On the Stationary Remaining Service Time in the Queueing Systems

Evsey Morozov[1,2,3] and Taisia Morozova[3]

[1] Institute of Applied Mathematical Research, Karelian Research Centre of the RAS, Petrozavodsk, Russia
[2] Moscow Center for Fundamental and Applied Mathematics, Moscow State University, Moscow 119991, Russia
[3] Petrozavodsk State University, Petrozavodsk, Russia, tiamorozova@mail.ru

Abstract. In this paper we study the remaining service time in a wide class of queueing systems. The remaining service time influences the behaviour of the customers and by this reason is often an important Quality of Service (QoS) indicator of the system. Our main result is intuitive and as follows: the stationary distribution of the remaining service time of a customer coincides with the stationary overshoot in the renewal process generated by service times multiplied by the stationary busy probability of the server. This distribution is also obtained for the non-stable system.

Keywords: Queueing System · Remaining Service Time · Retrial System · Busy Probability

1 Introduction

In this paper, we analyze the *stationary remaining service time* in a wide class of queueing systems. The remaining service time can be considered as an important QoS indicator of the system and may considerably influence the behaviour of the customers. It is worth mentioning that the remaining service time plays also an important role in the stability analysis based on the regenerative approach, see for instance [6]. However, in the stability analysis, rather the tightness of the remaining service time turns out to be essential [5] while a more detailed description, for instance, the explicit expression of the limiting distribution, plays a secondary role. At the first sight, the finding of the stationary distribution of the remaining service time seems to be a challenging problem. However, as this contribution shows, indeed this problem can be resolved, at least for a wide class of multiclass queueing systems described by a *regenerative process*, in which service times of a given class customers are independent identically distributed (iid) but *class-dependent* random variables. The regeneration property of the basic queueing process is critical for our analysis. By this reason, we focus on the systems with Poisson inputs of customers. An alternative setting allows a single renewal input where a class-i customer appears with a probability p_i regardless of the state of the system [6]. It follows from our analysis that the

service discipline and the *reliability* of the server is important for the analysis of the remaining service time while the property of input flow is less important, provided the regeneration property of the system takes place. To the best of our knowledge, there are a few papers only in which the remaining service time is the main object of research, see for instance [5,10,2,3,4]. In particular in the paper [5], the *tightness* of the remaining service time in various queueing systems is established.

Now we describe in brief the systems studying in this paper. We consider an m-server system $M/G/m$ with N independent Poisson inputs of customers belonging to different classes. The service times are assumed to be class-dependent, however iid for a given class. Such a system possesses a regeneration property at the arrival instants when a new customer meets fully empty system. More exactly, in this system the basic processes, the workload and queue-size, are *classically regenerative*, in which the *regeneration cycles* are iid [1].

To analyze this system we first assume that the system is stable (stationary). More precisely, we assume that the basic regenerative processes are positive recurrent that is the mean regeneration period of the system is finite [1,6]. Then we apply the coupling method to construct a modified system in which the aggregated busy time of a server during regeneration period remains unchanged. This modified system is more easy to be analyzed, and all limiting characteristics related to the busy time, including stationary remaining service time, are the same as in the original system. Then we apply the basic limit theorem for the regenerative process to deduce the target distribution. As we mentioned above, the functioning of the servers is critical for this analysis, if the service discipline is *work-conserving*, while it is less important which way the customers enter the servers, provided they stay with the same server until service is elapsed. This observation is critical for the analysis of the remaining service time in the system with non-reliable servers which will be presented in a future research.

The main result we obtain is intuitive and states that the stationary distribution of the remaining service time of a customer coincides with the stationary overshoot in the renewal process generated by the service times multiplied by the stationary busy probability of the server. This result holds even for non-identical servers, but the busy probability can be found explicitly for the identical servers only. We mention also that our results can be applied to the retrial systems with *state-dependent retrial rates*. Stability and performance analysis of such systems have been developed in the recent papers [8,7,9]. We note that the study of the remaining service time is especially motivated in the retrial systems in which, after each departure, there exists an idle time of the server. This idle time in general can be used by the orbital customers to control the rate of the retrial attempts. However a detailed analysis of the remaining service time in these system is assumed to be performed in a future research.

Thus the main contribution of this paper is as follows: the explicit expression for the stationary distribution of the remaining service time in the multiclass multiserver system both in stable and non-stable system.

The paper is organized as follows. In section 2, we describe the basic multi-server multiclass system. First, we define the regenerative structure of the system, and then construct a modified system which is easier to be investigated and in which the limit distributions related to the busy time process remain unchanged comparing with the original system. Then we obtain the stationary distribution of the remaining service time provided the system is *positive recurrent (stationary)*. In Section 3, we find the target distribution in the non-stationary system.

2 The Distribution of Remaining Service Time in Multiserver Multiclass Queueing System

We consider the classic multi-server system $M/G/m$ with m identical servers, N classes of customers which follow independent stationary Poisson inputs with rates λ_i, $i = 1, \ldots, N$, and we denote $\lambda = \sum_i \lambda_i$ the rate of the superposed (Poisson) input. Also we denote by $\{t_n,\, n \geq 1\}$ the instants of this aggregated input assuming $t_1 = 0$. It is assumed that service times of class-i customers, denoted by $\{S_n^{(i)},\, n \geq 1\}$, are iid with generic element $S^{(i)}$, service rate $\mu_i = 1/\mathsf{E}S^{(i)}$ and distribution function F_i, $i = 1, \ldots, N$. Let $Q(t)$ be the total number of customers in the system at instant t^-. First of all we describe the regenerative structure of the system. The regeneration instants $\{T_n\}$ of the process $\{Q(t),\, t \geq 0\}$ (and other related processes describing the dynamics of the system) are defined recursively as,

$$T_{n+1} = \inf_k (t_k > T_n : Q(t_k) = 0),\ n \geq 0,$$

where $T_0 := 0$ is the regeneration point provided zero initial condition holds: $Q(0) = 0$, $t_1 = 0$. In other words, zero initial condition means that the 1st customer arrives at the empty system at instant $t_1 = 0$. We denote by T the generic regeneration period, which is distributed as any distance $T_{n+1} - T_n$ between two adjacent regeneration points, $n \geq 0$. The regenerative process $\{Q(t)\}$ (and the underlying queueing system) is called *positive recurrent* if $\mathsf{E}T < \infty$. This requirement in fact is equivalent to *stability* of the process meaning the existence of the stationary distribution of $Q(t)$ as $t \to \infty$. (More on regenerative processes see in [1].)

Denote by $S(t)$ the remaining service time at instant t^-, assuming $S(t) = 0$ if the server is empty. Because all servers are identical, then our limiting results will not be dependent on the server index, and as a rule we omit index of server in our notation. Also denote by $S_i(t)$ the remaining service time at instant t^- of class-i customer. By definition $S_i(t) = 0$ if, at instant t, the server is empty or serves class-k customer, $k \neq i$. Then the following relation between indicator functions $\mathbf{1}(\cdot)$ holds, for each $x \geq 0$:

$$\mathbf{1}(S(t) > x) = \sum_{i=1}^{N} \mathbf{1}(S_i(t) > x),\ t \geq 0. \tag{1}$$

Thus, at each instant t, at most one indicator function in (1) equals one. Note that

$$\mathsf{P}(S_i(t) > x) = \mathsf{P}(S_i(t) > x | S_i(t) > 0)\mathsf{P}(S_i(t) > 0),$$

and thus (1) implies that

$$\mathsf{P}(S(t) > x) = \sum_{i=1}^{N} \mathsf{P}(S_i(t) > x | S_i(t) > 0)\mathsf{P}(S_i(t) > 0).$$

Our purpose is to find explicitly the limiting distribution of the remaining service time $S(t)$ as $t \to \infty$, when exists. By (1), it is enough to find the weak limit (limit in distribution)

$$S_i(t) \Rightarrow \mathbb{S}_i,\ t \to \infty,\ \ i = 1, \ldots, N,$$

where $\mathbb{S}_i$ denotes the stationary remaining service time of class-i customer. Denote

$$\rho_i = \lambda_i \mathsf{E}S^{(i)} \text{ and } \rho = \sum_i \rho_i,$$

and let $S(t) \Rightarrow \mathbb{S}$ be the stationary (unconditional) remaining service time (in an arbitrary server). Now we formulate and prove the following main result of this section.

Theorem 1. *Assume the following negative drift condition holds:*

$$\rho < m. \tag{2}$$

Then the system is positive recurrent and the distribution of the stationary remaining service time $\mathbb{S}$ has the following form

$$\mathsf{P}(\mathbb{S} \le x) = 1 - \sum_{i=1}^{N} \mathsf{P}_B^{(i)} \mu_i \int_x^{\infty} (1 - F_i(u))du, \tag{3}$$

where

$$\mathsf{P}_B^{(i)} = \frac{\rho_i}{m},\ i = 1, \ldots, N, \tag{4}$$

is the stationary probability that the server is occupied by a class-i customer.

Proof. To prove the positive recurrence, we note that a new customer arriving in the system is a class-i one with the probability $p_i := \lambda_i/\lambda$. Then our system becomes a single-class one in which the mean service time of an arbitrary customer equals

$$\mathsf{E}S = \sum_{i=1}^{N} p_i \mathsf{E}S^{(i)}.$$

Then assumption (2) becomes

$$\lambda \mathsf{E}S = \lambda \sum_{i=1}^{N} p_i \mathsf{E}S^{(i)} = \sum_{i=1}^{N} \rho_i < m,$$

and coincides with the well-known positive recurrence criterion of the $M/G/m$ system [1].

In what follows we consider an arbitrary but fixed server and introduce the process

$$B_i(t) = \int_0^t \mathbf{1}(S_i(u) > 0)du,\ t \geq 0,\ \ i = 1, \ldots, N, \tag{5}$$

which, for each i, equals the total time that the server is occupied by class-i customer in the interval $[0,\, t]$. Let

$$B_n^{(i)} = B_i(T_n) - B_i(T_{n-1}),\ i = 1, \ldots, N,\ n \geq 1,$$

be the increment of the process (5) in the n-th regeneration cycle of the system, that is in the interval $[T_{n-1},\, T_n)$, $n \geq 1$. Evidently, the random variables $\{B_n^{(i)},\, n \geq 1\}$ are iid with $B^{(i)}$ representing a generic increment during a generic period T.

We now slightly modify the system in the following way. First, we couple together all busy periods of the server when it serves class-i customers within each regeneration cycle of the system and then shift the resulting busy period, distributed as $B^{(i)}$, to the beginning of the corresponding regeneration cycle. Denote by $\tilde{S}_i(t)$ the remaining service time at instant t^- in this modified process, and put $\tilde{S}_i(t) = 0$ if no class-i customer in the server. We call the resulting process *i-modified* busy time process, and denote it by

$$\tilde{B}_i(t) = \int_0^t \mathbf{1}(\tilde{S}_i(u) > 0)du,\ t \geq 0. \tag{6}$$

Also we construct a zero-delayed *renewal process*, denoted by

$$\widehat{\mathbb{Z}}_i = \{\widehat{\mathbb{Z}}_n^{(i)} = S_1^{(i)} + \cdots + S_n^{(i)},\, n \geq 1\}, \tag{7}$$

which consists of the same service times $\{S_n^{(i)}\}$ that are used in the original busy time process $B_i(t)$ (and also in the process $\{\tilde{B}_i(t)\}$). Let

$$\widehat{S}_i(t) = \min_n(\widehat{\mathbb{Z}}_n^{(i)} - t : \widehat{\mathbb{Z}}_n^{(i)} - t \geq 0),$$

be the remaining renewal time (at instant t) in the process $\widehat{\mathbb{Z}}_i$. Note that the process $\widehat{\mathbb{Z}}_i$ *has no gaps* as opposed to the idle periods which occur after each busy period in the processes $\{\tilde{B}_i(t)\}$ and $\{B_i(t)\}$ in each regeneration cycle. Recall that F_i represents the distribution function of service time $S^{(i)}$ of class-i customer. It follows from the above construction, that *each service time* $S_n^{(i)}$, as

well as *each (composed) busy period* $\tilde{B}_n^{(i)}$, can be treated as a *regeneration period* of the remaining renewal time process $\{\widehat{S}_i(t)\}$. In particular, the renewal points $\hat{\mathbb{Z}}_n^{(i)}$, $n \geq 1$, are the regeneration instants of the process $\{\widehat{S}_i(t)\}$. Since both types of regeneration periods have finite means,

$$\mathsf{E}S^{(i)} < \infty, \;\; \mathsf{E}B^{(i)} < \infty, \;\; i = 1, \ldots, N,$$

then we obtain from the standard regenerative argument [1] that the following two equivalent representations of the same (with probability 1 (w.p.1)) limit exist, for each $x \geq 0$:

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t 1(\widehat{S}_i(u) > x)du = \frac{1}{\mathsf{E}B^{(i)}} \mathsf{E} \int_0^{B^{(i)}} 1(\widehat{S}_i(u) > x)du \;, \tag{8}$$

$$\begin{aligned} \lim_{t\to\infty} \frac{1}{t} \int_0^t 1(\widehat{S}_i(u) > x)du &= \frac{1}{\mathsf{E}S^{(i)}} \mathsf{E} \int_0^{S^{(i)}} 1(\widehat{S}_i(u) > x)du \\ &= \mu_i \int_x^\infty (1 - F_i(u))du, \end{aligned} \tag{9}$$

where, in order to obtain the second equality in (9), we have relied on the equality

$$\widehat{S}_i(u) = S^{(i)} - u \;\; \text{if} \;\; u \leq S^{(i)}.$$

Let us now again consider the original service process and make the following key observation: the stationary busy probability $\mathsf{P}_B^{(i)}$ in the original system equals the corresponding quantity in the system with the modified busy time process (6) because the total busy time when the server is occupied by class-i customers, within each regeneration period of the system, *remains unchanged*. Thus, due to this construction, we obtain the following relations

$$\begin{aligned} &\lim_{t\to\infty} \frac{1}{t} \int_0^t 1(S_i(u) > x)du = \lim_{t\to\infty} \frac{1}{t} \int_0^t 1(\tilde{S}_i(u) > x)du \\ &= \frac{1}{\mathsf{E}T} \mathsf{E} \int_0^T 1(\tilde{S}_i(u) > x)du = \frac{1}{\mathsf{E}T} \mathsf{E} \int_0^{B^{(i)}} 1(\tilde{S}_i(u) > x)du. \end{aligned} \tag{10}$$

In order to obtain the last equality, we have taken into account that, in the modified system,

$$\tilde{S}_i(t) = 0, \;\; t \in [B^{(i)}, \, T) \, .$$

On the other hand, it is easy to see that

$$\mathsf{E} \int_0^{B^{(i)}} 1(\tilde{S}_i(u) > x)du = \mathsf{E} \int_0^{B^{(i)}} 1(\widehat{S}_i(u) > x)du \;, \tag{11}$$

because we apply the *same service times* both in the original service process and in the renewal process $\{\widehat{Z}_i\}$. Since the input flow is Poisson then the regeneration

period length T is *non-lattice*, and it then follows from regenerative theory [1,6] that there exist the limits

$$\lim_{t\to\infty}\frac{B_i(t)}{t}=\frac{\mathsf{E}B^{(i)}}{\mathsf{E}T}=\mathsf{P}_B^{(i)}=\lim_{t\to\infty}\mathsf{P}(S_i(t)>0). \tag{12}$$

Then, from expressions (8)–(12), we find that, for each $x\geq 0$,

$$\begin{aligned}\lim_{t\to\infty}\mathsf{P}(S_i(t)>x)&=\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}(S_i(u)>x)du\\&=\mathsf{P}_B^{(i)}\mu_i\int_x^\infty(1-F_i(u))du,\ i=1,\ldots,N.\end{aligned} \tag{13}$$

Now it follows from (1) that the stationary distribution of unconditional remaining service time $\mathbb{S}$ exists and has the following explicit form:

$$\mathsf{P}(\mathbb{S}\leq x)=\lim_{t\to\infty}\mathsf{P}(S(t)\leq x)=1-\sum_{i=1}^N\mathsf{P}_B^{(i)}\mu_i\int_x^\infty(1-F_i(u))du,$$

and thus the basic result (3) is proved.

Now we find the probability $\mathsf{P}_B^{(i)}$ in an explicit form. To this end, we denote by $A_{ij}(t)$ the number of class-i arrivals in server j in the interval $[0,\,t)$, $i=1,\ldots,N$, $j=1,\ldots,m$. Because the servers are identical and the total number of class-i arrivals in the interval $[0,\,t)$,

$$A_i(t):=\sum_{j=1}^N A_{ij}(t),$$

satisfies, w.p.1,

$$\frac{A_i(t)}{t}\to\lambda_i,\ \ t\to\infty,$$

then we obtain that

$$\lim_{t\to\infty}\frac{A_{ij}(t)}{t}=\frac{\lambda_i}{m},\ \ j=1,\ldots,m. \tag{14}$$

Denote by $V_{ij}(t)$ the total work which is delivered to the server j by class-i arrivals in the interval $[0,\,t)$. Let $S_n^{(ij)}$ be the service time of the n-th class-i customer being served in server j. Of course, for each server j, the service time $S_n^{(ij)}$ is distributed as $S^{(i)}$, $i=1,\ldots,N$. Also denote by $B_{ij}(t)$ the busy time when server j is occupied by class-i customers in the interval $[0,\,t]$. Then we obtain the following balance equations

$$V_{ij}(t)=\sum_{n=1}^{A_{ij}(t)}S_n^{(ij)}=W_{ij}(t)+B_{ij}(t),\ \ j=1,\ldots,m;\ \ i=1,\ldots,N, \tag{15}$$

where $W_{ij}(t)$ denotes the remaining *class-i work* at instant t^- assigned for server j. It is well known [6] that, in the positive recurrent system,

$$W_{ij}(t) = o(t),\ t \to \infty \quad \text{w.p.1.}$$

Then it remains to take limit in both sides of the equation (15) as $t \to \infty$ and use the Strong Law of Large Numbers to obtain the required equality (4):

$$\lim_{t\to\infty} \frac{\sum_{n=1}^{A_{ij}(t)} S_n^{(ij)}}{A_{ij}(t)} \frac{A_{ij}(t)}{t} = \frac{\lambda_i \mathsf{E} S^{(i)}}{m} = \mathsf{P}_B^{(i)}, \tag{16}$$

where, by the symmetry, the limit is independent of the server index j. This result allows to present the limit distribution (3) as follows

$$\mathsf{P}(\mathbb{S} \le x) = 1 - \sum_{i=1}^{N} \frac{\lambda_i}{m} \int_x^\infty (1 - F_i(u))du, \tag{17}$$

and the proof of Theorem 1 is completed. □

Because

$$\mathsf{P}(S_i(t) > x) = \mathsf{P}(S_i(t) > x | S_i(t) > 0)\mathsf{P}(S_i(t) > 0),\ \ x \ge 0,$$

then it follows from (13) that the limit

$$\lim_{t\to\infty} \mathsf{P}(S_i(t) > x | S_i(t) > 0) = \mu_i \int_x^\infty (1 - F_i(u))du,$$

exists and is equal to the *conditional tail distribution* of the remaining service time provided the server is occupied by class-i customer, $i = 1, \ldots, N$.

3 Non-Stationary Case

The next problem we will address is the limiting distribution of the remaining service time provided the system is *not positive recurrent (not stationary)*. In other words, we assume that $\mathsf{E}T = \infty$, in which case, as it follows for instance from [6], the queue size $Q(t) \Rightarrow \infty$, that is

$$\lim_{t\to\infty} \mathsf{P}(Q(t) > k) = 1 \ \text{ for each } k \ge 0.$$

In this case, the condition

$$\rho = \sum_{i=1}^{N} \rho_i \ge m,$$

must be met. Because in this case all servers will be eventually full, then we expect that $\mathsf{P}(S(t) > 0) \to 1$. Because the service discipline is assumed to be FIFO (earlier we did not specify the discipline), then a new customer assigning

to a given server is class-i one with the probability $p_i = \lambda_i/\lambda$. Thus the limiting fraction of the *class-i work* $V_{ij}(t)$ (assigned for some server j) equals the limiting fraction of the total class-i work $V_i(t)$ among all the work $V(t)$ received in $[0,\, t)$, that is (see (14)-(16))

$$\lim_{t\to\infty}\frac{V_{ij}(t)}{V_i(t)} = \lim_{t\to\infty}\frac{V_i(t)}{V(t)} = \lim_{t\to\infty}\frac{\sum_{n=1}^{A_i(t)} S_n^{(i)}}{\sum_{i=1}^{N}\sum_{n=1}^{A_i(t)} S_n^{(i)}} = \frac{\rho_i}{\rho} =: \hat{\mathsf{P}}_B^{(i)},$$

where the ratio $\hat{\mathsf{P}}_B^{(i)}$ is the probability that, in the limit, an arbitrary server is occupied by a class-i customer. Note that if all customers have the same mean service time, that is $\mu_i \equiv \mu$, then $\hat{\mathsf{P}}_B^{(i)} = p_i$. Thus in this case we must replace in (13) the probability $\mathsf{P}_B^{(i)}$ by the probability $\hat{\mathsf{P}}_B^{(i)}$, implying

$$\lim_{t\to\infty}\mathsf{P}(S_i(t) > x) = \frac{\lambda_i}{\rho}\int_x^\infty (1 - F_i(u))du,\ \ i = 1, \ldots, N,$$

and thus finally we have, by analogy with (17),

$$\mathsf{P}(\mathbb{S} \le x) = 1 - \sum_{i=1}^{N}\frac{\lambda_i}{\rho}\int_0^x (1 - F_i(u))du,\ \ x \ge 0.$$

Note that in this (non-stationary case) the normalizing constant m in (17) is replaced by the constant ρ, that is the number of servers no more plays a role in this analysis.

4 Conclusion

In this work, we study the limiting distribution of the remaining service time in the buffered queueing multiserver systems with Poisson inputs of multiple classes of customers. Using the coupling method and the regeneration property of the systems, we find the target distribution in an explicit form.

5 Acknowledgement

This research is supported in part by Russian Foundation for Basic Research, projects No. 19-07-00303, 18-07-00156, 18-07-00147.

References

1. Asmussen, S.: Applied Probability and Queues. Wiley, N.Y. (1987)
2. Boer, P.T.D., Nicola, V.F., van Ommeren, J.K.C.: The remaining service time upon reaching a high level in m/g/1 queues. Questa, 39, 55–78 **23** (2001)
3. Kerner, Y.: The conditional distribution of the residual service time in the m n/g/1 queue. Models 24 (3), 364-375 (2008)

4. Kerner, Y.: Equilibrium joining probabilities for an m/g/1 queue. Games and Economic Behavior 71 (2), 521-526 (2011)
5. Morozov, E.: The tightness in the ergodic analysis of regenerative queueing processes. Queueing Syst. 27, 179–203 **23** (1997)
6. Morozov, E., Delgado, R.: Stability analysis of regenerative queues. automation and remote control. 1977–1991 (2009)
7. Morozov, E., Dimitriou, I.: Stability analysis of a multiclass retrial system with coupled orbit queues. Proceedings of 14th European Workshop, 73-90, EPEW 2017, Berlin, Germany, September 7-8, 85–98 **7** (2017). https://doi.org/10.1007/978-3-319-66583-2-6
8. Morozov, E., Morozova, T.: Analysis of a generalized system with coupled orbits. proceedings of FRUCT23 (2018)
9. Morozov, E., Morozova, T., Dimitriou, I.: Simulation of multiclass retrial system with coupled orbits. Proceedings of SMARTY18: First International Conference Stochastic Modeling and Applied Research of Technology Petrozavodsk, Russia **10** (2018)
10. Ross, S.M., Seshadri, S.: Hitting time in an m/g/1 queue. J. Appl. Prob, 36, 934-940 **6** (1999)