Adv. Appl. Prob. **22**, 929–956 (1990)
Printed in N. Ireland

ON THE SETTLING TIME OF THE CONGESTED $GI/G/1$ QUEUE

GEORGE D. STAMOULIS* AND
JOHN N. TSITSIKLIS,* *Massachusetts Institute of Technology*

Abstract

We analyze a stable $GI/G/1$ queue that starts operating at time $t=0$ with $N_0 \neq 0$ customers. First, we analyze the time T_{N_0} required for this queue to empty for the first time. Under the assumption that both the interarrival and the service time distributions are of the exponential type, we prove that $\lim_{N_0 \to \infty} T_{N_0}/N_0 \overset{a.s.}{=} 1/(\mu - \lambda)$, where λ and μ are the arrival and the service rates. Furthermore, assuming in addition that the interarrival time distribution is of the non-lattice type, we show that the settling time of the queue is essentially equal to $N_0/(\mu - \lambda)$; that is, we prove that

$$\lim_{N_0 \to \infty} d_{N_0}\left(\frac{N_0}{\mu - \lambda} c\right) = \begin{cases} 1, & \text{for } 0 < c < 1; \\ 0, & \text{for } c > 1, \end{cases}$$

where $d_{N_0}(t)$ is the total variation distance between the distribution of the number of customers in the system at time t and its steady-state distribution. Finally, we show that there is a similarity between the queue we analyze and a simple fluid model.

CONGESTED QUEUE; SETTLING TIME; TRANSIENT ANALYSIS

1. Introduction

In this paper, we analyze the settling time of a stable $GI/G/1$ queue, assuming that it is initially highly congested. Under certain assumptions on the distributions of the interarrival and the service times, we first prove that the time for the queue to empty is asymptotically proportional to the number of customers initially present at the queue. We then show that the time required for the queue to approach stationarity (settling time) is essentially equal to the time for it to empty.

We consider a $GI/G/1$ queue. The interarrival times are independent and identically distributed with moment generating function $A(s)$. The service times are independent and identically distributed with moment generating function $B(s)$; moreover, the service process is independent of the arrival process. The arrival and the service rates are denoted by λ and μ, respectively. The number of customers present at the system at time t (including the customer in service, if any) is denoted by $N(t)$ and it is taken to be right-continuous. The queueing system starts operating at time $t=0$; the arrival time of the first new customer has the interarrival time

Received 7 August 1989; revision received 30 November 1989.

* Postal address for both authors: Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Research supported by the NSF under Grant ECS-8552419, with matching funds from Bellcore Inc. and Du Pont, and the ARO under Grant DAAL03-86-K-0171.

distribution. The queue is said to be *stable* [6] if, as $k \to \infty$, the distribution of the waiting time of the kth customer to be served converges to a limiting function, which is the distribution of a proper random variable (i.e., a random variable that is finite with probability 1); this limiting function is independent of the initial number of customers. Except for the $D/D/1$ queue, a necessary condition for stability [10] is $\lambda < \mu$. In fact, stability is guaranteed [6] if $\lambda < \mu$ and the interarrival time distribution is of the non-lattice type. (A random variable Z is said to be of the lattice type if there exist constants a and b such that the only permissible values of Z are of the form $a + nb$, with n being integer.) Moreover, in this case, the stationary distribution $(\pi_k)_{k=0, \cdots}$ of the number of customers in the system exists; we have

$$\pi_k \stackrel{\text{def}}{=} \lim_{t \to \infty} \Pr[N(t) = k \mid N(0) = 1] \text{ for } k = 0, \cdots.$$

Henceforth, we restrict ourselves to stable queues with $\lambda < \mu$ and with interarrival time distributions of the non-lattice type. In addition, we shall always assume that both the interarrival and the service time distributions are of the exponential type. That is, we have $E[e^{sZ}] < \infty$ for some $s > 0$, where Z is a random variable distributed as the interarrival time; this implies that there exists some $s_A > 0$ such that $A(s)$ is defined for all s in the interval $(-\infty, s_A)$ (see Section 2.1). Similarly, we have $E[e^{\bar{s}Y}] < \infty$ for some $\bar{s} > 0$, where Y is a random variable distributed as the service time; it follows that there exists some $s_B > 0$ such that $B(s)$ is defined for all s in the interval $(-\infty, s_B)$. This assumption on the interarrival and the service time distributions is rather mild, since it holds for most of the distributions appearing in practical cases (e.g. Erlang, hyperexponential, etc.).

Let N_0 be the number of customers intially in the system; we shall always treat N_0 as a positive parameter. We denote by T_{N_0} the random variable corresponding to the time required for the queue to empty (for the first time), namely,

$$T_{N_0} \stackrel{\text{def}}{=} \inf_{t \geq 0} \{t : N(t) = 0\}.$$

It is known that the expected busy period duration is finite, i.e. $E[T_1] < \infty$ (see [9] and references therein). Moreover, Pollaczek [11] has derived the joint distribution of T_1 and the number of customers served during this period. Finally, some other results have been established for the case $N_0 = 1$ (e.g. see [12]).

In the context of the $M/M/1$ queue, using well-known results of queueing theory, T_{N_0} can be expressed as the sum of N_0 independent random variables all of which have the same distribution as the busy period duration T_1. Using this, it may be proved that $E[T_{N_0}] = N_0/(\mu - \lambda)$ and, if the random variables $T_1, T_2, \cdots$ are defined on the same probability space, then

$$\lim_{N_0 \to \infty} \frac{T_{N_0}}{N_0} \stackrel{\text{a.s.}}{=} \frac{1}{\mu - \lambda},$$

where a.s. stands for almost surely (i.e., with probability 1). Similar results hold for the $M/G/1$ queue with service time distribution of the exponential type. We briefly present these results in Section 3.

For the $GI/G/1$ queue under analysis, we prove that

$$\lim_{N_0 \to \infty} \frac{T_{N_0}}{N_0} \overset{\text{a.s.}}{=} \frac{1}{\mu - \lambda}. \tag{1}$$

It is worth noting that this result is in perfect agreement with intuition. Indeed, consider a pool that initially contains a quantity N_0 of fluid. If fluid is supplied at a constant rate λ and, at the same time, it is removed at a constant rate μ, with $\lambda < \mu$, then the pool empties in time $N_0/(\mu - \lambda)$. In fact, it will be shown (see Proposition 11 in Section 5.1) that this fluid analogy holds to a greater extent.

Next, we consider the settling time of the $GI/G/1$ queue under analysis. As shown in [1], there is a close relation between the time for a Markov process to reach stationarity and the hitting times of certain subsets of the state space. In our context, after establishing that the time until the queue empties (for the first time) is approximately equal to $N_0/(\mu - \lambda)$, we prove that it is also essentially the time for the queue to reach steady-state. Indeed, defining $\tau_{N_0} \overset{\text{def}}{=} N_0/(\mu - \lambda)$, we show that

$$\lim_{N_0 \to \infty} d_{N_0}(\tau_{N_0} c) = \begin{cases} 1, & \text{for } 0 < c < 1; \\ 0, & \text{for } c > 1, \end{cases} \tag{2}$$

where $d_{N_0}(t)$ denotes the total variation distance between the distribution of the number of customers in the system at time t (under the initial condition $N(0) = N_0$) and its stationary distribution; that is,

$$\begin{aligned} d_{N_0}(t) &\overset{\text{def}}{=} \max_{A \subseteq Z_+} |\Pr[N(t) \in \mathcal{A} \mid N(0) = N_0] - \pi(\mathcal{A})| \\ &= \tfrac{1}{2} \sum_{k=0}^{\infty} |\Pr[N(t) = k \mid N(0) = N_0] - \pi_k|, \end{aligned}$$

where π_k is the steady-state probability that the system contains k customers and $\pi(\mathcal{A}) = \sum_{k \in \mathcal{A}} \pi_k$ for $\mathcal{A} \subseteq Z_+$. (Intuitively, $d_{N_0}(t)$ is the 'distance' between the transient distribution of the number of customers in the system (at time t) and its stationary distribution.) Thus, τ_{N_0} (namely, $N_0/(\mu - \lambda)$) may be viewed as the settling time of the $GI/G/1$ queue under analysis when it initially contains a large number N_0 of customers.

To the best of our knowledge, both results in (1) and (2) are new. Results of the form (2) have been proved in [1] for 'rapidly mixing' Markov chains, and in [2] for the convergence to steady state of closed Jackson networks with a large number of customers. Interestingly enough, some of the results in [2] are in agreement with an

approximate fluid model. However, the difficulty with such an approach is that the validity of a fluid approximation is technically non-obvious. Thus, our work can be viewed as a step towards the justification of fluid approximations. We expect that our analysis can be substantially extended to cover more complex systems like networks of queues.

2. Background

2.1. *A note on random variables of the exponential type.* The discussion to follow is based on [3].

A random variable T is said to be of the *exponential type* if there exists some positive s such that $E[e^{sT}] < \infty$. The most straightforward example of such a random variable is one that is exponentially distributed with mean $1/\lambda$; for this random variable, we have $E[e^{sT}] < \infty$ for all $s < \lambda$.

The moment generating function of T is defined as follows: $G(s) \stackrel{\text{def}}{=} E[e^{sT}]$. If T is of the exponential type, then there exists some positive s_1 such that $G(s)$ is finite for all $0 \leqq s < s_1$. Moreover, for all $s^* \in (0, s_1)$ the following are true: $G(s)$ is strictly convex and continuous on $[0, s^*]$ and has derivatives of all orders on $(0, s^*)$; its first derivative is strictly increasing on $(0, s^*)$, provided that $\Pr[T = 0] \neq 1$. Furthermore, we have $E[T] < \infty$ and

$$E[T] = \lim_{s \downarrow 0} \frac{dG(s)}{ds}.$$

Henceforth, we restrict ourselves to random variables of the exponential type that satisfy in addition the following property: there exists some positive s_2 such that $E[e^{sT}] < \infty$ for all $s \in (-s_2, 0)$. Clearly, this property is satisfied by random variables that are either lower bounded (that is, there exists some finite constant t_0 such that $\Pr[T \geqq t_0] = 1$) or can be expressed as the difference of two lower bounded random variables of the exponential type that are independent.

The upper and lower tails of the distribution of a random variable of the exponential type may be upper bounded by using the Chernoff bound. Indeed, let t be a finite constant. We have

$$\Pr[T \geqq t] \leqq E[e^{sT}]e^{-st} = G(s)e^{-st}, \qquad \forall s \in (0, s_1). \tag{3}$$

In the case where $t > E[T]$, there exists some positive s (depending on t) such that $G(s)e^{-st} < 1$ for all $s \in (0, s')$. Similarly, we have

$$\Pr[T \leqq t] \leqq E[e^{-sT}]e^{st} = G(-s)e^{st}, \qquad \forall s \in (0, s_2).$$

In the case where $t < E[T]$, there exists some positive s'' (depending on t) such that $G(-s)e^{st} < 1$ for all $s \in (0, s'')$.

We apply the above results to the random variable $\sum_{i=1}^{N} X_i$, where $X_1, \cdots, X_N$ are independent random variables that have the distribution of the random variable

X, which is of the exponential type. Let δ be positive; we have

$$\Pr\left[\sum_{i=1}^{N} X_i \geqq E[X](1+\delta)N\right] \leqq \exp(-\phi_1(\delta)N), \tag{4}$$

where $\phi_1(\delta)$ is a positive constant depending on δ. Indeed, we have $E[\exp(s\sum_{i=1}^{N} X_i)] = [G(s)]^N$, where $G(s) = E[e^{sX}]$; applying (3) with $s = s^*$, where s^* is chosen to satisfy $G(s^*)\exp(-s^*E[X](1+\delta)) < 1$, and defining $\phi_1(\delta) \stackrel{\text{def}}{=} -\ln G(s^*) + s^*E[X](1+\delta)$, we obtain (4). Similarly, for any positive δ, we have

$$\Pr\left[\sum_{i=1}^{N} X_i \leqq E[X](1-\delta)N\right] \leqq \exp(-\phi_2(\delta)N), \tag{5}$$

where $\phi_2(\delta)$ is a positive constant depending on δ.

2.2. *A note on exponential convergence.* The discussion to follow is based on [7]. Let $(Z_N)_{N=1,\ldots}$ be a sequence of random variables (not necessarily defined on the same probability space) and let $(h_N)_{N=1,\ldots}$ be a sequence of positive numbers with $\lim_{N\to\infty} h_N = \infty$. The sequence $(Z_N/h_N)_{N=1,\ldots}$ of random variables is said to converge exponentially to the constant z, as $N \to \infty$, if for any positive δ there exists some $n(\delta) \geqq 1$ and some positive $\gamma(\delta)$ (both depending on δ) such that

$$\Pr\left[\left|\frac{Z_N}{h_N} - z\right| \geqq \delta\right] \leqq \exp(-\gamma(\delta)N), \qquad \forall N \geqq n(\delta).$$

Moreover, if the random variables $(Z_N)_{N=1,\ldots}$ are defined on the same probability space, then the inequality above implies almost sure convergence, namely $\lim_{N\to\infty}(Z_N/h_N) \stackrel{\text{a.s.}}{=} z$.

2.3. *A note on notation.* Throughout this paper, the notations $\Pr[\Gamma]$ and $E[X]$ stand for $\Pr[\Gamma \mid N(0) = N_0]$ and $E[X \mid N(0) = N_0]$, respectively. Similarly, $\Pr[\Gamma \mid \Delta]$ and $E[X \mid \Delta]$ stand for $\Pr[\Gamma \mid \Delta \text{ and } N(0) = N_0]$ and $E[X \mid \Delta \text{ and } N(0) = N_0]$, respectively, unless the event Δ is of the form $N(0) = n^*$.

Also, $\lfloor x \rfloor$ denotes the integer part of x, and $\lceil x \rceil$ denotes the smallest integer that is greater than or equal to x.

3. Results on the $M/M/1$ and the $M/G/1$ queues

In this section, we present some results concerning the time required for the stable $M/M/1$ queue to empty. Similar results hold for the stable $M/G/1$ queue with service time distribution of the exponential type, as well.

3.1. *The $M/M/1$ queue.* The proposition to follow suggests that, in the context of the $M/M/1$ queue, the random variable T_{N_0} can be expressed as the sum of N_0 independent and identically distributed random variables. This may be proved by decomposing T_{N_0} into sub-busy periods (e.g. see [10]).

Proposition 1. The following is true:

$$T_{N_0} \overset{\text{st}}{=} \sum_{i=1}^{N_0} V_i,$$

where $V_1, \cdots, V_{N_0}$ are independent random variables, all of which have the same distribution as the busy period T_1 (the notation $\overset{\text{st}}{=}$ denotes equality in distribution).

The moment generating function of T_1 is known in closed form (see [10]), namely

$$E[\exp(sT_1)] = \frac{\mu + \lambda - s - \sqrt{(\mu + \lambda - s)^2 - 4\mu\lambda}}{2\lambda}, \qquad \forall s \leqq 0; \tag{6}$$

this leads to the following result for the probability density function of T_1:

$$p_{T_1}(t) = \frac{e^{-(\lambda+\mu)t}}{t\sqrt{\frac{\lambda}{\mu}}} \cdot I_1(2t\sqrt{\lambda\mu}), \qquad \forall t \geqq 0,$$

where $I_1(\cdot)$ is the modified Bessel function of the first kind of order 1. According to [4], the integral $\int_0^\infty p_{T_1}(t)e^{st}\,dt$ is equal to the expression appearing on the right-hand side of (6) for all $s < (\sqrt{\lambda} - \sqrt{\mu})^2$. Therefore, the random variable T_1 is of the exponential type. Moreover, it follows from (6) that $E[T_1] = 1/(\mu - \lambda)$. Hence, using Proposition 1 (and the strong law of large numbers), we obtain the following results.

Proposition 2. The random variable T_{N_0} is of the exponential type and satisfies

$$E[T_{N_0}] = \frac{N_0}{\mu - \lambda} \overset{\text{def}}{=} \tau_{N_0} \quad \text{and} \quad \lim_{N_0 \to \infty} \frac{T_{N_0}}{N_0} \overset{\text{a.s.}}{=} \frac{1}{\mu - \lambda}.$$

3.2. *The $M/G/1$ queue.* Propositions 1 and 2 of Section 3.1 hold in the context of the stable $M/G/1$ queue with service time distribution of the exponential type, as well. Indeed, the decomposition of T_{N_0} into sub-busy periods is still applicable. Furthermore, it is well known (see [10]) that, in the context of the $M/G/1$ queue, the moment generating function $G(s)$ of the busy period duration T_1 satisfies the following functional equation: $G(s) = B(s - \lambda + \lambda G(s))$ for all $s \leqq 0$. This implies that $E[T_1] = 1/(\mu - \lambda)$; moreover, because of Proposition 1, we have

$$E[T_{N_0}] = \frac{N_0}{\mu - \lambda} \overset{\text{def}}{=} \tau_{N_0} \quad \text{and} \quad \lim_{N_0 \to \infty} \frac{T_{N_0}}{N_0} \overset{\text{a.s.}}{=} \frac{1}{\mu - \lambda}.$$

Finally, the fact that T_{N_0} is of the exponential type follows from Corollary 6 in Section 4.2.

4. Preliminary results on the $GI/G/1$ queue

In this section we present several results on the time required for the stable $GI/G/1$ queue to empty. As already mentioned in Section 1, it is assumed that both

the interarrival and the service time distributions are of the exponential type, that the interarrival time distribution is of the non-lattice type and that $\lambda < \mu$. A powerful result such as Proposition 1 does not hold in the case where the arrival process is not of the Poisson type. Thus, the derivation of (1) in the more general context of the $GI/G/1$ queue is considerably more complicated as compared to the proof of Proposition 2.

4.1. *Some preliminary results.* First, we establish a lower bound on $E[T_{N_0}]$. Let $\mathcal{N}$ denote the number of arrivals until the system is met empty for the first time. In other words, we have $\mathcal{N} = n$ if the arrival of the nth customer is the first to occur at a time larger than T_{N_0}. Let $\mathcal{Z}_{\mathcal{N}}$ denote the arrival time of the $\mathcal{N}$th customer. Clearly, the number of customers served until the queue empties equals $\mathcal{N} + N_0 - 1$. The following lemma is established in [12] by using Wald's equation. (In fact, only the case $N_0 = 1$ is considered there; however, the result may be easily extended to hold for $N_0 = 2, \cdots$.)

Lemma 3. The following are true:

$$E[\mathcal{N}] = \lambda E[\mathcal{Z}_{\mathcal{N}}] \quad \text{and} \quad E[\mathcal{N}] + N_0 - 1 = \mu E[T_{N_0}].$$

Moreover, if $E[\mathcal{N}] = \infty$, then $E[\mathcal{Z}_{\mathcal{N}}] = E[T_{N_0}] = \infty$.

In the case where $E[\mathcal{N}] < \infty$ (which will be shown to always be true for the type of queues we consider), Lemma 3 implies that the average arrival rate up to (and including) the time when the queue is met empty equals λ. Similarly, the average service rate up to (and including) the time when the queue empties equals μ. Based on this lemma, we prove the following result.

Proposition 4. The following is true:

$$\frac{N_0 - 1}{\mu - \lambda} < E[T_{N_0}], \quad \text{for} \quad N_0 = 1, \cdots.$$

Proof. The result is trivially true if $E[T_{N_0}] = \infty$. However, it will be shown later that this never occurs.

Assuming that $E[T_{N_0}] < \infty$, we prove the result as follows. Clearly, we have $T_{N_0} < \mathcal{Z}_{\mathcal{N}}$ with probability 1, which implies that $E[T_{N_0}] < E[\mathcal{Z}_{\mathcal{N}}]$. Combining this with Lemma 3, we obtain

$$E[T_{N_0}] < \frac{1}{\lambda}(\mu E[T_{N_0}] - N_0 + 1). \tag{7}$$

Rearranging terms in (7) and using the fact $\lambda < \mu$, we obtain the inequality in question.

4.2. *A bound on the upper tail of* T_{N_0}. In this subsection, we derive an upper bound on the upper tail of the distribution of T_{N_0}; we also prove some other results on T_{N_0} that follow from this bound.

Proposition 5. For any positive δ there exist positive numbers $C(\delta)$ and $\psi(\delta)$ such that

$$\Pr[T_{N_0} \geqq \tau_{N_0}(1+k\delta)] < C(\delta)\exp(-\psi(\delta)kN_0), \quad \text{for} \quad N_0, k = 1, \cdots,$$

where $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu-\lambda)$.

Proof. We fix a positive δ. Obviously, if the initially present customers and the ones to arrive during the time interval $[0, \tau_{N_0}(1+k\delta)]$ have a total service time that is smaller than $\tau_{N_0}(1+k\delta)$, then the queue is empty for some part of the time interval $[0, \tau_{N_0}(1+k\delta)]$, even though it may be non-empty at time $t = \tau_{N_0}(1+k\delta)$. Let Y_i be the service time of the ith customer, for $i = 1, \cdots$, and let $\mathcal{M}(\delta)$ be the number of arrivals during the time interval $[0, \tau_{N_0}(1+\delta)]$. Then,

$$\Pr[T_{N_0} \geqq \tau_{N_0}(1+k\delta)] \leqq \Pr\left[\sum_{i=1}^{\mathcal{M}(k\delta)+N_0} Y_i \geqq \tau_{N_0}(1+k\delta)\right]. \tag{8}$$

We define

$$q(\delta) \stackrel{\text{def}}{=} \lfloor \lambda\tau_{N_0}(1+\alpha\delta)\rfloor, \tag{9}$$

where α is a positive constant satisfying $1 < \alpha < \mu/\lambda$. Clearly,

$$\Pr\left[\sum_{i=1}^{\mathcal{M}(k\delta)+N_0} Y_i \geqq \tau_{N_0}(1+k\delta)\right] \leqq \Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \tau_{N_0}(1+k\delta)\right] + \Pr[\mathcal{M}(k\delta) \geqq q(k\delta)+1]. \tag{10}$$

In what follows, each of the two terms appearing in the right-hand side of (10) is appropriately upper bounded.

Since $\mathcal{M}(k\delta)$ is the number of arrivals during the interval $[0, \tau_{N_0}(1+k\delta)]$, we have

$$\begin{aligned}\Pr[\mathcal{M}(k\delta) \geqq q(k\delta)+1] &= \Pr\left[\sum_{i=1}^{q(k\delta)+1} Z_i \leqq \tau_{N_0}(1+k\delta)\right] \\ &= \Pr\left[\frac{1}{q(k\delta)+1}\left(\sum_{i=1}^{q(k\delta)+1} Z_i\right) \leqq \frac{\tau_{N_0}(1+k\delta)}{q(k\delta)+1}\right],\end{aligned} \tag{11}$$

where Z_i denotes the ith interarrival time. It follows from the definitions of $q(\delta)$ (in (9)) and τ_{N_0} that

$$\frac{\tau_{N_0}(1+k\delta)}{q(k\delta)+1} < \frac{1+k\delta}{\lambda(1+\alpha k\delta)}.$$

Furthermore, since $\alpha > 1$, we have

$$\frac{1+k\delta}{\lambda(1+\alpha k\delta)} < \frac{1+\delta}{\lambda(1+\alpha\delta)} = \frac{1}{\lambda}\left(1 - \frac{\delta(\alpha-1)}{1+\alpha\delta}\right)$$

for $k = 2, \cdots$. Combining these two inequalities with (11), we obtain

$$\text{(12)}\quad \Pr[\mathcal{M}(k\delta) \geqq q(k\delta)+1] \leqq \Pr\left[\sum_{i=1}^{q(k\delta)+1} Z_i \leqq \lambda^{-1}\left(1-\frac{\delta(\alpha-1)}{1+\alpha\delta}\right)(q(k\delta)+1)\right].$$

The random variables $(Z_i)_{i=1,\cdots,q(k\delta)+1}$ are independent and have the interarrival time distribution, which, by assumption, is of the exponential type; moreover, since $\alpha > 1$, we have $\delta(\alpha-1)/(1+\alpha\delta) > 0$. Thus, we may upper bound the right-hand quantity in (12) by applying the Chernoff bound; using (5) in Section 2.1, we have

$$\text{(13)}\quad \Pr\left[\sum_{i=1}^{q(k\delta)+1} Z_i \leqq \lambda^{-1}\left(1-\frac{\delta(\alpha-1)}{1+\alpha\delta}\right)(q(k\delta)+1)\right] \leqq \exp(-\varphi_1(\delta)(q(k\delta)+1)),$$

where $\varphi_1(\delta) > 0$. Since $\varphi_1(\delta)$ is positive, it follows from (9) that

$$\text{(14)}\quad \begin{aligned}\exp(-\varphi_1(\delta)(q(k\delta)+1)) &< \exp(-\varphi_1(\delta)\lambda\tau_{N_0}(1+\alpha k\delta)) \\ &< \exp(-\varphi_1(\delta)\lambda\tau_{N_0}\alpha k\delta).\end{aligned}$$

We define $\psi_1(\delta) \overset{\text{def}}{=} \varphi_1(\delta)\lambda\alpha\delta/(\mu-\lambda)$. Since $\lambda < \mu$ and $\varphi_1(\delta) > 0$, we have $\psi_1(\delta) > 0$. Combining the previous definition with (13), (14) and the fact $\tau_{N_0} \overset{\text{def}}{=} N_0/(\mu-\lambda)$, we obtain

$$\Pr\left[\sum_{i=1}^{q(k\delta)+1} Z_i \leqq \lambda^{-1}\left(1-\frac{\delta(\alpha-1)}{1+\alpha\delta}\right)(q(k\delta)+1)\right] < \exp(-\psi_1(\delta)kN_0).$$

Using (12) and the inequality above, we have

$$\text{(15)}\quad \Pr[\mathcal{M}(k\delta) \geqq q(k\delta)+1] < \exp(-\psi_1(\delta)kN_0).$$

The other term in the right-hand side of (10) may be upper bounded by reasoning similarly. First, we have

$$\text{(16)}\quad \Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \tau_{N_0}(1+k\delta)\right] = \Pr\left[\frac{1}{q(k\delta)+N_0}\left(\sum_{i=1}^{q(k\delta)+N_0} Y_i\right) \geqq \frac{\tau_{N_0}(1+k\delta)}{q(k\delta)+N_0}\right].$$

Using the definitions of $q(\delta)$ (in (9)) and τ_{N_0}, we obtain after some algebra

$$\frac{\tau_{N_0}(1+k\delta)}{q(k\delta)+N_0} \geqq \frac{1+k\delta}{\mu+\lambda\alpha k\delta}.$$

Furthermore, since $\alpha < \mu/\lambda$, we have

$$\frac{1+k\delta}{\mu+\lambda\alpha k\delta} > \frac{1+\delta}{\mu+\lambda\alpha\delta} = \frac{1}{\mu}\left(1+\frac{\delta(\mu-\lambda\alpha)}{\mu+\lambda\alpha\delta}\right) \quad \text{for} \quad k = 2, \cdots.$$

Using these two inequalities and (16), we obtain

$$\Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \tau_{N_0}(1+k\delta)\right] \leqq \Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \mu^{-1}\left(1+\frac{\delta(\mu-\lambda\alpha)}{\mu+\lambda\alpha\delta}\right)(q(k\delta)+N_0)\right]. \tag{17}$$

The random variables $(Y_i)_{i=1,\cdots,q(k\delta)+N_0}$ are independent and have the service time distribution, which, by assumption, is of the exponential type; moreover, since $\alpha<\mu/\lambda$, we have $\delta(\mu-\alpha\lambda)/(\mu+\lambda\alpha\delta)>0$. Thus, we may upper bound the right-hand quantity in (17) by applying the Chernoff bound; using (4) in Section 2.1, we obtain

$$\Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \mu^{-1}\left(1+\frac{\delta(\mu-\lambda\alpha)}{\mu+\lambda\alpha\delta}\right)(q(k\delta)+N_0)\right] \leqq \exp\left(-\varphi_2(\delta)(q(k\delta)+N_0)\right), \tag{18}$$

where $\varphi_2(\delta)>0$. Since $\varphi_2(\delta)$ is positive, it follows from (9) that

$$\begin{aligned}\exp\left(-\varphi_2(\delta)(q(k\delta)+N_0)\right) &< \exp\left(-\varphi_2(\delta)(\lambda\tau_{N_0}(1+\alpha k\delta)+N_0-1)\right)\\ &< \exp\left(\varphi_2(\delta)\right)\exp\left(-\varphi_2(\delta)\lambda\tau_{N_0}\alpha k\delta\right).\end{aligned} \tag{19}$$

We define

$$C_2(\delta)\overset{\text{def}}{=}\exp\left(\varphi_2(\delta)\right) \quad \text{and} \quad \psi_2(\delta)\overset{\text{def}}{=}\varphi_2(\delta)\alpha\delta/(\mu-\lambda). \tag{20}$$

Since $\lambda<\mu$ and $\alpha>0$, we have $\psi_2(\delta)>0$. Combining (18) with (19), (20) and the fact $\tau_{N_0}\overset{\text{def}}{=}N_0/(\mu-\lambda)$, we obtain

$$\Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \mu^{-1}\left(1+\frac{\delta(\mu-\alpha\lambda)}{\mu+\lambda\alpha\delta}\right)(q(k\delta)+N_0)\right] < C_2(\delta)\exp\left(-\psi_2(\delta)kN_0\right).$$

This together with (17) implies that

$$\Pr\left[\sum_{i=1}^{q(k\delta)+N_0} Y_i \geqq \tau_{N_0}(1+k\delta)\right] < C_2(\delta)\exp\left(-\psi_2(\delta)kN_0\right). \tag{21}$$

It follows from (8), (10), (15) and (21) that

$$\Pr\left[T_{N_0}\geqq\tau_{N_0}(1+k\delta)\right] < \exp\left(-\psi_1(\delta)kN_0\right)+C_2(\delta)\exp\left(-\psi_2(\delta)kN_0\right).$$

After defining $\psi(\delta)\overset{\text{def}}{=}\min\left\{\psi_1(\delta),\psi_2(\delta)\right\}$ and $C(\delta)\overset{\text{def}}{=}1+C_2(\delta)$, the result follows from the inequality above.

It is a consequence of the proposition above that the random variables $(T_n)_{n=1,\cdots}$ are of the exponential type; moreover, their moment generating functions have a common interval of definition in the positive axis. Indeed, we have the following result.

Corollary 6. There exists a positive s^* such that

$$E[\exp(sT_{N_0})] < \infty, \qquad \text{for } N_0 = 1, \cdots, \text{ and } \forall s \in (0, s^*).$$

Proof. We fix a positive δ. We have

$$\begin{aligned} E[\exp(sT_{N_0})] &= E[\exp(sT_{N_0}) \mid T_{N_0} < \tau_{N_0}(1+\delta)] \cdot \Pr[T_{N_0} < \tau_{N_0}(1+\delta)] \\ &\quad + \sum_{k=1}^{\infty} \{E[\exp(sT_{N_0}) \mid \tau_{N_0}(1+k\delta) \leqq T_{N_0} < \tau_{N_0}(1+(k+1)\delta)] \\ &\qquad \cdot \Pr[\tau_{N_0}(1+k\delta) \leqq T_{N_0} < \tau_{N_0}(1+(k+1)\delta)]\}. \end{aligned} \tag{22}$$

Obviously,

$$E[\exp(sT_{N_0}) \mid T_{N_0} < \tau_{N_0}(1+\delta)] < \exp(s\tau_{N_0}(1+\delta)), \qquad \forall s > 0,$$

and

$$E[\exp(sT_{N_0}) \mid \tau_{N_0}(1+k\delta) \leqq T_{N_0} < \tau_{N_0}(1+(k+1)\delta)] < \exp(s\tau_{N_0}(1+(k+1)\delta)),$$
$$\text{for } k = 1, \cdots, \text{ and } \forall s > 0.$$

Combining these inequalities with (22), we obtain

$$\begin{aligned} E[\exp(sT_{N_0})] &< \exp(s\tau_{N_0}(1+\delta)) + \sum_{k=1}^{\infty} \{\exp(s\tau_{N_0}(1+(k+1)\delta)) \\ &\qquad \cdot \Pr[\tau_{N_0}(1+k\delta) \leqq T_{N_0} < \tau_{N_0}(1+(k+1)\delta)]\} \\ &\leqq \exp(s\tau_{N_0}(1+\delta)) \\ &\quad + \exp(s\tau_{N_0}(1+\delta)) \sum_{k=1}^{\infty} \{\exp(s\tau_{N_0}k\delta) \cdot \Pr[T_{N_0} \geqq \tau_{N_0}(1+k\delta)]\}. \end{aligned} \tag{23}$$

Using Proposition 5, it follows from (23) that

$$\begin{aligned} E[\exp(sT_{N_0})] &< \exp(s\tau_{N_0}(1+\delta)) \\ &\quad + \exp(s\tau_{N_0}(1+\delta))C(\delta) \sum_{k=1}^{\infty} \exp(s\tau_{N_0}k\delta) \exp(-\psi(\delta)kN_0) \\ &= \exp(s\tau_{N_0}(1+\delta)) \\ &\quad + \exp(s\tau_{N_0}(1+\delta))C(\delta) \sum_{k=1}^{\infty} [\exp(s\tau_{N_0}\delta - \psi(\delta)N_0)]^k. \end{aligned} \tag{24}$$

Since $\psi(\delta) > 0$ and $\tau_{N_0} \overset{\text{def}}{=} N_0/(\mu - \lambda)$, we have $0 < \exp(s\tau_{N_0}\delta - \psi(\delta)N_0) < 1$ for all $s \in (0, s^*)$, where s^* is defined by $s^* \overset{\text{def}}{=} \psi(\delta)(\mu - \lambda)/\delta > 0$. Therefore, the geometric series in the lower part of (24) is convergent for all $s \in (0, s^*)$; this implies that $E[\exp(sT_{N_0})] < \infty$ for all $s \in (0, s^*)$. In particular, we have

$$E[\exp(sT_{N_0})] < \exp(s\tau_{N_0}(1+\delta)) + \exp(s\tau_{N_0}(1+\delta))C(\delta) \frac{1}{\exp\left(\left[\psi(\delta) - s\dfrac{\delta}{\mu - \lambda}\right]N_0\right) - 1}, \qquad \forall s \in (0, s^*).$$

Since T_{N_0} is of the exponential type, it follows from the discussion in Section 2.1 that $E[T_{N_0}]$ is finite for $N_0 = 1, \cdots$. In particular, reasoning as in the proof of Corollary 6, it can be shown that $E[T_{N_0}]$ is close to τ_{N_0} for sufficiently large N_0. Indeed, we have the following result.

Corollary 7. We have $E[T_{N_0}] < \infty$ for $N_0 = 1, \cdots$. Furthermore, for any positive δ, there exists some $n'(\delta) \geqq 1$ such that

$$E[T_{N_0}] \leqq \tau_{N_0}(1 + \delta), \qquad \forall N_0 \geqq n'(\delta),$$

where $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu - \lambda)$.

Using Proposition 4 and Corollary 7, it is easily established that

$$\lim_{N_0 \to \infty} \frac{E[T_{N_0}]}{N_0} = \frac{1}{\mu - \lambda}.$$

However, we are interested in a stronger result, namely

$$\lim_{N_0 \to \infty} \frac{T_{N_0}}{N_0} \stackrel{\text{a.s.}}{=} \frac{1}{\mu - \lambda}.$$

In order to prove this, we also need an upper bound on the lower tail of the distribution of T_{N_0}; such a result is presented in the next subsection.

4.3. *A bound on the lower tail of* T_{N_0}

Proposition 8. For any positive δ there exist some $n(\delta) \geqq 1$ and some positive $\xi(\delta)$ such that

$$\Pr[T_{N_0} \leqq \tau_{N_0}(1 - \delta)] < \exp(-\xi(\delta)N_0), \qquad \forall N_0 \geqq n(\delta),$$

where $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu - \lambda)$.

Proof. Clearly, it suffices to establish the result only for those δ in the interval $(0, 1)$.

We fix a δ satisfying $0 < \delta < 1$. Let $\mathcal{M}(\delta)$ denote the number of arrivals during the time interval $[0, \tau_{N_0}(1 - \delta)]$, where $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu - \lambda)$. Moreover, let Y_i denote the service time of the ith customer and Z_i denote the ith interarrival time. We define the sequence $(X_k)_{k=0,\ldots}$ of random variables as follows:

$$X_0 \stackrel{\text{def}}{=} \sum_{i=1}^{N_0-1} Y_i \quad \text{and} \quad X_k \stackrel{\text{def}}{=} \sum_{i=1}^{N_0+k-1} Y_i - \sum_{i=1}^{k} Z_i, \qquad \text{for } k = 1, \cdots. \tag{25}$$

Clearly, the k^*th customer to arrive is the first to meet an empty system upon arrival if and only if $X_1 \geqq 0, \cdots, X_{k^*-1} \geqq 0$ and $X_{k^*} < 0$ (for $k^* = 1$ the condition is $X_1 < 0$). Given the event $\mathcal{M}(\delta) = m$, the system is empty for some part of the time interval $[0, \tau_{N_0}(1 - \delta)]$ if and only if at least one of the first $m + 1$ customers to

arrive meets an empty system upon arrival. Therefore, we have

$$\Pr[T_{N_0} \leqq \tau_{N_0}(1-\delta) \mid \mathcal{M}(\delta) = m] = \Pr[\exists k \in \{1, \cdots, m+1\} : X_k < 0 \mid \mathcal{M}(\delta) = m], \qquad \text{for } m = 0, \cdots$$

or equivalently,

$$\Pr[T_{N_0} \leqq \tau_{N_0}(1-\delta) \mid \mathcal{M}(\delta) = m] = \Pr\left[\min_{1 \leqq k \leqq m+1} \{X_k\} < 0 \mid \mathcal{M}(\delta) = m\right], \qquad \text{for } m = 0, \cdots. \tag{26}$$

We define m^* by

$$m^* \stackrel{\text{def}}{=} \left\lceil \lambda \tau_{N_0}\left(1 - \frac{\delta}{2}\right) \right\rceil. \tag{27}$$

Using (26), we obtain after some algebra

$$\Pr[T_{N_0} \leqq \tau_{N_0}(1-\delta)] \leqq \sum_{m=0}^{m^*-1} \Pr\left[\min_{1 \leqq k \leqq m+1} \{X_k\} < 0 \text{ and } \mathcal{M}(\delta) = m\right] + \Pr[\mathcal{M}(\delta) \geqq m^*]. \tag{28}$$

Clearly, we have

$$\min_{1 \leqq k \leqq m+1} \{X_k\} \geqq \min_{1 \leqq k \leqq m^*} \{X_k\}, \qquad \text{for } m = 0, \cdots, m^* - 1.$$

Combining this with (28), we obtain

$$\begin{aligned} \Pr[T_{N_0} \leqq \tau_{N_0}(1-\delta)] &\leqq \sum_{m=0}^{m^*-1} \Pr\left[\min_{1 \leqq k \leqq m^*} \{X_k\} < 0 \text{ and } \mathcal{M}(\delta) = m\right] + \Pr[\mathcal{M}(\delta) \geqq m^*] \\ &= \Pr\left[\min_{1 \leqq k \leqq m^*} \{X_k\} < 0 \text{ and } \mathcal{M}(\delta) \leqq m^* - 1\right] + \Pr[\mathcal{M}(\delta) \geqq m^*] \\ &\leqq \Pr\left[\min_{1 \leqq k \leqq m^*} \{X_k\} < 0\right] + \Pr[\mathcal{M}(\delta) \geqq m^*]. \end{aligned} \tag{29}$$

In what follows, each of the two terms appearing in the lower part of (29) is appropriately upper bounded.

Since $\mathcal{M}(\delta)$ is the number of arrivals during the interval $[0, \tau_{N_0}(1-\delta)]$, we have

$$\Pr[\mathcal{M}(\delta) \geqq m^*] = \Pr\left[\sum_{i=1}^{m^*} Z_i \leqq \tau_{N_0}(1-\delta)\right] = \Pr\left[\frac{1}{m^*} \sum_{i=1}^{m^*} Z_i \leqq \frac{\tau_{N_0}(1-\delta)}{m^*}\right]. \tag{30}$$

Using the definition of m^* in (27) and that of τ_{N_0}, we have

$$\frac{\tau_{N_0}(1-\delta)}{m^*} \leqq \lambda^{-1} \frac{1-\delta}{1-\delta/2} < \lambda^{-1}\left(1 - \frac{\delta}{2}\right). \tag{31}$$

Combining (30) with (31), we have

$$\Pr[\mathcal{M}(\delta) \geqq m^*] \leqq \Pr\left[\sum_{i=1}^{m^*} Z_i \leqq \lambda^{-1}\left(1-\frac{\delta}{2}\right)m^*\right]. \tag{32}$$

The random variables $(Z_i)_{i=1,\cdots,m^*}$ are independent and have the interarrival time distribution, which, by assumption, is of the exponential type. Thus, since $\delta > 0$, we may upper bound the right-hand quantity in (32) by applying the Chernoff bound; using (5) in Section 2.1, we obtain

$$\Pr\left[\sum_{i=1}^{m^*} Z_i \leqq \lambda^{-1}\left(1-\frac{\delta}{2}\right)m^*\right] \leqq \exp\left(-\varphi_1(\delta)m^*\right), \tag{33}$$

where $\varphi_1(\delta) > 0$. Since $\varphi_1(\delta)$ is positive, it follows from (27) that

$$\exp\left(-\varphi_1(\delta)m^*\right) < \exp\left(-\varphi_1(\delta)\lambda\tau_{N_0}\left(1-\frac{\delta}{2}\right)\right). \tag{34}$$

We define

$$\phi_1(\delta) \stackrel{\text{def}}{=} \varphi_1(\delta)\frac{\lambda}{\mu-\lambda}\left(1-\frac{\delta}{2}\right).$$

Since $\lambda < \mu$ and $\varphi_1(\delta) > 0$, we have $\phi_1(\delta) > 0$. Combining the previous definitions with (33), (34) and the fact $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu-\lambda)$, we obtain

$$\Pr\left[\sum_{i=1}^{m^*} Z_i \leqq \lambda^{-1}\left(1-\frac{\delta}{2}\right)m^*\right] \leqq \exp\left(-\phi_1(\delta)N_0\right).$$

This together with (32) implies that

$$\Pr[\mathcal{M}(\delta) \geqq m^*] < \exp\left(-\phi_1(\delta)N_0\right). \tag{35}$$

We now consider the other term. We have

$$\begin{aligned}\Pr&\left[\min_{1\leq k\leq m^*}\{X_k\} < 0\right] \\ &= \Pr\left[\min_{1\leq k\leq m^*}\{X_k - X_0\} < -X_0\right] = E\left[\Pr\left[\min_{1\leq k\leq m^*}\{X_k - X_0\} < -X_0 \mid X_0\right]\right].\end{aligned} \tag{36}$$

Let α be some constant satisfying $0 < \alpha < \frac{1}{2}$. It follows from (36) that

$$\begin{aligned}\Pr&\left[\min_{1\leq k\leq m^*}\{X_k\} < 0\right] \\ &= \Pr\left[X_0 > N_0\mu^{-1}(1-\alpha\delta)\right] \\ &\quad \cdot E\left[\Pr\left[\min_{1\leq k\leq m^*}\{X_k - X_0\} < -X_0 \mid X_0\right] \,\middle|\, X_0 > N_0\mu^{-1}(1-\alpha\delta)\right] \\ &\quad + \Pr\left[X_0 \leqq N_0\mu^{-1}(1-\alpha\delta)\right] \\ &\quad \cdot E\left[\Pr\left[\min_{1\leq k\leq m^*}\{X_k - X_0\} < -X_0 \mid X_0\right] \,\middle|\, X_0 \leqq N_0\mu^{-1}(1-\alpha\delta)\right].\end{aligned} \tag{37}$$

We define the random variables $(V_k)_{k=0,\cdots,m^*}$ as follows: $V_k \stackrel{\text{def}}{=} -X_k + X_0$ for $k = 0, \cdots, m^*$. Using this definition and (25), we obtain

$$V_0 = 0 \quad \text{and} \quad V_k = \sum_{i=1}^{k} (-Y_{i+N_0-1} + Z_i), \qquad \text{for } k = 1, \cdots, m^*. \tag{38}$$

Obviously, we have

$$\Pr\left[\min_{1\leq k\leq m^*} \{X_k - X_0\} < -X_0 \mid X_0\right] = \Pr\left[\min_{1\leq k\leq m^*} \{-V_k\} < -X_0 \mid X_0\right] = \Pr\left[\max_{1\leq k\leq m^*} \{V_k\} > X_0 \mid X_0\right]. \tag{39}$$

However, it follows from (38) and (25) that the random variables $V_1, \cdots, V_{m^*}$ are independent of X_0. Using this and (39), we obtain

$$\Pr\left[\min_{1\leq k\leq m^*} \{X_k - X_0\} < -X_0 \mid X_0 = x_0\right] = \Pr\left[\max_{1\leq k\leq m^*} \{V_k\} > x_0\right], \qquad \forall x_0 \geq 0. \tag{40}$$

Furthermore, we have

$$\Pr\left[\max_{1\leq k\leq m^*} \{V_k\} > x_0\right] \leq \Pr\left[\max_{1\leq k\leq m^*} \{V_k\} > N_0\mu^{-1}(1-\alpha\delta)\right], \qquad \forall x_0 > N_0\mu^{-1}(1-\alpha\delta).$$

Combining this with (40), we obtain

$$E\left[\Pr\left[\min_{1\leq k\leq m^*} \{X_k - X_0\} < -X_0 \mid X_0\right] \,\middle|\, X_0 > N_0\mu^{-1}(1-\alpha\delta)\right] \leq \Pr\left[\max_{1\leq k\leq m^*} \{V_k\} > N_0\mu^{-1}(1-\alpha\delta)\right].$$

Using this and (37), we have

$$\Pr\left[\min_{1\leq k\leq m^*} \{X_k\} < 0\right] \leq \Pr\left[\max_{1\leq k\leq m^*} \{V_k\} > N_0\mu^{-1}(1-\alpha\delta)\right] + \Pr\left[X_0 \leq N_0\mu^{-1}(1-\alpha\delta)\right]. \tag{41}$$

In what follows, each of the terms appearing in the right-hand side of (41) is appropriately upper bounded.

We define $n_2(\delta) \stackrel{\text{def}}{=} \max\{2/(\alpha\delta) - 1, 2\}$. It follows that $N_0\mu^{-1}(1-\alpha\delta) \leq (N_0 - 1)\mu^{-1}(1-(\alpha/2)\delta)$ (and $N_0 - 1 > 0$) for all $N_0 \geq n_2(\delta), \cdots$. Therefore,

$$\Pr\left[X_0 \leq N_0\mu^{-1}(1-\alpha\delta)\right] \leq \Pr\left[X_0 \leq (N_0-1)\mu^{-1}\left(1-\frac{\alpha}{2}\delta\right)\right], \qquad \forall N_0 \geq n_2(\delta). \tag{42}$$

It follows from the definition of X_0 [see (25)] that the lower tail of its distribution may be upper bounded in the way presented in Section 2.1. (Recall that service

5. The main results on the $GI/G/1$ queue

The main results of this paper are presented in this section.

5.1. *Asymptotic linearity of T_{N_0} and the fluid analogy.*

Proposition 9. The sequence $(T_n/n)_{n=1,\dots}$ of random variables converges exponentially to $1/(\mu-\lambda)$, as $n\to\infty$. Furthermore, if these random variables are defined on the same probability space, then convergence holds in the 'a.s.' sense as well.

Proof. We fix some positive δ. Using Proposition 5 (applied with $k=1$) and Proposition 8 (and recalling that $\tau_{N_0}\stackrel{\text{def}}{=}N_0/(\mu-\lambda)$), we have

$$\Pr\left[\left|\frac{T_{N_0}}{N_0}-\frac{1}{\mu-\lambda}\right|\geqq\frac{\delta}{\mu-\lambda}\right]<C(\delta)\exp(-\psi(\delta)N_0)+\exp(-\xi(\delta)N_0),\qquad \forall N_0\geqq n(\delta).$$

Defining

$$\gamma(\delta)\stackrel{\text{def}}{=}\tfrac{1}{2}\min\{\psi(\delta),\xi(\delta)\}$$

and

$$n^*(\delta)\stackrel{\text{def}}{=}\max\left\{n(\delta),\frac{\ln(C(\delta)+1)}{\gamma(\delta)}\right\},$$

we obtain

$$\Pr\left[\left|\frac{T_{N_0}}{N_0}-\frac{1}{\mu-\lambda}\right|\geqq\frac{\delta}{\mu-\lambda}\right]<\exp(-\gamma(\delta)N_0),\qquad \forall N_0\geqq n^*(\delta). \tag{49}$$

This implies that the sequence $(T_n/n)_{n=1,\dots}$ of random variables converges exponentially to $1/(\mu-\lambda)$, as $n\to\infty$ (see Section 2.2). As already mentioned in Section 2.2, if the random variables $(T_n)_{n=1,\dots}$ are defined on the same probability space (which is always possible), then exponential convergence implies almost sure convergence.

Before proceeding to the other two main results, we establish a technical lemma. We define the random variable R_{N_0} as follows:

$$R_{N_0}\stackrel{\text{def}}{=}\inf_{t\geqq T_{N_0}}\{t:N(t)=1\}. \tag{50}$$

Clearly, R_{N_0} corresponds to the first regeneration point of the queue. It is intuitively clear that, for sufficiently large N_0, the upper tail of R_{N_0} behaves similarly to that of T_{N_0} (see Proposition 5). We now present this result; in order not to break continuity, we give the proof of this technical lemma in Section 5.3.

Lemma 10. For any positive δ there exists some $l(\delta)$ and some positive $\phi(\delta)$ such that

$$\Pr[R_{N_0}\geqq\tau_{N_0}(1+\delta)]\leqq\exp(-\phi(\delta)N_0),\qquad \forall N_0\geqq l(\delta).$$

In Section 1 we introduced a simple fluid model, namely a pool that initially contains a quantity N_0 of fluid; in this pool, fluid is supplied at a constant rate λ and,

at the same time, it is removed at a constant rate μ. As already pointed out in Section 1, the result in Proposition 9 is reminiscent of the fact that the pool empties in time $N_0/(\mu-\lambda)$. This analogy may be extended even further. Indeed, the aforementioned pool contains a quantity $(1-c)N_0$ of fluid at time $cN_0/(\mu-\lambda)$ for all $c \in [0, 1]$; moreover, it contains no fluid at time $cN_0/(\mu-\lambda)$ for all $c > 1$. The analogous results for the type of $GI/G/1$ queue under analysis are as follows.

Proposition 11. The following convergence results hold in the exponential sense:

$$\lim_{N_0 \to \infty} \frac{N(\tau_{N_0} c)}{N_0} = \begin{cases} 1-c, & \text{for } 0 \leqq c \leqq 1; \\ 0, & \text{for } c > 1, \end{cases}$$

where $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu-\lambda)$. Furthermore, if the random variables involved are defined on the same probability space, then the above results hold in the 'a.s.' sense of convergence, as well.

Outline of the proof. The result is obvious for $c = 0$.

(a) We fix some $c \in (0, 1]$. We have to show that for any positive ε there exists some $m(\varepsilon) \geqq 1$ and some positive $\phi(\varepsilon)$ (both of which may possibly depend on c) such that

$$\Pr[|N(\tau_{N_0} c) - (1-c)N_0| \geqq \varepsilon N_0] \leqq \exp(-\phi(\varepsilon)N_0), \qquad \forall N_0 \geqq m(\varepsilon).$$

Assuming that $N_0 \geqq 1/c$, we introduce the following priority scheme: we choose $\lceil (1-c)N_0 \rceil$ customers among those initially waiting to be served; these customers are assigned the lowest priority; priority assignment is irrelevant to the customers' service times. (In fact, no priority scheme is introduced for $c = 1$.) Therefore, the priority scheme introduced does not affect the statistics of the process $N(t)$ [14]. We define

$$\tilde{T} \stackrel{\text{def}}{=} \inf_{t \geqq 0} \{t : N(t) = \lceil (1-c)N_0 \rceil\}.$$

Since low-priority customers are 'transparent' to the ones with higher priority, we have

$$\tilde{T} \stackrel{\text{st}}{=} T_{\lfloor cN_0 \rfloor}. \tag{51}$$

(This is because $\lceil (1-c)N_0 \rceil + \lfloor cN_0 \rfloor = N_0$.) We fix a positive ε. For any positive δ, we have

$$\begin{aligned} \Pr[|N(\tau_{N_0} c) - (1-c)N_0| \geqq \varepsilon N_0] \\ &= \Pr[|N(\tau_{N_0} c) - (1-c)N_0| \geqq \varepsilon N_0 \text{ and } |\tilde{T} - \tau_{N_0} c| \geqq \delta N_0] \\ &\quad + \Pr[|N(\tau_{N_0} c) - (1-c)N_0| \geqq \varepsilon N_0 \text{ and } 0 \leqq \tilde{T} - \tau_{N_0} c < \delta N_0] \\ &\quad + \Pr[|N(\tau_{N_0} c) - (1-c)N_0| \geqq \varepsilon N_0 \text{ and } -\delta N_0 < \tilde{T} - \tau_{N_0} c < 0], \end{aligned}$$

which implies that

$$\begin{aligned}\Pr[|N(\tau_{N_0}c)-(1-c)N_0|\geqq\varepsilon N_0]&\leqq\Pr[|\tilde{T}-\tau_{N_0}c|\geqq\delta N_0]\\&\quad+\Pr[|N(\tau_{N_0}c)-(1-c)N_0|\geqq\varepsilon N_0 \text{ and } 0\leqq\tilde{T}-\tau_{N_0}c<\delta N_0]\\&\quad+\Pr[|N(\tau_{N_0}c)-(1-c)N_0|\geqq\varepsilon N_0|-\delta N_0<\tilde{T}-\tau_{N_0}c<0].\end{aligned}\tag{52}$$

Each of the three terms in the right-hand side of (52) will be upper bounded by some quantity that decays exponentially, as $N_0\to\infty$. This may easily be done for the first term. It suffices to combine (51) with (49).

Now, we consider the second term. Since the system contains more than $\lceil(1-c)N_0\rceil$ customers at any instant prior to $\tilde{T}$, we have

$$\begin{aligned}&\Pr[|N(\tau_{N_0}c)-(1-c)N_0|\geqq\varepsilon N_0 \text{ and } 0\leqq\tilde{T}-\tau_{N_0}c<\delta N_0]\\&\quad=\Pr[N(\tau_{N_0}c)-(1-c)N_0\geqq\varepsilon N_0 \text{ and } 0\leqq\tilde{T}-\tau_{N_0}c<\delta N_0]\\&\quad\leqq\Pr[\tilde{T}-\tau_{N_0}c<\delta N_0\,|\,N(\tau_{N_0}c)\geqq(\varepsilon+1-c)N_0 \text{ and } \tilde{T}\geqq\tau_{N_0}c].\end{aligned}\tag{53}$$

The quantity in the lower part of (53) equals the probability that the system contains exactly $\lceil(1-c)N_0\rceil$ customers at some time in the interval $(\tau_{N_0}c,\ \tau_{N_0}c+\delta N_0)$ even though there are at least $\lceil(\varepsilon+1-c)N_0\rceil$ customers present at time τ_{N_0}. Since $\lceil(\varepsilon+1-c)N_0\rceil-\lceil(1-c)N_0\rceil\geqq\lceil\varepsilon N_0\rceil-1$, the quantity in the lower part of (53) may be upper bounded by the probability that at least $\lceil\varepsilon N_0\rceil-1$ customers complete service during the time interval $[\tau_{N_0}c,\ \tau_{N_0}c+\delta N_0)$. Applying the Chernoff bound, it may be shown that the latter quantity decays exponentially, as $N_0\to\infty$, provided that $\delta<\varepsilon\mu^{-1}$.

Finally, we consider the third term in the right-hand side of (52). We have

$$\begin{aligned}&\Pr[|N(\tau_{N_0}c)-(1-c)N_0|\geqq\varepsilon N_0\,|-\delta N_0<\tilde{T}-\tau_{N_0}c<0]\\&\quad=\Pr[N(\tau_{N_0}c)\geqq(\varepsilon+1-c)N_0\,|-\delta N_0<\tilde{T}-\tau_{N_0}c<0]\\&\qquad+\Pr[N(\tau_{N_0}c)\leqq(-\varepsilon+1-c)N_0\,|-\delta N_0<\tilde{T}-\tau_{N_0}c<0].\end{aligned}\tag{54}$$

The first term on the right-hand side of (54) may be upper bounded by the probability that at least $\lceil\varepsilon N_0\rceil-1$ customers arrive during the time interval $(-\delta N_0+\tau_{N_0}c,\ \tau_{N_0}c)$. This decays exponentially, as $N_0\to\infty$, provided that $\delta<\varepsilon\lambda^{-1}$. (The arguments for this are similar to those in the previous paragraph.) As for the other term, it equals 0 in case of $\varepsilon>1-c$; in case of $\varepsilon\leqq1-c$, it may be upper bounded by the probability that at least $\lceil\varepsilon N_0\rceil$ customers complete service during the time interval $(-\delta N_0+\tau_{N_0}c,\ \tau_{N_0}c)$. This decays exponentially, as $N_0\to\infty$, provided that $\delta<\varepsilon\mu^{-1}$.

(b) We fix some $c>1$. We have to show that for any positive ε there exists some $m(\varepsilon)\geqq1$ and some positive $\phi(\varepsilon)$ such that

$$\Pr[N(\tau_{N_0}c)\geqq\varepsilon N_0]\leqq\exp(-\phi(\varepsilon)N_0),\qquad\forall N_0\geqq m(\varepsilon).\tag{55}$$

Let R_{N_0} be the random variable defined in (50). We fix a positive ε and a δ satisfying $0<\delta<c-1$. Reasoning as in similar cases, we obtain

$$\begin{aligned}\Pr[N(\tau_{N_0}c)\geqq\varepsilon N_0]&\leqq\Pr[R_{N_0}\geqq\tau_{N_0}(1+\delta)]\\&\quad+\Pr[N(\tau_{N_0}c)\geqq\varepsilon N_0\,|\,R_{N_0}<\tau_{N_0}(1+\delta)].\end{aligned}\tag{56}$$

The term $\Pr[R_{N_0} \geqq \tau_{N_0}(1+\delta)]$ decays exponentially, as $N_0 \to \infty$ (see Lemma 10). Furthermore, since $1+\delta < c$, and R_{N_0} is a regeneration point of the queue, we have

$$\Pr[N(\tau_{N_0}c) \geqq \varepsilon N_0 \mid R_{N_0} = t] = \Pr[N(\tau_{N_0}c - t) \geqq \varepsilon N_0 \mid N(0) = 1], \qquad \forall t \in [0, \tau_{N_0}(1+\delta)). \tag{57}$$

It has been established in [8] that the queue under analysis is geometrically stable under the initial condition $N(0) = 1$. That is, for any sufficiently small positive γ there exists a $D(\gamma)$ such that $\Pr[N(t) \geqq M \mid N(0) = 1] \leqq D(\gamma)e^{-\gamma M}$ for all $t \geqq 0$ and $M = 1, \cdots$. Applying this with $M = \lceil \varepsilon N_0 \rceil$ and using (57), it follows that the second term in the right-hand side of (56) decays exponentially, as $N_0 \to \infty$. Combining this with (56) (and the conclusion following it), we obtain (55) after some algebra.

5.2. *The settling time.* Proposition 11 implies that, for $c < 1$, the number $N(\tau_{N_0}c)$ of customers contained in the system at time $\tau_{N_0}c$ is asymptotically $\Theta(N_0)$ (i.e., of the same order of magnitude as N_0), with high probability. Therefore, for any fixed $c \in (0, 1)$, at time $c\tau_{N_0}$ the queue is still away from steady-state for all sufficiently large N_0. Thus, the settling time of the queue may not be asymptotically smaller than $\tau_{N_0}c$, for any $c \in (0, 1)$. On the other hand, applying Proposition 11 with $c = 1$, we see that the number $N(\tau_{N_0})$ of customers contained in the system at time τ_{N_0} is asymptotically $o(N_0)$ (i.e., of smaller order of magnitude than N_0), with high probability. Thus, it is reasonable to expect that, for sufficiently large N_0, the time required for the queue to approach stationarity, starting from the time instant τ_{N_0}, is negligible as compared to τ_{N_0}.

The above discussion implies that the settling time of the $GI/G/1$ queue under analysis is asymptotically equal to τ_{N_0}. This is established in the proposition to follow.

Proposition 12. The following is true:

$$\lim_{N_0 \to \infty} d_{N_0}(\tau_{N_0}c) = \begin{cases} 1, & \text{for } 0 < c < 1; \\ 0, & \text{for } c > 1, \end{cases}$$

where $d_{N_0}(t)$ is defined at the end of Section 1 and $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu - \lambda)$.

Proof. We consider the cases $0 < c < 1$ and $c > 1$ separately.

(a) Let c be a constant satisfying $0 < c < 1$. First, we show that $\lim_{N_0 \to \infty} \Pr[N(\tau_{N_0}c) \leqq D] = 0$ for all $D \geqq 0$. We fix some non-negative D and some ε^* that satisfies $0 < \varepsilon^* < 1 - c$. It has been established in Proposition 10 that there exist some $m(\varepsilon^*) \geqq 1$ and some positive $\phi(\varepsilon^*)$ such that

$$\Pr[|N(\tau_{N_0}c) - (1-c)N_0| \geqq \varepsilon^* N_0] \leqq \exp(-\phi(\varepsilon^*)N_0), \qquad \forall N_0 \geqq m(\varepsilon^*),$$

which implies that

$$\Pr[N(\tau_{N_0}c) \leqq (1 - c - \varepsilon^*)N_0] \leqq \exp(-\phi(\varepsilon^*)N_0), \qquad \forall N_0 \geqq m(\varepsilon^*). \tag{58}$$

Since $1 - c - \varepsilon^* > 0$, we have $D \leqq (1 - c - \varepsilon^*)N_0$ for all $N_0 \geqq D/(1 - c - \varepsilon^*)$.

Combining this with (58), we have

$$\Pr[N(\tau_{N_0}c) \leqq D] \leqq \exp(-\phi(\varepsilon^*)N_0), \qquad \forall N_0 \geqq M^*,$$

where

$$M^* \stackrel{\text{def}}{=} \max\left\{m(\varepsilon^*), \frac{D}{1-c-\varepsilon^*}\right\}.$$

Clearly, this implies that

$$\lim_{N_0 \to \infty} \Pr[N(\tau_{N_0}c) \leqq D] = 0. \tag{59}$$

Next, let $(\pi_k)_{k=0,\ldots}$ be the steady-state distribution of the number of customers in the queue. We fix an ε satisfying $0 < \varepsilon < 1$. Since the queue under analysis is stable, there exists some $D(\varepsilon) \geqq 0$ such that $\sum_{k \leqq D(\varepsilon)} \pi_k \geqq 1 - \varepsilon$. This together with (59) implies that there exists some $L(\varepsilon)$ such that

$$|\Pr[N(\tau_{N_0}c) \in \mathscr{A}_\varepsilon] - \pi(\mathscr{A}_\varepsilon)| > 1 - 2\varepsilon, \qquad \forall N_0 \geqq L(\varepsilon), \tag{60}$$

where $\mathscr{A}_\varepsilon \stackrel{\text{def}}{=} \{k : k \leqq D(\varepsilon)\}$. We now notice that

$$|\Pr[N(\tau_{N_0}c) \in \mathscr{A}_\varepsilon] - \pi(\mathscr{A}_\varepsilon)| \leqq d_{N_0}(\tau_{N_0}c) \leqq 1.$$

This together with (60) [which holds for all $\varepsilon \in (0, 1)$] proves that $\lim_{N_0 \to \infty} d_{N_0}(\tau_{N_0}c) = 1$.

(b) Let c be a constant satisfying $c > 1$ and let δ be defined as $\delta \stackrel{\text{def}}{=} (c-1)/2$; clearly, we have $\delta > 0$. Using the alternative expression for the total variation distance (see Section 1), we obtain

$$d_{N_0}(\tau_{N_0}c) = \tfrac{1}{2} \sum_{k=0}^{\infty} |\Pr[N(\tau_{N_0}(1+2\delta)) = k] - \pi_k|. \tag{61}$$

We have

$$\begin{aligned}\Pr[N(\tau_{N_0}(1+2\delta)) = k] = {} & \Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} \leqq \tau_{N_0}(1+\delta)] \\ & + \Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} > \tau_{N_0}(1+\delta)],\end{aligned}$$

where R_{N_0} is the random variable defined in (50) of Section 5.1. Using this and the triangle inequality, it follows from (61) that

$$\begin{aligned} d_{N_0}(\tau_{N_0}c) \leqq {} & \tfrac{1}{2} \sum_{k=0}^{\infty} |\Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} \leqq \tau_{N_0}(1+\delta)] - \pi_k| \\ & + \tfrac{1}{2} \sum_{k=0}^{\infty} \Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} > \tau_{N_0}(1+\delta)]. \end{aligned} \tag{62}$$

In what follows, each of the two terms in the right-hand side of (62) is appropriately upper bounded.

Starting with the last term, we have

$$\sum_{k=0}^{\infty} \Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} > \tau_{N_0}(1+\delta)] = \Pr[R_{N_0} > \tau_{N_0}(1+\delta)]. \tag{63}$$

Using Lemma 10, we obtain

$$\Pr[R_{N_0} > \tau_{N_0}(1+\delta)] \leqq \exp(-\phi(\delta)N_0), \tag{64}$$

where $\phi(\delta) > 0$.

Now we consider the first term in the right-hand side of (62). Since R_{N_0} is a regeneration point of the queue, we have

$$\begin{aligned}\Pr[N(\tau_{N_0}(1+2\delta)) = k \mid R_{N_0} = t^*] \\ = \Pr[N(\tau_{N_0}(1+2\delta) - t^*) = k \mid N(0) = 1], \qquad \forall t^* \in [0, \tau_{N_0}(1+\delta)].\end{aligned} \tag{65}$$

We define

$$f_k(t) \stackrel{\text{def}}{=} \Pr[N(t) = k \mid N(0) = 1] - \pi_k, \qquad \forall t \geqq 0. \tag{66}$$

Clearly, we have $\sum_{k=0}^{\infty} f_k(t) = 0$ for all $t \geqq 0$; thus, it follows that

$$\sum_{k=0}^{\infty} |f_k(t)| = 2 \sum_{k \in \mathscr{F}_t^+} f_k(t), \qquad \forall t \geqq 0,$$

where $\mathscr{F}_t^+ \stackrel{\text{def}}{=} \{k : f_k(t) \geqq 0\}$. Clearly, (66) implies that $\sum_{k \in \mathscr{A}} f_k(t) \leqq 1$ for any $\mathscr{A} \subseteq Z_+$. Combining the previous two results, we obtain

$$\sum_{k=0}^{\infty} |f_k(t)| \leqq 2, \qquad \forall t \geqq 0. \tag{67}$$

Since $\pi_k \stackrel{\text{def}}{=} \lim_{t \to \infty} \Pr[N(t) = k \mid N(0) = 1]$, it follows from (66) that $\lim_{t \to \infty} f_k(t) = 0$ for $k = 0, \cdots$. Due to (67), we can apply the dominated convergence theorem and obtain

$$\lim_{t \to \infty} \sum_{k=0}^{\infty} |f_k(t)| = 0.$$

Therefore, for any positive ε there exists some $D(\varepsilon) \geqq 0$ such that

$$\sum_{k=0}^{\infty} |f_k(t)| < \varepsilon, \qquad \forall t \geqq D(\varepsilon).$$

Using the inequality above and the fact $\tau_{N_0}(1+2\delta) - t^* \geqq \delta N_0/(\mu - \lambda)$ for all $t^* \in [0, \tau_{N_0}(1+\delta)]$, we obtain

$$\sum_{k=0}^{\infty} |f_k(\tau_{N_0}(1+2\delta) - t^*)| < \varepsilon, \qquad \forall t^* \in [0, \tau_{N_0}(1+\delta)] \quad \text{and} \quad \forall N_0 \geqq \frac{\mu - \lambda}{\delta} D(\varepsilon).$$

Combining this with (65) and (66), we have

$$\begin{aligned}\sum_{k=0}^{\infty} |\Pr[N(\tau_{N_0}(1+2\delta)) = k \mid R_{N_0} = t^*] - \pi_k| < \varepsilon, \\ \forall t^* \in [0, \tau_{N_0}(1+\delta)] \quad \text{and} \quad \forall N_0 \geqq M(\varepsilon),\end{aligned} \tag{68}$$

where $M(\varepsilon)$ is defined by $M(\varepsilon) \stackrel{\text{def}}{=} ((\mu - \lambda))D(\varepsilon)/\delta$. We also have

$$\begin{aligned}(69)\quad &\Pr[N(\tau_{N_0}(1+2\delta)) = k \,|R_{N_0} \leqq \tau_{N_0}(1+\delta)] \\ &\qquad = E[\Pr[N(\tau_{N_0}(1+2\delta)) = k| R_{N_0}] \,|\, R_{N_0} \leqq \tau_{N_0}(1+\delta)].\end{aligned}$$

Using (68) and (69) (and the fact $|E[X]| \leqq E[|X|]$ for any random variable X), we obtain

$$(70)\quad \sum_{k=0}^{\infty} |\Pr[N(\tau_{N_0}(1+2\delta)) = k \mid R_{N_0} \leqq \tau_{N_0}(1+\delta)] - \pi_k| < \varepsilon, \qquad \forall N_0 \geqq M(\varepsilon).$$

It also follows from (64) that $1 - \exp(-\phi(\delta)N_0) \leqq \Pr[R_{N_0} \leqq \tau_{N_0}(1+\delta)] \leqq 1$. Using this, we obtain (after some algebra)

$$\begin{aligned}&|\Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} \leqq \tau_{N_0}(1+\delta)] - \pi_k| \\ &\qquad \leqq |\Pr[N(\tau_{N_0}(1+2\delta)) = k \mid R_{N_0} \leqq \tau_{N_0}(1+\delta)] - \pi_k| \\ &\qquad\quad + \exp(-\phi(\delta)N_0) \cdot \Pr[N(\tau_{N_0}(1+2\delta)) = k \mid R_{N_0} \leqq \tau_{N_0}(1+\delta)].\end{aligned}$$

Combining this with (70), we have

$$\sum_{k=0}^{\infty} |\Pr[N(\tau_{N_0}(1+2\delta)) = k \quad \text{and} \quad R_{N_0} \leqq \tau_{N_0}(1+\delta)] - \pi_k| < \varepsilon + \exp(-\phi(\delta)N_0), \qquad \forall N_0 \geqq M(\varepsilon).$$

This together with (62), (63) and (64) implies that for any positive ε there exists some $M(\varepsilon)$ such that

$$0 \leqq d_{N_0}(\tau_{N_0} c) < \frac{\varepsilon}{2} + \exp(-\phi(\delta)N_0), \qquad \forall N_0 \geqq M(\varepsilon).$$

Since $\phi(\delta) > 0$, it follows that $\lim_{N_0 \to \infty} d_{N_0}(\tau_{N_0} c) = 0$.

5.3. *Proof of Lemma* 10. Let I be the random variable corresponding to the residual interarrival time at the random time instant T_{N_0}; moreover, let R_{N_0} be the random variable corresponding to the first regeneration point of the queue (see (50) of Section 5.1). We have

$$R_{N_0} = T_{N_0} + I.$$

Using this and the union bound, we have

$$\begin{aligned}(71)\quad &\Pr[R_{N_0} \geqq \tau_{N_0}(1+\delta)] = \Pr[T_{N_0} + I \geqq \tau_{N_0}(1+\delta)] \\ &\qquad \leqq \Pr\left[T_{N_0} \geqq \tau_{N_0}\left(1 + \frac{\delta}{2}\right)\right] + \Pr\left[I \geqq \tau_{N_0}\frac{\delta}{2}\right].\end{aligned}$$

In what follows, each of the two terms in the right-hand side of (71) is appropriately upper bounded.

Starting with the first term, we have

$$\Pr\left[T_{N_0} \geqq \tau_{N_0}\left(1+\frac{\delta}{2}\right)\right] < C'(\delta)\exp\left(-\psi'(\delta)N_0\right). \tag{72}$$

The result above follows from Proposition 5, applied with $k=1$ and with $\delta/2$ instead of δ.

We now consider the last term in the right-hand side of (71). Let $\mathcal{N}$ be the random variable corresponding to the number of arrivals until the system is met empty for the first time (see also the discussion preceding Lemma 3). Moreover, let Y_i denote the service time of the ith customer and let Z_i denote the ith interarrival time. We have

$$\mathcal{N} \stackrel{\text{def}}{=} \min_{k \geqq 1}\left\{k: \sum_{i=1}^{N_0+k-1} Y_i < \sum_{i=1}^{k} Z_i\right\}. \tag{73}$$

Since I is the residual interarrival time at the time instant T_{N_0}, we have $I < Z_{\mathcal{N}}$ with probability 1; this implies that

$$\Pr\left[I \geqq \tau_{N_0}\frac{\delta}{2}\right] \leqq \Pr\left[Z_{\mathcal{N}} \geqq \tau_{N_0}\frac{\delta}{2}\right]. \tag{74}$$

We define n^* as follows:

$$n^* \stackrel{\text{def}}{=} \lceil \lambda\tau_{N_0}(1+\varepsilon)\rceil, \tag{75}$$

where ε is some positive constant. We have

$$\begin{aligned}\Pr\left[Z_{\mathcal{N}} \geqq \tau_{N_0}\frac{\delta}{2}\right] &= \Pr\left[Z_{\mathcal{N}} \geqq \tau_{N_0}\frac{\delta}{2} \text{ and } \mathcal{N} \leqq n^*\right] \\ &\quad + \Pr\left[Z_{\mathcal{N}} \geqq \tau_{N_0}\frac{\delta}{2} \text{ and } \mathcal{N} > n^*\right].\end{aligned} \tag{76}$$

Clearly, we have

$$\Pr\left[Z_{\mathcal{N}} \geqq \tau_{N_0}\frac{\delta}{2} \text{ and } \mathcal{N} \leqq n^*\right] \leqq \Pr\left[\max_{i=1,\cdots,n^*}\{Z_i\} \geqq \tau_{N_0}\frac{\delta}{2}\right].$$

Combining this with (76), we obtain

$$\Pr\left[Z_{\mathcal{N}} \geqq \tau_{N_0}\frac{\delta}{2}\right] \leqq \Pr\left[\max_{i=1,\cdots,n^*}\{Z_i\} \geqq \tau_{N_0}\frac{\delta}{2}\right] + \Pr[\mathcal{N} > n^*]. \tag{77}$$

Each of the two terms on the right-hand side of (77) will be upper bounded by an exponentially decaying quantity.

Starting with the last term, we have (due to (73))

$$\begin{aligned}\Pr[\mathcal{N} > n^*] &= \Pr\left[\sum_{i=1}^{N_0+k-1} Y_i \geqq \sum_{i=1}^{k} Z_i \;\forall k \in \{1,\cdots,n^*\}\right] \\ &\leqq \Pr\left[\sum_{i=1}^{N_0+n^*-1} Y_i \geqq \sum_{i=1}^{n^*} Z_i\right].\end{aligned} \tag{78}$$

We define

$$\mathscr{X} \stackrel{\text{def}}{=} \sum_{i=1}^{N_0+n^*-1} Y_i - \sum_{i=1}^{n^*} Z_i.$$

Clearly, we have $E[e^{s\mathscr{X}}] = [B(s)]^{N_0+n^*-1}[A(-s)]^{n^*}$ for all $s \in (-s_A, s_B)$; thus, applying the Chernoff bound (see (3) in Section 2.1), we obtain

$$\text{(79)} \quad \Pr\left[\sum_{i=1}^{N_0+n^*-1} Y_i \geqq \sum_{i=1}^{n^*} Z_i\right] = \Pr[\mathscr{X} \geqq 0] \leqq [B(s)]^{N_0+n^*-1}[A(-s)]^{n^*}, \qquad \forall s \in (0, s_B).$$

Since $B(s) > 1$ and $A(-s) < 1$ for all $s \in (0, s_B)$, it follows from (75) (and from the fact $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu - \lambda)$) that

$$\text{(80)} \quad [B(s)]^{N_0+n^*-1}[A(-s)]^{n^*} < ([B(s)]^{\lambda(1+\varepsilon)/(\mu-\lambda)+1}[A(-s)]^{\lambda(1+\varepsilon)/(\mu-\lambda)})^{N_0}, \qquad \forall s \in (0, s_B).$$

Defining

$$f(s) \stackrel{\text{def}}{=} [B(s)]^{\lambda(1+\varepsilon)/(\mu-\lambda)+1}[A(-s)]^{\lambda(1+\varepsilon)/(\mu-\lambda)},$$

we have $f(0) = 1$ and

$$\lim_{s \downarrow 0} \frac{df(s)}{ds} = \frac{1}{\mu}\left(\frac{\lambda}{\mu-\lambda}(1+\varepsilon)+1\right) - \frac{1}{\lambda}\left(\frac{\lambda}{\mu-\lambda}(1+\varepsilon)\right) = -\frac{\varepsilon}{\mu} < 0.$$

Therefore, there exists some $s_1 \in (0, s_B)$ such that $f(s_1) < 1$. Applying (79) and (80) with $s = s_1$, we obtain

$$\Pr\left[\sum_{i=1}^{N_0+n^*-1} Y_i \geqq \sum_{i=1}^{n^*} Z_i\right] < \exp(-\varphi_1 N_0),$$

where $\varphi_1 \stackrel{\text{def}}{=} -\ln f(s_1) > 0$. This together with (78) implies that

$$\text{(81)} \quad \Pr[\mathscr{N} > n^*] \leqq \exp(-\varphi_1 N_0).$$

We now consider the other term in the right-hand side of (77). Applying the union bound and using the fact that the random variables $(Z_i)_{i=1,\cdots,n^*}$ are identically distributed, we obtain

$$\text{(82)} \quad \Pr\left[\max_{i=1,\cdots n^*}\{Z_i\} \geqq \tau_{N_0}\frac{\delta}{2}\right] \leqq \sum_{i=1}^{n^*} \Pr\left[Z_i \geqq \tau_{N_0}\frac{\delta}{2}\right] = n^* \Pr\left[Z \geqq \tau_{N_0}\frac{\delta}{2}\right],$$

where Z is a random variable having the interarrival time distribution. Applying the Chernoff bound (see (3) in Section 2.1), we have

$$\text{(83)} \quad \Pr\left[Z \geqq \tau_{N_0}\frac{\delta}{2}\right] \leqq A(s)\exp\left(-s\tau_{N_0}\frac{\delta}{2}\right), \qquad \forall s \in (0, s_A).$$

We also have (due to (75) and the fact that $\tau_{N_0} \stackrel{\text{def}}{=} N_0/(\mu - \lambda)$) that

$$n^* < \lambda\tau_{N_0}(1+\varepsilon) + 1 \leqq \frac{\lambda N_0}{\mu - \lambda}(1+2\varepsilon)$$

for all $N_0 \geqq l^*$, where $l^* \stackrel{\text{def}}{=} (\mu - \lambda)/(\lambda\varepsilon)$. We fix some $s \in (0, s_A)$. Using the previous

inequality, (82) and (83), we obtain

$$(84)\quad \Pr\left[\max_{i=1,\cdots,n^*}\{Z_i\}\geqq\tau_{N_0}\frac{\delta}{2}\right]<\frac{\lambda N_0}{\mu-\lambda}(1+2\varepsilon)A(s)\exp\left(-\frac{s\delta}{2(\mu-\lambda)}N_0\right),\qquad \forall N_0\geqq l^*.$$

We define

$$\varphi_2(\delta)\stackrel{\text{def}}{=}\frac{s\delta}{4(\mu-\lambda)}\quad\text{and}\quad D_2(\delta)\stackrel{\text{def}}{=}\frac{4\lambda}{s\delta}(1+2\varepsilon)A(s).$$

Using these definitions and the inequality $x<(1/\alpha)e^{\alpha x}$ (for $\alpha>0$), it follows from (84) (after some algebra) that

$$\Pr\left[\max_{i=1,\cdots,n^*}\{Z_i\}\geqq\tau_{N_0}\frac{\delta}{2}\right]<D_2(\delta)\exp\left(-\varphi_2(\delta)N_0\right),\qquad \forall N_0\geqq l^*.$$

This together with (77) and (81) implies that

$$\Pr\left[I\geqq\tau_{N_0}\frac{\delta}{2}\right]\leqq\exp\left(-\varphi_1 N_0\right)+D_2(\delta)\exp\left(-\varphi_2(\delta)N_0\right),\qquad \forall N_0\geqq l^*.$$

Combining this with (72) and (71), we obtain

$$\Pr\left[R_{N_0}\geqq\tau_{N_0}(1+\delta)\right]<C'(\delta)\exp\left(-\psi'(\delta)N_0\right)+\exp\left(-\varphi_1 N_0\right)+D_2(\delta)\exp\left(-\varphi_2(\delta)N_0\right),\qquad \forall N_0\geqq l^*.$$

After defining

$$\phi(\delta)\stackrel{\text{def}}{=}\tfrac{1}{2}\min\left\{\psi'(\delta),\ \varphi_1,\ \varphi_2(\delta)\right\}$$

and

$$l(\delta)\stackrel{\text{def}}{=}\max\left\{l^*,\ \frac{\ln\left[C'(\delta)+1+D_2(\delta)\right]}{\phi(\delta)}\right\},$$

the result follows from the inequality above.

It is worth noting the following. Since $T_{N_0}<R_{N_0}$ with probability 1, it follows from Proposition 8 that for any positive δ we have $\Pr\left[R_{N_0}\leqq\tau_{N_0}(1-\delta)\right]\leqq\exp\left(-\xi(\delta)N_0\right)$ for all $N_0\geqq n(\delta)$. ($\xi(\delta)$ and $n(\delta)$ are the same as in Proposition 8.) Combining this result with Lemma 10, one may easily prove that

$$\lim_{N_0\to\infty}\frac{R_{N_0}}{N_0}\stackrel{\text{a.s.}}{=}\frac{1}{\mu-\lambda}$$

(see also Proposition 9).

6. Conclusions

In this paper we have analyzed a certain type of stable $GI/G/1$ queue, namely that with $\lambda<\mu$, with the service time distribution being of the exponential type and with the interarrival time distribution being of the exponential and of the non-lattice types. This type of queue fits most practical cases. Assuming that such a queue

initially contains N_0 customers, with $N_0 \neq 0$, we proved that the time T_{N_0} required for the queue to empty is asymptotically proportional to N_0, namely

$$\lim_{N_0 \to \infty} \frac{T_{N_0}}{N_0} \overset{\text{a.s.}}{=} \frac{1}{\mu - \lambda}.$$

Using properties of the random variable T_{N_0}, we proved that the settling time of the queue is asymptotically equal to τ_{N_0} (i.e., $N_0/(\mu - \lambda)$). Finally, we proved that after scaling both time and the number of customers in the system by N_0, as N_0 increases, the queue asymptotically behaves as if customers were arriving at a constant rate λ and, at the same time, were departing at a constant rate μ, as in a simple fluid model.

An interesting direction for further research is to consider the asymptotic behaviour of T_{N_0} in the context of a $G/G/1$ queue where the interarrival and service times form a stationary and ergodic sequence $(Y_i, Z_i)_{i=1,\ldots}$. It is reasonable to conjecture that most of our results are still valid in this more general context, and that their proofs would include coupling and stochastic monotonicity arguments. (This point was suggested by the referee and by F. Baccelli.) Moreover, we believe that the results established in this paper may be extended to queueing networks. In such systems, it is the dependence among arrivals that makes our analysis not directly applicable. Thus, one has to show that, after scaling time by N_0, this dependence becomes unimportant. Such a result has been established in [13] for the simple case of a stable tandem of exponential servers with Poisson arrivals.

References

[1] ALDOUS, D. (1983) Random walks on finite groups and rapidly mixing Markov chains. *Séminaire de Probabilités* XVII, Lecture Notes in Mathematics **986**, Springer-Verlag, Berlin, 243–297.

[2] ANANTHARAM, V. (1988) The settling time of a closed network. Unpublished.

[3] BAHADUR, R. R. (1971) *Some Limit Theorems in Statistics*. SIAM, Philadelphia.

[4] BATEMAN MANUSCRIPT PROJECT (1954) *Tables of Integral Transforms*, Vol. 1. McGraw-Hill, New York.

[5] BOROVKOV, A. A. (1976) *Stochastic Processes in Queueing Theory*. Springer-Verlag, Berlin.

[6] COHEN, J. W. (1969) *The Single-Server Queue*. North-Holland, Amsterdam.

[7] ELLIS, R. S. (1985) *Entropy, Large Deviations, and Statistical Mechanics*. Springer-Verlag, Berlin.

[8] HAJEK, B. (1982) Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Prob.* **14**, 502–525.

[9] HEYMAN, D. P. AND SOBEL, M. J. (1982) *Stochastic Methods in Operations Research*, Vol. 1. McGraw-Hill, New York.

[10] KLEINROCK, L. (1975) *Queueing Systems*, Vol. 1: *Theory*. Wiley, New York.

[11] POLLACZEK, F. (1952) Sur la répartition des périodes d'occupation ininterrompue d'un guichet. *C.R. Acad. Sci. Paris* **234**, 2042–2044.

[12] PRABU, N. U. (1980) *Stochastic Storage Processes*. Springer-Verlag, Berlin.

[13] STAMOULIS, G. D. (1988) Transient Analysis of Some Open Queueing Systems. S.M. Thesis. Technical Report LIDS-TH-1976, MIT.

[14] TAKÁCS, L. (1962) *Introduction to the Theory of Queues*. Oxford University Press, New York.