e^{ju}-1\right)\left\{\lambda\mathbf{E}_1\mathbf{H}(u,t) + e^{-ju}j\sigma\frac{\partial}{\partial u}\left[\mathbf{E}_2\mathbf{H}(u,t)\right]\right\}. \end{gathered} \tag{10}$$

This system of equations is the main for the study of the current paper. We will solve it using the method of asymptotic diffusion analysis [16,25] in the next sections.

4. First Stage of Asymptotic Analysis

Let us find the solution of system (10) using the method of asymptotic analysis [16,25] under asymptotic condition of long delay in the orbit: $\sigma \to 0$. To do this, let us make the following substitutions:

$$\sigma = \varepsilon, \quad \tau = \varepsilon t, \quad u = \varepsilon w, \quad \mathbf{H}(u,t) = \mathbf{F}(w,\tau,\varepsilon).$$

Then we obtain the following system:

$$\begin{gathered} \varepsilon\frac{\partial \mathbf{F}(w,\tau,\varepsilon)}{\partial \tau} = \left(\mathbf{A}+e^{j\varepsilon w}\lambda\mathbf{B}\right)\mathbf{F}(w,\tau,\varepsilon) + \left(\mathbf{I}_0 - e^{-j\varepsilon w}\mathbf{I}_1\right) j\frac{\partial \mathbf{F}(w,\tau,\varepsilon)}{\partial w}, \\ \varepsilon\frac{\partial}{\partial \tau}\left[\mathbf{E}\mathbf{F}(w,\tau,\varepsilon)\right] = \left(e^{j\varepsilon w}-1\right)\left\{\lambda\mathbf{E}_1\mathbf{F}(w,\tau,\varepsilon) + je^{-j\varepsilon w}\frac{\partial}{\partial w}\left[\mathbf{E}_2\mathbf{F}(w,\tau,\varepsilon)\right]\right\}. \end{gathered} \tag{11}$$

We can prove the following statement.

Theorem 1. *Under the asymptotic condition $\sigma \to 0$, the following equality holds.*

$$\lim_{\sigma\to 0} \mathrm{E}\left\{e^{jw\sigma i\left(\frac{\tau}{\sigma}\right)}\right\} = e^{jwx(\tau)}. \tag{12}$$

Here the scalar function $x(\tau)$ is a solution of ordinary differential equation

$$x'(\tau) = a(x) = \lambda\mathbf{E}_1\mathbf{R} - x\mathbf{E}_2\mathbf{R}, \tag{13}$$

where $\mathbf{R} = \mathbf{R}(x)$ *is a left-top triangle matrix which is a solution of homogeneous linear system*

$$[\mathbf{A} + \lambda\mathbf{B} + x(\mathbf{I}_1 - \mathbf{I}_0)]\,\mathbf{R} = 0 \tag{14}$$

and satisfies the normalization condition

$$\mathbf{ER} = \sum_{n_1+n_2\le N} R(n_1, n_2) = 1. \tag{15}$$

Proof. Denoting $\lim\limits_{\varepsilon\to 0}\mathbf{F}(w,\tau,\varepsilon) = \mathbf{F}(w,\tau)$ and making asymptotic transition $\varepsilon \to 0$ in (11), we derive

$$\begin{gathered}(\mathbf{A} + \lambda\mathbf{B})\mathbf{F}(w,\tau) + (\mathbf{I}_0 - \mathbf{I}_1)j\frac{\partial\mathbf{F}(w,\tau)}{\partial w} = 0,\\ \frac{\partial}{\partial\tau}\left[\mathbf{EF}(w,\tau)\right] = jw\left\{\lambda\mathbf{E}_1\mathbf{F}(w,\tau) + j\frac{\partial}{\partial w}\left[\mathbf{E}_2\mathbf{F}(w,\tau)\right]\right\}.\end{gathered} \tag{16}$$

We find the solution of this system in the form

$$\mathbf{F}(w,\tau) = e^{jwx(\tau)}\mathbf{R}, \tag{17}$$

where $\mathbf{R}$ is a top-left triangle matrix with non-zero entries $R(n_1, n_2)$ for $n_1 + n_2 \le N$, and $x(\tau)$ is a scalar function which has a meaning of asymptotic (while $\sigma \to 0$) value of the normalized number of customers in the orbit $\sigma i\left(\frac{\tau}{\sigma}\right)$. Entries $R(n_1, n_2)$ have a meaning of asymptotic (while $\sigma \to 0$) probabilities that n_1 servers are working at the first phase and n_2 servers are working at the second phase. Substituting (17) into (16), we obtain equalities (14) and (13). Because entries $R(n_1, n_2)$ of matrix $\mathbf{R}$ are probabilities, they satisfy (15). Since the scalar function $x(\tau)$ is an asymptotic value of the normalized number of customers in the orbit $\sigma i\left(\frac{\tau}{\sigma}\right)$, equality (12) is true. The theorem is proved. □

Probability distribution $\mathbf{R}$ is a solution of system of Equation (14) whose coefficients depend on x, then $\mathbf{R}$ is a matrix function: $\mathbf{R} = \mathbf{R}(x)$. Due to this, we can rewrite (13) as the following expression:

$$a(x) = \lambda\mathbf{E}_1\mathbf{R}(x) - x\mathbf{E}_2\mathbf{R}(x), \tag{18}$$

which determines the function $a(x)$. This function is very important for the study due to the following two reasons. The first one is that we showed that $x'(\tau) = a(x)$. Therefore, function $a(x)$ characterizes dynamic properties of the process $x(\tau)$. The second reason is that $a(x)$ will be used in Section 7 as a drift coefficient for the diffusion process which determines the asymptotic number of customers in the orbit.

Remark 1. *For* $\sigma \to 0$ *(the mean retrial time* $1/\sigma \to \infty$*), the number of customers in the orbit explosively increases, Theorem 1 shows that* $i(t)/(1/\sigma)$ *converges to a deterministic process independent of* σ *and thus has the same order as* $1/\sigma$*. Theorem 1 can be interpreted as the law of large numbers.*

5. Existence of Steady-State Regime

Considering function $a(x)$ which is determined as (18), let us derive a condition of steady-state regime existence.

Due to (18), $a(0) > 0$ and values of the process $x(\tau)$ are growing in the neighborhood of point $x = 0$. If $a(x) > 0$ for all x, then steady-state regime does not exist in the considered retrial queue. The inequality $\lim\limits_{x\to\infty} a(x) < 0$ is the necessary condition for the existence of the steady-state regime.

Let us prove the following statement.

Theorem 2. *Steady-state regime in the considered retrial queue exists if and only if*

$$\lambda < N\mu_2 \frac{(N-1)\mu_2 + \mu_1}{N\mu_2 + \mu_1}. \tag{19}$$

Proof. Sufficiency of condition (19) was proved in [8]. Let us prove its necessity. We rewrite (18) in the form:

$$a(x) = \lambda \sum_{n_1+n_2=N} R(n_1, n_2, x) - x \sum_{n_1+n_2 \le N-1} R(n_1, n_2, x). \tag{20}$$

To determine value of $\lim_{x\to\infty} a(x)$, let us find limit properties of probabilities $R(n_1, n_2, x)$ under $x \to \infty$. Transition intensities for states (n_1, n_2) are depicted in Figure 1 (for simplicity, we draw the graph for partial case of $N = 4$). In the considered system, we have the following possible transitions:

- from state (n_1, n_2) to state $(n_1 + 1, n_2)$ with intensity $(\lambda + x)$, where λ is an intensity of arrivals and x is an asymptotic intensity of retrials from orbit,
- from state (n_1, n_2) to state $(n_1 - 1, n_2 + 1)$ with intensity $n_1\mu_1$ when one of n_1 customers completes service at the first phase and goes to the second one,
- from state (n_1, n_2) to state $(n_1 - 1, n_2)$ with intensity $n_2\mu_2$ for $n_1 > 0$ when one of n_2 customers completes its service at the second phase and leaves the system but one customer served at the first phase instantly goes to the second phase,
- from state $(0, n_2)$ to state $(0, n_2 - 1)$ with intensity $n_2\mu_2$ when one of n_2 customers completes its service at the second phase and leaves the system.

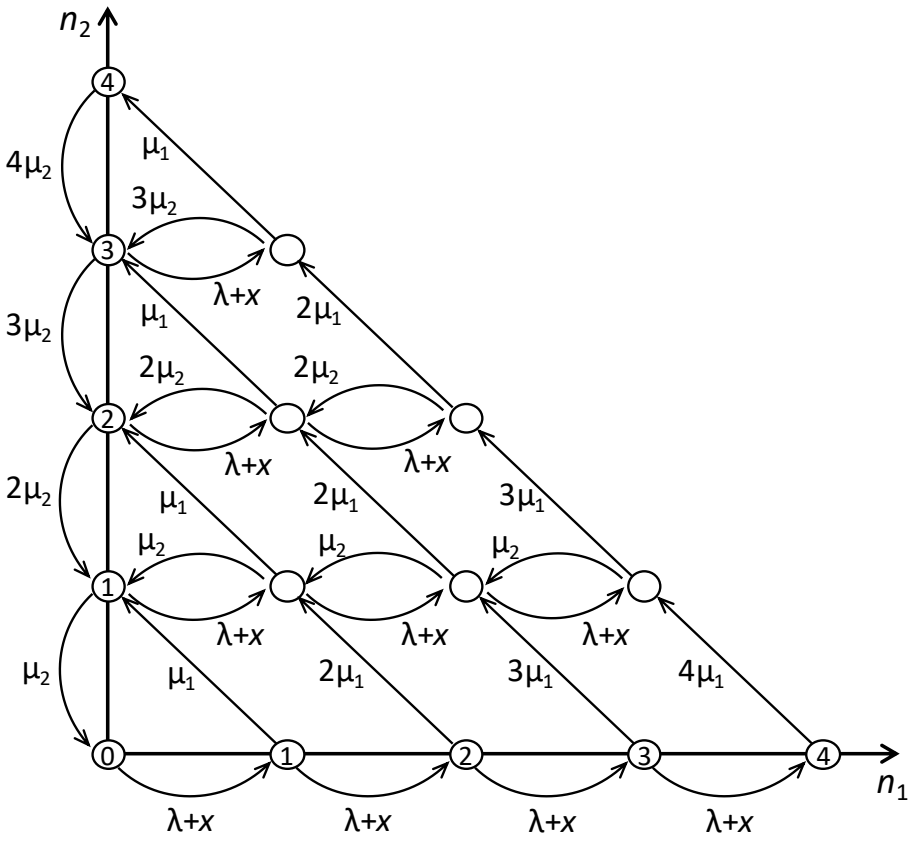

Figure 1. Transition for states (n_1, n_2).

Using the method cut of graph, we derive equations for probabilities $R(n_1, n_2, x)$. We can derive them from system of (14) and (15) but it is clearer how they are obtained if using cut of graph. It should be noted that the system of (14) and (15) is equivalent to the system of equations for finding the

stationary distribution of the Markov chain with the transition diagram as in Figure 1. This Markov chain represents the M/M/N/N queueing system with setup time with the arrival rate given by $\lambda + x$.

For diagonal cuts when $n_1 + n_2 = n \leq N - 1$ we can write

$$(\lambda + x) \sum_{n_1+n_2=n} R(n_1, n_2, x) = \mu_2 \sum_{n_1=1}^{n+1} n_2 R(n + 1 - n_2, n_2, x). \tag{21}$$

Therefore, under the limit condition $x \to \infty$ all the probabilities $R(n_1, n_2, x)$ are equal to zero when $n_1 + n_2 \leq N - 1$. Then due to the normalization requirement, we have

$$\lim_{x\to\infty} \sum_{n_1+n_2=N} R(n_1, n_2, x) = 1, \tag{22}$$

and, so, all non-zero probabilities $R(n_1, n_2)$ are located on the diagonal $n_1 + n_2 = N$. Moreover, from (21), it follows that all the sums $\sum_{n_1+n_2=n} R(n_1, n_2, x)$ are infinitesimals of order $\left(\frac{1}{x}\right)^{N-n}$ for $n \leq N - 1$. For horizontal cuts we can write the following equalities:

$$N\mu_2 R(0, N, x) = \mu_1 R(1, N - 1, x),$$
$$n\mu_2 R(0, n_2, x) = \mu_1 \sum_{n_1=1}^{N-(n_2-1)} n_1 R(n_1, n_2 - 1, x) \quad \text{for } n_2 \leq N - 1.$$

Points $(0, n_2)$ for $n_2 \leq N - 1$ lay on diagonals of the graph, therefore, all $R(0, n_2, x)$ for $n_2 \leq N - 1$ are infinitesimals of order $\left(\frac{1}{x}\right)^{N-n_2}$. Then all the sums $\sum_{n_1=1}^{N-(n_2-1)} R(n_1, n_2 - 1, x)$ are infinitesimals of order $\left(\frac{1}{x}\right)^{N-n_2}$ too. Hence, for $n_2 = N - 1$, the sum $R(1, N - 2, x) + R(2, N - 2, x)$ and probability $R(2, N - 2, x)$ have the infinitesimal order of $\frac{1}{x}$. For $n_2 \leq N - 2$, all the sums $\sum_{n_1=1}^{N-(n_2-1)} R(n_1, n_2 - 1, x)$ and probabilities $R(N - (n_2 - 1), n_2 - 1, x)$ are infinitesimals of order $\left(\frac{1}{x}\right)^2$ or higher. Therefore, under the limit $x \to \infty$, only two probabilities $R(0, N)$ and $R(1, N - 1)$ are not equal to zero. Due to the normalization condition, we can write the following:

$$N\mu_2 R(0, N) = \mu_1 R(1, N - 1), \qquad R(0, N) + R(1, N - 1) = 1.$$

Solution of this system is as follows:

$$R(0, N) = \frac{\mu_1}{N\mu_2 + \mu_1}, \qquad R(1, N - 1) = \frac{N\mu_2}{N\mu_2 + \mu_1}. \tag{23}$$

Let us consider the sum $R(1, N - 2, x) + R(2, N - 2, x)$. As we find before, it is an infinitesimal of order $\frac{1}{x}$. Let us write the equation for the probability $R(2, N - 2, x)$:

$$-\left[2\mu_1 + (N - 2)\mu_2\right] R(2, N - 2, x) + (\lambda + x)R(1, N - 2, x) + 3\mu_1 R(3, N - 3, x) = 0. \tag{24}$$

Here $R(3, N - 3, x)$ is an infinitesimal of order $\left(\frac{1}{x}\right)^2$, hence, the probability $R(1, N - 2, x)$, which is multiplied here by x, also, is an infinitesimal of order $\left(\frac{1}{x}\right)^2$ because only in this case

equality (24) is true and sum $R(1, N-2, x) + R(2, N-2, x)$ is an infinitesimal of order $\frac{1}{x}$. Therefore, for $n_1 + n_2 \leq N-1$, only one probability $R(0, N-1, x)$ has infinitesimal order $\frac{1}{x}$, all other probabilities $R(n_1, n_2, x)$ have higher infinitesimal order. Let us write equation for the probability $R(0, N-1, x)$ as follows:

$$[\lambda + x + (N-1)\mu_2]\, R(0, N-1, x) =$$
$$N\mu_2 R(0, N, x) + (N-1)\mu_2 R(1, N-1, x) + \mu_1 R(1, N-2, x).$$

Taking here the limit $x \to \infty$, we obtain

$$\lim_{x\to\infty} xR(0, N-1, x) = N\mu_2 R(0, N) + (N-1)\mu_2 R(1, N-1).$$

Due to (23), we can rewrite it as follows:

$$\lim_{x\to\infty} xR(0, N-1, x) = N\mu_2 \frac{(N-1)\mu_2 + \mu_1}{N\mu_2 + \mu_1}. \tag{25}$$

Let us come back to the function $a(x) = x'(\tau)$ which determines derivative of the process $x(\tau) = \sigma i(\frac{\tau}{\sigma})$ under asymptotic condition $\sigma \to 0$. Let us find limit of the function $a(x)$ from (20) under condition $x \to \infty$. Due to (22) and (25), we can write

$$\lim_{x\to\infty} a(x) = \lambda - N\mu_2 \frac{(N-1)\mu_2 + \mu_1}{N\mu_2 + \mu_1}.$$

Because $\lim\limits_{x\to\infty} a(x) < 0$ is necessary for the existence of steady-state regime, the condition (19) is true. The theorem is proved. □

Remark 2. *In [6], the M/M/N/Setup model with infinite buffer is analyzed and the stability condition is given $\lambda < N\mu_2$. Theorem 2 shows that the stability condition for the corresponding retrial model is more strict than that of the infinite buffer counterpart.*

6. Second Stage of Asymptotic Analysis

In Section 4, we made the first stage of the asymptotic analysis and, similar to the law of large numbers, we obtain equality (12) which determines the convergence of characteristic function of the process $\sigma i\left(\frac{\tau}{\sigma}\right)$ to the deterministic function $x(\tau)$ under condition $\sigma \to 0$. Now, let us perform the second stage of the asymptotic analysis to obtain more detailed parameters of the convergence.

Let us make the following substitution in system (10)

$$\mathbf{H}(u,t) = \mathbf{H}^{(1)}(u,t) e^{j\frac{u}{\sigma}x(\sigma t)}. \tag{26}$$

We obtain

$$\begin{gathered}
\frac{\partial \mathbf{H}^{(1)}(u,t)}{\partial t} + jua(x)\mathbf{H}^{(1)}(u,t) = \\
\left[\mathbf{A} + e^{ju}\lambda\mathbf{B} + x\left(e^{-ju}\mathbf{I}_1 - \mathbf{I}_0\right)\right]\mathbf{H}^{(1)}(u,t) + \left(\mathbf{I}_0 - e^{-ju}\mathbf{I}_1\right) j\sigma\frac{\partial \mathbf{H}^{(1)}(u,t)}{\partial u}, \\
\frac{\partial}{\partial t}\left[\mathbf{E}\mathbf{H}^{(1)}(u,t)\right] + jua(x)\mathbf{E}\mathbf{H}^{(1)}(u,t) = \\
\left(e^{ju} - 1\right)\left\{\left(\lambda\mathbf{E}_1 - xe^{-ju}\mathbf{E}_2\right)\mathbf{H}^{(1)}(u,t) + j\sigma e^{-ju}\frac{\partial}{\partial u}\left[\mathbf{E}_2\mathbf{H}^{(1)}(u,t)\right]\right\}.
\end{gathered} \tag{27}$$

Substitution (26) is made to go to a centered process for $i(t)$: the function $\mathbf{H}^{(1)}(u,t)$ is a matrix characteristic function of the centered process $i(t) - \frac{1}{\sigma}x(\sigma t)$, where function $x(\tau)$ have been obtained at the first stage of the asymptotic analysis in Section 4.

Denoting $\sigma = \varepsilon^2$ and making in (27) substitutions

$$\tau = \varepsilon^2 t, \quad u = \varepsilon w, \quad \mathbf{H}^{(1)}(u,t) = \mathbf{F}^{(1)}(w,\tau,\varepsilon), \tag{28}$$

we derive

$$\begin{gathered}
\varepsilon^2 \frac{\partial \mathbf{F}^{(1)}(w,\tau,\varepsilon)}{\partial \tau} + j\varepsilon w a(x)\mathbf{F}^{(1)}(w,\tau,\varepsilon) = \\
\left[\mathbf{A} + e^{j\varepsilon w}\lambda\mathbf{B} + x\left(e^{-j\varepsilon w}\mathbf{I}_1 - \mathbf{I}_0\right)\right]\mathbf{F}^{(1)}(w,\tau,\varepsilon) + j\varepsilon\left(\mathbf{I}_0 - e^{-j\varepsilon w}\mathbf{I}_1\right)\frac{\partial \mathbf{F}^{(1)}(w,\tau,\varepsilon)}{\partial w}, \\
\varepsilon^2 \frac{\partial}{\partial \tau}\left[\mathbf{E}\mathbf{F}^{(1)}(w,\tau,\varepsilon)\right] + j\varepsilon w a(x)\mathbf{E}\mathbf{F}^{(1)}(w,\tau,\varepsilon) = \\
\left(e^{j\varepsilon w} - 1\right)\left\{\left[\lambda\mathbf{E}_1 - xe^{-j\varepsilon w}\mathbf{E}_2\right]\mathbf{F}^{(1)}(w,\tau,\varepsilon) + e^{-j\varepsilon w} j\varepsilon \frac{\partial}{\partial w}\left[\mathbf{E}_2\mathbf{F}^{(1)}(w,\tau,\varepsilon)\right]\right\}.
\end{gathered} \tag{29}$$

We can prove the following statement.

Theorem 3. *Let $\Phi(w,\tau)$ be an asymptotic (as $\sigma \to 0$) characteristic function of normalized and centered process $\sqrt{\sigma}\left[i\left(\frac{\tau}{\sigma}\right) - \frac{1}{\sigma}x(\tau)\right]$:*

$$\Phi(w,\tau) = \lim_{\sigma\to 0} \mathrm{E}\exp\left\{jw\sqrt{\sigma}\left[i\left(\frac{\tau}{\sigma}\right) - \frac{1}{\sigma}x(\tau)\right]\right\}. \tag{30}$$

Then it satisfies the equation

$$\frac{\partial \Phi(w,\tau)}{\partial \tau} = a'(x)w\frac{\partial \Phi(w,\tau)}{\partial w} + b(x)\frac{(jw)^2}{2}\Phi(w,\tau), \tag{31}$$

where function $a(x)$ is determined by (18), $b(x)$ has the form

$$b(x) = a(x) + 2\left[(\lambda + x)\mathbf{E}_2\mathbf{g}(x) + x\mathbf{E}_2\mathbf{R}(x)\right], \tag{32}$$

and top-left triangle matrix $\mathbf{g}(x)$ is a particular solution of the system of equations

$$\left[\mathbf{A} + \lambda\mathbf{B} + x\left(\mathbf{I}_1 - \mathbf{I}_0\right)\right]\mathbf{g}(x) = a(x)\mathbf{R}(x) + \left(x\mathbf{I}_1 - \lambda\mathbf{B}\right)\mathbf{R}(x), \tag{33}$$

and satisfies additional condition

$$\mathbf{E}\mathbf{g}(x) = 0. \tag{34}$$

Proof. We rewrite the first equation of (29) with precision up to infinitesimals of order ε^2:

$$\begin{gathered}
j\varepsilon w a(x)\mathbf{F}^{(1)}(w,\tau,\varepsilon) = \left[\mathbf{A} + \lambda\mathbf{B} + x\left(\mathbf{I}_1 - \mathbf{I}_0\right) + j\varepsilon w\left(\lambda\mathbf{B} - x\mathbf{I}_1\right)\right]\mathbf{F}^{(1)}(w,\tau,\varepsilon) \\
+ j\varepsilon\left(\mathbf{I}_0 - \mathbf{I}_1\right)\frac{\partial \mathbf{F}^{(1)}(w,\tau,\varepsilon)}{\partial w} + O\left(\varepsilon^2\right),
\end{gathered} \tag{35}$$

and let us write its solution in the form of expansion:

$$\mathbf{F}^{(1)}(w,\tau,\varepsilon) = \Phi(w,\tau)\left[\mathbf{R}(x) + j\varepsilon w\mathbf{f}(x)\right] + O\left(\varepsilon^2\right), \tag{36}$$

where $\mathbf{f}(x)$ is some matrix function whose expression we will obtain later. Substituting (36) into (35), we obtain

$$j\varepsilon w a(x)\mathbf{R}(x) = [\mathbf{A} + \lambda\mathbf{B} + x(\mathbf{I}_0 - \mathbf{I}_1)][\mathbf{R}(x) + j\varepsilon w\mathbf{f}(x)]$$
$$+j\varepsilon w(\lambda\mathbf{B} - x\mathbf{I}_1)\mathbf{R}(x) + (\mathbf{I}_0 - \mathbf{I}_1)\mathbf{R}(x)j\varepsilon\frac{\partial\Phi(w,\tau)/\partial w}{\Phi(w,\tau)} + O\left(\varepsilon^2\right).$$

Taking into account (14) and dividing by $j\varepsilon w$, under the asymptotic condition $\varepsilon \to 0$, we obtain the following equation for the matrix function $\mathbf{f}(x)$:

$$a(x)\mathbf{R}(x) = [\mathbf{A} + \lambda\mathbf{B} + x(\mathbf{I}_1 - \mathbf{I}_0)]\mathbf{f}(x) + (\lambda\mathbf{B} - x\mathbf{I}_1)\mathbf{R}(x) + (\mathbf{I}_0 - \mathbf{I}_1)\mathbf{R}(x)\frac{\partial\Phi(w,\tau)/\partial w}{w\Phi(w,\tau)}. \quad (37)$$

Applying the superposition principle, we can write a solution of this equation in the form

$$\mathbf{f}(x) = C\mathbf{R}(x) + \mathbf{g}(x) - \boldsymbol{\varphi}(x)\frac{\partial\Phi(w,\tau)/\partial w}{w\Phi(w,\tau)}. \quad (38)$$

Substituting it into (37), we obtain equations

$$[\mathbf{A} + \lambda\mathbf{B} + x(\mathbf{I}_1 - \mathbf{I}_0)]\mathbf{g}(x) = a(x)\mathbf{R}(x) + (x\mathbf{I}_1 - \lambda\mathbf{B})\mathbf{R}(x), \quad (39)$$

$$[\mathbf{A} + \lambda\mathbf{B} + x(\mathbf{I}_1 - \mathbf{I}_0)]\boldsymbol{\varphi}(x) = (\mathbf{I}_0 - \mathbf{I}_1)\mathbf{R}(x). \quad (40)$$

Differentiating (14) by x, we obtain equation

$$[\mathbf{A} + \lambda\mathbf{B} + x(\mathbf{I}_1 - \mathbf{I}_0)]\frac{d\mathbf{R}(x)}{dx} + (\mathbf{I}_1 - \mathbf{I}_0)\mathbf{R}(x) = 0$$

which coincides with (40). Therefore, its solution $\boldsymbol{\varphi}(x)$ may be presented in the form

$$\boldsymbol{\varphi}(x) = \frac{d\mathbf{R}(x)}{dx}. \quad (41)$$

Notice that an additional condition $\mathbf{E}\boldsymbol{\varphi}(x) \equiv 0$ is true due to the normalization condition. Because matrix function $\mathbf{g}(x)$ is a particular solution of non-homogeneous system (39), we choose an additional equation for it in the form (34) so that $\mathbf{g}(x)$ is uniquely defined. Let us write the second equation of system (29) with the precision up to $O\left(\varepsilon^3\right)$:

$$\varepsilon^2\frac{\partial}{\partial\tau}\left[\mathbf{E}\mathbf{F}^{(1)}(w,\tau,\varepsilon)\right] + j\varepsilon w a(x)\mathbf{E}\mathbf{F}^{(1)}(w,\tau,\varepsilon) =$$
$$j\varepsilon w\left\{(\lambda\mathbf{E}_1 - x\mathbf{E}_2 + j\varepsilon w x\mathbf{E}_2)\mathbf{F}^{(1)}(w,\tau,\varepsilon) + j\varepsilon\frac{\partial}{\partial w}\left[\mathbf{E}_2\mathbf{F}^{(1)}(w,\tau,\varepsilon)\right]\right\}$$
$$+\frac{(j\varepsilon w)^2}{2}(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\mathbf{F}^{(1)}(w,\tau,\varepsilon) + O\left(\varepsilon^3\right).$$

Substituting (36) here, we obtain

$$\varepsilon^2\frac{\partial\Phi(w,\tau)}{\partial\tau} + j\varepsilon w a(x)\Phi(w,\tau)[1 + j\varepsilon w\mathbf{E}\mathbf{f}(x)] =$$
$$j\varepsilon w\Big\{(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\Phi(w,\tau)[\mathbf{R}(x) + j\varepsilon w\mathbf{f}(x)] + j\varepsilon w x\mathbf{E}_2\Phi(w,\tau)\mathbf{R}(x)$$
$$+j\varepsilon\frac{\partial\Phi(w,\tau)}{\partial w}\mathbf{E}_2\mathbf{R}(x)\Big\} + \frac{(j\varepsilon w)^2}{2}\Phi(w,\tau)(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\mathbf{R}(x) + O\left(\varepsilon^3\right).$$

Taking into account (13), combining similar terms, dividing by ε and taking the limit $\varepsilon \to 0$, we derive

$$\frac{\partial \Phi(w,\tau)}{\partial \tau} + (jw)^2 a(x)\Phi(w,\tau)\mathbf{E}\mathbf{f}(x) = \frac{(jw)^2}{2}\Phi(w,\tau)a(x)$$
$$+(jw)^2\Phi(w,\tau)\left\{(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\,\mathbf{f}(x) + x\mathbf{E}_2\mathbf{R}(x) + \frac{\partial\Phi(w,\tau)/\partial w}{w\Phi(w,\tau)}\mathbf{E}_2\mathbf{R}(x)\right\}.$$

Substituting (38) here and taking into account that $\mathbf{E}\mathbf{R}(x) \equiv 1$, $\mathbf{E}\mathbf{g}(x) \equiv 0$, $\mathbf{E}\boldsymbol{\varphi}(x) \equiv 0$ and $\mathbf{E}\mathbf{f}(x) \equiv 0$, we obtain the following equation:

$$\begin{aligned}\frac{\partial \Phi(w,\tau)}{\partial \tau} &= w\frac{\partial \Phi(w,\tau)}{\partial w}\left[(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\,\boldsymbol{\varphi}(x) - \mathbf{E}_2\mathbf{R}(x)\right]\\ &+\frac{(jw)^2}{2}\Phi(w,\tau)\left\{a(x) + 2\left[(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\,\mathbf{g}(x) + x\mathbf{E}_2\mathbf{R}(x)\right]\right\}.\end{aligned} \tag{42}$$

Considering (41) and differentiating (13), we obtain

$$a'(x) = (\lambda\mathbf{E}_1 - x\mathbf{E}_2)\frac{\partial\mathbf{R}(x)}{\partial x} - \mathbf{E}_2\mathbf{R}(x) = (\lambda\mathbf{E}_1 - x\mathbf{E}_2)\,\boldsymbol{\varphi}(x) - \mathbf{E}_2\mathbf{R}(x).$$

This is the coefficient in the first term in the right part of (42). As for the coefficient in the second term, we denote as $b(x)$:

$$b(x) = a(x) + 2\left[(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\,\mathbf{g}(x) + x\mathbf{E}_2\mathbf{R}(x)\right]. \tag{43}$$

We rewrite equality (42) in the form

$$\frac{\partial \Phi(w,\tau)}{\partial \tau} = a'(x)w\frac{\partial \Phi(w,\tau)}{\partial w} + b(x)\frac{(jw)^2}{2}\Phi(w,\tau)$$

which coincides with (31). Due to $\mathbf{E}\mathbf{g}(x) \equiv 0$, we can write the following:

$$(\lambda\mathbf{E}_1 - x\mathbf{E}_2)\,\mathbf{g}(x) = \lambda\mathbf{E}_1\mathbf{g}(x) + x\mathbf{E}_1\mathbf{g}(x) = (\lambda + x)\mathbf{E}_1\mathbf{g}(x).$$

So, (43) is coincident with (32). Equality (30) is true due to substitutions (26), (28), expansion (36), the limit $\varepsilon = \sqrt{\sigma} \to 0$ and $\mathbf{E}\mathbf{R}(x) \equiv 1$. Thus, the theorem is proved. □

Remark 3. *Theorem 3 shows that* $\left[i\left(\frac{\tau}{\sigma}\right) - \frac{1}{\sigma}x(\tau)\right]/\sqrt{1/\sigma}$ *converges to a diffusion process. This can be regarded as a central limit theorem for the process* $i(t)$*, where the mean is* $\frac{1}{\sigma}x(\tau)$ *and the variance is* $\sqrt{1/\sigma}$*.*

7. Method of Asymptotic Diffusion Analysis

In this section, we consider an implementation of the method of asymptotic diffusion analysis for obtaining probability distribution of the process $i(t)$ of the number of customers in the orbit under asymptotic condition $\sigma \to 0$ in considered retrial queue. In what follows, we use y and $y(\tau)$ interchangeably.

Let us make the following inverse Fourier transform in (31):

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jwy}\Phi(w,\tau)dw = P(y,\tau).$$

We obtain the equation

$$\frac{\partial P(y,\tau)}{\partial \tau} = -\frac{\partial}{\partial y}\left\{a'(x)yP(y,\tau)\right\} + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left\{b(x)P(y,\tau)\right\}. \tag{44}$$

Here $y(\tau) = \lim\limits_{\sigma\to 0}\sqrt{\sigma}\left[i(\frac{\tau}{\sigma}) - \frac{1}{\sigma}x(\tau)\right]$ is a limit for the normalized and centered process $\sqrt{\sigma}\left[i\left(\frac{\tau}{\sigma}\right) - \frac{1}{\sigma}x(\tau)\right]$; $\Phi(w,\tau)$ is its characteristic function and $P(y,\tau)$ is its probability density function (p.d.f.).

Equation (44) is the Fokker – Planck equation for p.d.f. $P(y,\tau)$, therefore, the process $y(\tau)$ is a diffusion process with a drift coefficient equal to $a'(x)y$ and a diffusion coefficient equal to $b(x)$.

Diffusion process $y(\tau)$ is a solution of the stochastic differential equation

$$dy(\tau) = a'(x)yd\tau + \sqrt{b(x)}dw(\tau), \tag{45}$$

where $w(\tau)$ is the Wiener process. *This equation is difficult to solve directly.*

We rewrite the ordinary differential Equation (13) in the form:

$$dx(\tau) = a(x)d\tau. \tag{46}$$

Consider the following stochastic process:

$$z(\tau) = x(\tau) + \varepsilon y(\tau), \tag{47}$$

where $\varepsilon = \sqrt{\sigma}$ as in the previous section. It is easy to confirm that the process $z(\tau)$ is the limit of the normalized process $\sigma i\left(\frac{\tau}{\sigma}\right)$ under condition $\sigma \to 0$. Thus, $z(\tau)$ has a more direct relation with $i(t)$ which is the quantity of interest.

We derive a stochastic differential equation for $z(\tau)$.

Differentiating (47) and taking into account (45), (46), we obtain

$$dz(\tau) = dx(\tau) + \varepsilon dy(\tau) = \left[a(x) + \varepsilon y a'(x)\right]d\tau + \varepsilon\sqrt{b(x)}dw(\tau).$$

Let us rewrite coefficients in this equality in the form using the first order Taylor series expansion with respect to ϵ.

$$a(x) + \varepsilon y a'(x) = a(x+\varepsilon y) + O\left(\varepsilon^2\right) = a(z) + O\left(\varepsilon^2\right),$$
$$\varepsilon\sqrt{b(x)} = \varepsilon\sqrt{b(x+\varepsilon y) + O(\varepsilon)} = \varepsilon\sqrt{b(z)} + O\left(\varepsilon^2\right).$$

So, we can rewrite the equality with precision up to $O\left(\varepsilon^2\right)$ as follows:

$$dz(\tau) = a(z)d\tau + \varepsilon\sqrt{b(z)}dw(\tau).$$

Due to $\varepsilon = \sqrt{\sigma}$, we can write the equation

$$dz(\tau) = a(z)d\tau + \sqrt{\sigma b(z)}dw(\tau)$$

which is a stochastic differential equation and its solution $z(\tau)$ is a diffusion process with drift coefficient $a(z)$ and diffusion coefficient $\sigma b(z)$.

Stationary p.d.f. $\pi(z)$ of the process $z(\tau)$ is a solution of the Fokker–Planck equation

$$-\left[a(z)\pi(z)\right]' + \frac{\sigma}{2}\left[b(z)\pi(z)\right]'' = 0.$$

We find a solution of this differential equation in the form.

$$-\left[a(z)\pi(z)\right] + \frac{\sigma}{2}\left[b(z)\pi(z)\right]' = 0.$$

The general solution for this differential equation is given in the form.

$$\pi(z) = \frac{C}{b(z)} \exp\left\{\frac{2}{\sigma}\int_0^z \frac{a(x)}{b(x)}dx\right\}, \tag{48}$$

where C is an arbitrary constant. Because, we are looking for a solution which is a probability density function with support in $[0, \infty]$, we choose C as follows.

$$C = \left\{\int_0^\infty \frac{1}{b(z)} \exp\left[\frac{2}{\sigma}\int_0^z \frac{a(x)}{b(x)}dx\right] dz\right\}^{-1}.$$

Therefore, we found asymptotic probability density function $\pi(z)$ of the normalized number of customers in the orbit in the steady state, i.e., $\sigma i(\frac{\tau}{\sigma})$ as $\sigma \to 0$ and then $\tau \to \infty$. In the next section, we further use it to build a discrete distribution of the number of customers in the orbit.

8. Approximations of Steady-State Distributions

The goal of the work is to find enough precise approximations for discrete probability distribution $P(n_1, n_2)$ of the number of servers working in the first and in the second phases and for probability distribution $P(i)$ of the number of customers in the orbit in steady-state regime of considered multi-server retrial queue with setup time.

Two-dimensional probability distribution $P(n_1, n_2)$ can be approximated by entries $R(n_1, n_2, x)$ of the top-left triangle matrix $\mathbf{R}(x)$ if we choose appropriate value x of function $x(\tau)$ which is determined by the differential equation $x'(\tau) = a(x)$. To do this, we choose $x = \kappa$ where κ is a positive root of the equation

$$a(x) = 0.$$

For such value of x we have $x'(\tau) = 0$ that means the steady-state regime.

Algorithm for constructing of the approximations $R(n_1, n_2)$ and $R(n)$ for $R(n_1, n_2, x)$ and $R(n, x)$ is as follows. Here $R(n)$ and $R(n, x)$ are the probabilities that the number of busy servers (in both phases) is n.

1. Let us rewrite Equation (14) for entries $R(n_1, n_2, x)$ of the matrix $\mathbf{R}(x)$:

$$-(\lambda + x)R(0, 0, x) + \mu_2 R(0, 1, x) = 0 \quad \text{for } n_1 + n_2 = 0,$$

$$\begin{gathered}-(\lambda + x + n_1\mu_1 + n_2\mu_2)R(n_1, n_2, x) + (\lambda + x)R(n_1 - 1, n_2, x) \\ +(n_1 + 1)\mu_1 R(n_1 + 1, n_2 - 1, x) + n_2\mu_2 R(n_1 + 1, n_2, x) = 0 \quad \text{for } 1 \le n_1 + n_2 \le N,\end{gathered}$$

$$\begin{gathered}-(n_1\mu_1 + n_2\mu_2)R(n_1, n_2, x) + (\lambda + x)R(n_1 - 1, n_2, x) \\ +(n_1 + 1)\mu_1 R(n_1 + 1, n_2 - 1, x) = 0 \quad \text{for } n_1 + n_2 = N\end{gathered}$$

and let us find a solution $R(n_1, n_2, x)$ of the system which satisfies the normalization requirement $\sum_{n_1+n_2\le N} R(n_1, n_2, x) \equiv 1$ for given N, λ, μ_1 and μ_2.

2. From equality (13), we derive

$$a(x) = \lambda \sum_{n_1+n_2=N} R(n_1, n_2, x) - x \sum_{n_1+n_2\le N-1} R(n_1, n_2, x). \tag{49}$$

Using notation

$$r(x) = \sum_{n_1+n_2=N} R(n_1, n_2, x) \tag{50}$$

and applying the normalization requirement

$$1 = \sum_{n_1+n_2\le N} R(n_1,n_2,x) = \sum_{n_1+n_2=N} R(n_1,n_2,x) + \sum_{n_1+n_2\le N-1} R(n_1,n_2,x),$$

we can rewrite (49) in the form

$$a(x) = (\lambda + x)r(x) - x.$$

3. Using numerical methods, we find a solution $x = \kappa$ of the equation

$$a(x) = 0.$$

4. Substituting $x = \kappa$ into $R(n_1,n_2,x)$, we obtain approximation $R(n_1,n_2) = R(n_1,n_2,\kappa)$ of two-dimensional probability distribution of the number of servers working in the first and in the second phases of service.

5. Using expression

$$R(n) = \sum_{n_1=0}^{n} R(n_1, n-n_1), \tag{51}$$

we find approximation $R(n)$ of the probability distribution of the number of busy servers in the considered retrial queue.

To find an approximation for discrete steady-state probability distribution $P(i)$ of the number of customers in the orbit, we apply p.d.f. $\pi(z)$ of continuous limit process normalized by σ. To do this, we apply expression (48) and find values of the following function:

$$G(i) = \frac{1}{b(\sigma i)} \exp\left\{\frac{2}{\sigma}\int_0^{\sigma i} \frac{a(x)}{b(x)}dx\right\}. \tag{52}$$

We obtain an approximation $P_A(i)$ for the distribution $P(i)$ in the form

$$P_A(i) = G(i) \Big/ \sum_{i=0}^{\infty} G(i). \tag{53}$$

It should be noted that this approximation is not unique and we might choose other alternative distribution also [26].

Algorithm of constructing of the approximation $P_A(i)$ is as follows.

1. Let us rewrite Equation (33) for each non-zero entry $g(n_1,n_2,x)$ of the matrix $\mathbf{g}(x)$:

$$-(\lambda+x)g(0,0,x) + \mu_2 g(0,1,x) = a(x)R(0,0,x) \quad \text{for } n_1+n_2 = 0,$$

$$\begin{gathered} -(\lambda + x + n_1\mu_1 + n_2\mu_2)g(n_1,n_2,x) \\ +(\lambda+x)g(n_1-1,n_2,x) + (n_1+1)\mu_1 g(n_1+1,n_2-1,x) + n_2\mu_2 g(n_1+1,n_2,x) = \\ a(x)R(n_1,n_2,x) + xR(n_1-1,n_2,x) \quad \text{for } 1 \le n_1+n_2 \le N-1, \end{gathered}$$

$$\begin{gathered} -(n_1\mu_1 + n_2\mu_2)g(n_1,n_2,x) + (\lambda+x)g(n_1-1,n_2,x) + (n_1+1)\mu_1 g(n_1+1,n_2-1,x) = \\ [a(x)-\lambda]\,R(n_1,n_2,x) + xR(n_1-1,n_2,x) \quad \text{for } n_1+n_2 = N. \end{gathered}$$

Then we find solution $g(n_1,n_2,x)$ of this system which satisfies additional requirement $\sum_{n_1+n_2\le N} g(n_1,n_2,x) \equiv 0$ for given N, λ, μ_1 and μ_2.

2. From (32), we can write

$$b(x) = a(x) + 2\left[(\lambda+x)\sum_{n_1+n_2=N} g(n_1,n_2,x) + x\sum_{n_1+n_2\le N-1} R(n_1,n_2,x)\right].$$

Here $\sum_{n_1+n_2\le N-1} R(n_1,n_2,x) = 1 - \sum_{n_1+n_2=N} R(n_1,n_2,x) = 1 - r(x)$, therefore, we can write function $b(x)$ in the form

$$b(x) = a(x) + 2\left[(\lambda + x)\sum_{n_1+n_2=N} g(n_1,n_2,x) + x(1-r(x))\right].$$

3. We substitute obtained functions $a(x)$ and $b(x)$ into (52) to calculate values $G(i)$.

4. Applying (53), we obtain probability distribution $P_A(i)$ which approximates the goal distribution $P(i)$ for small values of σ.

9. Applicability Area of Obtained Results

To estimate precision and applicability area of obtained approximations, we compare the results with empiric probabilities obtained on the base of simulation results. For numerical estimation of the error, we traditionally use Kolmogorov distance

$$\Delta = \max_{0\le i<\infty}\left|\sum_{k=0}^{i}[P_1(k) - P_2(k)]\right|, \tag{54}$$

where $P_1(k)$ and $P_2(k)$ are discrete distributions that should be compared. In our case, one of them is approximation (53) and another one is an empiric distribution of the number of customers in the orbit obtained from simulation results.

We use

$$\Delta \le 0.05 \tag{55}$$

as a precision criterion for the approximation to be applicable. To avoid simulation error influence, we choose such sample sizes in the simulation experiments to empiric distributions have error $\Delta \le 0.001$ (to estimate it, we use comparisons between several simulation results). You may choose any level of the error in the form of the Kolmogorov distance. Due to the definition of the Kolmogorov distance, the value 0.05 may be interpreted as a 5% error and this fact has no other meaning.

For simulation experiments, the following parameters of the retrial queue with setup time were chosen: the number of servers $N = 5$, setup time distribution parameter $\mu_1 = 2$, main service distribution parameter $\mu_2 = 1$. Intensity of arrivals λ were chosen to satisfy steady-state existence requirement (19). Therefore, if we denote the system load parameter by $\rho \in [0;1)$, then we can choose value of λ as follows:

$$\lambda = \rho N\mu_2 \frac{(N-1)\mu_2 + \mu_1}{N\mu_2 + \mu_1}.$$

This can be rewritten as

$$\rho = \frac{\lambda}{N\mu_2}\frac{N\mu_2 + \mu_1}{(N-1)\mu_2 + \mu_1},$$

representing the traffic intensity of our system.

We will vary parameters ρ and σ such that (55) is satisfied. Results of such analysis for various values of parameters ρ and σ are presented in Table 1. We highlight by boldface font values that satisfy condition (55). Therefore, we can conclude that obtained approximation (53) becomes more precise when system load parameter ρ grows and delay in the orbit increases (σ becomes less). For values $\rho \ge 0.9$ and for values $\sigma \le 0.2$ it becomes applicable in a wide range.

Table 1. Values of Kolmogorov distance (Δ) of approximation distribution P_A given by (53) and its exact one (by simulation) for various values of σ (retrial rate) and ρ (traffic intensity). The cases with $\Delta \leq 0.05$ are in boldface.

σ	**2**	**1**	**0.5**	**0.2**	**0.1**
$\rho = 0.6$	0.0578	0.0962	0.0907	**0.0403**	**0.0143**
$\rho = 0.7$	0.0842	0.0810	**0.0471**	**0.0122**	**0.0070**
$\rho = 0.8$	0.0650	**0.0396**	**0.0154**	**0.0035**	**0.0027**
$\rho = 0.9$	**0.0247**	**0.0103**	**0.0020**	**0.0012**	**0.0010**

Distributions of the number of customers in the orbit obtained by simulation experiments and by the approximation for $\rho = 0.8$ are shown in Figure 2. As one can see, these distributions are very close to each other and especially for the cases ($\sigma = 1, 0.5, 0.2$) where $\Delta \leq 0.05$, and this fact proves the high accuracy of the obtained approximation. Just for large values of σ, we have a jump at the starting point $n = 0$ in real distribution which we cannot approximate properly.

An important observation is that the approximation becomes more accurate as the traffic intensity increases. This is because, we approximate the scaled number of customers in the orbit by a truncated distribution of the one (stationary distribution of the diffusion process) which might have negative support. When the traffic intensity, the mass for negative part is small and thus this approximation is appropriate.

Results on errors in mean and variance for these distributions are shown in Table 2.

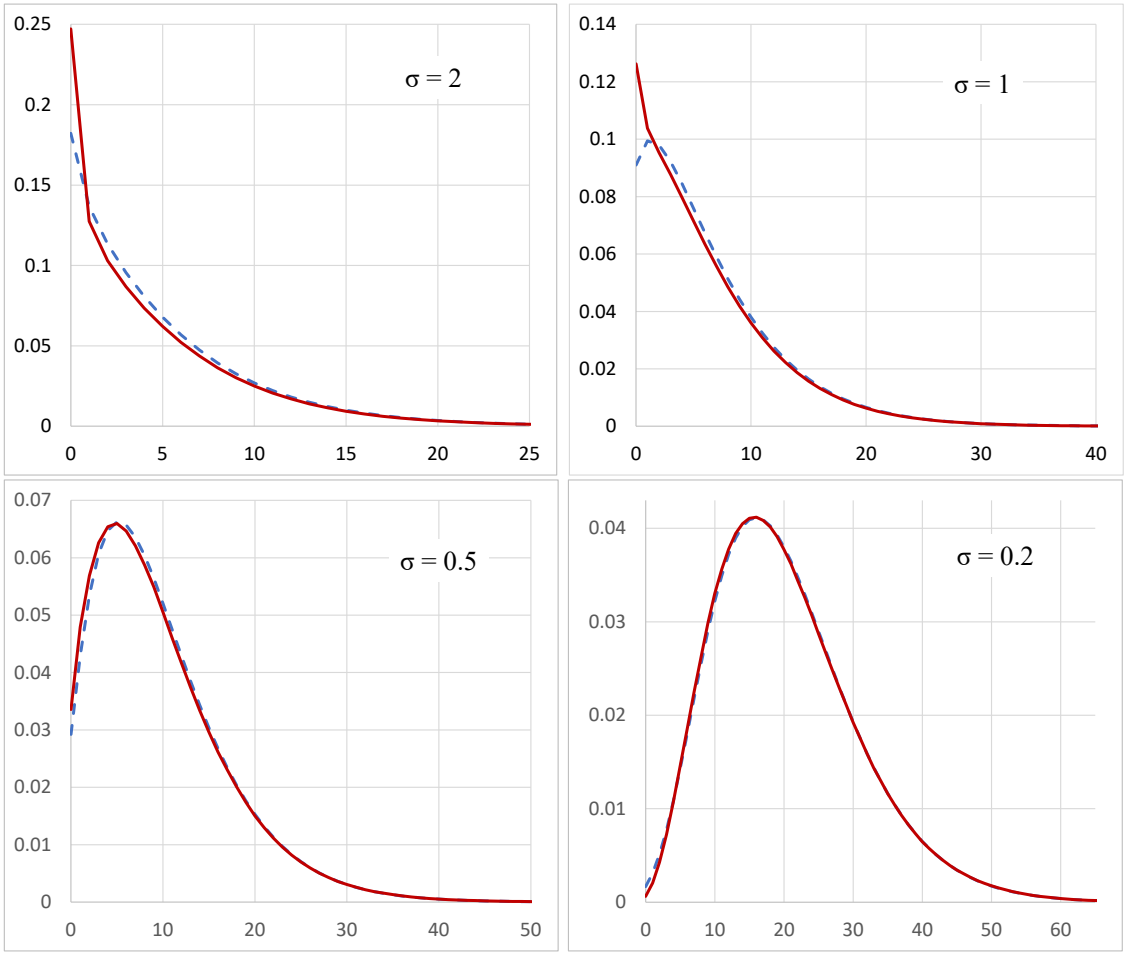

Figure 2. Comparisons of the approximation (dashed line) and the simulation results (solid line) obtained for the probability distribution of the number of customers in the orbit for $\rho = 0.8$ and various values of σ.

Table 2. Errors in mean and variance of approximation distribution P_A given by (53) for $\rho = 0.8$ and various values of σ.

σ	**2**	**1**	**0.5**	**0.2**
Mean error	8%	5%	1.5%	0.002%
Variance error	0.5%	0.3%	0.04%	<0.0001%

In Figure 3, we compare the mean power consumption of our model in the limitting regime $\sigma \to 0$ (denoted by Asymptotic) with that of the buffered model [6], i.e., instead of retrial, blocked customers join the buffer in front of the servers ($\sigma \to \infty$). Figure 3 shows the total power consumption by setup servers (phase 1) and active servers (phase 2) against the arrival rate λ. We assume that both the setup server and the active one consume an energy unit per unit time. Thus, the power consumption is equal to the mean of the sum of the number of setup servers and that of active ones. The mean number of setup servers and that of active ones in the limit regime is calculated using the distribution $R(n_1, n_2)$ and is given by $\sum_{n_1+n_2=1}^{N} R(n_1, n_2) \times (n_1 + n_2)$. For a reference, we also plot the power consumption of the ON-IDLE model (M/M/c queue), where the mean number of active servers and that of idle servers are given by λ/μ_2 and $(c - \lambda/\mu_2)$, respectively. Here we assume that power concumption of an idle server is of 60% that of an active one. As a result, power consmption for hte ON-IDLE model is simply given by $\lambda/\mu_2 + 0.6 \times (c - \lambda/\mu_2)$. Fixed parameters are given by $\mu_2 = 1, c = 5$ for both models. We observe that the power consumption of the asymptotic regime is larger than that of the buffered model. This is intuitive because in the buffered model, once a server completes a service (phase 2), a waiting customer can immediately occupy the server without setup (phase 1). Furthermore, ON-OFF models are more power-saving than the ON-IDLE model when the arrival rate is relatively small, and the opposite trend is observed when the arrival rate is large enough.

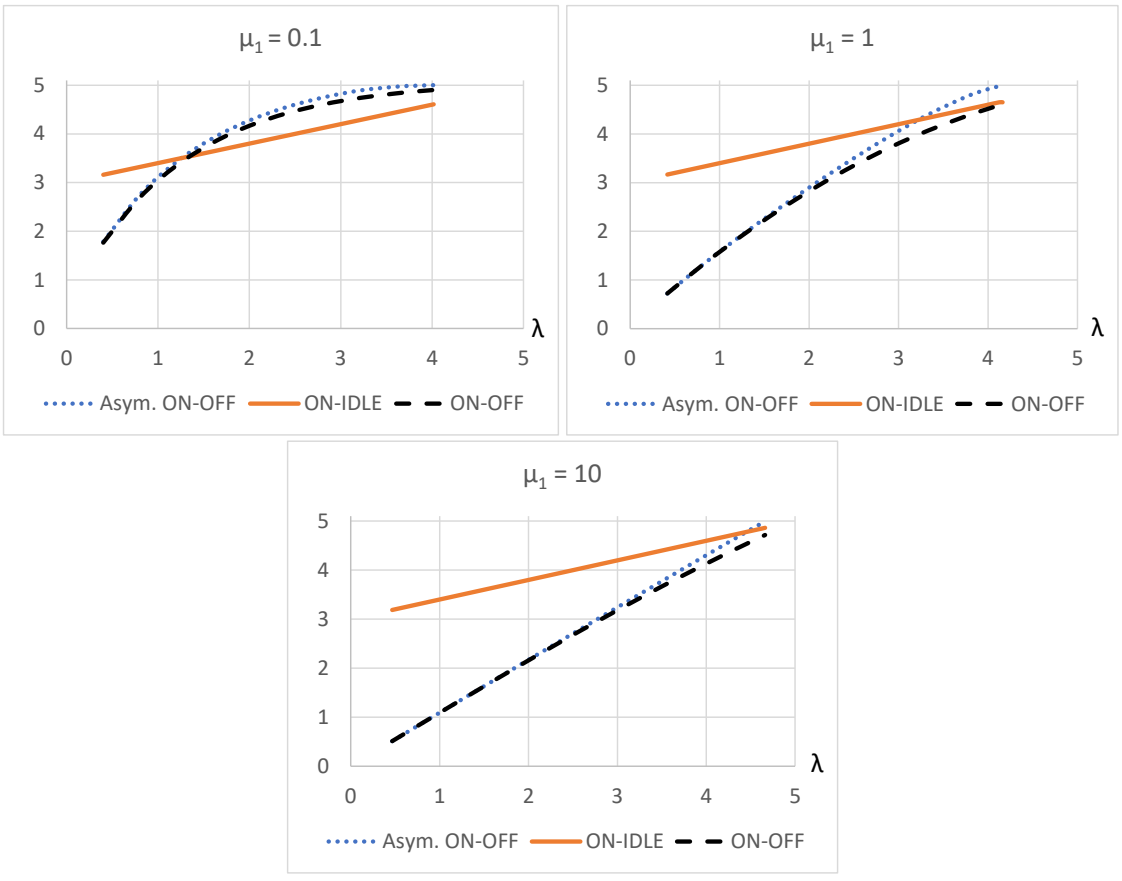

Figure 3. Power consumption against the arrival rate λ.

10. Conclusions

In the paper, we considered a multiserver retrial queue with setup time. Using methods of asymptotic analysis under the condition of long delay in the orbit, we derived necessary condition for the existence of the steady-state regime. Using the method of asymptotic diffusion analysis, we obtained diffusion process whose probability density function is used an approximation for the probability distribution of the number of customers in the orbit. Using simulations and numerical experiments, we showed that the approximation is good for a wide enough range of parameters. Further studies may be made for more complex models with setup time, e.g., models with impatient customers and/or with outgoing calls. The asymptotic regime where the number of servers tends to infinity [18,27] might also be worth for investigation. In our paper, although we proved that the scaled version of the number of customers in the orbit converges to a diffusion process which has a stationary distribution. In the future work, we plan to give a rigorous proof of the convergence of the stationary scaled number of customers in the orbit to the stationary distribution of the diffusion process.

Author Contributions: Conceptualization, A.N., T.P.-D.; formal analysis, A.M., T.P.-D.; investigation, A.M., A.N., S.P., T.P.-D.; methodology, A.N.; software, A.M.; writing–original draft, S.P.; writing–review & editing, A.M., T.P.-D. All authors have read and agreed to the published version of the manuscript.

Funding: The third author T.P. was supported in part by JSPS Kakenhi No. 18K18006.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ren, Y.; Phung-Duc, T.; Chen, J.C.; Yu, Z.W. Dynamic auto scaling algorithm (dasa) for 5g mobile networks. In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–6.
2. Phung-Duc, T.; Ren, Y.; Chen, J.C.; Yu, Z.W. Design and analysis of deadline and budget constrained autoscaling (DBCA) algorithm for 5G mobile networks. In Proceedings of the 2016 IEEE international conference on cloud computing technology and science (CloudCom), Luxembourg, 12–15 December 2016; pp. 94–101.
3. Gandhi, A.; Harchol-Balter, M.; Adan, I. Server farms with setup costs. *Perform. Eval.* **2010**, *67*, 1123–1138. [CrossRef]
4. Gandhi, A.; Doroudi, S.; Harchol-Balter, M.; Scheller-Wolf, A. Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *ACM Sigmetrics Perform. Eval. Rev.* **2013**, *41*, 153–166. [CrossRef]
5. Maccio, V.J.; Down, D.G. Structural properties and exact analysis of energy-aware multiserver queueing systems with setup times. *Perform. Eval.* **2018**, *121*, 48–66. [CrossRef]
6. Phung-Duc, T. Exact solutions for M/M/c/setup queues. *Telecommun. Syst.* **2017**, *64*, 309–324. [CrossRef]
7. Maccio, V.J.; Down, D.G. On optimal control for energy-aware queueing systems. In Proceedings of the 2015 27th International Teletraffic Congress, Ghent, Belgium, 8–10 September 2015; pp. 98–106.
8. Phung-Duc, T.; Kawanishi, K. Multiserver retrial queue with setup time and its application to data centers. *J. Ind. Manag. Optim.* **2019**, *15*, 15–35. [CrossRef]
9. Phung-Duc, T. Single server retrial queues with setup time. *J. Ind. Manag. Optim.* **2017**, *13*, 1329–1345. [CrossRef]
10. Phung-Duc, T.; Masuyama, H.; Kasahara, S.; Takahashi, Y. M/M/3/3 and M/M/4/4 retrial queues. *J. Ind. Manag. Optim.* **2009**, *5*, 431–451. [CrossRef]
11. Phung-Duc, T.; Masuyama, H.; Kasahara, S.; Takahashi, Y. State-dependent M/M/c/c+ r retrial queues with Bernoulli abandonment. *J. Ind. Manag. Optim.* **2010**, *6*, 517–540. [CrossRef]
12. Cohen, J.W. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommun. Rev.* **1957**, *18*, 49–100.
13. Falin, G.I.; Templeton, J.G.C. *Retrial Queues*; Chapman & Hall: London, UK, 1997.

14. Danilyuk, E.Y.; Moiseeva, S.P.; Sztrik, J. Asymptotic Analysis of Retrial Queueing System M/M/1 with Impatient Customers, Collisions and Unreliable Server. *J. Sib. Fed. Univ. Math. Phys.* **2020**, *13*, 218–230 [CrossRef]
15. Fedorova, E.; Nazarov, A.; Moiseev, A. Asymptotic Analysis Methods for Multi-Server Retrial Queueing Systems. In *Applied Probability and Stochastic Processes*; Springer: Singapore, 2020; pp. 159–177.
16. Moiseev, A.; Nazarov, N.; Paul, S. Asymptotic Diffusion Analysis of Multi-Server Retrial Queue with Hyper-Exponential Service. *Mathematics* **2020**, *8*, 531. [CrossRef]
17. Artalejo, J.R.; Gomez-Corral, A. *Retrial Queueing Systems*; Springer: Berlin/Heidelberg, Germany, 2008
18. Halfin, S.; Whitt, W. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **1981**, *29*, 567–588. [CrossRef]
19. Mandelbaum, A.; Massey, W.A.; Reiman, M.I. Strong approximations for Markovian service networks. *Queueing Syst.* **1998**, *30*, 149–201. [CrossRef]
20. Robert, P. *Stochastic Networks and Queues*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 52.
21. Whitt, W. On the heavy-traffic limit theorem for GI/G/∞ queues. *Adv. Appl. Probab.* **1982**, *14*, 171–190. [CrossRef]
22. Whitt, W. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer Science & Business Media: Berlin/Heidelberg, Germany, 2002.
23. Whitt, W. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Manag. Sci.* **2004**, *50*, 1449–1461. [CrossRef]
24. Pender, J.; Phung-Duc, T. A law of large numbers for M/M/c/delayoff-setup queues with nonstationary arrivals. In *Analytical and Stochastic Modelling Techniques and Applications*; LNCS 9845; Springer: Cham, Switzerland, 2016; pp. 253–268.
25. Nazarov, A.; Phung-Duc, T.; Paul, S.; Lizura, O. Asymptotic-Diffusion Analysis for Retrial Queue with Batch Poisson Input and Multiple Types of Outgoing Calls. In *International Conference on Distributed Computer and Communication Networks*; LNCS 11965; Springer: Cham, Switzerland, 2019; pp. 207–222.
26. Jantschi, L. A Test Detecting the Outliers for Continuous Distributions Based on the Cumulative Distribution Function of the Data Being Tested. *Symmetry* **2019**, *11*, 835. [CrossRef]
27. Maccio, V.J.; Down, D.G. Asymptotic performance of an energy-aware G/G/C queue with general setup times. *Queueing Model. Service Manag.* **2020**, *3*, 111–135.