Heavy Tails in Multi-Server Queue[1]

SERGUEI FOSS foss@ma.hw.ac.uk
Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, Scotland
DMITRY KORSHUNOV korshunov@math.nsc.ru
Sobolev Institute of Mathematics, Novosibirsk 630090, Russia

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.

1. Introduction

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

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

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

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

As usual we say that a distribution G on $\mathbf{R}^+$ is subexponential (belongs to the class $\mathscr{S}$) if $\overline{G * G}(x) \sim 2\overline{G}(x)$ 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

[1]Supported by EPSRC grant No. R58765/01, INTAS Project No. 00-265 and RFBR grant No. 02-01-00358

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 ρ in the system. More precisely, it depends on ρ via the value of $k \in \{0, 1, \dots, s-1\}$ for which $k \le \rho < k+1$. In particular, Whitt conjectured that

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

"where γ and η 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 $R(w) = (R_1(w), \dots, R_s(w))$ be the operator on $\mathbf{R}^s$ which orders the coordinates of $w \in \mathbf{R}^s$ in ascending order, i.e., $R_1(w) \le \dots \le R_s(w)$. Then the residual work vector $W_n = (W_{n1}, \dots, W_{ns})$ which the nth customer observes just upon its arrival satisfies the celebrated Kiefer–Wolfowitz recursion: $W_1 = i \cdot 0$,

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

here $e_1 = (1, 0, \dots, 0)$, $i = (1, \dots, 1)$ and $w^+ = (\max(0, w_1), \dots, \max(0, w_s))$. 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 [12, 4]. 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 $x \to \infty$,*

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

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,

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

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 $\overline{G}$) is *regularly varying* at infinity with index $\gamma > 0$ (belongs to the class $\mathscr{RV}$), if $\overline{G}(x) > 0$ for all x and, for any fixed $c > 0$, $\overline{G}(cx)/\overline{G}(x) \to c^{-\gamma}$ as $x \to \infty$.

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

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

where

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

Recall definitions of a number of classes of heavy-tailed distributions. A distribution G is *long-tailed* (belongs to the class $\mathscr{L}$) if $\overline{G}(x) > 0$ for all x and, for any fixed t,

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

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

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

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*

$$\mathbf{P}\{W > x\} \sim \frac{1}{2a-b} \overline{B}_I\Big(\frac{b}{b-a} x\Big) \qquad \text{as } x \to \infty.$$

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 $\overline{G}(x) > 0$ for all x and $\liminf_{x\to\infty} \overline{G}(2x)/\overline{G}(x) > 0$. 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

$$W'_{n+1} = R(W'_n + e_1\sigma_n - ia')^+.$$

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

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

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

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

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

First, we use induction to prove the inequality

$$W_n \le W'_n + iM_n \quad \text{a.s.} \tag{2}$$

Indeed, for $n = 1$ we have $0 \le 0 + iM_1$. Assume the inequality is proved for some n; we prove it for $n+1$. Indeed,

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

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

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

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:

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

Since the random variable ξ_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,

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

The inequality (2) implies that $W \le_{\rm st} W'+M$, where W' and M are independent. Let a random variable η have distribution G and be independent of M. Since $\eta \ge_{\rm st} W'$, we have $W \le_{\rm st} \eta + M$. Therefore, for any $h > 0$,

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

by (3). Integrating by parts yields

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

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

$$\overline{G}(x-1) \le \overline{G}(x) e^{\varepsilon}$$

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

$$\overline{G}(x-y) \le c(\varepsilon)\overline{G}(x) e^{\varepsilon y}$$

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

$$\int_h^x \overline{G}(x-y) e^{-\beta y} dy \le c(\varepsilon)\overline{G}(x) \int_h^x e^{-\beta y/2} dy \le \frac{c(\varepsilon)}{\beta/2} \overline{G}(x) e^{-\beta h/2}.$$

Hence,

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

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

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

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 times $\tau'_n \equiv a'$. Let W' be a stationary waiting time in this auxiliary system.

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

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

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

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

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

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

Indeed, by induction arguments,

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

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

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

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 $W \ge_{\rm st} W' - M$ where W' and M are independent. Therefore, for any $h > 0$,

$$\mathbf{P}\{W > x\} \ge \mathbf{P}\{W' > x + h\}\mathbf{P}\{M \le h\} \ge \overline{G}(x+h)\mathbf{P}\{M \le h\}.$$

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

$$\liminf_{x\to\infty} \frac{\mathbf{P}\{W > x\}}{\overline{G}(x)} \ge \mathbf{P}\{M \le h\}.$$

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* (ξ_n, η_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

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

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

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

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 $U_n \ge 0$, 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*

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

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

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

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

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

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

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

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*

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

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

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

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 $x \to \infty$,*

$$\mathbf{P}\{W > x\} \geq \frac{1+o(1)}{a(2a-b)}\Big[(\overline{B}_I(x))^2 + b\int_0^\infty \overline{B}_I(x+ya)\overline{B}(x+y(a-b))dy\Big]. \quad (6)$$

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

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

R e m a r k 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

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

We start with some auxiliary results. The proof of the theorem is given in subsection 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 α and β, $\alpha > \beta$,*

$$J \equiv \int_0^\infty\int_0^\infty f(\alpha y+\beta z)f(\beta y+\alpha z)dydz$$
$$= \frac{(f_I(0))^2}{\alpha^2-\beta^2} - \frac{2\beta}{\alpha^2-\beta^2}\int_0^\infty f_I(\alpha u)f(\beta u)du.$$

P r o o f. Put $u = \alpha y+\beta z$ and $v = \beta y+\alpha z$. Then

$$\begin{aligned} J &= \frac{1}{\alpha^2-\beta^2}\int_0^\infty f(u)du\int_{\beta u/\alpha}^{\alpha u/\beta} f(v)dv \\ &= \frac{1}{\alpha^2-\beta^2}\int_0^\infty f(u)f_I(\beta u/\alpha)du - \frac{1}{\alpha^2-\beta^2}\int_0^\infty f(u)f_I(\alpha u/\beta)du \\ &= \frac{\alpha}{\alpha^2-\beta^2}\int_0^\infty f(\alpha u)f_I(\beta u)du - \frac{\beta}{\alpha^2-\beta^2}\int_0^\infty f(\beta u)f_I(\alpha u)du. \end{aligned}$$

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

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

and

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

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 $k < l \le n$,

$$\mathbf{P}\{\overline{C}_{nkl}\} \le \sum_{j=0}^{\infty} \mathbf{P}\{\sigma_1 > x + j(a-b')\} \to 0 \quad \text{as } x \to \infty. \tag{7}$$

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 $x \to \infty$,*

$$\lim_{n\to\infty} \sum_{\substack{k,l=1\\ k<l}}^{n-N} \mathbf{P}\{A_{nkl}\} \sim \frac{1}{a(2a-b')}\left[(\overline{B}_I(x))^2 + b'\int_0^\infty \overline{B}_I(x+ya)\overline{B}(x+y(a-b'))dy\right].$$

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

$$\lim_{n\to\infty} \sum_{\substack{k,l=1\\ k<l}}^{n-N} \mathbf{P}\{A_{nkl}\} = \lim_{n\to\infty} \sum_{l=N}^{n-1} \sum_{k=1}^{n-l-1} \mathbf{P}\{A'_{kl}\} = \sum_{l=N}^{\infty} \sum_{k=1}^{\infty} \mathbf{P}\{A'_{kl}\}. \tag{8}$$

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

$$I_- \equiv \int_N^\infty \int_1^\infty \mathbf{P}\{A(y,z)\}dydz \le \sum_{l=N}^{\infty} \sum_{k=1}^{\infty} \mathbf{P}\{A'_{kl}\}\\ \le \int_0^\infty \int_0^\infty \mathbf{P}\{A(y,z)\}dydz \equiv I_+. \tag{9}$$

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

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

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

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

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

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

Further,

$$\begin{aligned}
&\mathbf{P}\{A(y,z)\} \\
&= \overline{B}(x+ya+za)\overline{B}(x+z(a-b')) \\
&\quad + \mathbf{P}\Big\{x+ya+z(a-b') < \sigma_1 \le x+ya+za,\ \sigma_2 > x+z(a-b'), \\
&\qquad\qquad\qquad \sigma_1+\sigma_2 > 2x+ya+z(2a-b')\Big\} \\
&= \overline{B}(x+ya+za)\overline{B}(x+z(a-b')) \\
&\quad + \mathbf{P}\Big\{x+ya+z(a-b') < \sigma_1 \le x+ya+za,\ \sigma_1+\sigma_2 > 2x+ya+z(2a-b')\Big\} \\
&\equiv \overline{B}(x+ya+za)\overline{B}(x+z(a-b')) + Q(y,z),
\end{aligned}$$

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

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

By the total probability formula,

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

The integration against y leads to the equalities

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

Integrating against z, we obtain:

$$\begin{aligned}\int_0^\infty\int_0^\infty Q(y,z)dydz &= \frac{b'}{a}\int_0^\infty\int_0^{z} \overline{B}(x+za-tb')\overline{B}(x+z(a-b')+tb')dtdz\\ &= \frac{b'}{a}\int_0^\infty\int_t^{\infty} \overline{B}(x+za-tb')\overline{B}(x+z(a-b')+tb')dzdt\\ &= \frac{b'}{a}\int_0^\infty\int_0^{\infty} \overline{B}(x+za+t(a-b'))\overline{B}(x+z(a-b')+ta)dzdt.\end{aligned}$$

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

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

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

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

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

3.3. Proof of Theorem 3. If $\overline{B}_I(x)$ is long-tailed, then the function in x

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

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

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

Integrating by parts we get the equality for RHS integral

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

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 τ_n be deterministic, i.e., $\tau_n \equiv a$. Then the event A_{nkl} implies the event

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

which implies

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

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:

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

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

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

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:

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

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

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

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 $x \to \infty$,*

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

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 τ_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 σ_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

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

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 α_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 σ_n by induction. Indeed, α_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 α_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 ith 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 ith queue. By monotonicity, with probability 1,

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

for any $n \geq 1$. Hence,

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

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,

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

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

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

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

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

Introduce the events, for $k < n$,

$$\begin{aligned} A_{nk}^{(1)} &= \{\sigma_k^{(1)} > x + (n-k)(a-b)\}, \\ A_{nk}^{(2)} &= \{\sigma_k^{(2)} > x + (n-k)(a-b)\}. \end{aligned}$$

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

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

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

$$C_{nk}^{(1)} = \Big\{\Big\{\sigma_k^{(1)} > x + (n-k)(a-b+\delta)\Big\} \cap \bigcap_{\substack{j=1 \\ j\ne k}}^{n-1} \Big\{\sigma_j^{(1)} \le x + (n-j)(a-b)\Big\}\Big\}.$$

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

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

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

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

The events $C^{(1)}_{nk}$, $k \le n-M$, are disjoint, hence,

$$\begin{aligned}\mathbf{P}\Big\{W^{(1)}_n > x,\ \bigcup_{k=1}^{n-M} C^{(1)}_{nk}\Big\} &= \sum_{k=1}^{n-M} \mathbf{P}\{W^{(1)}_n > x,\ C^{(1)}_{nk}\} \\ &\ge (1-\delta)^2 \sum_{k=M}^{n-1} \mathbf{P}\{\sigma^{(1)}_k > x + k(a-b+\delta)\}.\end{aligned}$$

The latter implies the following lower bound:

$$\begin{aligned}\liminf_{n\to\infty} \mathbf{P}\Big\{W^{(1)}_n > x,\ \bigcup_{k=1}^{n-M} C^{(1)}_{nk}\Big\} &\ge (1-\delta)^2 \sum_{k=M}^{\infty} \overline{B}(x + k(a-b+\delta)) \\ &\sim \frac{(1-\delta)^2}{a-b+\delta}\overline{B}_I(x)\end{aligned}$$

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

$$\liminf_{n\to\infty} \mathbf{P}\Big\{W^{(1)}_n > x,\ \bigcup_{k=1}^{n-N} A^{(1)}_{nk}\Big\} \ge \frac{1+o(1)}{a-b}\overline{B}_I(x) \quad \text{as } x \to \infty.$$

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

P r o o f of Theorem 4 continued. Estimate (14) and Lemma 6 imply

$$\begin{aligned}\mathbf{P}\Big\{W_n > x,\ \bigcap_{k=1}^{n-N} \overline{A^{(1)}_{nk}} \cup \bigcap_{l=1}^{n-N} \overline{A^{(2)}_{nl}}\Big\} &\le \mathbf{P}\Big\{W^{(1)}_n > x,\ W^{(2)}_n > x,\ \bigcap_{k=1}^{n-N} \overline{A^{(1)}_{nk}} \cup \bigcap_{l=1}^{n-N} \overline{A^{(2)}_{nl}}\Big\} \\ &\le \mathbf{P}\Big\{W^{(1)}_n > x,\ \bigcap_{k=1}^{n-N} \overline{A^{(1)}_{nk}}\Big\}\mathbf{P}\{W^{(2)}_n > x\} \\ &\quad + \mathbf{P}\{W^{(1)}_n > x\}\mathbf{P}\Big\{W^{(2)}_n > x,\ \bigcap_{l=1}^{n-N} \overline{A^{(2)}_{nl}}\Big\}.\end{aligned}$$

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

$$\limsup_{n\to\infty} \mathbf{P}\Big\{W_n > x,\ \bigcap_{k=1}^{n-N} \overline{A^{(1)}_{nk}} \cup \bigcap_{l=1}^{n-N} \overline{A^{(2)}_{nl}}\Big\} = o((\overline{B}_I(x))^2).$$

Since

$$\bigcap_{k=1}^{n-N} \overline{A^{(1)}_{nk}} \cup \bigcap_{l=1}^{n-N} \overline{A^{(2)}_{nl}} = \bigcap_{k,l=1}^{n-N} \Big(\overline{A^{(1)}_{nk}} \cup \overline{A^{(2)}_{nl}}\Big) = \overline{\bigcup_{k,l=1}^{n-N} \Big(A^{(1)}_{nk} \cap A^{(2)}_{nl}\Big)},$$

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

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

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

$$\begin{aligned} D_{nk}^{(1)} &= \{\sigma_k^{(1)} > x + (l-k)a + (n-l)(a-b')\}, \\ D_{nl}^{(2)} &= \{\sigma_l^{(2)} > x + (n-l)(a-b')\}, \\ D_{nkl} &= \{\sigma_k^{(1)} + \sigma_l^{(2)} > 2x + (l-k)a + (n-l)(2a-b')\}. \end{aligned}$$

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

$$\begin{aligned} D_{nk}^{(1)} &= \{\sigma_k^{(1)} > x + (n-k)(a-b')\}, \\ D_{nl}^{(2)} &= \{\sigma_l^{(2)} > x + (k-l)a + (n-k)(a-b')\}, \\ D_{nkl} &= \{\sigma_k^{(1)} + \sigma_l^{(2)} > 2x + (k-l)a + (n-k)(2a-b')\}. \end{aligned}$$

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:

$$\begin{aligned} &\mathbf{P}\{W_n > x\} \\ &\le \mathbf{P}\Big\{W_n > x,\ \bigcup_{k,l=1}^{n-N} F_{nkl}\Big\} + \mathbf{P}\Big\{W_n > x,\ \overline{\bigcup_{k,l=1}^{n-N} F_{nkl}},\ \bigcup_{k,l=1}^{n-N} \Big(A_{nk}^{(1)} \cap A_{nl}^{(2)}\Big)\Big\} \\ &\quad + \mathbf{P}\Big\{W_n > x,\ \overline{\bigcup_{k,l=1}^{n-N} \Big(A_{nk}^{(1)} \cap A_{nl}^{(2)}\Big)}\Big\} \\ &\equiv P_{n1} + P_{n2} + P_{n3}. \end{aligned} \tag{19}$$

Here the first term is not greater than

$$\begin{aligned} P_{n1} &\le \mathbf{P}\Big\{W_n > x,\ \bigcup_{\substack{k,l=1 \\ k<l}}^{n-1} F_{nkl}\Big\} + \mathbf{P}\Big\{W_n > x,\ \bigcup_{\substack{k,l=1 \\ k>l}}^{n-1} F_{nkl}\Big\} + \mathbf{P}\Big\{W_n > x,\ \bigcup_{k=1}^{n-1} F_{nkk}\Big\} \\ &\equiv P_{n11} + P_{n12} + P_{n13}. \end{aligned} \tag{20}$$

The third probability is negligible in the sense that

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

$$\begin{aligned}
&\le \overline{B}(x)\sum_{k=1}^{\infty}\overline{B}(x+k(a-b-\varepsilon))\\
&\le \overline{B}(x)\overline{B}_I(x)/(a-b-\varepsilon)=o((\overline{B}_I(x))^2) \qquad (21)
\end{aligned}$$

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

$$\begin{aligned}
P_{n11} &\le \sum_{k=1}^{n-1}\mathbf{P}\Big\{W_n>x,\ D_{nk}^{(1)},\alpha_k=1,\ \bigcup_{l=k+1}^{n-1}(D_{nl}^{(2)}\cap D_{nkl})\Big\}\\
&\quad+\sum_{k=1}^{n-1}\mathbf{P}\Big\{W_n>x,\ D_{nk}^{(1)},\alpha_k=2,\ \bigcup_{l=k+1}^{n-1}(D_{nl}^{(2)}\cap D_{nkl})\Big\}\equiv\Sigma_1+\Sigma_2.
\end{aligned}$$

For Σ_1, we have the following inequality and equalities:

$$\begin{aligned}
\Sigma_1 &\le \sum_{\substack{k,l=1\\k<l}}^{n-1}\mathbf{P}\Big\{D_{nk}^{(1)},\alpha_k=1,D_{nl}^{(2)},D_{nkl}\Big\}\\
&=\sum_{\substack{k,l=1\\k<l}}^{n-1}\mathbf{P}\{\alpha_k=1\}\mathbf{P}\Big\{D_{nk}^{(1)},D_{nl}^{(2)},D_{nkl}\Big\}=\frac{1}{2}\sum_{\substack{k,l=1\\k<l}}^{n-1}\mathbf{P}\{F_{nkl}\}, \qquad (22)
\end{aligned}$$

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 Σ_2 is not greater than

$$\begin{aligned}
\Sigma_2 &\le \sum_{k=1}^{n-1}\mathbf{P}\Big\{W_n>x,\ D_{nk}^{(1)},\alpha_k=2\Big\}\\
&=\sum_{k=1}^{n-1}\mathbf{P}\{D_{nk}^{(1)}\}\mathbf{P}\Big\{W_n>x,\ \alpha_k=2\Big\}\le\mathbf{P}\{W_n>x\}\sum_{k=1}^{n-1}\mathbf{P}\{D_{nk}^{(1)}\}.
\end{aligned}$$

Hence, $\Sigma_2=o(\mathbf{P}\{W_n>x\})$ as $x\to\infty$ uniformly in n. Combining the latter fact with estimate (22) for Σ_1, we get

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

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

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

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

$$\limsup_{n\to\infty} P_{n1} \le \frac{1+o(1)}{a(2a-b')}\Big[(\overline{B}_I(x))^2 + b'\int_0^\infty \overline{B}_I(x+ya')\overline{B}(x+y(a-b'))dy\Big]. \tag{24}$$

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

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

We have

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

Thus,

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

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

$$\begin{aligned}\mathbf{P}\{W_n > x \mid A^{(1)}_{nk}, A^{(2)}_{nl}, \overline{F}_{nkl}\} \le{}& \mathbf{P}\{W_n > x \mid W_{k1} \le w, W_{l2} \le w, A^{(1)}_{nk}, A^{(2)}_{nl}, \overline{F}_{nkl}\}\\ &+\mathbf{P}\{W_{k1} > w\} + \mathbf{P}\{W_{l2} > w\}.\end{aligned}$$

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 $k \ge 0$ and $l \ge 0$,

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

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

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

Combining these estimates we obtain from (26),

$$P_{n2} \le 3\varepsilon \sum_{k,l=1}^{n-N} \mathbf{P}\{A^{(1)}_{nk} \cap A^{(2)}_{nl}\} = 3\varepsilon\Big(\sum_{k=1}^{n-N} \mathbf{P}\{A^{(1)}_{nk}\}\Big)^2.$$

Hence,

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

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 $\delta > 0$:*

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

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 $\delta > 0$,

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

Theorem 5 implies the following

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

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

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 $c > b/a$,

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

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 τ_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

$$\begin{aligned} A_{nk} &= \{\sigma_k > 2x + (2a - b')(n - k)\}, \\ C_{nk} &= \bigcap_{\substack{l=1 \\ l \neq k}}^{n} \{\sigma_l \leq 2x + (2a - b')(n - l)\}. \end{aligned}$$

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

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

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

$$\begin{aligned}\mathbf{P}\{W_n > x\} &\ge \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x, A_{nk}, C_{nk}\} \\ &\ge \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x, A_{nk}\} - \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}, \overline{C}_{nk}\}.\end{aligned}$$

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

$$\begin{aligned}\mathbf{P}\{W_n > x\} &\ge \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x, A_{nk}\} - \sup_{k\le n} \mathbf{P}\{\overline{C}_{nk}\} \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}\} \\ &= \sum_{k=1}^{n-N} \mathbf{P}\{W_n > x \mid A_{nk}\}\mathbf{P}\{A_{nk}\} - o(1) \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}\} \qquad (30)\end{aligned}$$

as $x \to \infty$ uniformly in $n \ge 1$, 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 $k \le n - N$. Therefore, we can derive from (30) the estimate

$$\begin{aligned}\mathbf{P}\{W > x\} = \lim_{n\to\infty} \mathbf{P}\{W_n > x\} &\ge (1-\varepsilon) \lim_{n\to\infty} \sum_{k=1}^{n-N} \mathbf{P}\{A_{nk}\} \\ &= (1-\varepsilon) \sum_{k=N}^{\infty} \overline{B}(2x + (2a-b')k),\end{aligned}$$

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

$$\begin{aligned}\sum_{k=N}^{\infty} \overline{B}(2x + (2a-b')k) &\sim \frac{1}{2a-b'} \overline{B}_I(2x + (2a-b')N) \\ &= \frac{1}{2a-b'} \overline{B}_I\Big(\frac{b'}{b'-a}x\Big) = \frac{1}{2a-b'} \overline{B}_I\Big(\frac{b+\delta}{b-a}x\Big)\end{aligned}$$

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 $x \to \infty$:*

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

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

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

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 $M_{0,n} \geq \sigma_0 + \ldots + \sigma_{n-1} - na$. 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 $M_{0,n}$,

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

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,

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

Thus, the following inequalities are valid for any n:

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

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

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

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

$$\sigma_0 + \ldots + \sigma_{L-1} - La \le \widehat{\sigma}_0 \le \sigma_0 + \ldots + \sigma_{L-1} \quad \text{a.s.} \tag{35}$$

Since the distribution of σ_1 is assumed to be subexponential, the asymptotics for the lower and upper bounds in the latter inequalities are the same: as $x \to \infty$,

$$\mathbf{P}\Big\{\sum_0^{L-1} \sigma_i - La > x\Big\} \sim \mathbf{P}\Big\{\sum_0^{L-1} \sigma_i > x\Big\} \sim L\overline{B}(x). \tag{36}$$

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

$$\overline{G}_I(x) = \int_x^\infty \overline{G}(y)dy \sim L\overline{B}_I(x). \tag{37}$$

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 $x \to \infty$:

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

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

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

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

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

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

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

$$\overline{B}_I(2x) = O\Big(\overline{B}_I\Big(\frac{b}{b-a}x\Big)\Big). \tag{40}$$

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

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

Therefore, from (35) and (36),

$$\mathbf{P}\Big\{\{\widehat{\sigma}_k > x\} \setminus \bigcup_{i=0}^{L-1}\{\sigma_{kL+i} > x + La + (L-i)c\}\Big\} = o(\overline{B}(x)). \tag{41}$$

Take $c = (\widehat{a} - \widehat{b})/L$. 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

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

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

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

$$\le \lim_{n\to\infty}\sum_{j=1}^{nL}\mathbf{P}\{W_{nL,1}>x,\sigma_{nL-j}>2x+j(2a-b+\varepsilon)\}+\varepsilon O(\overline{B}_I(2x))+o(\overline{B}_I(2x))$$
$$=\lim_{n\to\infty}\Bigg(\sum_{j=1}^{N(1-\varepsilon)}+\sum_{j=N(1-\varepsilon)}^{nL}\Bigg)+\varepsilon O(\overline{B}_I(2x))\equiv\lim_{n\to\infty}(\Sigma_1+\Sigma_2)+\varepsilon O(\overline{B}_I(2x)),$$

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

$$\begin{aligned}\Sigma_2 &\le \sum_{j=N(1-\varepsilon)}^{\infty}\mathbf{P}\{\sigma>2x+j(2a-b)\}\\ &\sim \frac{1}{2a-b}\overline{B}_I(2x+N(1-\varepsilon)(2a-b))=\frac{1}{2a-b}\overline{B}_I\Big(\frac{b}{b-a}x-\varepsilon\frac{2a-b}{b-a}x\Big).\end{aligned}$$

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

$$\Sigma_2\le\frac{1}{2a-b}\overline{B}_I\Big(\frac{b}{b-a}x\Big)+\delta\overline{B}_I(2x),$$

which coincides with the lower bound in Theorem 5.

Now consider the first term Σ_1. Since the queue is stable, one can choose $K>0$ such that $\mathbf{P}\{W_{n,2}\le K\}\ge 1-\varepsilon$ for all k. Then

$$\begin{aligned}\Sigma_1 &\le \sum_{j=1}^{N(1-\varepsilon)}\mathbf{P}\{W_{nL-j,2}>K,\sigma_{nL-j}>2x+(2a-b+\varepsilon)j\}\\ &\quad+\sum_{j=1}^{N(1-\varepsilon)}\mathbf{P}\{W_{nL,1}>x,W_{nL-j,2}\le K,\sigma_{nL-j}>2x+(2a-b+\varepsilon)j\}\\ &\equiv \Sigma_{1,1}+\Sigma_{1,2}.\end{aligned}$$

We have

$$\begin{aligned}\Sigma_{1,1} &= \sum_{j=1}^{N(1-\varepsilon)}\mathbf{P}\{W_{nL-j,2}>K\}\mathbf{P}\{\sigma_1>2x+(2a-b+\varepsilon)j\}\\ &\le \varepsilon\sum_{j=1}^{\infty}\mathbf{P}\{\sigma_1>2x+(2a-b)j\}\le\frac{\varepsilon}{2a-b}\overline{B}_I(2x).\end{aligned}$$

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

$$\begin{aligned}\Sigma_{1,2} &\le \sum_{j=1}^{N(1-\varepsilon)}\mathbf{P}\{\sigma_{nL-j}>2x+(2a-b)j,K+M_{nL-j+1,j-1}>x\}\\ &= \sum_{j=1}^{N(1-\varepsilon)}\mathbf{P}\{\sigma_{nL-j}>2x+(2a-b)j\}\mathbf{P}\{K+M_{0,j-1}>x\}.\end{aligned}$$

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

$$\begin{aligned}\Sigma_{1,2} &\le \mathbf{P}\{K + M_{0,N(1-\varepsilon)} > x\} \sum_{j=1}^{\infty} \mathbf{P}\{\sigma_1 > 2x + (2a-b)j\} \\ &\le \mathbf{P}\{M_{0,N(1-\varepsilon)} > x - K\} \frac{1}{2a-b} \overline{B}_I(2x).\end{aligned}$$

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)

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

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

$$(\varepsilon+\delta) O(B_I(2x)) \le (\varepsilon+\delta) O\Big(B_I\Big(\frac{bx}{b-a}\Big)\Big)$$

due to (40). Since $\varepsilon > 0$ and $\delta > 0$ may be chosen as small 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, $y \to \infty$, $x \le y$,*

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

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

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

P r o o f. 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

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

and

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

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 Subsection 3.3 imply the estimate, as x, $y \to \infty$,

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

Hence,

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

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

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

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

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

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*

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

P r o o f. 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 $y \to \infty$,*

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

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 $N = 1, 2, \ldots,$

$$\begin{aligned}\mathbf{P}\{W_1^0 > x, W_2^0 > y\} &= \lim_{n\to\infty}\mathbf{P}\{W_{n,1} > x, W_{n,2} > y\} \\ &\ge \lim_{n\to\infty}\mathbf{P}\Big\{W_{n-N,2} > y + Na, \sum_{j=n-N}^{n-1}(\sigma_j - \tau_j) > x\Big\} \\ &= \lim_{n\to\infty}\mathbf{P}\{W_{n-N,2} > y + Na\}\mathbf{P}\Big\{\sum_{j=1}^{N}(\sigma_j - \tau_j) > x\Big\}.\end{aligned}$$

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

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

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 $B_I \in \mathscr{IRV}$,

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

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 $K = \inf_t \overline{B}_I(ct)/\overline{B}_I(t) > 0$, 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 $\varepsilon \in (0, 1 - 1/\sqrt{c})$,

$$\begin{aligned}
&\mathbf{P}\{W_1^0 > x, W_2^0 > y\} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n-N_x(1-\varepsilon)} \mathbf{P}\{W_{n,1} > x, W_{n,2} > y, \sigma_i > 2x + (n-i)(2a-b)\} + \varepsilon O(\overline{B}_I(2x)) \\
&= \lim_{n\to\infty} \Bigg(\sum_{i=1}^{n-N_y(1+\varepsilon)} + \sum_{i=n-N_y(1+\varepsilon)}^{n-N_x(1-\varepsilon)} \Bigg) + \varepsilon O(\overline{B}_I(2x)) \equiv (\Sigma_1 + \Sigma_2) + \varepsilon O(\overline{B}_I(2x)).
\end{aligned}$$

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

$$\Sigma_2 = \lim_{n\to\infty} \sum_{i=n-N_y(1+\varepsilon)}^{n-N_x(1-\varepsilon)} \mathbf{P}\{W_{i,2} \le K, W_{n,1} > x, W_{n,2} > y, \sigma_i > 2x + (n-i)(2a-b)\} \\ + \varepsilon O(\overline{B}_I(2x)).$$

From Lemma 2 and its Corollaries,

$$\begin{aligned}
\Sigma_2 &= (1 + o(1)) \sum_{j=N_x(1-\varepsilon)}^{N_y(1+\varepsilon)} \mathbf{P}\{\sigma_1 > y + ja\} + \varepsilon O(\overline{B}_I(x)) \\
&= \frac{1+o(1)}{a}\Big(\overline{B}_I\Big(y + \frac{x(1-\varepsilon)a}{b-a}\Big) - \overline{B}_I\Big(\frac{y(b+\varepsilon a)}{b-a}\Big)\Big) + \varepsilon O(\overline{B}_I(x)) \\
&= \frac{1+o(1)}{a}\Big(\overline{B}_I\Big(y\Big(1 + \frac{a}{c(b-a)}\Big)\Big) - \overline{B}_I\Big(\frac{yb}{b-a}\Big)\Big) + (\varepsilon + \delta) O(\overline{B}_I(x)),
\end{aligned}$$

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,

$$\begin{aligned}
\Sigma_1 &= (1 + o(1)) \sum_{j=N_y(1+\varepsilon)}^{\infty} \mathbf{P}\{\sigma_1 > 2y + j(2a-b)\} + \varepsilon O(\overline{B}_I(x)) \\
&= \frac{1+o(1)}{2a-b}\overline{B}_I\Big(\frac{yb}{b-a}\Big) + \varepsilon O(\overline{B}_I(x)).
\end{aligned}$$

Summing up the terms and letting ε 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, $T_n = \tau_1 + \ldots + \tau_n$, 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, τ_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],

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

Acknowledgment

The authors gratefully acknowledge helpful discussions with Onno Boxma and Bert Zwart, and comments from Daryl Daley.

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 multiserver 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.