# Heavy Tails in Multi-Server Queue

**Serguei Foss · Dmitry Korshunov**

**Abstract**   In this paper, the asymptotic behaviour of the distribution tail of the stationary waiting time $W$ in the $GI/GI/2$ FCFS queue is studied. Under subexponential-type assumptions on the service time distribution, bounds and sharp asymptotics are given for the probability $\mathbf{P}\{W > x\}$. We also get asymptotics for the distribution tail of a stationary two-dimensional workload vector and of a stationary queue length. These asymptotics depend heavily on the traffic load.

**Keywords**   FCFS multi-server queue · Stationary waiting time · Large deviations · Long tailed distribution · Subexponential distribution

**AMS subject classification:** 60K25

## 1. Introduction

It is well known (see, for example, [1,15,18]) that in the stable single server *first-come-first-served* queue $GI/GI/1$ with typical interarrival time $\tau$ and typical service time $\sigma$ the tail of stationary waiting time $W$ is related to the service time distribution tail $\bar{B}(x) = \mathbf{P}\{\sigma > x\}$ via the equivalence

$$\mathbf{P}\{W > x\} \sim \frac{1}{\mathbf{E}\tau - \mathbf{E}\sigma} \int_x^\infty \bar{B}(y)\,dy \quad \text{as } x \to \infty, \qquad (1)$$

S. Foss
Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, Scotland
e-mail: foss@ma.hw.ac.uk

D. Korshunov
Sobolev Institute of Mathematics, Novosibirsk 630090, Russia
e-mail: korshunov@math.nsc.ru

provided the *subexponentiality* of the *integrated tail distribution $B_I$* defined by its tail

$$\bar{B}_I(x) \equiv \min\left(1, \int_x^\infty \bar{B}(y)\,dy\right), \quad x > 0.$$

As usual we say that a distribution $G$ on $\mathbf{R}^+$ is subexponential (belongs to the class $\mathscr{S}$) if $\overline{G * G}(x) \sim 2\bar{G}(x)$ as $x \to \infty$. The converse assertion is also true, that is, the equivalence (1) implies the subexponentiality of $B_I$, see [15, Theorem 1] for the case of Poisson arrival stream and [14, Theorem 1] for the general case.

In this paper we consider the *GI/GI/s* FCFS queue which goes back to Kiefer and Wolfowitz [13]. We have $s$ identical servers, i.i.d. interarrival times $\{\tau_n\}$ with finite mean $a = \mathbf{E}\tau_1$, and i.i.d. service times $\{\sigma_n\}$ with finite mean $b = \mathbf{E}\sigma_1$. The sequences $\{\tau_n\}$ and $\{\sigma_n\}$ are mutually independent. The system is assumed to be *stable*, i.e., $\rho \equiv b/a \in (0, s)$. We are interested in the asymptotic tail behaviour of the stationary waiting time distribution $\mathbf{P}\{W > x\}$ as $x \to \infty$.

It was realized recently (see, for example, existence results for moments in [16,17]; an asymptotic hypothesis in [19]; asymptotic results for fluid queues fed by heavy-tailed on-off flows in [5]) that the heaviness of the stationary waiting time tail depends substantially on the load $\rho$ in the system. More precisely, it depends on $\rho$ via the value of $k \in \{0, 1, \ldots, s - 1\}$ for which $k \le \rho < k + 1$. In particular, Whitt conjectured that

$$\mathbf{P}\{W > x\} \sim \gamma \left(\int_{\eta x}^\infty \bar{B}(y)\,dy\right)^{s-k} \quad \text{as } x \to \infty,$$

"where $\gamma$ and $\eta$ are positive constants (as functions of $x$)" [sic, [19]]. In the present paper we show that, in general, the tail behaviour of $W$ is more complicated.

Let $R(w) = (R_1(w), \ldots, R_s(w))$ be the operator on $\mathbf{R}^s$ which orders the coordinates of $w \in \mathbf{R}^s$ in ascending order, i.e., $R_1(w) \leq \cdots \leq R_s(w)$. Then the residual work vector $W_n = (W_{n1}, \ldots, W_{ns})$ which the $n$th customer observes just upon its arrival satisfies the celebrated Kiefer-Wolfowitz recursion: $W_1 = i \cdot 0$,

$$
\begin{aligned}
W_{n+1} &= R((W_{n1} + \sigma_n - \tau_{n+1})^+, (W_{n2} - \tau_{n+1})^+, \\
&\quad \ldots, (W_{ns} - \tau_{n+1})^+) \\
&= R(W_n + e_1\sigma_n - i\tau_{n+1})^+,
\end{aligned}
$$

here $e_1 = (1, 0, \ldots, 0), i = (1, \ldots, 1)$ and $w^+ = (\max(0, w_1), \ldots, \max(0, w_s))$. The value of $W_{n1}$ is the delay which customer $n$ experiences. In particular, the stationary waiting time $W$ is a weak limit for $W_{n1}$.

The process $W_n$ is a Markov chain in $\mathbf{R}^s$. It is well known that, for general multi-dimensional Markov chains, large deviation problems are very difficult to solve even for stationary distributions. Usually they can be solved in low dimensions only, 2 or 3 at most, see [4,12]. Almost all known results are derived for so-called Cramér case which corresponds to light-tailed distributions of jumps. In the heavy-tailed case almost nothing is known for general multi-dimensional Markov chains.

The process $W_n$ presents a particular but very important example of a Markov chain in $\mathbf{R}^s$, even if we are interested in the first component $W_{n1}$. As follows from our analysis, the case $s = 2$ can be treated in detail. The stability condition for this particular case is $b < 2a$. One of the following cases can occur:

(i) the maximal stability case when $b < a$;
(ii) the intermediate case when $b = a$;
(iii) the minimal stability case when $b \in (a, 2a)$.

We find the exact asymptotics for $\mathbf{P}\{W > x\}$ in the maximal and minimal cases. We also describe the most probable way for the occurrence of large deviations. In the intermediate case, we only provide upper and lower bounds. Then we study the asymptotics for the tail of the distribution of a stationary two-dimensional workload vector and give comments on the tail asymptotics of the stationary queue length.

For $s > 2$, the stability condition is $b < sa$. We hope that, for $s > 2$, direct modifications of our arguments may lead to exact asymptotics in two particular cases when either $b < a$ (the maximal stability) or $b \in ((s-1)a, sa)$ (the minimal stability). However, one has to overcome many extra technicalities for that. Insofar as the case $b \in [a, (s-1)a]$ is

concerned, we are extremely sceptical on the possibility to get any sharp tail asymptotics in explicit form.

For the two-server queue, in the maximal stability case, we prove the following:

**Theorem 1.** Let $s = 2$ and $b < a$. When the integrated tail distribution $B_I$ is subexponential, the tail of the stationary waiting time satisfies the asymptotic relation, as $x \to \infty$,

$$
\mathbf{P}\{W > x\} \sim \frac{1}{a(2a - b)}\left[(\bar{B}_I(x))^2 \right.
$$
$$
\left. + b \int_0^\infty \bar{B}_I(x + ya)\bar{B}(x + y(a - b))\, dy\right].
$$

The proof follows by combining the lower bound given in Theorem 3 (Section 3) and the upper bound given in Theorem 4 (Section 4). Simpler lower and upper bounds for $\mathbf{P}\{W > x\}$ are given in the following

**Corollary 1.** Under the conditions of Theorem 1,

$$
\frac{2a + b}{2a^2(2a - b)} \leq \liminf_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{(\bar{B}_I(x))^2} \leq \limsup_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{(\bar{B}_I(x))^2}
$$
$$
\leq \frac{1}{2a(a - b)}.
$$

In our opinion, in Theorem 1 it is possible to obtain a compact expression for the tail asymptotics of $\mathbf{P}(W > x)$ only in the regularly varying case. A distribution $G$ (or its tail $\bar{G}$) is *regularly varying* at infinity with index $\gamma > 0$ (belongs to the class $\mathscr{RV}$), if $\bar{G}(x) > 0$ for all $x$ and, for any fixed $c > 0$, $\bar{G}(cx)/\bar{G}(x) \to c^{-\gamma}$ as $x \to \infty$.

**Corollary 2.** Let $b < a$ and the tail distribution $\bar{B}$ of service time be regularly varying with index $\gamma > 1$. Then

$$
\mathbf{P}\{W > x\} \sim c'(\bar{B}_I(x))^2,
$$

where

$$
c' = \frac{1}{a(2a - b)}\left[1 + \frac{b}{\gamma - 1}\right.
$$
$$
\left. \times \int_0^\infty \frac{dz}{(1 + za)^{\gamma - 1}(1 + z(a - b))^\gamma}\right].
$$

Recall definitions of a number of classes of heavy-tailed distributions. A distribution $G$ is *long-tailed* (belongs to the

class $\mathscr{L}$ if $\bar{G}(x) > 0$ for all $x$ and, for any fixed $t$,

$$\frac{\bar{G}(x + t)}{\bar{G}(x)} \to 1 \quad \text{as } x \to \infty.$$

A distribution $G$ belongs to the class $\mathscr{IRV}$ of *intermediate regularly varying distributions* if $\bar{G}(x) > 0$ for all $x$ and

$$\lim_{c \downarrow 1} \liminf_{x \to \infty} \frac{\bar{G}(cx)}{\bar{G}(x)} = 1.$$

Clearly, $\mathscr{RV} \subset \mathscr{IRV}$.

In the minimal stability case, we prove the following

**Theorem 2.** Let $s = 2$ and $a < b < 2a$, $B \in \mathscr{S}$ and $B_I \in \mathscr{IRV}$. Then

$$\mathbf{P}\{W > x\} \sim \frac{1}{2a - b} \bar{B}_I \left( \frac{b}{b - a} x \right) \quad as\ x \to \infty.$$

The proof is given in Section 7 and is based on the lower and upper bounds stated in Theorems 5 and 6 respectively.

One can provide simple sufficient conditions for $B \in \mathscr{S}$ and $B_I \in \mathscr{IRV}$. Let $\mathscr{D}$ be the class of all distributions $G$ on $\mathbf{R}^+$ such that $\bar{G}(x) > 0$ for all $x$ and $\liminf_{x \to \infty} \bar{G}(2x)/\bar{G}(x) > 0$. Then the following are known: (i) $\mathscr{RV} \subset \mathscr{IRV} \subset (\mathscr{L} \bigcap \mathscr{D}) \subset \mathscr{S}$; (ii) if $B \in \mathscr{D}$ has a finite first moment, then $B_I \in \mathscr{IRV}$ (see e.g. [6]). Therefore, if $B \in \mathscr{L} \bigcap \mathscr{D}$ and has a finite first moment, then $B$ satisfies the conditions of Theorem 2. Note that the converse is not true, in general: there exists a distribution $B \in \mathscr{S}$ with a finite first moment such that $B_I \in \mathscr{IRV}$ but $B \notin \mathscr{L} \bigcap \mathscr{D}$ (see Example 2 in [9, Section 6]).

The paper is organized as follows. Section 2 contains some auxiliary results. In Section 3, we formulate and prove a result concerning a lower bound for $\mathbf{P}\{W > x\}$ in the maximal stability case. The corresponding upper bound is given in Section 4. Sections 5, 6, and 7 deal, respectively, with lower bounds, upper bounds, and asymptotics for $\mathbf{P}\{W > x\}$ in the minimal stability case. In Section 8, we prove further results related to the joint distribution of a stationary workload vector. Comments on the asymptotics for a stationary queue length distribution may be found in Section 9.

A number of upper and lower bounds for $\mathbf{P}\{W > x\}$ in s-server queue are proposed in Remarks 2, 3, 4, and 5.

## 2. Preliminaries

### 2.1. Reduction to deterministic input stream case in assertions associated with upper bounds

Consider a general *GI/GI/s* queue. Take any $a' \in (b/s, a)$. Consider an auxiliary *D/GI/s* system with the same service times $\{\sigma_n\}$ and deterministic interarrival times $\tau_n' \equiv a'$: $W_1' = 0$ and

$$W_{n+1}' = R(W_n' + e_1 \sigma_n - i a')^+.$$

Let $W'$ be a stationary waiting time in this auxiliary system.

**Lemma 1.** If $\mathbf{P}\{W' > x\} \leq \bar{G}(x)$ for some long-tailed distribution $G$, then

$$\limsup_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{\bar{G}(x)} \leq 1.$$

*Proof.* Denote $\xi_n = a' - \tau_n$. Put $M_0 = 0$ and, for $n \geq 1$,

$$M_n = \max\{0, \xi_n, \xi_n + \xi_{n-1}, \dots, \xi_n + \cdots + \xi_1\}$$
$$= \max(0, \xi_n + M_{n-1}) = (\xi_n + M_{n-1})^+.$$

First, we use induction to prove the inequality

$$W_n \leq W_n' + i M_n \quad \text{a.s.} \tag{2}$$

Indeed, for $n = 1$ we have $0 \leq 0 + i M_1$. Assume the inequality is proved for some $n$; we prove it for $n + 1$. Indeed,

$$\begin{aligned} W_{n+1} &= R(W_n + e_1 \sigma_n - i \tau_{n+1})^+ \\ &\leq R(W_n' + i M_n + e_1 \sigma_n - i \tau_{n+1})^+ \\ &= R(W_n' + e_1 \sigma_n - i a' + i(M_n + \xi_{n+1}))^+. \end{aligned}$$

Since $(u + v)^+ \leq u^+ + v^+$,

$$\begin{aligned} W_{n+1} &\leq R(W_n' + e_1 \sigma_n - i a')^+ + i(M_n + \xi_{n+1})^+ \\ &\equiv W_{n+1}' + i M_{n+1}, \end{aligned}$$

and the proof of (2) is complete.

Let $M$ be the weak limit for $M_n$ which exists due to $\mathbf{E}\xi_1 = a' - a < 0$ and Strong Law of Large Numbers. The following stochastic equality holds:

$$M =_{\text{st}} \max\{0, \xi_1, \xi_1 + \xi_2, \dots, \xi_1 + \cdots + \xi_n, \dots\}.$$

Since the random variable $\xi_1$ is bounded from above (by $a'$), there exists $\beta > 0$ such that $\mathbf{E}e^{\beta \xi_1} = 1$. Then by Cramér estimate (see, for example, [8, Section 5]), for any $x$,

$$\mathbf{P}\{M > x\} \leq e^{-\beta x}. \tag{3}$$

The inequality (2) implies that $W \leq_{\text{st}} W' + M$, where $W'$ and $M$ are independent. Let a random variable $\eta$ have distribution

$G$ and be independent of $M$. Since $\eta \geq_{st} W'$, we have $W \leq_{st} \eta + M$. Therefore, for any $h > 0$,

$$
\begin{aligned}
\mathbf{P}\{W > x\} &\leq \int_0^{x-h} \mathbf{P}\{M > x - y\}\mathbf{P}\{\eta \in dy\} \\
&\quad + \mathbf{P}\{\eta > x - h\} \\
&\leq \int_0^{x-h} e^{-\beta(x-y)}G(dy) + \bar{G}(x - h),
\end{aligned}
$$

by (3). Integrating by parts yields

$$
\begin{aligned}
\int_0^{x-h} e^{-\beta(x-y)}G(dy) &= -e^{-\beta(x-y)}\bar{G}(y)\Big|_0^{x-h} \\
&\quad + \beta \int_0^{x-h} \bar{G}(y)e^{-\beta(x-y)}dy \\
&\leq e^{-\beta x} + \beta \int_h^x \bar{G}(x-y)e^{-\beta y}dy.
\end{aligned}
$$

The distribution $G$ is long-tailed, thus, for any $\varepsilon > 0$ there exists $x(\varepsilon)$ such that

$$
\bar{G}(x - y) \leq \bar{G}(x)e^{\varepsilon}
$$

for any $x \geq x(\varepsilon)$. Hence, there exists $c(\varepsilon) < \infty$ such that

$$
\bar{G}(x - y) \leq c(\varepsilon)\bar{G}(x)e^{\varepsilon y}
$$

for any $x \geq x(\varepsilon)$. Take $\varepsilon = \beta/2$. Then

$$
\begin{aligned}
\int_h^x \bar{G}(x-y)e^{-\beta y}dy &\leq c(\varepsilon)\bar{G}(x) \int_h^x e^{-\beta y/2}dy \\
&\leq \frac{c(\varepsilon)}{\beta/2}\bar{G}(x)e^{-\beta h/2}.
\end{aligned}
$$

Hence,

$$
\mathbf{P}\{W > x\} \leq e^{-\beta x} + 2c(\varepsilon)\bar{G}(x)e^{-\beta h/2} + \bar{G}(x - h).
$$

Taking into account also that $\bar{G}(x - h) \sim \bar{G}(x)$ for any fixed $h > 0$, we obtain

$$
\limsup_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{\bar{G}(x)} \leq 2c(\varepsilon)e^{-\beta h/2} + 1.
$$

Letting $h \to \infty$ yields the conclusion of the Lemma.

## 2.2. Reduction to deterministic input stream case in assertions associated with lower bounds

Take any $a' > a$. As in the previous subsection, consider an auxiliary $D/GI/s$ system with the same service times $\{\sigma_n\}$ and deterministic interarrival time $\tau'_n \equiv a'$. Let $W'$ be a stationary waiting time in this auxiliary system.

**Lemma 2.** If $\mathbf{P}\{W' > x\} \geq \bar{G}(x)$ for some long-tailed distribution $G$, then

$$
\liminf_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{\bar{G}(x)} \geq 1.
$$

*Proof.* Put $\xi_n = \tau_n - a'$, $M_0 = 0$ and

$$
\begin{aligned}
M_n &= \max\{0, \xi_n, \xi_n, +\xi_{n-1}, \ldots, \xi_n + \cdots + \xi_1\} \\
&= (M_{n-1} + \xi_n)^+.
\end{aligned}
$$

For any $n \geq 1$, the following inequality holds:

$$
W_n \geq W'_n - iM_n. \tag{4}
$$

Indeed, by induction arguments,

$$
\begin{aligned}
W_{n+1} &= R(W_n + e_1\sigma_n - i\tau_{n+1})^+ \\
&\geq R(W'_n - iM_n + e_1\sigma_n - i\tau_{n+1})^+ \\
&= R(W'_n + e_1\sigma_n - ia' - i(M_n + \xi_{n+1}))^+.
\end{aligned}
$$

Since $(u - v)^+ \geq u^+ - v^+$,

$$
\begin{aligned}
W_{n+1} &\geq R(W'_n + e_1\sigma_n - ia')^+ - i(M_n + \xi_{n+1})^+ \\
&\equiv W'_{n+1} - iM_{n+1},
\end{aligned}
$$

and the proof of (4) is complete.

Let $M$ be the weak limit for $M_n$ which exists due to $\mathbf{E}\xi_1 = a - a' < 0$ and the Strong Law of Large Numbers. The inequality (4) implies that $W \geq_{st} W' - M$ where $W'$ and $M$ are independent. Therefore, for any $h > 0$,

$$
\begin{aligned}
\mathbf{P}\{W > x\} &\geq \mathbf{P}\{W' > x + h\}\mathbf{P}\{M \leq h\} \\
&\geq \bar{G}(x + h)\,\mathbf{P}\{M \leq h\}.
\end{aligned}
$$

The distribution $G$ is long-tailed, thus $\bar{G}(x + h) \sim \bar{G}(x)$ for any fixed $h > 0$ and

$$
\liminf_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{\bar{G}(x)} \geq \mathbf{P}\{M \leq h\}.
$$

Letting $h \to \infty$, we obtain the desired estimate from below.

### 2.3. Adapted versions of the Law of Large Numbers

It is well known that obtaining lower bounds for systems under assumptions of heavy tails usually requires some variant of the Law of Large Numbers. Here we provide such a tool for the two-server queue.

**Lemma 3.** Let $(\xi_n, \eta_n)$, $n = 1, 2, \ldots$, be independent identically distributed pairs of random variables. Let the two-dimensional Markov chain $V_n = (V_{n1}, V_{n2})$, $n = 1, 2, \ldots$, be defined in the following way: $V_1$ has an arbitrary distribution and

$$V_{n+1} = \begin{cases} V_n + (\xi_n, \eta_n), & \text{if } V_{n1} \leq V_{n2}, \\ V_n + (\eta_n, \xi_n), & \text{if } V_{n1} > V_{n2}. \end{cases}$$

If $\mathbf{E}\eta_1 < \mathbf{E}\xi_1$, then the following convergence in probability holds:

$$\frac{V_n}{n} \to \left( \frac{\mathbf{E}\xi_1 + \mathbf{E}\eta_1}{2}, \frac{\mathbf{E}\xi_1 + \mathbf{E}\eta_1}{2} \right) \quad \text{as } n \to \infty.$$

*Proof.* Since $V_{n+1,1} + V_{n+1,2} = V_{n1} + V_{n2} + \xi_n + \eta_n$, by the Law of Large Numbers

$$\frac{V_{n1} + V_{n2}}{n} \to \mathbf{E}\xi_1 + \mathbf{E}\eta_1 \quad \text{as } n \to \infty. \tag{5}$$

Define a Markov chain $U_n = V_{n2} - V_{n1}$. If $U_n \geq 0$, then $U_{n+1} - U_n = \eta_n - \xi_n$, while if $U_n < 0$, then $U_{n+1} - U_n = \xi_n - \eta_n = -(\eta_n - \xi_n)$, so, $U_n$ is the oscillating random walk. Since $\mathbf{E}\xi_1 > \mathbf{E}\eta_1$, the mean drift of the chain $U_n$ is negative on the positive half-line and is positive on the negative half-line. Therefore, for any sufficiently large $A$, the set $[-A, A]$ is positive recurrent for this Markov chain. In particular, the distributions of $U_n$ are tight. Hence, $U_n/n \to 0$ in probability as $n \to \infty$. Together with (5), it implies the desired assertion of Lemma.  □

The classical Law of Large Numbers and Lemma 3 imply the following

**Corollary 3.** Let $\mathbf{E}\eta_1 < \mathbf{E}\xi_1 < 0$ and $\varepsilon > 0$. Then

$$\mathbf{P}\{V_{n1} > 0, V_{n2} > 0 \mid V_1 = (v_1, v_2)\} \to 1$$

as $N \to \infty$ uniformly in $n \geq N$ and in $(v_1, v_2)$ on the set

$$\{v_1, v_2 > n(|\mathbf{E}\xi_1| + \varepsilon), v_1 + v_2 > n(|\mathbf{E}\xi_1| + |\mathbf{E}\eta_1| + \varepsilon)\}.$$

**Corollary 4.** Let $\mathbf{E}\eta_1 < \mathbf{E}\xi_1 < 0$ and $\varepsilon > 0$. Then

$$\mathbf{P}\{V_{n1} > 0, V_{n2} > 0 \mid V_1 = (v_1, v_2)\} \to 0$$

as $N \to \infty$ uniformly in $n \geq N$ and in $(v_1, v_2)$ on the complementary set

$$\{v_1 > n(|\mathbf{E}\xi_1| - \varepsilon), v_2 > n(|\mathbf{E}\xi_1| - \varepsilon), v_1 + v_2 > n(|\mathbf{E}\xi_1| + |\mathbf{E}\eta_1| - \varepsilon)\}.$$

**Corollary 5.** Let $\mathbf{E}\eta_1 < 0$, $\mathbf{E}\xi_1 > 0$, $\mathbf{E}\eta_1 + \mathbf{E}\xi_1 < 0$ and $\varepsilon > 0$. Then

$$\mathbf{P}\{V_{n1} > x, V_{n2} > x \mid V_1 = (v_1, v_2)\} \to 1$$

as $x, N \to \infty$ uniformly in $n \geq N$ and in $(v_1, v_2)$ on the set

$$\{v_1 > x - n(\mathbf{E}\xi_1 - \varepsilon), v_2 > 2x + n(|\mathbf{E}\xi_1 + \mathbf{E}\eta_1| + \varepsilon)\}.$$

### 3. The maximal stability case: A lower bound

**Theorem 3.** Assume $b \in (0, a)$. Let the integrated service time distribution $B_I$ be long-tailed. Then the tail of the stationary waiting time $W$ admits the following estimate from below: as $x \to \infty$,

$$\mathbf{P}\{W > x\} \geq \frac{1 + o(1)}{a(2a - b)} \left[ (\bar{B}_I(x))^2 + b \int_0^\infty \bar{B}_I(x + ya) \right.$$
$$\left. \times \bar{B}(x + y(a - b)) \, dy \right]. \tag{6}$$

*Remark 1.* From (6), one can get the lower bound in Corollary 1. Namely, replace $\bar{B}(x + y(a - b))$ by a smaller term $\bar{B}(x + ya)$ in the integral in the RHS of (6). Then the new integral is equal to $b(\bar{B}_I(x))^2/2a$, and the lower bound follows since

$$\frac{1}{a(2a - b)} \left( 1 + \frac{b}{2a} \right) = \frac{2a + b}{2a^2(2a - b)}.$$

*Remark 2.* By use of Strong Law of Large Numbers, one can get the following result for $s$-server queue, $s \geq 2$. If $b < a$, then there exists a constant $K \equiv K(a, b, s)$ such that

$$\mathbf{P}\{W > x\} \geq (K + o(1))(\bar{B}_I(x))^s.$$

We start with some auxiliary results. The proof of the theorem is given in Section 3.4.

### 3.1. An integral equality

**Lemma 4.** Let $f(y)$ be an integrable function. Put $f_I(y) \equiv \int_y^\infty f(z)dz$. Then, for any positive $\alpha$ and $\beta$, $\alpha > \beta$,

$$J \equiv \int_0^\infty \int_0^\infty f(\alpha y + \beta z) f(\beta y + \alpha z)\, dy\, dz$$

$$= \frac{(f_I(0))^2}{\alpha^2 - \beta^2} - \frac{2\beta}{\alpha^2 - \beta^2} \int_0^\infty f_I(\alpha u) f(\beta u)\, du.$$

*Proof.* Put $u = \alpha y + \beta z$ and $v = \beta y + \alpha z$. Then

$$J = \frac{1}{\alpha^2 - \beta^2} \int_0^\infty f(u)\, du \int_{\beta u/\alpha}^{\alpha u/\beta} f(v)\, dv$$

$$= \frac{1}{\alpha^2 - \beta^2} \int_0^\infty f(u) f_I(\beta u/\alpha)\, du$$

$$- \frac{1}{\alpha^2 - \beta^2} \int_0^\infty f(u) f_I(\alpha u/\beta)\, du$$

$$= \frac{\alpha}{\alpha^2 - \beta^2} \int_0^\infty f(\alpha u) f_I(\beta u)\, du$$

$$- \frac{\beta}{\alpha^2 - \beta^2} \int_0^\infty f(\beta u) f_I(\alpha u)\, du.$$

$\square$

Integration by parts yields

$$\int_0^\infty f_I(\beta u) f(\alpha u)\, du$$

$$= \frac{(f_I(0))^2}{\alpha} - \frac{\beta}{\alpha} \int_0^\infty f_I(\alpha u) f(\beta u)\, du.$$

By substituting this equality into the previous one, we arrive at the conclusion of the Lemma.

### 3.2. Some calculations with two big service times

Fix $\varepsilon > 0$ and put $b' = b - \varepsilon$. For $k$ and $l$, $k < l \le n$, define the events $A_{nkl}$ and $C_{nkl}$ by the equalities

$$A_{nkl} = \{\sigma_K > x + (l-k)a + (n-l)(a-b'),$$

$$\sigma_l > x + (n-l)(a-b'),$$

$$\sigma_k + \sigma_l > 2x + (l-k)a + (n-l)(2a-b')\}$$

and

$$C_{nkl} = \bigcap_{\substack{j=1 \\ j \ne k,l}}^{n} \{\sigma_j \le x + (n-j)(a-b')\}.$$

Note that the events $A_{nkl} \cap C_{nkl}$ are disjoint for different pairs $(k, l)$. Due to the existence of $\mathbf{E}\sigma$, uniformly in $n \ge 1$ and $k < l \le n$,

$$\mathbf{P}\{\bar{C}_{nkl}\} \le \sum_{j=0}^\infty \mathbf{P}\{\sigma_1 > x + j(a-b')\} \to 0 \quad \text{as } x \to \infty.$$

$$(7)$$

**Lemma 5.** Assume $b \in (0, a)$. Let the integrated tail distribution $B_I$ be long-tailed. Then, for any fixed $N \ge 1$ and for any $\varepsilon > 0$, as $x \to \infty$,

$$\lim_{n \to \infty} \sum_{\substack{k,l=1 \\ k<l}}^{n-N} \mathbf{P}\{A_{nkl}\} \sim \frac{1}{a(2a-b')}$$

$$\times \left[ (\bar{B}_I(x))^2 + b' \int_0^\infty \bar{B}_I(x + ya) \bar{B}(x + y(a-b'))\, dy \right].$$

*Proof.* Put

$$A'_{kl} = \{\sigma_1 > x + ka + l(a-b'), \sigma_2 > x + l(a-b'), \sigma_1 + \sigma_2 > 2x + ka + l(2a-b')\},$$

so that $\mathbf{P}\{A_{nkl}\} = \mathbf{P}\{A'_{l-k,n-l}\}$ and

$$\lim_{n \to \infty} \sum_{\substack{k,l=1 \\ k<l}}^{n-N} \mathbf{P}\{A_{nkl}\} = \lim_{n \to \infty} \sum_{l=N}^{n-1} \sum_{k=1}^{n-l-1} \mathbf{P}\{A'_{kl}\}$$

$$= \sum_{l=N}^\infty \sum_{k=1}^\infty \mathbf{P}\{A'_{kl}\}.$$

$$(8)$$

$\square$

Consider also the events

$$A(y, z) = \{\sigma_1 > x + ya + z(a-b'), \sigma_2 > x + z(a-b'),$$

$$\sigma_1 + \sigma_2 > 2x + ya + z(2a-b')\},$$

which satisfy $A(k, l) = A'_{kl}$. Since the probability $\mathbf{P}\{A(y, z)\}$ is non-increasing in $y$ and $z$, we have the inequalities

$$I_- \equiv \int_N^\infty \int_1^\infty \mathbf{P}\{A(y, z)\}\, dy\, dz \le \sum_{l=N}^\infty \sum_{k=1}^\infty \mathbf{P}\{A'_{kl}\}$$

$$\le \int_0^\infty \int_0^\infty \mathbf{P}\{A(y, z)\}\, dy\, dz \equiv I_+.$$

$$(9)$$

The values of integrals $I_-$ and $I_+$ are close to each other in the following sense:

$$I_+ - I_- \leq \int_0^N \int_0^\infty \mathbf{P}\{A(y, z)\}\, dy\, dz$$
$$+ \int_0^\infty \int_0^1 \mathbf{P}\{A(y, z)\}\, dy\, dz$$
$$\leq N\mathbf{P}\{\sigma_2 > x\} \int_0^\infty \mathbf{P}\{\sigma_1 > x + ya\}dy$$
$$+ \mathbf{P}\{\sigma_1 > x\} \int_0^\infty \mathbf{P}\{\sigma_1 > x + z(a - b')\}\, dz.$$

Recall that the distribution $\bar{B}_I(x)$ is long tailed, which is equivalent to $\bar{B}(x) = o(\bar{B}_I(x))$. Therefore, as $x \to \infty$,

$$I_+ - I_- \leq \frac{N + 1}{a - b'} \bar{B}(x)\bar{B}_I(x) = o((\bar{B}_I(x))^2).$$

Now it follows from (9) that, as $x \to \infty$,

$$\sum_{l=N}^\infty \sum_{k=1}^\infty \mathbf{P}\{A'_{kl}\} = \int_0^\infty \int_0^\infty \mathbf{P}\{A(y, z)\}dy\, dz$$
$$+ o((\bar{B}_I(x))^2). \qquad (10)$$

Further,

$$\mathbf{P}\{A(y, z)\}$$
$$= \bar{B}(x + ya + za)\bar{B}(x + z(a - b'))$$
$$+ \mathbf{P}\{x + ya + z(a - b') < \sigma_1 \leq x + ya + za,$$
$$\sigma_2 > x + z(a - b'), \sigma_1 + \sigma_2 > 2x$$
$$+ ya + z(2a - b')\}$$
$$= \bar{B}(x + ya + za)\bar{B}(x + z(a - b'))$$
$$+ \mathbf{P}\{x + ya + z(a - b') < \sigma_1 \leq x + ya + za,$$
$$\sigma_1 + \sigma_2 > 2x + ya + z(2a - b')\}$$
$$\equiv \bar{B}(x + ya + za)\bar{B}(x + z(a - b')) + Q(y, z),$$

since the event $\{\sigma_1 \leq x + ya + za, \sigma_1 + \sigma_2 > 2x + ya + z(2a - b')\}$ implies the event $\{\sigma_2 > x + z(a' - b)\}$. Consequently integrating over $y$ and $z$, we obtain

$$\int_0^\infty \int_0^\infty \bar{B}(x + ya + za)\bar{B}(x + z(a - b'))\, dy\, dz$$
$$= \frac{1}{a} \int_0^\infty \bar{B}_I(x + za)\bar{B}(x + z(a - b'))\, dz.$$

By the total probability formula,

$$Q(y, z) = \int_0^{zb'} \mathbf{P}\{\sigma_1 \in x + ya + z(a - b') + dt\}$$
$$\times \mathbf{P}\{\sigma_2 > x + za - t\}$$
$$= \int_0^{zb'} \bar{B}(x + za - t)B(x + ya + z(a - b') + dt).$$

The integration against $y$ leads to the equalities

$$\int_0^\infty Q(y, z)\, dy = \frac{1}{a} \int_0^{zb'} \bar{B}(x + za - t)$$
$$\times B_I(x + z(a - b') + dt)$$
$$= \frac{1}{a} \int_0^{zb'} \bar{B}(x + za - t)$$
$$\times \bar{B}(x + z(a - b') + t)\, dt$$
$$= \frac{b'}{a} \int_0^z \bar{B}(x + za - tb')$$
$$\times \bar{B}(x + z(a - b') + tb')\, dt.$$

Integrating against $z$, we obtain:

$$\int_0^\infty \int_0^\infty Q(y, z)dy\, dz = \frac{b'}{a} \int_0^\infty \int_0^z \bar{B}(x + za - tb')$$
$$\times \bar{B}(x + z(a - b') + tb')\, dt\, dz$$
$$= \frac{b'}{a} \int_0^\infty \int_t^\infty \bar{B}(x + za - tb')$$
$$\times \bar{B}(x + z(a - b') + tb')\, dz\, dt$$
$$= \frac{b'}{a} \int_0^\infty \int_0^\infty \bar{B}(x + za + t(a - b'))$$
$$\times \bar{B}(x + z(a - b') + ta)\, dz\, dt.$$

By Lemma 4 with $f(y) = \bar{B}(x + y), \alpha = a,$ and $\beta = a - b'$, the latter integral is equal to

$$\frac{1}{a(2a - b')}(\bar{B}_I(x))^2 - \frac{2(a - b')}{a(2a - b')}$$
$$\times \int_0^\infty \bar{B}_I(x + ya)\bar{B}(x + y(a - b'))\, dy.$$

Putting everything together into (10), we obtain the following equivalence, as $x \to \infty$:

$$\sum_{l=1}^\infty \sum_{k=1}^\infty \mathbf{P}\{A'_{kl}\} \sim \frac{1}{a(2a - b')}(\bar{B}_I(x))^2$$
$$+ \frac{b'}{a(2a - b')} \int_0^\infty \bar{B}_I(x + ya)\bar{B}(x + y(a - b'))\, dy,$$

which due to (8) completes the proof of the Lemma.

### 3.3. Proof of Theorem 3

If $\bar{B}_I(x)$ is long-tailed, then the function in $x$

$$(\bar{B}_I(x))^2 + b \int_0^\infty \bar{B}_I(x+ya)\bar{B}(x+y(a-b))\,dy$$

is long-tailed as well. Indeed, for any fixed $t$, we have, as $x \to \infty$,

$$\int_0^\infty \bar{B}_I(x+t+ya)\bar{B}(x+t+y(a-b))\,dy$$
$$\sim \int_0^\infty \bar{B}_I(x+ya)\bar{B}(x+t+y(a-b))\,dy.$$

Integrating by parts we get the equality for RHS integral

$$-\frac{1}{a-b}\bar{B}_I(x+ya)\bar{B}_I(x+t+y(a-b))\Big|_0^\infty$$
$$-\int_0^\infty \bar{B}(x+ya)\bar{B}_I(x+t+y(a-b))\,dy$$
$$\sim \frac{1}{a-b}(\bar{B}_I(x))^2 - \int_0^\infty \bar{B}(x+ya)\bar{B}_I(x+y(a-b))\,dy$$
$$= \int_0^\infty \bar{B}_I(x+ya)\bar{B}(x+y(a-b))\,dy.$$

So, we can apply Lemma 2, and it is sufficient to prove the lower bound of Theorem 3 for the queueing system $D/GI/2$ with deterministic input stream. Let the interarrival times $\tau_n$ be deterministic, i.e., $\tau_n \equiv a$. Then the event $A_{nkl}$ implies the event

$$\{W_{k+1,2} > x + (l-k)a + (n-l)(a-b') - a,$$
$$W_{l+1,1} > x + (n-l)(a-b') - a, \; W_{l+1,1} + W_{k+1,2}$$
$$> 2x + (l-k)a + (n-l)(2a-b') - 2a\},$$

which implies

$$\{W_{l+1,2}, W_{l+1,1} > x + (n-l)(a-b') - a,$$
$$W_{l+1,1} + W_{l+1,2} > 2x + (n-l)(2a-b') - 2a\}.$$

Thus, by Corollary 3 (with $\xi = \sigma - \tau$, and $\eta = -\tau$) there exists $N$ such that

$$\mathbf{P}\{W_n > x \mid A_{nkl}\} \geq 1 - \varepsilon \tag{11}$$

for any $n > N$ and $k < l < n - N$.

Taking into account that the events $A_{nkl} \cap C_{nkl}$ are disjoint for distinct pairs $(k, l)$, we obtain the following estimates:

$$\mathbf{P}\{W_n > x\} \geq \sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{W_n > x, A_{nkl}, C_{nkl}\}$$
$$\geq \sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{W_n > x, A_{nkl}\}$$
$$- \sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{A_{nkl}, \bar{C}_{nkl}\}.$$

Since the events $A_{nkl}$ and $C_{nkl}$ are independent,

$$\mathbf{P}\{W_n > x\} \geq \sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{W_n > x, A_{nkl}\}$$
$$- \sup_{kl}\mathbf{P}\{C_{nkl}\}\sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{A_{nkl}\}$$
$$= \sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{W_n > x \mid A_{nkl}\}\mathbf{P}\{A_{nkl}\}$$
$$- o(1)\sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{A_{nkl}\}$$

as $x \to \infty$ uniformly in $n$, by (7). Together with (11) it implies that, for sufficiently large $x$ and $n > N$,

$$\mathbf{P}\{W_n > x\} \geq (1-2\varepsilon)\sum_{k=1}^{n-N}\sum_{l=k+1}^{n-N}\mathbf{P}\{A_{nkl}\}.$$

Letting now $n \to \infty$, we derive from Lemma 5 the following lower bound, for all sufficiently large $x$:

$$\mathbf{P}\{W > x\} \geq \frac{1-3\varepsilon}{a(2a-b')}\left[(\bar{B}_I(x))^2 + b'\int_0^\infty \bar{B}_I(x+ya)\right.$$
$$\left.\times \bar{B}(x+y(a-b'))\,dy\right].$$

Note that, for any $b' < b < a$,

$$\int_0^\infty \bar{B}_I(x+ya)\bar{B}(x+y(a-b'))\,dy$$
$$\geq \frac{a-b}{a-b'}\int_0^\infty \bar{B}_I(x+ya)\bar{B}(x+y(a-b))\,dy.$$

We complete the proof of the Theorem by letting $\varepsilon \downarrow 0$.

## 4. The maximal stability case: An upper bound

**Theorem 4.**  Assume $b \in (0, a)$. Suppose that the distribution $B_I$ is subexponential. Then, as $x \to \infty$,

$$\mathbf{P}\{W > x\} \leq \frac{1 + o(1)}{a(2a - b)} \left[ (\bar{B}_I(x))^2 + b \int_0^\infty \bar{B}_I(x + ya) \right.$$

$$\left. \times \bar{B}(x + y(a - b)) \, dy \right].$$

By Lemma 1, it is sufficient to prove this upper bound for the queueing system $D/GI/2$ with deterministic input stream. So, let the interarrival times $\tau_n$ be deterministic, i.e., $\tau_n \equiv a$. Let $\sigma_n^{(1)}$ and $\sigma_n^{(2)}$, $n \geq 1$, be independent random variables with common distribution $B$. In this Section, define the service times $\sigma_n$ recursively. For that, we have to associate workloads with servers. Put $U_1 = (U_{1,1}, U_{1,2}) = (0, 0)$ and introduce the recursion

$$U_{n+1} = \left( U_n + e_{\alpha_n}\sigma_n - ia \right)^+ \tag{12}$$

where $\alpha_n = 1$ if $U_{n,1} < U_{n,2}$ and $\alpha_n = 2$ if $U_{n,1} > U_{n,2}$. If $U_{n,1} = U_{n,2}$, then $\alpha_n$ takes values 1 and 2 with equal probabilities independently of everything else. Note that $W_n = R(U_n)$ a.s. for any $n = 1, 2, \ldots$.

Now we can define $\sigma_n$ by induction. Indeed, $\alpha_0$ is chosen at random from the set $\{1, 2\}$. Put $\sigma_0 = \sigma_0^{(\alpha_0)}$. Then $U_1$ is defined by recursion (12) with $n = 0$. Assume that $U_n$ is defined for some $n > 0$. Then $\alpha_n$ is defined, too. Put $\sigma_n = \sigma_n^{(\alpha_n)}$ and determine $U_{n+1}$ by (12).

Due to the symmetry, for any $n$,

$$\mathbf{P}\{\alpha_n = 1\} = \mathbf{P}\{\alpha_n = 2\} = 1/2. \tag{13}$$

Consider two auxiliary $D/GI/1$ queueing systems which work in parallel: at any time instant $T_n = na$, $n = 1, 2, \ldots$, one customer arrives in the first queue and one in the second. Service times in queue $i = 1, 2$ are equal to $\sigma_n^{(i)}$. Denote by $W_n^{(i)}$, $i = 1, 2$, the waiting times in the $i$th queue and put $W_n^{(1)} = W_n^{(2)} = 0$. Since $b < a$, both queues are stable. Let $W^{(i)}$ be a stationary waiting time in the $i$th queue. By monotonicity, with probability 1,

$$W_n \leq \min\left(W_n^{(1)}, W_n^{(2)}\right) \tag{14}$$

for any $n \geq 1$. Hence,

$$W \leq \min\left\{W^{(1)}, W^{(2)}\right\}. \tag{15}$$

**Lemma 6.**  The waiting times $\{W_n^{(1)}\}$ and $\{W_n^{(2)}\}$ are independent.

*Proof* follows from the observation that the input (deterministic) stream and service times in the first queue do not depend on the input (also deterministic) stream and service times in the second one.

Provided $B_I$ is a subexponential distribution,

$$\mathbf{P}\left\{W^{(i)} > x\right\} \sim \frac{1}{a - b} \bar{B}_I(x) \quad \text{as } x \to \infty. \tag{16}$$

Then Lemma 6 together with (15) implies the following simple upper bound:

$$\limsup_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{(\bar{B}_I(x))^2} \leq \frac{1}{(a - b)^2}. \tag{17}$$

*Remark 3.*  For a $GI/GI/s$ queue with $a < b$ and subexponential distribution $B_I$, similar arguments lead to

$$\limsup_{x \to \infty} \frac{\mathbf{P}\{W > x\}}{(\bar{B}_I(x))^s} \leq \frac{1}{(a - b)^s}.$$

Introduce the events, for $k < n$,

$$A_{nk}^{(1)} = \left\{\sigma_k^{(1)} > x + (n - k)(a - b)\right\},$$

$$A_{nk}^{(2)} = \left\{\sigma_k^{(2)} > x + (n - k)(a - b)\right\}.$$

**Lemma 7.**  (See also [3, Theorem 5]). Provided the distribution $B_I$ is subexponential, for any fixed N,

$$\limsup_{n \to \infty} \mathbf{P}\left\{W_n^{(1)} > x, \bigcap_{k=1}^{n-N} \overline{A_{nk}^{(1)}}\right\} = o(\bar{B}_I(x)) \quad \text{as } x \to \infty.$$

*Proof.*  For any $\delta > 0$, consider the disjoint events

$$C_{nk}^{(1)} = \left\{ \left\{\sigma_k^{(1)} > x + (n - k)(a - b + \delta)\right\} \right.$$

$$\left. \cap \bigcap_{\substack{j=1 \\ j \neq k}}^{n-1} \left\{\sigma_j^{(1)} \leq x + (n - j)(a - b)\right\} \right\}.$$

Due to the Law of Large Numbers, there exists $M > N$ such that

$$\mathbf{P}\left\{W_n^{(1)} > x \mid C_{nk}^{(1)}\right\} \geq 1 - \delta$$

for any $n \geq M$ and $k \leq n - M$ and, by the limit at (7),

$$\mathbf{P}\{C_{nk}^{(1)}\} \geq (1 - \delta)\mathbf{P}\{\sigma_k^{(1)} > x + (n - k)(a - b + \delta)\}.$$

The events $C_{nk}^{(1)}, k \leq n - M$, are disjoint, hence,

$$\mathbf{P}\left\{W_n^{(1)} > x, \bigcup_{k=1}^{n-M} C_{nk}^{(1)}\right\} = \sum_{k=1}^{n-M} \mathbf{P}\{W_n^{(1)} > x, C_{nk}^{(1)}\}$$

$$\geq (1 - \delta)^2 \sum_{k=M}^{n-1} \mathbf{P}\{\sigma_k^{(1)} > x$$

$$+ k(a - b + \delta))\}.$$

The latter implies the following lower bound:

$$\liminf_{n \to \infty} \mathbf{P}\left\{W_n^{(1)} > x, \bigcup_{k=1}^{n-M} C_{nk}^{(1)}\right\}$$

$$\geq (1 - \delta)^2 \sum_{k=m}^{\infty} \bar{B}(x + k)(a - b + \delta)) \sim \frac{(1 - \delta)^2}{a - b + \delta} \bar{B}_I(x)$$

as $x \to \infty$. Since $A_{nk}^{(1)} \supseteq C_{nk}^{(1)}$ and since $M > N$ and $\delta > 0$ can be chosen arbitrarily,

$$\liminf_{n \to \infty} \mathbf{P}\left\{W_n^{(1)} > x, \bigcup_{k=1}^{n-N} A_{nk}^{(1)}\right\} \geq \frac{1 + o(1)}{a - b} \bar{B}_I(x)$$

$$\text{as } x \to \infty.$$

Together with (16), it implies the assertion of the Lemma.

*Proof of Theorem* 4 continued. Estimate (14) and Lemma 6 imply

$$\mathbf{P}\left\{W_n > x, \bigcap_{k=1}^{n-N} \overline{A_{nk}^{(1)}} \cup \bigcap_{l=1}^{n-N} \overline{A_{nl}^{(2)}}\right\}$$

$$\leq \mathbf{P}\left\{W_n^{(1)} > x, W_n^{(2)} > x, \bigcap_{k=1}^{n-N} \overline{A_{nk}^{(1)}} \cup \bigcap_{l=1}^{n-N} \overline{A_{nl}^{(2)}}\right\}$$

$$\leq \mathbf{P}\left\{W_n^{(1)} > x, \bigcap_{k=1}^{n-N} \overline{A_{nk}^{(1)}}\right\} \mathbf{P}\{W_n^{(2)} > x\}$$

$$+ \mathbf{P}\{W_n^{(1)} > x\} \mathbf{P}\left\{W_n^{(2)} > x, \bigcap_{l=1}^{n-N} \overline{A_{nl}^{(2)}}\right\}.$$

Applying now Lemma 7 and relation (16), we conclude that, as $x \to \infty$,

$$\liminf_{n \to \infty} \mathbf{P}\left\{W_n > x, \bigcap_{k=1}^{n-N} \overline{A_{nk}^{(1)}} \cup \bigcap_{l=1}^{n-N} \overline{A_{nl}^{(2)}}\right\} = o((\bar{B}_I(x))^2).$$

Since

$$\bigcap_{k=1}^{n-N} \overline{A_{nk}^{(1)}} \cup \bigcap_{l=1}^{n-N} \overline{A_{nl}^{(2)}} = \bigcap_{k,l=1}^{n-N} \left(\overline{A_{nk}^{(1)}} \cup \overline{A_{nl}^{(2)}}\right)$$

$$= \overline{\bigcup_{k,l=1}^{n-N} \left(A_{nk}^{(1)} \cap A_{nl}^{(2)}\right)},$$

we obtain the equivalent relation, as $x \to \infty$,

$$\liminf_{n \to \infty} \mathbf{P}\left\{W_n > x, \overline{\bigcup_{k,l=1}^{n-N} \left(A_{nk}^{(1)} \cap A_{nl}^{(2)}\right)}\right\} = o((\bar{B}_I(x))^2). \quad (18)$$

Fix $\varepsilon > 0$ and put $b' = b + \varepsilon$. For any $n$ and $k \leq l \leq n$, define

$$D_{nk}^{(1)} = \left\{\sigma_k^{(1)} > x + (l - k)a + (n - l)(a - b')\right\},$$

$$D_{nl}^{(2)} = \left\{\sigma_l^{(2)} > x + (n - l)(a - b')\right\},$$

$$D_{nkl} = \left\{\sigma_k^{(1)} + \sigma_l^{(2)} > 2x + (l - k)a + (n - l)(2a - b')\right\}.$$

For any $n$ and $l \leq k \leq n$, define

$$D_{nk}^{(1)} = \left\{\sigma_k^{(1)} > x + (n - k)(a - b')\right\},$$

$$D_{nl}^{(2)} = \left\{\sigma_l^{(2)} > x + (k - l)a + (n - k)(a - b')\right\},$$

$$D_{nkl} = \left\{\sigma_k^{(1)} + \sigma_l^{(2)} > 2x + (k - l)a + (n - k)(2a - b')\right\}.$$

Denote

$$F_{nkl} = D_{nk}^{(1)} \cap D_{nl}^{(2)} \cap D_{nkl}.$$

We can derive an upper bound on the probability of the event $\{W_n > x\}$ as follows:

$$\mathbf{P}\{W_n > x\} \leq \mathbf{P}\left\{W_n > x, \bigcup_{k,l=1}^{n-N} F_{nkl}\right\}$$

$$+ \mathbf{P}\left\{W_n > x, \overline{\bigcup_{k,l=1}^{n-N} F_{nkl}}, \bigcup_{k,l=1}^{n-N} \left(A_{nk}^{(1)} \cap A_{nl}^{(2)}\right)\right\}$$

$$+ \mathbf{P}\left\{W_n > x, \overline{\bigcup_{k,l=1}^{n-N} \left(A_{nk}^{(1)} \cap A_{nl}^{(2)}\right)}\right\}$$

$$\equiv P_{n1} + P_{n2} + P_{n3}. \quad (19)$$

Here the first term is not greater than

$$
P_{n1} \leq \mathbf{P}\left\{ W_n > x, \bigcup_{\substack{k,l=1 \\ k<l}}^{n-1} F_{nkl} \right\} + \mathbf{P}\left\{ W_n > x, \bigcup_{\substack{k,l=1 \\ k>l}}^{n-1} F_{nkl} \right\}
$$

$$
+ \mathbf{P}\left\{ W_n > x, \bigcup_{k=1}^{n-1} F_{nkk} \right\}
$$

$$
\equiv P_{n11} + P_{n12} + P_{n13}. \tag{20}
$$

The third probability is negligible in the sense that

$$
P_{n13} \leq \mathbf{P}\left\{ \bigcup_{k=1}^{n-1} \left( D_{nk}^{(1)} \cap D_{nk}^{(2)} \right) \right\} \leq \sum_{k=1}^{n-1} \mathbf{P}\{D_{nk}^{(1)}\}\mathbf{P}\{D_{nk}^{(2)}\}
$$

$$
\leq \bar{B}(x) \sum_{k=1}^{\infty} \bar{B}(x + k(a - b - \varepsilon))
$$

$$
\leq \bar{B}(x)\bar{B}_I(x)/(a - b - \varepsilon) = o((\bar{B}_I(x))^2) \tag{21}
$$

as $x \to \infty$, since $\bar{B}(x) = o(\bar{B}_I(x))$. The first probability in (20) admits the following upper bound:

$$
P_{n11} \leq \sum_{k=1}^{n-1} \mathbf{P}\left\{ W_n > x, D_{nk}^{(1)}, \alpha_k = 1, \bigcup_{l=k+1}^{n-1} \left( D_{nl}^{(2)} \cap D_{nkl} \right) \right\}
$$

$$
+ \sum_{k=1}^{n-1} \mathbf{P}\left\{ W_n > x, D_{nk}^{(1)}, \alpha_k = 2, \bigcup_{l=k+1}^{n-1} \left( D_{nl}^{(2)} \cap D_{nkl} \right) \right\}
$$

$$
\equiv \Sigma_1 + \Sigma_2.
$$

For $\Sigma_1$, we have the following inequality and equalities:

$$
\Sigma_1 \leq \sum_{\substack{k,l=1 \\ k<l}}^{n-1} P\{D_{nk}^{(1)}, \alpha_k = 1, D_{nl}^{(2)}, D_{nkl}\}
$$

$$
= \sum_{\substack{k,l=1 \\ k<l}}^{n-1} \mathbf{P}\{\alpha_k = 1\}\mathbf{P}\{D_{nk}^{(1)}, D_{nl}^{(2)}, D_{nkl}\}
$$

$$
= \frac{1}{2} \sum_{\substack{k,l=1 \\ k<l}}^{n-1} \mathbf{P}\{F_{nkl}\}, \tag{22}
$$

by independence of the event $\{\alpha_k = 1\}$ from $D_{nk}^{(1)}$, $D_{nl}^{(2)}$ and $D_{nkl}$ and by the symmetry (13). The sum $\Sigma_2$ is not greater than

$$
\Sigma_2 \leq \sum_{k=1}^{n-1} \mathbf{P}\{W_N > x, D_{nk}^{(1)}, \alpha_k = 2\}
$$

$$
= \sum_{k=1}^{n-1} \mathbf{P}\{D_{nk}^{(1)}\}\mathbf{P}\{W_n > x, \alpha_k = 2\}
$$

$$
\leq \mathbf{P}\{W_n > x\} \sum_{k=1}^{n-1} \mathbf{P}\{D_{nk}^{(1)}\}.
$$

Hence, $\Sigma_2 = o(\mathbf{P}\{W_n > x\})$ as $x \to \infty$ uniformly in $n$. Combining the latter fact with estimate (22) for $\Sigma_1$, we get

$$
P_{n11} \leq \frac{1}{2} \sum_{\substack{k,l=1 \\ k<l}}^{n-1} \mathbf{P}\{F_{nkl}\} + o(\mathbf{P}\{W_n > x\}). \tag{23}
$$

Taking into account the equality $P_{n11} = P_{n12}$, we obtain from (20), (21) and (23) the following estimate:

$$
P_{n1} \leq \sum_{\substack{k,l=1 \\ k<l}}^{n-1} \mathbf{P}\{F_{nkl}\} + o((\bar{B}_I(x))^2)
$$

as $x \to \infty$ uniformly in $n$. Now applying the calculations of Section 3.3 we can write down the following estimate, as $x \to \infty$:

$$
\limsup_{n \to \infty} P_{n1} \leq \frac{1 + o(1)}{a(2a - b')} \left[ (\bar{B}_I(x))^2 + b' \int_0^{\infty} \bar{B}_I(x + ya') \right.
$$

$$
\left. \times \bar{B}(x + y(a - b')) \, dy \right]. \tag{24}
$$

It is proved in (18) that, uniformly in $n$,

$$
P_{n3} = o\left((\bar{B}_I(x))^2\right) \quad \text{as } x \to \infty. \tag{25}
$$

We have

$$
\overline{\bigcup_{k,l=1}^{n-N} F_{nkl}} \cap \bigcup_{k,l=1}^{n-N} \left( A_{nk}^{(1)} \cap A_{nl}^{(2)} \right) \subseteq \bigcup_{k,l=1}^{n-N} \left( A_{nk}^{(1)} \cap A_{nl}^{(2)} \cap \bar{F}_{nkl} \right).
$$

Thus,

$$
P_{n2} \leq \sum_{k,l=1}^{n-N} \mathbf{P}\{W_n > x \mid A_{nk}^{(1)}, A_{nl}^{(2)}, \bar{F}_{nkl}\}\mathbf{P}\{A_{nk}^{(1)} \cap A_{nl}^{(2)}\}. \tag{26}
$$

Conditioning on $W_{nk}$ and $W_{nl}$ yields, for any $w > 0$;

$$
\mathbf{P}\{W_n > x \mid A_{nk}^{(1)}, A_{nl}^{(2)}, \bar{F}_{nkl}\}
$$

$$
\leq \mathbf{P}\{W_n > x \mid W_{k1} \leq w, W_{l2} \leq w, A_{nk}^{(1)}, A_{nl}^{(2)}, \bar{F}_{nkl}\}
$$

$$
+ \mathbf{P}\{W_{k1} > w\} + \mathbf{P}\{W_{l2} > w\}.
$$

Since $b < 2a$, the two-server queue is stable and, in particular, the sequence of distributions of random variables $(W_{n1}, W_{n2})$ is tight. It means that, for any fixed $\varepsilon > 0$, there exists $w$ such that, for any $k \geq 0$ and $l \geq 0$,

$$
\mathbf{P}\{W_{k1} > w\} \leq \varepsilon \quad \text{and} \quad \mathbf{P}\{W_{l2} > w\} \leq \varepsilon.
$$

Also, from the stability and from Corollary 4, for any fixed $\varepsilon > 0$, and $w > 0$, there exists $N$ such that, for any $n \geq N$ and $k, l \leq n - N$,

$$\mathbf{P}\{W_n > x \mid W_{k1} \leq w, W_{l2} \leq w, A_{nk}^{(1)}, A_{nl}^{(2)}, \bar{F}_{nkl}\} \leq \varepsilon.$$

Combining these estimates we obtain from (26),

$$P_{n2} \leq 3\varepsilon \sum_{k,l=1}^{n-N} \mathbf{P}\{A_{nk}^{(1)} \cap A_{nl}^{(2)}\} = 3\varepsilon \left(\sum_{k=1}^{n=N} \mathbf{P}\{A_{nk}^{(1)}\}\right)^2.$$

Hence,

$$P_{n2} \leq 3\varepsilon \left(\sum_{k=1}^{\infty} \bar{B}(x + k(a - b'))\right)^2 \leq \frac{3\varepsilon}{(a - b')^2}(\bar{B}_I(x))^2. \tag{27}$$

Since the choice of $\varepsilon > 0$ is arbitrary, relations (19)–(25) and (27) imply the conclusion of Theorem 4.

## 5. The minimal stability case: Lower bounds

**Theorem 5.** Let $b \in (a, 2a)$ and the integrated tail distribution $B_I$ be long tailed. Then the tail of the stationary waiting time satisfies the following inequality, for any fixed $\delta > 0$:

$$\mathbf{P}\{W > x\} \geq \frac{1 + o(1)}{2a - b} \bar{B}_I\left(\frac{b + \delta}{b - a}x\right) \text{ as } x \to \infty.$$

Notice that if $b \in (a, 2a)$ then $\frac{b}{b-a} > 2$.

*Remark 4.* By use of similar arguments, one can get the following result for an $s$-server queue, $s \geq 2$: if the integrated distribution $B_I$ is long tailed and $b \in ((s - 1)a, sa)$ then, for any $\delta > 0$,

$$\mathbf{P}\{W > x\} \geq \frac{1 + o(1)}{sa - b} \bar{B}_I\left(\frac{(s - 1)b - s(s - 2)a + \delta}{b - (s - 1)a}\right) \text{ as } x \to \infty.$$

Theorem 5 implies the following

**Corollary 6.** Assume that $B_I \in \mathscr{IRV}$. Then, as $x \to \infty$,

$$\mathbf{P}\{W > x\} \geq \frac{1 + o(1)}{2a - b} \bar{B}_I\left(\frac{b}{b - a}x\right).$$

In the case $b \in [a, 2a)$ one can also derive a lower bound which is similar to (6). More precisely, assume $b \in [a, 2a)$. Then introduce another two-server queue with the same service times and with inter-arrival times $\tau_n' = c\tau_n$, where $c > b/a$. For this queue, denote by $W'$ a stationary waiting time of a typical customer. Due to monotonicity, $\mathbf{P}\{W' > x\} \leq \mathbf{P}\{W > x\}$ for all $x$. Applying Theorem 3 and Remark 1, we get the following lower bound for the case $b \in [a, 2a)$: if the integrated tail distribution $B_I$ is long-tailed, then, for any $c > b/a$,

$$\mathbf{P}\{W > x\} \geq (1 + o(1))\frac{2ca + b}{2c^2a^2(2ca - b)}(\bar{B}_I(x))^2. \tag{28}$$

*Proof of Theorem 5.* By Lemma 2, it is sufficient to prove the lower bound for the queueing system $D/GI/2$ with deterministic input stream. Let the inter-arrival times $\tau_n$ be deterministic, i.e., $\tau_n \equiv a$. For any $\delta > 0$, set $\varepsilon = \frac{\delta(b-a)}{a+\delta}$. Put $b' = b - \varepsilon$ and $N = \frac{x}{b'-a}$. For any $k \in [1, n - N]$ consider the events

$$A_{nk} = \{\sigma_k > 2x + (2a - b')(n - k)\},$$

$$C_{nk} = \bigcap_{\substack{l=1 \\ l \neq k}}^{n} \{\sigma_l \leq 2x + (2a - b')(n - l)\}.$$

Since $\mathbf{E}\sigma$ is finite,

$$\mathbf{P}\{\bar{C}_{nk}\} \leq \sum_{l=1}^{\infty} \mathbf{P}\{\sigma_1 > 2x + (2a - b')l\} = O(\bar{B}_I(2x)) \to 0 \tag{29}$$

as $x \to \infty$ uniformly in $n \geq 1$ and $k \leq n$. Since the events $A_{nk} \cap C_{nk}, k \in [1, n]$, are disjoint, we obtain

$$\mathbf{P}\{W_n > x\} \geq \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x, A_{nk}, C_{nk}\}$$

$$\geq \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x, A_{nk}\} - \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}, \bar{C}_{nk}\}.$$

Since the events $A_{nk}$ and $C_{nk}$ are independent, we get

$$\mathbf{P}\{W_n > x\} \geq \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x, A_{nk}\}$$

$$- \sup_{k \leq n} \mathbf{P}\{\bar{C}_{nk}\} \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}\}$$

$$= \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x \mid A_{nk}\}\mathbf{P}\{A_{nk}\}$$

$$- o(1) \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}\} \tag{30}$$

as $x \to \infty$ uniformly in $n \geq 1$, by (29). The event $A_{nk}$ implies the event

$$W_{k+1,2} > 2x + (2a - b')(n - k) - a.$$

Thus, it follows from Corollary 5 that

$$\mathbf{P}\{W_n > x \mid A_{nk}\} \to 1$$

as $x \to \infty$ uniformly in $n$ and $k \leq n - N$. Therefore, we can derive from (30) the estimate

$$\mathbf{P}\{W > x\} = \lim_{n \to \infty} \mathbf{P}\{W_n > x\} \geq (1 - \varepsilon) \lim_{n \to \infty} \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}\}$$

$$= (1 - \varepsilon) \sum_{k=N}^{\infty} \bar{B}(2x + (2a - b')k),$$

which is valid for all sufficiently large $x$. Since the tail $\bar{B}_I(v)$ is long-tailed,

$$\sum_{k=N}^{\infty} \bar{B}(2x + (2a - b')k) \sim \frac{1}{2a - b'} \bar{B}_I(2x + (2a - b')k)N)$$

$$= \frac{1}{2a - b'} \bar{B}_I\left(\frac{b'}{b' - a}x\right) = \frac{1}{2a - b'} \bar{B}_I\left(\frac{b + \delta}{b - a}x\right)$$

as $x \to \infty$. The proof is complete. □

## 6. The minimal stability case: An upper bound

**Theorem 6.** Assume $b \in [a, 2a)$. Let both $B$ and $B_I$ be subexponential distributions. Then the tail of the stationary waiting time satisfies the following inequality, as $x \to \infty$:

$$\mathbf{P}\{W \geq x\} \leq \frac{1 + o(1)}{2a - b} \bar{B}_I(2x).$$

*Remark 5.* By use of the same arguments, one can get the following result for any $s$-server queue, $s \geq 2$: if $B_I \in \mathscr{S}$ and $b < sa$, then

$$\mathbf{P}\{W \geq x\} \leq \frac{1 + o(1)}{sa - b} \bar{B}_I(sx) \quad \text{as } x \to \infty$$

provided that either (i) $\sigma_1 \geq (s - 1)a$ a.s., or (ii) $B \in \mathscr{S}$.

*Remark 6.* For an $s$-server queue, Foss and Chernova [10] have proposed another way of obtaining upper bounds; it is based on comparison with a queue with the so-called *cyclic* service discipline.

*Proof of Theorem* 6. From Lemma 3, it is sufficient to consider the case of constant interarrival times $\tau_n \equiv a$ only. Put $M_{n,0} = 0$ and

$$M_{n,i+1} = (M_{n,i} + \sigma_{n+i} - a)^+.$$

Since $b > a$, $M_{0,n} \to \infty$ a.s. as $n \to \infty$ and, due to the Law of Large Numbers,

$$\frac{M_{0,n}}{n} \to b - a \quad \text{a.s.} \tag{31}$$

and in mean. Note that $\mathbf{E}M_{0,n} \geq n(b - a)$, since $M_{0,n} \geq \sigma_0 + \cdots + \sigma_{n-1} - na$. For any given $\varepsilon > 0$, choose an integer $L > 0$ such that

$$\frac{\mathbf{E}M_{0,L}}{L} \in [b - a, b - a + \varepsilon). \tag{32}$$

Consider any initial workload vector $W_0 = (W_{0,1}, W_{0,2}) \geq 0$. Put $Z_n = W_{n,1} + W_{n,2}$ Since the increments of the minimal coordinate of the waiting time vector is not greater than the increments of $M_{o,n}$,

$$W_{1,n} - W_{1,0} \leq M_{0,n} \quad \text{for any } n.$$

Hence, provided $W_{n,2} \geq a$, we have the inequality

$$Z_{n+1} - Z_n \leq M_{0,n+1} - M_{0,n} - a.$$

If $Z_0 \geq 2aL$, then $W_{0,2} \geq aL$ and, for $n = 0, \ldots, L - 1$, $W_{n,2} \geq a(L - n) \geq a$. Therefore, if $Z_0 \geq 2aL$, then

$$Z_L \leq Z_0 + M_{0,L} - aL.$$

Monotonicity implies, for any initial vector $W_0$,

$$Z_L \leq \max\{2aL, Z_0\} + M_{0,L} - aL.$$

Thus, the following inequalities are valid for any $n$:

$$Z_{(n+1)L} \leq \max\{2aL, Z_{nL}\} + M_{nL,L} - aL. \tag{33}$$

Consider a single-server queue with i.i.d. service times $\hat{\sigma}_n = M_{nL,L}$ and constant inter-arrival times $\hat{\tau}_n = La$ and denote by $\widehat{W}_n$ a waiting time of $n$th customer. This queue is stable since $\hat{b} \equiv \mathbf{E}\hat{\sigma}_1 < aL \equiv \hat{a}$. Put $\widehat{W}_0 = 0$. Assuming that $Z_0 = 0$, we can derive from (33) the following bounds: for all $n = 0, 1, \ldots$,

$$Z_{nL} \leq 2aL + \widehat{W}_n \quad \text{a.s.} \tag{34}$$

Denote $\bar{G}(x) = \mathbf{P}\{\hat{\sigma}_0 > x\}$. We show that the integrated tail distribution $G_I$ is a subexponential one. We need to consider only the case $L > 1$. Note first that

$$\sigma_0 + \cdots + \sigma_{L-1} - La \le \hat{\sigma}_0 \le \sigma_0 + \cdots + \sigma_{L-1} \quad \text{a.s.} \quad (35)$$

Since the distribution of $\sigma_1$, is assumed to be subexponential, the asymptotics for the lower and upper bounds in the latter inequalities are the same: as $x \to \infty$,

$$\mathbf{P}\left\{\sum_0^{L-1} \sigma_i - La > x\right\} \sim \mathbf{P}\left\{\sum_0^{L-1} \sigma_i > x\right\} \sim L\bar{B}(x). \quad (36)$$

Therefore, the tail $\bar{G}(x)$ has the same asymptotics and $G$ is a subexponential distribution. Thus,

$$\bar{G}_I(x) = \int_x^\infty \bar{G}(y)\,dy \sim L\bar{B}_I(x), \quad (37)$$

and $G_I$ is a subexponential distribution, too. Thus, by classic result (1) for the single server queue, the steady state distribution of the waiting time $\widehat{W}_n$ satisfies the following relations, as $x \to \infty$:

$$\lim_{n\to\infty} \mathbf{P}\{\widehat{W}_n > x\} \sim \frac{1}{\hat{a} - \hat{b}}\bar{G}_I(x) \le \frac{1}{(2a - b - \varepsilon)L}\bar{G}_I(x)$$
$$\sim \frac{1}{2a - b - \varepsilon}\bar{B}_I(x), \quad (38)$$

by (32) and (37). Since $Z_n = W_{n,1} + W_{n,2} \ge 2W_{n,1}$,

$$\mathbf{P}\{W > x\} = \mathbf{P}\{2W > 2x\} \le \lim_{n\to\infty} \mathbf{P}\{Z_{nL} > 2x\}.$$

Now it follows from (34) and (38) that

$$\mathbf{P}\{W > x\} \le \lim_{n\to\infty} \mathbf{P}\{\widehat{W}_n > 2x - 2aL\}$$
$$\le \frac{1 + o(1)}{2a - b - \varepsilon}\bar{B}_I(2x - 2aL)$$
$$\sim \frac{1}{2a - b - \varepsilon}\bar{B}_I(2x),$$

since $B_I$ is long-tailed. Letting $\varepsilon \downarrow 0$ concludes the proof. $\blacksquare$

## 7. The minimal stability case: Exact asymptotics

In this Section, we prove Theorem 2. First note that, as follows from (28), the tail $\mathbf{P}\{W > x\}$ may be heavier than that in Theorem 2, in general. For instance, this happens if

$$\bar{B}_I\left(\frac{b}{b-a}x\right) = o(\bar{B}_I^2(x)) \quad \text{as } x \to \infty. \quad (39)$$

Assume $b \in (a, 2a)$ and consider, for example, a service time distribution with the Weibull integrated tail $\bar{B}_I(x) = e^{-x^\beta}$, $\beta \in (0, 1)$. Then (39) holds if $(\frac{b}{b-a})^\beta > 2$.

*Proof of Theorem 2.* Since $B_I \in \mathscr{IRV}$, both the lower bound in Theorem 5 and the upper bound in Theorem 6 are of the same order,

$$\bar{B}_I(2x) = O\left(\bar{B}_I\left(\frac{b}{b-a}x\right)\right). \quad (40)$$

We use the notation from the previous Section. In particular, we fix $\varepsilon > 0$ and choose $L$ satisfying (32). For any constant $c \ge 0$, (35) implies

$$\bigcup_{i=0}^{L-1}\{\sigma_{kL+i} > x + La + (L - i)c\}$$
$$\subseteq \left\{\sum_{i=0}^{L-1}\sigma_{kL+i} - La > x\right\} \subseteq \{\hat{\sigma}_k > x\}.$$

Therefore, from (35) and (36),

$$\mathbf{P}\left\{\{\hat{\sigma}_k > x\} \setminus \bigcup_{i=0}^{L-1}\{\sigma_{kL+i} > x + La + (L - i)c\}\right\}$$
$$= o(\bar{B}(x)). \quad (41)$$

Take $c = (\hat{a} - \hat{b})/L$. By (34)

$$\mathbf{P}\{W > x\} = \lim_{n\to\infty} \mathbf{P}\{W_{nL,1} > x\}$$
$$= \lim_{n\to\infty} \mathbf{P}\{W_{nL,1} > x, \widehat{W}_n > 2x - 2aL\}.$$

Standard arguments concerning how large deviations in the single server queue $\widehat{W}_n$ occur imply the relation

$$\mathbf{P}\{W > x\} = \lim_{n\to\infty}\sum_{k=0}^{n-1}\mathbf{P}\{W_{nL,1} > x,$$
$$\hat{\sigma}_k > 2x + (n - k)(\hat{a} - \hat{b})\} + o(\bar{B}_I(2x))$$
$$= \lim_{n\to\infty}\sum_{i=0}^{nL-1}\mathbf{P}\{W_{nL,1} > x, \sigma_i > 2x + (n - i)c\}$$
$$+ o(\bar{B}_I(2x)),$$

by (41). Now it follows from (32) that

$$\mathbf{P}\{W > x\} = \lim_{n\to\infty}\sum_{k=0}^{nL-1}\mathbf{P}\{W_{nL,1} > x,$$
$$\sigma_i > 2x + (n - i)(2a - b - \varepsilon)\} + o(\bar{B}_I(2x))$$

$$\leq \lim_{n\to\infty} \sum_{j=1}^{nL} \mathbf{P}\{W_{nL,1} > x,$$

$$\sigma_{nL-j} > 2x + j(2a - b + \varepsilon)\} + \varepsilon O(\bar{B}_I(2x))$$

$$+ o(\bar{B}_I(2x)) = \lim_{n\to\infty} \left( \sum_{j=1}^{N(1-\varepsilon)} + \sum_{j=N(1-\varepsilon)}^{nL} \right)$$

$$+ \varepsilon O(\bar{B}_I(2x))$$

$$\equiv \lim_{n\to\infty} (\Sigma_1 + \Sigma_2) + \varepsilon O(\bar{B}_I(2x)),$$

where $N = x/(b - a)$. The second term admits the following estimate

$$\Sigma_2 \leq \sum_{j=N(1-\varepsilon)}^{\infty} \mathbf{P}\{\sigma > 2x + j(2a - b)\}$$

$$\sim \frac{1}{2a - b} \bar{B}_I(2x + N(1-\varepsilon)(2a - b))$$

$$= \frac{1}{2a - b} \bar{B}_I\left( \frac{b}{b - a} x - \varepsilon \frac{2a - b}{b - a} x \right).$$

It follows from $B_I \in \mathscr{RV}$ that, for any $\delta > 0$, there exists $\varepsilon > 0$ such that

$$\Sigma_2 \leq \frac{1}{2a - b} \bar{B}_I\left( \frac{b}{b - a} x \right) + \delta \bar{B}_I(2x),$$

which coincides with the lower bound in Theorem 5.

Now consider the first term $\Sigma_1$. Since the queue is stable, one can choose $K > 0$ such that $\mathbf{P}\{W_{n,2} \leq K\} \geq 1 - \varepsilon$ for all $K$. Then

$$\Sigma_1 \leq \sum_{j=1}^{N(1-\varepsilon)} \mathbf{P}\{W_{nL-j,2} > K, \sigma_{nL-j} > 2x + (2a - b + \varepsilon)j\}$$

$$+ \sum_{j=1}^{N(1-\varepsilon)} \mathbf{P}\{W_{nL,1} > x, W_{nL-j,2} \leq K,$$

$$\sigma_{nL-j} > 2x + (2a - b + \varepsilon)j\}$$

$$\equiv \Sigma_{1,1} + \Sigma_{1,2}.$$

We have

$$\Sigma_{1,1} \leq \sum_{j=1}^{N(1-\varepsilon)} \mathbf{P}\{W_{nL-j,2} > K\} \mathbf{P}\{\sigma_1 > 2x + (2a - b + \varepsilon)j\}$$

$$\leq \varepsilon \sum_{j=1}^{\infty} \mathbf{P}\{\sigma_1 > 2x + (2a - b)j\} \leq \frac{\varepsilon}{2a - b} \bar{B}_I(2x).$$

Note that if $W_{nL-j,2} \leq K$, then $W_{nL,1} \leq K + M_{nL-j+1,j-1}$. Therefore,

$$\Sigma_{1,2} \leq \sum_{j=1}^{N(1-\varepsilon)} \mathbf{P}\{\sigma_{nL-j} > 2x + (2a - b)j,$$

$$K + M_{nL-j+1,j-1} > x\}$$

$$= \sum_{j=1}^{N(1-\varepsilon)} \mathbf{P}\{\sigma_{nL-j} > 2x + (2a - b)j\}$$

$$\times \mathbf{P}\{K + M_{0,j-1} > x\}.$$

Since the sequence $M_{0,j}$ stochastically increases,

$$\Sigma_{1,2} \leq \mathbf{P}\{K + M_{0,N(1-\varepsilon)} > x\} \sum_{j=1}^{\infty} \mathbf{P}\{\sigma_1 > 2x + (2a - b)j\}$$

$$\leq \mathbf{P}\{M_{0,N(1-\varepsilon)} > x - K\} \frac{1}{2a - b} \bar{B}_I(2x).$$

Since

$$\frac{x - K}{N(1-\varepsilon)} \to \frac{b - a}{1 - \varepsilon} > b - a \quad \text{as } x \to \infty,$$

we have by (31)

$$\mathbf{P}\{M_{0,N(1-\varepsilon)} > x - K\} = \mathbf{P}\left\{ \frac{M_{N(1-\varepsilon)}}{N(1-\varepsilon)} > \frac{x - K}{N(1-\varepsilon)} \right\} \to 0.$$

Thus, we have shown that the upper bound for $\mathbf{P}\{W > x\}$ is not bigger than the lower bound in Theorem 5 plus a term of order

$$(\varepsilon + \delta) O(B_I(2x)) \leq (\varepsilon + \delta) O\left( B_I\left( \frac{bx}{b - a} \right) \right)$$

due to (40). Since $\varepsilon > 0$ and $\delta > 0$ may be chosen as small as we please, the proof of Theorem 2 is complete.

## 8. Tail asymptotics for the two-dimensional workload vector

Denote by $W^0 = (W_1^0, W_2^0)$ a weak limit for the vectors $W_n$ as $n \to \infty$. Clearly, $W = W_1^0$.

### 8.1. Maximal stability case

First, we obtain simple lower and upper bounds which are equivalent up to some constant. Second, we give (without a proof) a result related to the exact asymptotics.

**Theorem 7.** Let $b < a$ and $B_I \in \mathscr{L}$ Then, as $x, y \to \infty, x \le y$,

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \ge \frac{1 + o(1)}{a^2} \bar{B}_I(x)\bar{B}_I(y).$$

If, in addition, $B_I \in \mathscr{S}$, then

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \le \frac{2 + o(1)}{(a - b)^2} \bar{B}_I(x)\bar{B}_I(y).$$

*Proof.* Fix $\varepsilon > 0$ and put $a' = a + \varepsilon$. For $k, l \le n, k \ne l$, define the events $A_{nkl}$ and $C_{nkl}$ by the equalities

$$A_{nkl} = \{\sigma_k > x + (n - k)a', \sigma_l > y + (n - l)a'\}$$

and

$$C_{nkl} = \bigcap_{\substack{j=1 \\ j \ne k,l}}^{n} \{\sigma_j \le x + (n - j)a'\}.$$

Note that the events $A_{nkl} \cap C_{nkl}$ are disjoint for different pairs $(k, l)$ and

$$\mathbf{P}\{W_{n1} > x, W_{n2} > y\} \ge \sum_{k=1}^{n} \sum_{l=k+1}^{n} \mathbf{P}\{W_{n1} > x,$$
$$W_{n2} > y, A_{nkl}, C_{nkl}\}.$$

Then the same calculations as in Section 3.3 imply the estimate, as $x, y \to \infty$,

$$\mathbf{P}\{W_{n1} > x, W_{n2} > y\}$$
$$\ge (1 + o(1)) \sum_{k=1}^{n} \sum_{l=k+1}^{n} \bar{B}(x + (n - k)a')\bar{B}(y + (n - l)a')$$
$$= (1 + o(1)) \sum_{k=1}^{n-1} \sum_{l=1}^{n-k-1} \bar{B}(x + ka')\bar{B}(y + la').$$

Hence,

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \ge (1 + o(1)) \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \bar{B}(x + ka')$$
$$\times \bar{B}(y + la') \sim \bar{B}_I(x)\bar{B}_I(y)/a'^2$$

and the lower bound is proved.

Proceed to the upper bound. Due to construction of the majorant $(W_n^{(1)}, W_n^{(2)})$ in Section 4, we have the inequality

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \le \lim_{n \to \infty} \left[\mathbf{P}\{W_n^{(1)} > x, W_n^{(2)} > y\}\right.$$
$$+ \mathbf{P}\{W_n^{(1)} > y, W_n^{(2)} > x\}\right]$$
$$= 2 \lim_{n \to \infty} \mathbf{P}\{W_n^{(1)} > x\}\mathbf{P}\{W_n^{(2)} > y\}.$$

Together with (16) it implies the desired upper bound. Theorem 7 is proved.

Turn now to the exact asymptotics. Below is the result. The proof is rather complicated and will be presented in another paper. Denote

$$R(x, y) = \bar{B}_I(x)\bar{B}_I(y) + b \int_0^{\infty} \bar{B}_I(y + za)$$
$$\times \bar{B}(x + x(a - b)) \, dz.$$

Recall that Theorem 1 states that $\mathbf{P}\{W_1^0 > x\} \sim R(x, x)/a(2a - b)$ given $B_I \in \mathscr{S}$.

**Theorem 8.** Assume $b < a$ and $B_I \in \mathscr{S}$. Let $x, y \to \infty, x \le y$. Then

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\}$$
$$\sim \frac{1}{a(2a - b)} R(y, y) + \frac{1}{a^2}(R(x, y) - R(y, y)).$$

### 8.2. Minimal stability case

We prove the following

**Theorem 9.** Assume $a < b < 2a$, $B \in \mathscr{S}$, and $B_I \in \mathscr{IRV}$. Let $x, y, \to \infty$ in such a way that $y/x \to c \in [1, \infty]$. Then

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \sim \frac{1}{a}\bar{B}_I\left(y\left(1 + \frac{a}{c(b - a)}\right)\right)$$
$$+ \frac{b - a}{a(2a - b)}\bar{B}_I\left(y\frac{b}{b - a}\right).$$

*Proof.* Start with the case $c = \infty$. From Theorem 10 in [3], one can get the following:

**Corollary 7.** Assume $b \in (a, 2a)$. If $B \in \mathscr{S}$ and $B_I \in \mathscr{S}$, then, as $y \to \infty$,

$$\mathbf{P}\{W_2^0 > y\} \sim \frac{1}{a}\bar{B}_I(y) + \frac{b - a}{a(2a - b)}\bar{B}_I\left(y\frac{b}{b - a}\right).$$

It is clear that

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \le \mathbf{P}\{W_2^0 > y\}.$$

On the other hand, for any $N = 1, 2, \ldots,$

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\} = \lim_{n \to \infty} \mathbf{P}\{W_{n,1} > x, W_{n,2} > y\}$$

$$\ge \lim_{n \to \infty} \mathbf{P}\left\{W_{n-N,2} > y + Na, \sum_{j=n-N}^{n-1}(\sigma_j - \tau_j) > x\right\}$$

$$= \lim_{n \to \infty} \mathbf{P}\{W_{n-N,2} > y + Na\}\mathbf{P}\left\{\sum_{j=1}^{N}(\sigma_j - \tau_j) > x\right\}.$$

Fix $\varepsilon > 0$. Put $N = N(x) = x(1+\varepsilon)/(b-a)$. Then by LLN

$$\mathbf{P}\left\{\sum_{j=1}^{N}(\sigma_j - \tau_j) > x\right\} \ge 1 - \varepsilon$$

for all sufficiently large $x$ and, as $n \to \infty$,

$$\mathbf{P}\{W_{n-N,2} > y + Na\} \to \mathbf{P}\{W_2^0 > y + Na\}.$$

Since $B_I \in \mathscr{IRV}$,

$$\mathbf{P}\{W_2^0 > y + Na\} \sim \mathbf{P}\{W_2^0 > y\} \quad \text{as } y \to \infty.$$

By letting $\varepsilon \to 0$, we get the result.

Now consider the case $c < \infty$. If $c = 1$, then the result follows from Theorem 2. Let $c \in (1, \infty)$. We give here only a sketch of the proof, by making links to the proof of Theorem 2.

Since

$$\mathbf{P}\{W_1^0 > y\} \le \mathbf{P}\{W_1^0 > x, W_2^0 > y\} \le \mathbf{P}\{W_1^0 > x\}$$

and

$$\mathbf{P}\{W_1^0 > y\} \sim \mathbf{P}\{W_1^0 > cx\} \ge (K + o(1))\mathbf{P}\{W_1^0 > x\}$$

where $K = \inf_t \bar{B}_I(ct)/\bar{B}_I(t) > 0$, one can get from the proof of Theorem 2 the following equivalences: for $N_x = x/(b-a)$, $N_y = y/(b-a)$, and for $\varepsilon \in (0, 1 - 1/\sqrt{c})$,

$$\mathbf{P}\{W_1^0 > x, W_2^0 > y\}$$

$$= \lim_{n \to \infty} \sum_{i=1}^{n-N_x(1-\varepsilon)} \mathbf{P}\{W_{n,1} > x, W_{n,2} > y,$$

$$\sigma_i > 2x + (n-i)(2a-b)\} + \varepsilon O(\bar{B}_I(2x))$$

$$= \lim_{n \to \infty}\left(\sum_{i=1}^{n-N_y(1+\varepsilon)} + \sum_{i=n-N_y(1+\varepsilon)}^{n-N_x(1-\varepsilon)}\right) + \varepsilon O(\bar{B}_I(2x))$$

$$\equiv (\Sigma_1 + \Sigma_2) + \varepsilon O(\bar{B}_I(2x)).$$

Choose $K > 0$ such that $\mathbf{P}\{W_{n,2} > K\} \le \varepsilon$ for all $n$. Then

$$\Sigma_2 = \lim_{n \to \infty} \sum_{i=n-N_y(1+\varepsilon)}^{n-N_x(1-\varepsilon)} \mathbf{P}\{W_{i,2} \le K, W_{n,1} > x, W_{n,2} > y,$$

$$\sigma_i > 2x + (n-i)(2a-b)\} + \varepsilon O(\bar{B}_I(2x)).$$

From Lemma 2 and its Corollaries,

$$\Sigma_2 = (1 + o(1)) \sum_{j=N_x(1-\varepsilon)}^{N_y(1+\varepsilon)} \mathbf{P}\{\sigma_1 > y + ja\} + \varepsilon O(\bar{B}_I(x))$$

$$= \frac{1 + o(1)}{a}\left(\bar{B}_I\left(y + \frac{x(1-\varepsilon)a}{b-a}\right) - \bar{B}_I\left(\frac{y(b+\varepsilon a)}{b-a}\right)\right)$$

$$\quad + \varepsilon O(\bar{B}_I(x))$$

$$= \frac{1 + o(1)}{a}\left(\bar{B}_I\left(y\left(1 + \frac{a}{c(b-a)}\right)\right) - \bar{B}_I\left(\frac{yb}{b-a}\right)\right)$$

$$\quad + (\varepsilon + \delta)O(\bar{B}_I(x)),$$

due to $B_I \in \mathscr{IRV}$. From Lemma 3 and its Corollaries, one can also conclude that, for $i < n - N_y(1 + \varepsilon)$, if $\sigma_i < 2y + (n-i)(2a - b - \varepsilon)$ and $W_{i,2} \le K$, then, with probability close to one, both coordinates of the vector $(W_{n,1}, W_{n,2})$ take values less then $y$ for all sufficiently large $n$. From the other side, if $\sigma_i > 2y + (n-i)(2a - b + \varepsilon)$ then, with probability close to one, $y < W_{n,1} \le W_{n,2}$. Therefore,

$$\Sigma_1 = (1 + o(1)) \sum_{j=N_y(1+\varepsilon)}^{\infty} \mathbf{P}\{\sigma_1 > 2y + j(2a - b)\}$$

$$\quad + \varepsilon O(\bar{B}_I(x))$$

$$= \frac{1 + o(1)}{2 - b}\bar{B}_I\left(\frac{yb}{b-a}\right) + \varepsilon O(\bar{B}_I(x)).$$

Summing up the terms and letting $\varepsilon$ and $\delta \to 0$ concludes the proof.

## 9. Comments on stationary queue length

Let $Q_n$ be a queue length viewed by an arriving customer $n$, and $Q$ its stationary version in discrete time (i.e. Palm-stationary). Due to the distributional Little's law,

$$\mathbf{P}\{Q > n\} = \mathbf{P}\{W > T_n\}$$

where $W$ is the stationary waiting time, $T_n = \tau_1 + \cdots + \tau_n$, and $W$, and $T_n$ do not depend on each other. When a distribution of $W$ is long-tailed, the asymptotics for $\mathbf{P}\{W > T_n\}$, $n \to \infty$, have been found in [2] and in [11]. If, in addition, $\tau_n$ has a non-lattice distribution, there exists a stationary

distribution $G$ for a queue length in continuous time. Then, from Lemma 1 in [11],

$$\bar{G}(n) \sim \mathbf{P}\{Q > n\} \quad \text{as } n \to \infty.$$

## References

1. S. Asmussen, *Applied Probability and Queues*, 2nd ed. (Springer, New York, 2003).
2. S. Asmussen, C. Kluppelberg and K. Sigman, Sampling at subexponential times, with queueing applications, Stoch. Process. Appl 79 (1999) 265–286.
3. F. Baccelli and S. Foss, Moments and tails in monotone-separable stochastic networks, Ann. Appl. Probab 14 (2004) 612–650.
4. A.A. Borovkov and A.A. Mogul'skii, Large deviations for Markov chains in the positive quadrant, Russ. Math. Surv 56 (2001) 803–916.
5. S. Borst, M. Mandjes and A.P. Zwart, Exact asymptotics for fluid queues fed by heavy-tailed On-Off flows, Ann. Appl. Probab 14 (2004) 903–957.
6. O.J. Boxma, S.G. Foss, J.-M. Lasgouttes and R. Nunez Queija, Waiting time asymptotics in the single server queue with service in random order, Queueing Systems 46 (2004) 35–73.
7. O.J. Boxma, Q. Deng and A.P. Zwart, Waiting-time asymptotics for the $M/G/2$ queue with heterogeneous servers, Queueing Systems 40 (2002) 5–31.
8. H. Cramér, *Collective Risk Theory* (Esselte, Stockholm, 1955).
9. D. Denisov, S. Foss and D. Korshunov, Tail asymptotics for the supremum of a random walk when the mean is not finite, Queueing Systems 46 (2004) 15–33.
10. S.G. Foss and N.I. Chernova, On optimality of FCFS discipline in multi-channel queueing systems and networks, Siberian Math. J 42 (2001) 372–385.
11. S. Foss and D. Korshunov, Sampling at a random time with a heavy-tailed distribution, Markov Processes and Related Fields 6 (2000) 643–658.
12. I.A. Ignatyuk, V. Malyshev and V. Scherbakov, Boundary effects in large deviation problems, Russ. Math. Surv 49 (1994) 41–99.
13. J. Kiefer and J. Wolfowitz, On the theory of queues with many servers, Tran. Amer. Math. Soc 78 (1955) 1–18.
14. D. Korshunov, On distribution tail of the maximum of a random walk, Stochastic Process. Appl 72 (1997) 97–103.
15. A.G. Pakes, On the tails of waiting-time distribution, J. Appl. Probab 12 (1975) 555–564.
16. A. Scheller-Wolf, Further delay moment results for FIFO multi-server queues, Queueing Systems 34 (2000) 387–400.
17. A. Scheller-Wolf and K. Sigman, Delay moments for FIFO $GI/GI/s$ queues, Queueing Systems 25 (1997) 77–95.
18. N. Veraverbeke, Asymptotic behavior of Wiener-Hopf factors of a random walk, Stochastic Process. Appl 5 (1977) 27–37.
19. W. Whitt, The impact of a heavy-tailed service-time distribution upon the $M/GI/s$ waiting-time distribution, Queueing Systems 36 (2000) 71–87.