NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Tkachenko Andrey

MULTICHANNEL QUEUING SYSTEMS WITH BALKING AND REGENERATIVE INPUT FLOW

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: SCIENCE, TECHNOLOGY AND INNOVATION
WP BRP 14/STI/2013

Tkachenko Andrey,[1]

Multichannel queuing systems with balking and regenerative input flow [2]

Motivated by the application to telephone call centers this paper is focused on the multichannel queueing system with heterogeneous servers, regenerative input flow and balking. Servers times are random variables but not necessary exponential. If a new customer encountering j other customers in the system then it stays for service with probability f_j and gets rejection with probability $1 - f_j$. For such a system an ergodicity condition is established and functional heavy traffic limit theorems are proved under critically loaded ($\rho \uparrow 1$) and overloaded conditions ($\rho \geq 1$).

JEL Classification: C6.
Key words: multichannel queueing system, regenerative input flow, ergodicity, heavy traffic limits, balking, reneging.

[1]National Research University Higher School of Economics (Moscow, Russia). Institute for Industrial and Market Studies. The Researcher; E-mail:tkachenko_av@hse.ru

[2]I would like to thank Professor Afanasyeva Larisa Grigoryevna for considerable discussion of the problem and significant comments. This paper is partially supported by grant RFBR 13-01-00653.

1 Introduction

Motivated by the application to telephone call centers and more general customer contact centers, this paper focuses on the queueing systems with impatient customers. Recently there has been great interest in multiserver queueing systems with impatient customer, see [27], [29]. Most call centers can be classified into two types: revenue-generating and service-oriented. The revenue-generating call centers typically perform sales functions. For example, they may take customer orders, and have the opportunity to sell customers more goods. In contrast, the service-oriented call centers typically provide customer service, and generate only minimal revenue. For example, they may provide technical support. An important aspect for modeling of service-oriented call centers is the impatience behavior of the customers. Two common modes in which customers display their impatience are balking and reneging. A call in customer who cannot be helped immediately by a human server might be told how long a wait or how many people he/she faces before an operator is available. Then the customer might hang up (i.e., balk) or decide to hold. This is the balking behavior: a customer refuses to enter the queue if the wait is too long or the queue is too big. On the other hand, a customer who is waiting for an operator might hang up (i.e., renege) before getting served if the wait in line becomes too long. This is the reneging behavior. Of course, there can be a combination of the two. It is acknowledged that customers impatience is significant in practice and modeling call centers (cf. [18], [12], [26]).

Queuing systems with balking and reneging have been studied by many researchers, e.g., see the review [24] and literature survey in [17]). For Markovian multichannel queueing systems $M|M|r+M$ with heterogeneous servers different basic characteristics of the system have been analyzed, such as stationary distribution of the number of customers in the system, the mean number of abandoned customers on a fixed period of time and other, e.g., see [6], [20], [8] ,[5]. For the non-Markovian multichannel systems $M|GI|r+D$ and $M|GI|r+GI$ with identical servers stationary distribution of the virtual queueing time and workload, the mean time for system to be idle, probability for customer to abandon and other characteristics have been obtained, e.g., see [28], [17], [19], [9]. In the paper [27] the system $GI|GI|r+GI$ is approximated by a fluid model as $r \to \infty$. On the one hand, this fluid system yields quite accurate approximation for basic model if it is overloaded. On the other hand, for the fluid system it is easier to obtain basic characteristics and provide numerical examples. In [15] the mean waiting time of service have been investigated for system $GI|GI|r+GI$ if the system is overloaded and $r \to \infty$. In [10] for the system $G|GI|r+GI$ have been established convergence of the normalized queue length process and abandonment process to the diffusion processes as $r \to \infty$ and relationship between them in Halfin-Whitt regime was obtained. In [14] the multidimensional diffusion processes was exploited to approximate the dynamics of a queue with customer reneging served by many parallel servers. One of the main features of all mentioned investigated systems is their unconditional stable behavior because of construction of balking and reneging.

In the present paper we focus on the multichannel queueing system with r heterogeneous servers regenerative input flow and balking. Service times of the servers are generally distributed. If a new customer encountering j other customers in the system stays for service with probability f_j and gets rejection with probability $1 - f_j$. We suppose that $f_j \to f$ and $f \geq 0$. Such multichannel system with identical servers have been studied in [7] and there have been established necessary and sufficient condition of its ergodicity. The same system with single server have been analyzed in [2], where was obtained an convergence of the stationary distribution of number of customers in the system and virtual waiting time to the exponential distribution in a critically loaded regime.

Firstly, in the paper a system is established which is likely to be not ergodic if $f > 0$ and necessary and sufficient ergodic condition is obtained. Secondly, functional heavy traffic limit theorems are provided in a critically loaded ($\rho \uparrow 1$) and overloaded systems ($\rho \geq 1$). In order to prove these theorems we introduce some majorizing systems that bound the number of customer in the basic system both above and below. Then we prove that difference between these majorizing processes is stochastically bounded. Moreover, the behavior of these processes is close to the one of the classical systems ($f_j = 1$ for all j) and we can employ some previous results [3], [16] and [25].

The article is organised as follows. In the second section we provide definition of the regeneration flow and discuss its properties. In the third section we describe the model in detail. The next section is devoted to the two lemmas, which are useful for the following section but has there own interest as well. Basing on these lemmas we establish necessary and sufficient condition of the system ergodicity in the fifth section. In the final section we obtain functional heavy traffic limits both if the system overloaded and critically loaded.

2 Regenerative flow and its properties

Assume an integer-valued stochastic process $\{X(t), t \geq 0\}$ to be defined on the probability space $(\Omega, \mathfrak{F}, P)$. The process has non-decreasing right-continuous trajectories and $X(0) = 0$. Such a process is called a stochastic flow. Usually such a flow describes the number of customers arriving at a queueing system during the time interval [0; t]. We follow [2] to define the regeneration flow.

Definition 1. *A stochastic flow $X(t)$ is regenerative if there is an increasing sequence of the random variables $\{\theta_i,\ i \geq 0\}, \theta_0 = 0$ such that the sequence $\{\varkappa_i\}_{i=1}^{\infty} = \{X(\theta_{i-1} + t) - X(\theta_{i-1}), \theta_i - \theta_{i-1}, t \in [0, \theta_i - \theta_{i-1})\}_{i=1}^{\infty}$ consists of independent identically distributed random elements on $(\Omega, \mathfrak{F}, P)$.*

The random variable θ_i is ith regeneration moment of the $X(t)$ and $\tau_i = \theta_i - \theta_{i-1}$ is its ith regeneration period ($i = 1, 2, \ldots$). Let $\xi_i = X(\theta_i) - X(\theta_{i-1})$ be the number of customers arrived at the system during the ith regeneration period. Assume that $\tau = E\tau_i < \infty$, $a = E\xi_i < \infty$. Therefore there is a limit $\lambda = \lim_{t\to\infty} \frac{X(t)}{t} = \frac{a}{\tau}$ a.s. and λ is the input flow intensity.

The class of regenerative processes is quite broad and it includes all fundamental flows which are exploited in queueing theory. Firstly, the doubly stochastic Poisson process [13] with stochastic regenerative intensity is the regenerative one. There are many other examples of the regenerative flows e.g. semi-markovian, Markov-modulated, Markov-arrival processes and others [4]. Some of them are in general not doubly stochastic Poisson processes. Note that definition 1 does not imply the regeneration points θ_i to be arrival moments. Thus definition 1 is more general then the one is given in [21].

The regenerative flow has several fundamental properties and we provide some of them that we use later.

1. If $E\xi_1^{2+\delta} < \infty$, $E\tau_1^{2+\delta} < \infty$ for some $\delta > 0$, then the normalized process

$$\widehat{X}(t) = \frac{X(tT) - \lambda tT}{\sigma_X \sqrt{T}} \tag{2.1}$$

C-converges (e.g. see [1]) on any finite interval $[0, v]$ as $T \to \infty$ to a Brownian motion. Here

$$\sigma_X^2 = \frac{\sigma_\xi^2}{\tau} + \frac{a^2\sigma_\tau^2}{\tau^3} - \frac{2ar_{\xi\tau}}{\tau^2}, \tag{2.2}$$

$$\sigma_\xi^2 = D\xi_1, \quad \sigma_\tau^2 = D\tau_1, \quad r_{\xi\tau} = cov(\xi_1, \tau_1).$$

Moreover there is C-convergence $(X(tT) - \lambda tT)T^{-1} \to 0$ as $T \to \infty$.

2. Let $p \in (0, 1]$ and $X^p(t)$ is a process constructed from $X(t)$ by thinning. For the ordinary flow $X(t)$ this means that every its jump independently of each other with probability p is the jump of $X^p(t)$. If $X(t)$ has a jump of size m at any moment then the jump of $X^p(t)$ at this moment has binomial distribution with parameters m and p. It is quite obvious that if $X(t)$ is a regenerative flow then $X^p(t)$ is a regenerative one as well and moments of there regeneration are coincide. Moreover

$$\sigma_{X^p}^2 = \frac{p(1-p)a + p^2\sigma_\xi^2}{\tau} + \frac{(pa)^2\sigma_\tau^2}{\tau^3} - \frac{2ap^2r_{\xi\tau}}{\tau^2}.$$

3 Multichannel System with Balking

Let S be a queueing system with r servers and regenerative input flow $X(t)$. Arriving customer directs to any idle server if it exists according to the non-idling, work-conserving, non-preemptive[3] service discipline. Service times are independent random variables that do

[3]This means that when a customer arrives at the server then the server works with the customer without interruption

not depend on input flow. The service time η_n^i of the n-th customer by the i-th server has distribution function $B_i(x)$ with finite mean β_i^{-1}, $(i = \overline{1,r})$. Let us denote $\beta = \sum_{i=1}^{r} \beta_i$.

Let $\{f_j\}_{j=0}^{\infty}$, $f_j \in [0,1]$ be the non-increasing sequence. If a new customer encountering j other customers in the system stays for service with probability f_j and gets rejection with probability $1 - f_j$. If all servers are occupied then customer joins the common queue with infinite capacity. Note if $f_j = 1$ as $j \leq m$ for some m and $f_j = 0$ as $j > m$ then it is a system with finite capacity but if $m = 0$ it is a system with abandonment. In case of $f_j = 1$ for all $j \geq 0$ it is a classical multichannel queueing model with infinite capacity. We assume the following conditions to be fulfilled

Condition 1. *For some* $i = \overline{1,r}$

$$P\{\xi_1 = 0\} + P\{\xi_1 = 1,\ \tau_1 - t_1 > \eta_1^i\} > 0. \tag{3.1}$$

Condition 2. *The distribution of the regeneration period* τ_n *of the input flow has absolutely continuous component.*

Condition 3. *The sequence* f_j *and its limit* $f \geq 0$ *satisfy the following inequality*

$$\sum_{j=0}^{\infty} (f_j - f) < \infty. \tag{3.2}$$

Let $Q(t)$ be a number of customers in the system S at time t. We consider an embedded process $Q_n = Q(\theta_n - 0)$ with discrete time. The Condition 1 provides a return of the process Q_n to zero state from any bounded set with non-zero probability, i.e. for any $j > 0$ there exists $m(j) \geq 0$ such that $P\{Q_{n+m(j)} = 0 | Q_n \leq j\} > 0$. This condition is a sufficient one for aperiodicity of the Q_n as well , i.e. $P\{Q_{n+1} = 0 | Q_n = 0\} > 0$. Thus the processes $Q(t)$ and Q_n are regenerative and we are justified to exploit Smith's stability theorem [22]. The regeneration points of the $Q(t)$ are the moments of regeneration of the input flow $\{\theta_i\}_{i=1}^{\infty}$ which satisfy the equality $Q(\theta_i - 0) = 0$. Let us denote these points by $\{\theta_{\alpha_k}\}_{k=1}^{\infty}$. So the regenerative points of the process Q_n form the sequence $\{\alpha_k\}_{k=1}^{\infty}$.

Note if $f > 0$ then some new arriving customers always justified to join the system even when the queue is large. Thus if the intensity of the input flow is high this system may be not ergodic. The aim of the research is to ascertain necessary and sufficient condition for the system to be ergodic and establish behaviour of the processes Q_n and $Q(t)$ if the system overloaded and critically loaded. In order to achieve it we need to prove an auxiliary lemmas, but we believe they represent their own interest as well.

4 Majorizing lemma

We consider an auxiliary system S^+ with the same input flow and servers but it differ from S by the additional service rule. We divide arriving customers at the systems S^+ into two types. If a new customer encountering j other customers in the system then it joins the single queue and settle to be the first type customer with probability f and the second type customer with probability $f_j - f$. It does not join the queue with probability $1 - f_j$. Let us assume not to serve second type customers in the S^+ but they form the queue and do not make barrier for first type customers. Denote $Q_i^+(t)$ to be the number of customers of ith type at the moment t in the system S^+ $(i = 1, 2)$.

Lemma 1. *If* $Q(0) = Q_1^+(0)$ *and* $Q_2^+(0) = 0$ *then the following stochastic inequality is fulfilled for every* $t > 0$

$$Q_1^+(t) - r \leq Q(t) \leq Q_1^+(t) + Q_2^+(t) + r. \tag{4.1}$$

Proof. We prove only the second inequality of the lemma, because the first one is proved by the same way.

We organize operating of the systems in such a way that inequality (4.1) is fulfilled for all trajectories and stochastic properties of the processes $Q_1^+(t)$, $Q_2^+(t)$ (i.e. their finite dimensional distributions) are preserved. Such an approach is generally called a unique probability space approach. Namely, we construct a probability space, where S and S^+ operate according to the following rules.

(a) *The rule of arrival.* Systems S and S^+ deal with the common realization of the input $X(t)$. When there is a jump of $X(t)$ the customer directs to both systems.

(b) *The rule of joining.* Let $\{v_n\}_{n=1}^{\infty}$ be a sequence of independent random variables uniformly distributed in $[0, 1]$. Let the nth customer encountered $j(k)$ other customers in the system $S(S^+)$ at the arriving moment. If $v_n \leq f_j$ then it joins the system S and balks otherwise. In case of $v_n \leq f$ this customer joins the system S^+ and called first type customer. If $f < v_n \leq f_k$ then it joins the system S^+ and called the second type customer and in the case of $v_n > f_k$ it gets rejection from the S^+.

(c) *The rule of service.* For every k $(k = \overline{1, r})$ there are two independent sequences $\{e_m^k\}_{m=1}^{\infty}$ and $\{\widehat{e}_m^k\}_{m=1}^{\infty}$ of independent identically distributed random variables with distribution function $B_k(x)$. The service time is assigned when the customer arrives at a server (not in the system) for both systems. For the system S service time is chosen only from $\{e_m^k\}_{m=1}^{\infty}$ if customer is served by kth server. For the system S^+ service time of the customer which is served by the kth server is chosen according to the following rules

- If customers arrive at the servers k simultaneously in both systems S and S^+ then their service times are the same and they are an element from $\{e_m^k\}_{m=1}^{\infty}$.

- If at arriving moment of customer to the kth server of the system S^+ there is a customer in the kth server of the system S with service time e_m^k then the service time of the arrived customer in S^+ equals e_m^k .
- If at arriving moment of customer to the kth server of the S^+ there is no customer in the kth server of the S then the service time of the arrived customer equals an ordered element from sequence $\{\widehat{e}_m^k\}_{m=1}^{\infty}$.

Every customer in S that have the same service time as a customer in S^+ corresponding to some element from $\{e_m^k\}_{m=1}^{\infty}(k = \overline{1,r})$ is called a coupled customer. The corresponding customer from S^+ is said to be a coupler.

By $\zeta(t)$ and $\zeta^+(t)$ denote the number of occupied servers in S and S^+ at the moment t respectively. Let $\gamma(t)$ be the number of coupled customers at time t in S. We assume that the described processes have right-continuous trajectories. Introduce the process

$$Y(t) = \big(Q(t), Q^+(t), \delta(t)\big) ,$$

where

$$\delta(t) = [\zeta(t) - \zeta^+(t)]^+ + \gamma(t). \tag{4.2}$$

Obviously $\delta(t) \le r$. Actually in case of $\zeta(t) < \zeta^+(t)$ we have $\delta(t) = \eta(t) \le r$. Note that number of coupled customers is not exceeding $\zeta^+(t)$. Therefore if $\zeta(t) \ge \zeta^+(t)$ then we obtain $\delta(t) \le \zeta(t) \le r$.

Introduce the set of states

$$\mathfrak{L} = \{(i, j, \delta),\ i \ge 0,\ j \ge 0,\ 0 \le \delta \le r\ \ i \le j + \delta\}.$$

If $Y(t) \in \mathfrak{L}$ then at time t the inequality (4.1) is fulfilled. Let us show that if the process $Y(t)$ hits $\mathfrak{L}$ then it stays there forever. Since $Y(0) \in \mathfrak{L}$ then it proves the inequality (4.1) for all t. Let $\{t_n\}_{n=1}^{\infty}$, $t_0 = 0$ be the sequence of the moments of state change of $Y(t)$. It is sufficient to show that if $Y(t_n) \in \mathfrak{L}$, then $Y(t_{n+1}) \in \mathfrak{L}$. Table 1 shows events that lead to the change of state of $Y(t)$, the states that the process can be situated with positive probability, and possible transitions from these states. Table 1 below contains only those events that can occur with positive probability in the continuous case when interarrival intervals and service times have absolutely continuous distribution. In the discrete case the events not represented in Table 1 can be considered as the sequence of events that have occurred. For definiteness we assume that all the events that lead to an exit of the customers from the system occur before the arrival events.

Table 1: Transitions

Event		Initial state	New state
Customer arrives in S and S^+	1	(i,j,δ), $i \le j+\delta$	$(i+1,j+1,\delta)$ $(i+1,j+1,\delta+1)$
Customer arrives in S	2	(i,j,δ), $i<j$	$(i+1,j,\delta)$ $(i+1,j,\delta+1)$
Customer arrives in S^+	3	(i,j,δ), $i>j$, $i\le j+\delta$	$(i,j+1,\delta)$ $(i,j+1,\delta-1)$
Customer leaves S	4	(i,j,δ), $i<j+\delta$	$(i-1,j,\delta)$ $(i-1,j,\delta-1)$ $(i-1,j,\delta-2)$
	5	$(i,i-\delta,\delta)$	$(i-1,i-\delta,\delta)$ $(i-1,i-\delta,\delta-1)$
Customer leaves S^+	6	(i,j,δ), $i<j$, $\zeta(t_n-0)<\zeta^+(t_n-0)$	$(i,j-1,\delta)$ $(i,j-1,\delta+1)$
	7	(i,j,δ), $i\le j+\delta$, $\zeta(t_n-0)\ge\zeta^+(t_n-0)$	$(i,j-1,\delta+1)$
Customer leaves S and S^+	8	(i,j,δ), $i<j+\delta$	$(i-1,j-1,\delta)$ $(i-1,j-1,\delta-1)$
	9	$(i,i-\delta,\delta)$	$(i-1,i-\delta-1,\delta)$

We need some comments for corresponding rows of the Table 1.

1. a) Assume all servers of S and S^+ are occupied. Hence new customers join the queues in both systems and new state of the process $Y(t_n)=(i+1,j+1,\delta)$.

b) Consider the case when customer directs to an idle server only in the system S. Note that the number of coupled customers does not change in this case. If $\zeta(t_n-0)\ge\zeta^+(t_n-0)$, then $\delta(t_n)=\delta(t_n-0)+1$ and new state of the process $Y(t_n)=(i+1,j+1,\delta+1)$. If $\zeta(t_n-0)<\zeta^+(t_n-0)$, then $\delta(t_n)=\delta(t_n-0)$ and new state $Y(t_n)=(i+1,j+1,\delta)$.

c) Assume a new customer directs to an idle server only in the system S^+. Therefore all servers in the S are occupied. Hence the arrived customer couple some customer in the S and $\delta(t_n)=\delta(t_n-0)$. Therefore the new state of the $Y(t_n)=(i+1,j+1,\delta)$.

d) If new customers directs to idle servers in both systems then $\zeta(t_n)-\zeta^+(t_n)=\zeta(t_n-0)-\zeta^+(t_n-0)$. Moreover the number of coupled customers may not change or increase to the 1. Thus $Y(t_n)=(i+1,j+1,\delta)$ or $Y(t_n)=(i+1,j+1,\delta+1)$.

2. Since the customer joins the system S only, so according to the rule (b) with regarding that sequence $\{f_n\}_{n=0}^{\infty}$ is not increasing we conclude that $i<j$. Note that number of coupled customers remains unchanged. Therefore $\delta(t_n)=\delta(t_n-0)+1$ if $\zeta(t_n)>\zeta^+(t_n)$ and $\delta(t_n)=\delta(t_n-0)$ otherwise. Thus $Y(t_n)=(i+1,j,\delta+1)$ or $Y(t_n)=(i+1,j,\delta)$ and the process

$Y(t_n)$ remains in the set $\mathfrak{L}$ because of $i < j$.

3. If the arriving customer is a coupler and $[\zeta(t_n) - \zeta^+(t_n)]^+ = [\zeta(t_n - 0) - \zeta^+(t_n - 0)]^+$ then the new state $Y(t_n) = (i, j+1, \delta+1)$ and $Y(t_n) = (i, j+1, \delta)$ otherwise.

4. a) Assume that the departing customer is not coupled one then $\gamma(t_n) = \gamma(t_n - 0)$. If $\zeta(t_n - 0) \le \zeta^+(t_n - 0)$ then the process $Y(t_n)$ change its state for $(i-1, j, \delta)$. In the case of $\zeta(t_n - 0) > \zeta^+(t_n - 0)$ if queue of the S is not empty then $\delta(t_n) = \delta(t_n - 0)$ and $\delta(t_n) = \delta(t_n - 0) - 1$ otherwise. Therefore $Y(t_n) = (i-1, j, \delta)$ or $Y(t_n) = (i-1, j, \delta - 1)$.

b) Assume that the departing customer is coupled one then $\gamma(t_n) = \gamma(t_n - 0) - 1$. By analogy from the previous item the process $Y(t_n)$ may have transitions to $(i-1, j, \delta-1)$ and $(i-1, j, \delta-2)$. The initial state such that $i < j + \delta$, so we conclude that new states are in the set $\mathfrak{L}$.

5. a) Assume that at the departure moment of the customer the queue of S is not empty, i.e. $i > r$. Then $\zeta(t_n) = \zeta(t_n - 0)$. If the departing customer is coupled one then $Y(t_n) = (i-1, i-\delta, \delta-1)$ and $Y(t_n) = (i-1, i-\delta, \delta)$ otherwise.

b) Assume $i \le r$. Let k of $i - \delta$ customers are the second type customers in the S^+. Then $\zeta^+(t_n - 0) = i - \delta - k$, $\zeta(t_n - 0) = i$ and at the moment $t_n - 0$ the following equality is fulfilled $\delta = i - (i - \delta - k) + \gamma = \delta + k + \gamma$, where γ is a number of coupled customers. Therefore $k = \gamma = 0$. Thus the departing customer is not coupled one and the new state $Y(t_n) = (i-1, i-\delta, \delta-1)$.

6. Since $\zeta^+(t_n - 0) > \zeta(t_n - 0)$ then $i < j$. If a new customer arrived to the server instead of departing one is a coupler then the new state $Y(t_n) = (i, j-1, \delta+1)$ and $Y(t_n) = (i, j-1, \delta)$ otherwise. Inasmuch as $i < j$, then process $Y(t)$ remains in the set $\mathfrak{L}$.

7. Note if there is a new first type customer arrived to the server instead of the departing one in the system S^+, then this customer is a coupler. Actually, in this case $\zeta^+(t_n - 0) = r$ and because of the initial state $Y(t_n - 0)$ we have $\zeta(t_n) = \zeta(t_n - 0) = r$ and according to the rule (c) we conclude it. Therefor if there is a first type customer in the queue then $\gamma(t_n) = \gamma(t_n - 0) + 1$ and $\zeta^+(t_n) = \zeta^+(t_n - 0) - 1$ otherwise. In any case $Y(t_n) = (i, j-1, \delta+1)$.

8. Since we assumed continuity of service and interarrival times then the departing customer is a coupled one.

a) If there are some customers waiting for service in the systems S and S^+ then the number of couples remains unchanged and $Y(t_n) = (i-1, j-1, \delta)$.

b) Assume that there is a customer in the queue of the S and there is no first type customer in queue of the S^+. Then $\zeta(t_n) = \zeta(t_n - 0) = r$, $\zeta^+(t_n) = \zeta^+(t_n - 0) - 1$ and $Y(t_n) = (i-1, j-1, \delta)$.

c) The case when there is a first type customer in the queue of the S^+ and there is no one customer in the queue of the S yields $\zeta(t_n) < r$, $\zeta^+(t_n) = r$ and $\delta(t_n) = \delta(t_n - 0) - 1$. Thus $Y(t_n) = (i-1, j-1, \delta-1)$.

d) If there are no customers waiting for service in both systems then $Y(t_n) = (i-1, j-1, \delta-1)$.

Note that on account of the initial state $Y(t_n - 0)$ the last two cases c) and d) remain the process $Y(t)$ in the set $\mathfrak{L}$.

9. In this case the departing customer is coupled one as well as in the previous case.

a) If there are customers waiting for service in both systems S and S^+ then the number of coupled customers remains unchanged and $Y(t_n) = (i-1, i-\delta-1, \delta)$.

b) Assume that there is a customer in the queue of the S and there is no first type customer in queue of the S^+. Thus $\zeta(t_n) = \zeta(t_n - 0) = r$, $\zeta^+(t_n) = \zeta^+(t_n - 0) - 1$ and $Y(t_n) = (i-1, j-1, \delta)$.

c) The case when there is a customer waiting for service in the S^+ and there is no one in the S is impossible. Actually, if we assume the contrary then it yields $i < r$ and $i - \delta > r$.

d) The case when there is no customer waiting for service in both systems is impossible. Let us assume the contrary then $i < r$. Let k of $i-\delta$ customers are the second type customers in S^+. Then $\zeta^+(t_n - 0) = i - \delta - k$, $\zeta(t_n - 0) = i$ and the relationship (4.2) for moment $t_n - 0$ yields $\delta = [i - (i - \delta - k)]^+ + \gamma = \delta + k + \gamma$, where $\gamma = \gamma(t_n - 0)$. Therefore $k = \gamma = 0$, i.e. the departing customer is not coupled one but it is a contrary to the notion in the beginning of the item 9.

All considered cases prove the lemma.■

For the following statements we need a definition.

Definition 2. *Stochastic process $\{Y(t), t \geq 0\}$ with values in $\mathbb{R}_+$ is stochastically bounded if for every $\varepsilon > 0$ there exists $y < \infty$ such that for all $t \geq 0$*

$$P\{Y(t) < y\} > 1 - \varepsilon.$$

Now we formulate the statement which is similar to the lemma 2 in [2].

Lemma 2. *If $f_j \downarrow f$ and (3.2) is fulfilled then the process $Q_2^+(t)$ is stochastically bounded.*

Proof. Since $Q_2^+(t)$ is not decreasing on t then it is sufficient to prove that the random variable $Q_2^+(\infty) = \lim_{t\to\infty} Q_2^+(t)$ is finite with probability one. Let us estimate the probability of the event

$$\{Q_2^+(\infty) = \infty\} = \bigcap_{k=1}^{\infty} \{Q_2^+(\infty) > k\} = \bigcap_{k=1}^{\infty} \bigcup_{j>k} \{Q_2^+(\infty) = j\}.$$

Let D_n be an event that the nth customer joins the system S^+ and it has the second type. Since

$$\{Q_2^+(\infty) = j\} = \bigcup_{n=1}^{\infty} \{Q_2^+(t_n - 0) = j - 1\} \cap D_n,$$

we have

$$\{Q_2^+(\infty) = \infty\} = \bigcap_{k=1}^{\infty} \bigcup_{j>k} \bigcup_{n=1}^{\infty} \{Q_2^+(t_n - 0) = j - 1\} \cap D_n.$$

Denote $\Lambda_j = \bigcup_{n=1}^{\infty} \{Q_2^+(t_n - 0) = j-1\} \cap D_n$. Then

$$\{Q_2^+(\infty) = \infty\} \subseteq \limsup_{j\to\infty} \Lambda_j \tag{4.3}$$

Note that $\mathbf{P}\{\Lambda_j\} \le f_{j-1} - f$. The Borel-Cantelli lemma [11] and (3.2) yield $\mathbf{P}\{\limsup_{j\to\infty} \Lambda_j\} = 0$. This fact assembled with (4.3) proves the lemma. ■

5 Ergodic theorem

In this section we establish necessary and sufficient condition for ergodicity of the system S by exploiting majorizing lemma for the process $Q(t)$ and ergodic theorem for classical multichannel system with a regenerative input flow and heterogeneous servers [3].

Definition 3. *We say that point $y \in \mathbb{R}_+$ is reachable from zero by process $Y(t)$ if for every $\varepsilon > 0$ there exists neighborhood $\Delta_\varepsilon(y) = \{x : \ |x-y| < \varepsilon\}$ and $t_\varepsilon \ge 0$ such that $P\{Y(t_\varepsilon) \in \Delta_\varepsilon(y) | Y(0) = 0\} > 0$. Let $B_0(Y(t))$ be a set of all reachable points from zero by the process $Y(t)$.*

The set of reachable points from zero for process $Y(t) \in \mathbb{Z}_+$ is defined the same way where set $\Delta_\varepsilon(y)$ equals y.

Definition 4. *Process $\{Y(t),\ t \ge 0\}$ is ergodic if for every initial state $Y(0) = y \in B_0(Y(t))$ there exists a limit*

$$\lim_{t\to\infty} \mathbf{P}\{Y(t) \le x\} = F(x),$$

where $F(x)$ is a distribution function and it does not depend on y.

Definition 5. *Stochastic process $Y(t)$ is strongly stochastically unbounded if for every $\varepsilon > 0$ and $y < \infty$ there exists $t_0 < \infty$ such that for all $t > t_0$*

$$\mathbf{P}\{Y(t) \ge y\} > 1 - \varepsilon.$$

Theorem 1. *1. If conditions 1, 3 are fulfilled and traffic coefficient*

$$\rho = f\lambda\beta^{-1} < 1, \tag{5.1}$$

then the process Q_n is ergodic. Moreover if additionally condition 2 is fulfilled then the process $Q(t)$ is ergodic as well.

2. If $\rho > 1$ or $\rho = 1$ and additional assumptions are fulfilled

$$E\tau_1^{2+\delta} < \infty, \quad E\xi_1^{2+\delta} < \infty, \quad E(\eta_1^i)^2 < \infty, \quad i = \overline{1, r} \tag{5.2}$$

for some $\delta > 0$ then this process is not ergodic.

Proof. We introduce the system S^- with the same input flow and servers but it has the property that all second type customers get rejection upon arrival. Thus only first type customers can join the system. Let $Q^-(t)$ be a number of customers in the system S^- at time t. We note that $Q^-(t) =^d Q_1^+(t)$. The intensity of the input flow for the S^- equals $f\lambda$, so the traffic coefficient satisfies (5.1) and S^- is a classical multichannel queueing system. From the theorem 2 in [3] for this system we conclude that if $\rho \geq 1$ and (5.2) is fulfilled then the process $Q^-(t)$ is strongly stochastically unbounded. According to the lemma 1 we deduce that $Q(t)$ is strongly stochastically unbounded as well and certainly is not ergodic.

Let $\rho < 1$. Basing on the theorem 2 in [3] we conclude that the process $Q^-(t)$ is stochastically bounded. Since $Q^-(t) =^d Q_1^+(t)$ from lemma 2 we deduce that the process $Q^+(t) = Q_1^+(t) + Q_2^+(t)$ has the same property. Applying the lemma 1 we conclude stochastic boundedness of the process $Q(t)$ and its ergodicity as well (see theorem 1 in [3]). ■

Remark 1. *Condition 3 in the theorem 1 can be omitted. However if we exclude it then the lemma 2 will be incorrect and we have to construct another majorizing system* $\widehat{S}$. *In the system* $\widehat{S}$ *for every* $\varepsilon > 0$ *an arriving customer has first type with probability* $\max(f + \varepsilon, 1)$. *Second type customers join the system but they are not served and leave the system in finite deterministic time* V_ε *(e.g. see [7]). We do not exploit this system because it is seemed to be useless for the following functional limit theorems.*

6 Functional limit theorems

In this section we investigate functional limits of the normalized process $Q(t)$ in a critically loaded ($\rho \uparrow 1$) and overloaded ($\rho \geq 1$) system S. The proofs of theorems are based on (4.1) and functional approximations for classical multichannel systems without balking. We assume the conditions 1, 2 and 3 to be fulfilled.

6.1 Overloaded systems

Let traffic coefficient $\rho \geq 1$. Since system S is not ergodic under this condition therefore system is called overloaded. For S, S^+ S^- we introduce the following normalized processes

$$\widehat{Q}_n(t) = \frac{Q(nt) - \beta(\rho - 1)nt}{\widehat{\sigma}\sqrt{n}},$$

$$\widehat{Q}_n^-(t) = \frac{Q^-(nt) - \beta(\rho - 1)nt}{\widehat{\sigma}\sqrt{n}},$$

$$\widehat{Q}_n^+(t) = \frac{Q^+(nt) - \beta(\rho - 1)nt}{\widehat{\sigma}\sqrt{n}}. \tag{6.1}$$

Here

$$\widehat{\sigma}^2 = \sigma_X^2 + \sigma_\beta^2,\ \sigma_X^2 = \frac{f(1-f)a + f^2\sigma_\xi^2}{\tau} + \frac{(fa)^2\sigma_\tau^2}{\tau^3} - \frac{2af^2 cov(\xi_1, \tau_1)}{\tau^2}, \tag{6.2}$$

$$\sigma_\beta^2 = \sum_{i=1}^{r} \sigma_i^2 \beta_i^3,\ \sigma_\tau^2 = Var(\tau_1),\ \sigma_\xi^2 = Var(\xi_1),\ \sigma_i^2 = Var(\eta_1^i),\ i = \overline{1,r}.$$

In [16] for system without balking ($f_j = 1$ for all j) and recurrent input flow it is shown that if $\rho > 1$ ($\rho = 1$) then the normalized process $\widehat{Q}_n(t)$ weakly converge on any finite interval $[0,t]$ to standard Brownian motion (absolute value of standard Brownian motion) as $n \to \infty$. We prove similar result for system with balking and regenerative input flow.

Theorem 2. *If $\rho > 1$ ($\rho = 1$) and for some $\delta > 0$ condition* (5.2) *is fulfilled then the process $\widehat{Q}_n(t)$ weakly converges on any finite interval $[0,t]$ to a standard Brownian motion (absolute value of standard Brownian motion) as $n \to \infty$.*

Proof. We consider two cases.

1) For $f = 1$ the proof is similar to the theorem 3.1 in [16]. The only one aspect requires an attention. In [16] authors exploit recurrent input flow in order to prove its convergence to the Brownian motion. However the same is true for regenerative process as well if (5.2) is fulfilled (property 1).

2) Let $f < 1$. We employ (4.1) for normalized processes (6.1)

$$\widehat{Q}_n^-(t) - \frac{r}{\widehat{\sigma}\sqrt{n}} \le \widehat{Q}_n(t) \le \widehat{Q}_n^-(t) + \frac{Q_2^+(nt)}{\widehat{\sigma}\sqrt{n}} + \frac{r}{\widehat{\sigma}\sqrt{n}}. \tag{6.3}$$

Here we use that $Q_1^+(t) =^d Q^-(t)$. From lemma 2 we conclude that $Q_2^+(nt)$ is stochastically bounded as $n \to \infty$, so the second term in the right part of (6.3) tends to 0 as $n \to \infty$. Note that S^- is a classical multichannel system without balking and thinned regenerative input flow where variance equals (6.2) (property 2 of regenerative flow). From item 1) we conclude that left and right parts of the expression (6.3) weakly converge to the common limit. Therefore the process $\widehat{Q}_n(t)$ tends to the same limit that completes the proof. ■

6.2 Critically loaded systems

In this section we investigate the behavior of the system S when ρ tends to 1 below. Firstly, we construct the sequence of systems S_n such that the traffic coefficient for S_n equals ρ_n and $\rho_n \uparrow 1$. Therefore we consider the time compression asymptotic. Namely we introduce the sequence of input flows

$$X_n(t) = X\left(\rho^{-1}\left(1 - \frac{1}{\sqrt{n}}\right)t\right)$$

such that X_n is an input flow in system S_n. Assume that the sequences of probabilities of joining the queue $\{f_j\}_{j=1}^\infty$ and service times $\{\eta_k^i\}_{k=1}^\infty, i = \overline{1,r}$ of customers are the same for all S_n. Note that in such definition the traffic coefficient of the system S_n equals $\rho_n = f\beta^{-1}\lim_{t\to\infty}\frac{X_n(t)}{t} = 1 - \frac{1}{\sqrt{n}}$. For every system S_n we introduce corresponding majorizing systems S_n^- and S_n^+ which deal with the common realization of the input flow $X_n(t)$. Let $Q_n(t)$, $Q_n^-(t)$ and $Q_n^+(t)$ be a number of customers in the systems S_n, S_n^- and S_n^+ at time t respectively. Obviously that these variables satisfy stochastic inequality (4.1), i.e.

$$Q_n^-(t) - r \le Q_n(t) \le Q_{1,n}^+(t) + Q_{2,n}^+(t) + r, \tag{6.4}$$

where $Q_{1,n}^+(t)$ ($Q_{2,n}^+(t)$) is a number of first (second) type customers in the S_n^+ and $Q_n^+(t) = Q_{1,n}^+(t) + Q_{2,n}^+(t)$. For system S_n^+ the lemma 2 is fulfilled as well, so $Q_{2,n}^+(t)$ is stochastically bounded as $t \to \infty$. We introduce normalized processes

$$\widetilde{Q}_n(t) = \frac{Q_n(nt)}{\sqrt{n}},$$

$$\widetilde{Q}_n^-(t) = \frac{Q_n^-(nt)}{\sqrt{n}},$$

$$\widetilde{Q}_{i,n}^+(t) = \frac{Q_{i,n}^+(nt)}{\sqrt{n}}, \quad i = 1, 2. \tag{6.5}$$

Theorem 3. *If for some $\delta > 0$ the conditions* (5.2) *is fulfilled then the normalized process $\widetilde{Q}_n(t)$ weakly converge on any finite interval* $[0, t]$ *to the reflected Brownian motion with drift coefficient $-\beta$ and diffusion coefficient $\widetilde{\sigma}^2$ as $n \to \infty$. Here $\widetilde{\sigma}^2 = \beta^3\sigma_\beta^2 + \frac{\sigma_X^2}{\rho}$.*

Proof. In case of $f = 1$ the proof is a corollary of the theorem 5.7.1 [25] where we exploit the property 1 of the regenerative flow. Let $f < 1$. We apply the inequality (6.4) for normalized processes

$$\widetilde{Q}_n^-(t) - \frac{r}{\sqrt{n}} \le \widetilde{Q}_n(t) \le \widetilde{Q}_{1,n}^+(t) + \widetilde{Q}_{2,n}^+(t) + \frac{r}{\sqrt{n}}. \tag{6.6}$$

Let us remind that $Q_{1,n}^+(t) =^d Q_n^-(t)$ and the sequence of processes $Q_{2,n}^+(t)$ is stochastically bounded. Thus left and right parts of the expression (6.6) tends to the same limit. Note that $Q_n^-(t)$ is the queue length process for system without balking and thinned regenerative input flow. Therefore exploiting theorem 5.7.1 of [25] completes the proof. ■

Remark 2. *We noticed in remark 1 that condition 3 is not necessary for theorem 1. However it is not true for theorem 2, so we can not omit it.*

7 Summary and conclusion

In this paper we have considered the multichannel queueing system with heterogeneous servers, balking and regenerative input flow. In order to investigate this model we have introduced two other systems that bound the number of customers in the basic one above and below (Lemma 1). Moreover, the difference between queue length processes of these systems is stochastically bounded (Lemma 2). Exploiting these lemmas and theorems for classical multichannel system [3] we have established necessary and sufficient conditions for ergodicity (Theorem 1). We have also proved functional limit theorems for normalized queue-length process under critically loaded (Theorem 2) and overloaded (Theorem 3) conditions exploiting Lemma 1, Lemma 2 and results of the papers [16] and [25]. There are many further research topics worth pursuing. First, the conjecture that the steady-state distribution of the limit process $Q(\infty) = \lim_{t\to\infty} Q(t)$ converge to the exponential law as traffic coefficient tends to the 1 below remains to be proved, see [2] for single server case. Second, stochastic-process limits for waiting times remain to be investigated, see, e.g. [23] for many servers systems. Third, the model could be generalized for the one where customers may renege during their waiting time as well as balk at the arriving moment.

References

[1] L. G. Afanas'eva and E. E. Bashtova. Coupling method for asymptotic analysis of queues with regenerative input and unreliable server. Queueing Systems, In press, 2014.

[2] L. G. Afanas'eva and T. N. Belorusov. Limit theorems for systems with impatient customers under high load conditions. *Theory of Probability & Its Applications*, 56(4):674–682, January 2012.

[3] L. G. Afanas'eva and A. V. Tkachenko. Multichannel queueing systems with regenerative input flow. Theory of Probability & Its Applications, In press, June 2014.

[4] Larisa Afanasyeva, Elena Bashtova, and Ekaterina Bulinskaya. Limit theorems for semi-markov queues and their applications. *Communications in Statistics-Simulation and Computation*, 41(6):688–709, 2012.

[5] R.O. Al-Seedy, A.A. El-Sherbiny, S.A. El-Shehawy, and S.I. Ammar. Transient solution of the M/M/c queue with balking and reneging. *Computers & Mathematics with Applications*, 57(8):1280–1285, April 2009.

[6] C. J. Ancker and A. V. Cafarian. Queuing with reneging and multiple heterogeneous servers. *Naval Research Logistics Quarterly*, 10(1):125–149, 1963.

[7] T. N. Belorusov. Ergodicity of a multichannel queueing system with balking. *Theory of Probability & Its Applications*, 56(1):120–126, January 2012.

[8] Nam Kyoo Boots and Henk Tijms. AnM/M/c queue with impatient customers. *Top*, 7(2):213–220, December 1999.

[9] Nam Kyoo Boots and Henk Tijms. A multiserver queueing system with impatient customers. *Management Science*, 45(3):444–448, March 1999.

[10] J. G. Dai and Shuangchi He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, May 2010.

[11] Willliam Feller. *An introduction to probability theory and its applications*. Wiley, New York, NY, USA, 2 edition, 1957.

[12] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, June 2002.

[13] J. Grandell. *Doubly Stochastic Poisson Processes*. Springer, 1976 edition, August 1976.

[14] Shuangchi He and J. G. Dai. Many-server queues with customer abandonment: numerical analysis of their diffusion models. *arXiv preprint arXiv:1104.0347*, 2011.

[15] Rouba Ibrahim and Ward Whitt. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*, 55(10):1729–1742, October 2009.

[16] Donald L. Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, 2(1):150 – 177, 1970.

[17] Foad Iravani and Baris Balcioglu. Approximations for the M/GI/N+GI type call center. *Queueing Systems*, 58(2):137–153, February 2008.

[18] Ger Koole and Avishai Mandelbaum. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1-4):41–59, July 2002.

[19] Liqiang Liu and Vidyadhar G. Kulkarni. Balking and reneging in M/G/s systems, exact analysis and approximations. *Probability in the Engineering and Informational Sciences*, 22(03):355–371, 2008.

[20] A. Montazer-Haghighi, J. Medhi, and S.G. Mohanty. On a multiserver markovian queueing system with balking and reneging. *Computers & Operations Research*, 13(4):421–425, 1986.

[21] Evsey Morozov. Weak regeneration in modeling of queueing processes. *Queueing Systems*, 46(3-4):295–315, March 2004.

[22] W. L. Smith. Regenerative stochastic processes. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 232(1188):6–31, October 1955.

[23] Rishi Talreja and Ward Whitt. Heavy-traffic limits for waiting times in many-server queues with abandonment. *The Annals of Applied Probability*, 19(6):2137–2175, 2009.

[24] Kangzhou Wang, Na Li, and Zhibin Jiang. Queueing system with impatient customers: A review. In *2010 IEEE International Conference on Service Operations and Logistics and Informatics (SOLI)*, pages 82–87, 2010.

[25] Ward Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, January 2002.

[26] Ward Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, February 2005.

[27] Ward Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1):37–54, January 2006.

[28] Wei Xiong and Tayfur Altiok. An approximation for multi-server queues with deterministic reneging times. *Annals of Operations Research*, 172(1):143–151, November 2009.

[29] Sergey Zeltyn and Avishai Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems*, 51(3-4):361–402, December 2005.

Authors: Tkachenko Andrey, National Research University Higher School of Economics (Moscow, Russia). Institute for Industrial and Market Studies. The Researcher; E-mail: tkachenko_av@hse.ru, tel.8-903-187-90-48 (mob.)