Sojourn time tails in the single server queue with heavy-tailed service times

Onno Boxma*, Denis Denisov†

January 8, 2010

Abstract

We consider the $GI/GI/1$ queue with regularly varying service requirement distribution of index $-\alpha$. It is well known that, in the $M/G/1$ FCFS queue, the sojourn time distribution is also regularly varying, of index $1-\alpha$, whereas in the case of LCFS or Processor Sharing, the sojourn time distribution is regularly varying of index $-\alpha$. That raises the question whether there exist service disciplines that give rise to a regularly varying sojourn time distribution with any index $-\gamma \in [-\alpha, 1-\alpha]$. In this paper that question is answered affirmatively.

Keywords: $GI/GI/1$ queue, regular variation, sojourn time tail.

1 Introduction

Traffic in high-speed communication networks exhibits burstiness on a wide range of time scales. This manifests itself in phenomena like long-range dependence and self-similarity. An explanation for these phenomena is found in heavy-tailed characteristics of the underlying activity patterns, like connection times, scene lengths and file sizes.
Heavy-tailed traffic characteristics have a dramatic effect on flow-level delays experienced by users. Hence there is much interest in the influence of scheduling and priority mechanisms on these delays. In [2] a survey is presented concerning the impact of the service discipline in single-server queues on delay asymptotics, for the case of the $M/G/1$ queue with regularly varying service requirement distribution. A distribution $F(\cdot)$ on $[0,\infty)$ is called regularly varying of index $-\alpha$ if

$$1 - F(x) = x^{-\alpha} L(x), \qquad x \geq 0, \tag{1}$$

where $L(\cdot)$ is a slowly varying function, i.e., $\lim_{x\to\infty} L(\eta x)/L(x) = 1$, $\eta > 1$. Let W_{FCFS}, $W_{LCFS-PR}$ and W_{PS} be the stationary sojourn times for the First-Come-First-Served

*EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

†School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK (D.Denisov@ma.hw.ac.uk); present address: Cardiff University

(FCFS), Last-Come-First-Served Preemptive Resume (LCFS-PR) and Processor Sharing (PS) single server queues. In the sequel, $g(x) \sim h(x)$ denotes $\lim_{x\to\infty} g(x)/h(x) = 1$. It is known that, cf. [2], in the $M/G/1$ case, when B is regularly varying of index $-\alpha$,

$$\mathbf{P}\{W_{FCFS} > x\} \sim C_{FCFS} x \mathbf{P}\{B > x\}, \tag{2}$$

$$\mathbf{P}\{W_{LCFS-PR} > x\} \sim C_{LCFS-PR} \mathbf{P}\{B > x\}, \tag{3}$$

$$\mathbf{P}\{W_{PS} > x\} \sim C_{PS} \mathbf{P}\{B > x\}, \tag{4}$$

as $x \to \infty$. More precisely, W_{FCFS} is regularly varying of index $1-\alpha$, whereas $W_{LCFS-PR}$ and W_{PS} are regularly varying of index $-\alpha$. The former also is shown to hold for LCFS non-preemptive, whereas the latter is shown to hold for Foreground-Background Processor Sharing (FBPS) and Shortest Remaining Processing Time First (SRPTF). In Section 7 of [2] this raised the question whether (given that the service requirement distribution is regularly varying of index $-\alpha$), $-\alpha$ and $1-\alpha$ are the only possible indices of the sojourn time distribution in a work-conserving single server queue. The main goal of the present paper is to show that any index between $-\alpha$ and $1-\alpha$ can occur. We do this by devising a work-conserving service discipline with a particular parameter, and by showing that various choices of that parameter lead to any index $-\gamma \in [-\alpha, 1-\alpha]$.

There are various trade-offs involved with the choice of a service discipline. We just mention a few aspects, referring to the *Special issue on new perspectives in scheduling* [4] for extensive discussions of concepts like fairness, tails in scheduling and scheduling classifications. One classical fact is that, in the $M/G/1$ queue, the waiting time variance is minimal for FCFS [7] and maximal for LCFS [9], among all service disciplines which do not affect the distribution of the number of customers in the system at any time. Another good property of FCFS is that it is logarithmically optimal for light-tailed service times [8].

On the other hand, (2) reveals a bad property of FCFS. While FCFS is good in terms of mean and variance of the sojourn time, there is an essential drawback: if a customer with an excessively large service demand arrives to the queue, then many subsequent customers might suffer significant delays. This is not a very fair system from the point of view of other customers. In addition this system encourages big customers to ask for an even bigger amount of service.

Several types of service disciplines can be designed to avoid such a situation. In some respects the fairest one is the Round Robin service discipline (and Processor Sharing as the limit). In this service discipline every customer receives a quantum of service and then returns to the end of the queue and waits for its turn again until eventually it gets served completely. In this service discipline a customer with big service demand will not affect other customers as strongly as in the FCFS queue. However, a possible drawback is that the system wastes its resources switching from one customer to another.

LCFS-PR also has this drawback of (possibly often) interrupting the service of a customer, which might lead to a waste of resources. The class of service disciplines that we shall introduce, parameterized by a parameter $\beta \in [0, 1]$, ranges from FCFS to LCFS-PR. Different choices of β make the system behave more like FCFS or more like LCFS-PR. One might thus try to have the best of both worlds:

- Infrequent service interrupts,
- Customers do not suffer significant delays because of one big customer.

A global description of the discipline is as follows (see Section 2 for a detailed specification). Every customer in the system is either green or red. Initially, all customers are green. All green customers have the highest priority and are served in FCFS manner. Now assume that a customer that is currently served has been served for a *very long time*, to be specified later. Then this customer is declared red. All red customers have the lowest priority. If there are only red customers in the system they are served as FCFS. If a green customer arrives to the system then the service of the red customer in service is interrupted and service of the green customer starts.
A similar system is implemented by many Internet Service Providers. Some of them specify a maximum amount of data which can be downloaded during the peak hours. After this amount is exceeded, the rate of downloading is halved for several hours. If a user still continues active downloading then its rate is halved again.
Actually, we initially had another service discipline in mind, that would give rise to a regularly varying sojourn time distribution with any index $\gamma \in [-\alpha, 1-\alpha]$: a job of size $x > 1$ is split into $x^{1-\beta_0}$ pieces of size x^{β_0}, $0 < \beta_0 < 1$, and when one piece of a job is served, the remainder of the job moves back to the end of the queue. Any value $-\gamma$ of the index of the sojourn time between $-\alpha$ and $1-\alpha$ may be obtained by choosing $\beta_0 = \alpha/(1-\gamma)$. The intuition behind this is the following: the most likely way to experience a long delay is to arrive during a long service piece. Each piece follows a power law with exponent $-\alpha/\beta_0$, and the residual of a piece follows a power law with exponent $\gamma = 1 - \alpha/\beta_0$. That proposed service discipline is actually close to a service discipline that was introduced by Kherani and Kumar [5, 6] in studying TCP, the Transmission Control Protocol for the Internet. Kherani and Kumar consider an Internet link carrying http-like traffic. The file transfers are controlled by an AWP (Adaptive Window Protocol); an example of such a protocol is TCP. They study the AWP-controlled traffic feeding into the link buffer. The contents of the link buffer comprise the windows of each of the active flows. The windows are served in a round-robin manner. The service process resembles serving a job in pieces, instead of serving the full job in one piece. Kherani and Kumar are, a.o., interested in the tail behaviour of the link buffer content, for heavy-tailed file size distributions.

One may give a similar asymptotic analysis of the above-mentioned ('pieces') discipline as the analysis we give for the red-green discipline in the next few sections. However, the analysis for the above-described discipline would be somewhat more technical, and that has determined our choice for the red-green discipline.

The paper is organized as follows. In Section 2 the service policy of this paper is described in detail, and subsequently the main result is formulated (Theorem 2.1): the sojourn time asymptotics for the $GI/GI/1$ queue with that particular service policy. Section 3 provides a lower bound for this sojourn time tail, and Section 4 an upper bound – which coincides with the lower bound, thus proving the theorem. Section 5 contains asymptotics for the length of the busy period in the $GI/GI/1$ queue and for the maximum workload in a busy period. These results are used in the proof of the theorem.

2 Service policy and main results

We consider a GI/GI/1 single server queue. The interarrival times $\{A_i\}_{-\infty}^{+\infty}$ are i.i.d. random variables. Let $\{B_i\}_{i=-\infty}^{\infty}$ be the i.i.d. service demands of the customers. We assume that

the tail

$$\mathbf{P}\{B > t\} \sim t^{-\alpha} L(t), \quad t \to \infty,$$

is regularly varying with parameter $\alpha > 1$. Throughout A and B are random variables with the same distribution as A_i and B_i respectively. We assume that $a := \mathbf{E}(A - B) > 0$, which ensures the stability of the system.

Let $\beta \in (0, 1)$ be a fixed constant. Let B_i be the service demand of customer i. Also, let M_i be the maximum service demand of the customers who arrived earlier than customer i and belong to the same busy period. We compare B_i with $M_i^{1-\beta}$. If $B_i \leq M_i^{1-\beta}$, then customer i will always be green. Otherwise, if $B_i > M_i^{1-\beta}$ the following happens: customer i waits for its service, then it is served for $B_i^{1-\beta}$ amount of time. After that this customer is declared red and its service may be interrupted by a green customer.

All the red customers are served as in the LCFS-PR queue with respect to each other. If there is a green customer in the system then the service of the red customer is interrupted and the service of the green customer starts. Green customers are served as in the FCFS service discipline with respect to each other.

Let W_k be the sojourn time of the kth customer and let W denote a stationary sojourn time. The following theorem is the main result of the paper.

Theorem 2.1. *Consider a GI/GI/1 queue with the above-defined service discipline. Assume that $\alpha > 1$ and $\beta \in (0, 1/(\alpha + 1))$. Then the following asymptotics hold*

$$\mathbf{P}\{W > x\} \sim \frac{1}{a} \int_x^\infty P\{B > y^{1/(1-\beta)}\} dy, \quad x \to \infty.$$

In particular, W is regularly varying with parameter $\gamma := \frac{\alpha}{1-\beta} - 1$.

Proof of Theorem 2.1 The first statement of the theorem follows from the lower bound given in Lemma 3.1 in Section 3 and the upper bounds given by Lemma 4.2 and Lemma 4.3 in Section 4. Furthermore, it follows from the properties of regularly varying functions that

$$\frac{1}{a} \int_x^\infty P\{B > y^{1/(1-\beta)}\} dy$$

is regularly varying with parameter $\gamma = \alpha/(1 - \beta) - 1$. Clearly, when β varies from 0 to $1/(\alpha + 1)$ the index γ varies continuously from $\alpha - 1$ to α.

3 Lower bound

In this section we are going to obtain a lower bound. Let customer 0 arrive at time epoch 0 with service demand B_0. Customer 1 arrives at the time instant A_1 with the service demand B_1. In general, customer l arrives at $A_1 + \cdots + A_l$ with the service demand B_l. Let $\xi_i = B_{i-1} - A_i$ and $S_n = \sum_{i=1}^n \xi_i$. Then,

$$\nu := \min\{n \geq 1 : S_n \leq 0\}$$

is the number of customers served during the first busy period.

Let W_k be the sojourn time of the kth customer. Then we have the following representation, see e.g. [1, Corollary 1.4, page 171]:

$$\begin{aligned}\mathbf{P}\{W > x\} &= \frac{\mathbf{E}\#\{0 \le k < \nu : W_k > x\}}{\mathbf{E}\nu} \\ &= \frac{1}{\mathbf{E}\nu}\sum_{k=0}^{\infty}\mathbf{P}\{\nu > k, W_k > x\}. \end{aligned} \tag{5}$$

Using representation (5) it is not difficult to give an accurate lower bound.

Lemma 3.1. *(Lower bound) Consider a GI/GI/1 queue with the service discipline defined above. Assume that $\alpha > 1$ and $\beta \in (0, 1/(\alpha+1))$. Then the following asymptotic lower bound holds:*

$$\mathbf{P}\{W > x\} \ge \frac{1+o(1)}{a}\int_x^{\infty} P\{B > y^{1/(1-\beta)}\}dy, \quad x \to \infty.$$

Proof of Lemma 3.1. Fix $\varepsilon > 0, R > 0$ and $N \ge 1$. Let $M_{[i,j]} = \max_{i \le l \le j} B_l$ for $i < j$. Further, define the event

$$G_{[i,j]}(k,x) = \{M_{[i,j]} \le R + x + k(a+\varepsilon)\}.$$

Then

$$\mathbf{P}\{\nu > k, W_k > x\} \ge \sum_{i=0}^{N}\mathbf{P}\{\nu > k; G_{[0,i-1]}(k,x), G_{[i+1,k]}(k,x), B_i^{1-\beta} > R+x+k(a+\varepsilon); W_k > x\}.$$

First note that in the event $G_{[0,i-1]}(k,x)$ customer i becomes red since

$$B_i^{1-\beta} > R + x + k(a+\varepsilon) \ge \max(B_0, \dots, B_{i-1}) \tag{6}$$

and, consequently (since $0 < \beta < 1$) we also have $B_i > \max(B_0, \dots, B_{i-1})^{1-\beta}$. Second note that in the event $G_{[i+1,k]}(k,x)$,

$$B_i^{1-\beta} > R + x + k(a+\varepsilon) \ge \max(B_{i+1}, \dots, B_k).$$

Equivalently, for $j : i+1 \le j \le k$,

$$B_j < B_i^{1-\beta}.$$

Now one should note that due to (6), we have $B_i^{1-\beta} = \max(B_0, \dots, B_i)^{1-\beta}$ and, thus,

$$B_j < \max(B_0, \dots, B_i)^{1-\beta}.$$

By the definition of the service discipline the latter inequality implies that customer $j : i+1 \le j \le k$ stays green.

Summing up we can see that in the event $G_{[0,i-1]}(k,x) \cap G_{[i+1,k]}(k,x)$ customer i becomes red and there is no red customer in the interval $[i+1,k]$. Therefore, the waiting time W_k of customer k in this event satisfies

$$W_k > B_i^{1-\beta} + \widehat{S}_{[i+1,k-1]},$$

where $\widehat{S}_{[i+1,k-1]} = \sum_{l=i+1}^{k-1}(B_l - A_l)$.

Next note that the event

$$\{\nu > i, B_i^{1-\beta} > R + x + k(a+\varepsilon), \min_{j=i+1,\dots,k-1} \widehat{S}_{[i+1,j]} > -R - k(a+\varepsilon)\}$$
$$\subset \{\nu > k, B_i^{1-\beta} > R + x + k(a+\varepsilon)\}.$$

Therefore,

$$\begin{aligned}
&\mathbf{P}\{\nu > k, W_k > x\} \\
&\geq \sum_{i=0}^{N} \mathbf{P}\{\nu > k; G_{[0,i-1]}(k,x), G_{[i+1,k]}(k,x), B_i^{1-\beta} > R + x + k(a+\varepsilon); W_k > x\} \\
&\geq \sum_{i=0}^{N} \mathbf{P}\{\nu > i, G_{[0,i-1]}(k,x)\}\mathbf{P}\Big\{ B_i^{1-\beta} > R + x + k(a+\varepsilon), \\
&G_{[i+1,k]}(k,x), \min_{j=i+1,\dots,k-1} \widehat{S}_{[i+1,j]} > -R - k(a+\varepsilon), B_i^{1-\beta} + \widehat{S}_{[i+1,k-1]} > x \Big\} \\
&= \mathbf{P}\{B^{1-\beta} > R + x + k(a+\varepsilon)\} \sum_{i=0}^{N} \mathbf{P}\{\nu > i, G_{[0,i-1]}(k,x)\}\mathbf{P}\Big\{ G_{[i+1,k]}(k,x), \\
&\qquad\qquad \min_{j=i+1,\dots,k-1} \widehat{S}_{[i+1,j]} > -R - k(a+\varepsilon) \Big\}.
\end{aligned}$$

Next by the Law of Large Numbers for the partial maxima, for $k \to \infty$,

$$\frac{\min_{j=i+1,\dots,k-1} \sum_{l=i+1}^{j}(B_l - A_l)}{k} \to -a, \quad \text{a.s.}$$

Therefore, for fixed N, taking R sufficiently large we have the following bound

$$\mathbf{P}\left\{ \min_{j=i+1,\dots,k-1} \widehat{S}_{[i+1,j]} > -R - k(a+\varepsilon) \right\} \geq 1 - \varepsilon,$$

uniformly in $k > i$ and $i \leq N$. Also, for sufficiently large R,

$$\mathbf{P}(\overline{G}_{[i+1,k]}(k,x)) \leq \sum_{l=i+1}^{k} \mathbf{P}\{B_l > R + x + k(a+\varepsilon)\} \leq k\mathbf{P}\{B > R + x + k(a+\varepsilon)\} \leq \varepsilon,$$

where we use the fact that $\mathbf{E}B < \infty$. Thus taking R sufficiently large we can guarantee that

$$\begin{aligned}
&\mathbf{P}\Big\{ G_{[i+1,k]}(k,x), \min_{j=i+1,\dots,k-1} \widehat{S}_{[i+1,j]} > -R - k(a+\varepsilon) \Big\} \\
&\qquad \geq \mathbf{P}\left\{ \min_{j=i+1,\dots,k-1} \widehat{S}_{[i+1,j]} > -R - k(a+\varepsilon) \right\} - \mathbf{P}(\overline{G}_{[i+1,k]}(k,x)) \geq 1 - 2\varepsilon.
\end{aligned}$$

Now for sufficiently large N and x,

$$\sum_{i=0}^{N} \mathbf{P}\{\nu > i, M_{i-1} \le x + k(a-\varepsilon)\} \ge (1-\varepsilon) \sum_{i=0}^{N} \mathbf{P}\{\nu > i\} \ge (1-2\varepsilon)\mathbf{E}\nu.$$

This gives,

$$\begin{aligned}\mathbf{P}\{\nu > k, W_k > x\} &\ge \mathbf{P}\{B^{1-\beta} > R + x + k(a+\varepsilon)\}(1-\varepsilon)\mathbf{E}\nu(1-2\varepsilon) \\ &\ge (1-3\varepsilon)\mathbf{E}\nu\mathbf{P}\{B^{1-\beta} > R + x + k(a+\varepsilon)\} \ge (1-4\varepsilon)\mathbf{E}\nu\mathbf{P}\{B^{1-\beta} > x + k(a+\varepsilon)\},\end{aligned}$$

for sufficiently large R and x.

Summing everything up and using (5) we obtain for sufficiently large x,

$$\mathbf{P}\{W > x\} \ge \frac{1}{\mathbf{E}\nu} \sum_{k=0}^{\infty} (1-4\varepsilon)\mathbf{E}\nu\mathbf{P}\{B^{1-\beta} > x + ka\} \sim \frac{1-4\varepsilon}{a} \int_x^{\infty} P\{B > y^{1/(1-\beta)}\} dy,$$

as $x \to \infty$. Since $\varepsilon > 0$ is arbitrary the latter implies the required lower bound. The proof of Lemma 3.1 is complete.

4 Upper bound

In this section we are going to provide an upper bound. For that we are going to use again the representation (5). Let $G_k = \{\text{customer } k \text{ is always green}\}$. Split the sum in (5) in two parts:

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbf{P}\{\nu > k, W_k > x\} &= \sum_{k=0}^{\infty} \mathbf{P}\{\nu > k, W_k > x, G_k\} \\ &+ \sum_{k=0}^{\infty} \mathbf{P}\{\nu > k, W_k > x, \overline{G}_k\} = P^{green}(x) + P^{red}(x).\end{aligned}$$

In the event G_k the waiting time of customer k depends only on his service demand and the demands of the preceding customers. In the event $\overline{G}_k$ the waiting time of the kth customer may depend on the subsequent customers. This is the reason why we consider these situations separately.

4.1 Upper bound for $P^{green}(x)$

In this subsection we give an upper bound for $P^{green}(x)$.

Lemma 4.1. *Let $\delta > 0$ be a constant and let*

$$\mu = \min\{i \ge 0 : B_i^{1-\beta} > \delta(x + ak)\}.$$

Then, for any $\varepsilon_0 > 0$ there exists $x_0 > 0$ such that for $x > x_0$,

$$\mathbf{P}\{\nu > k, \mu \le k, W_k > x, G_k\} \le (\mathbf{E}\nu + \varepsilon_0)\mathbf{P}\{B^{1-\beta} > x + ak\}. \tag{7}$$

PROOF OF LEMMA 4.1. Pick $\varepsilon > 0$ such that

$$\mathbf{E}\nu\mathbf{P}\{B^{1-\beta} > (1-\varepsilon)(x+ak)\} \le (\mathbf{E}\nu + \varepsilon_0/4)\mathbf{P}\{B^{1-\beta} > x+ak\}, \tag{8}$$

for sufficiently large x. This is possible since $\mathbf{P}(B > x)$ is a regularly varying function.

Then,

$$\begin{aligned}\mathbf{P}\{\nu > k, \mu \le k, W_k > x, G_k\} \le \mathbf{P}\{\nu > k, \mu \le k, B_\mu^{1-\beta} > (1-\varepsilon)(x+ak)\} \\ + \mathbf{P}\{\nu > k, \mu \le k, B_\mu^{1-\beta} \le (1-\varepsilon)(x+ak), W_k > x, G_k\}. \end{aligned}\tag{9}$$

Clearly, the first term in the RHS of (9) is majorized as follows:

$$\begin{aligned}\mathbf{P}\{\nu > k, \mu \le k, B_\mu^{1-\beta} > (1-\varepsilon)(x+ak)\} &\le \sum_{i=0}^{k} \mathbf{P}\{\nu > i, B_i^{1-\beta} > (1-\varepsilon)(x+ak)\} \\ &= \sum_{i=0}^{k} \mathbf{P}\{\nu > i\}\mathbf{P}\{B_i^{1-\beta} > (1-\varepsilon)(x+ak)\} \\ &\le \mathbf{E}\nu\mathbf{P}\{B^{1-\beta} > (1-\varepsilon)(x+ak)\} \\ &\le (\mathbf{E}\nu + \varepsilon_0/4)\mathbf{P}\{B^{1-\beta} > x+ak\}, \end{aligned}\tag{10}$$

where the latter inequality holds for sufficiently large x due to (8). Let $\widehat{\varepsilon} > 0$ be a constant which we define later. Split the second term in the RHS of (9):

$$\begin{aligned}&\mathbf{P}\{\nu > k, \mu \le k, B_\mu^{1-\beta} \le (1-\varepsilon)(x+ak), W_k > x, G_k\} \\ &\le \mathbf{P}\{\nu > k, \mu \le k, \max(M_{[0,\mu-1]}, M_{[\mu+1,k]}) > \widehat{\varepsilon}(x+ak)\} \\ &+\mathbf{P}\{\nu > k, \mu \le k, B_\mu^{1-\beta} \le (1-\varepsilon)(x+ak), \max(M_{[0,\mu-1]}, M_{[\mu+1,k]}) \le \widehat{\varepsilon}(x+ak), W_k > x, G_k\}. \end{aligned}\tag{11}$$

Bound for the first term in the RHS of (11):

$$\begin{aligned}&\mathbf{P}\{\nu > k, \mu \le k, \max(M_{[0,\mu-1]}, M_{[\mu+1,k]}) > \widehat{\varepsilon}(x+ak)\} \\ &\le \sum_{i=0}^{k} \mathbf{P}\{\nu > i, B_i^{1-\beta} > \delta(x+ak)\}\mathbf{P}\{M_{[i+1,k]} > \widehat{\varepsilon}(x+ak)\} \\ &+ \sum_{i=0}^{k} \mathbf{P}\{\nu > i, M_{[0,i-1]} > \widehat{\varepsilon}(x+ak)\}\mathbf{P}\{B_i^{1-\beta} > \delta(x+ak)\} \\ &\le 2\mathbf{E}\nu\mathbf{P}\{B^{1-\beta} > \delta(x+ak)\}k\mathbf{P}\{B > \widehat{\varepsilon}(x+ak)\} \le \varepsilon_0/4\mathbf{P}\{B^{1-\beta} > x+ak\}, \end{aligned}\tag{12}$$

where the latter inequality holds for large x since $\mathbf{P}\{B^{1-\beta} > x\}$ is a regularly varying function.

Bound for the second term in the RHS of (11):

Note that in the event $\{\mu \le k, M_{[\mu+1,k]} \le \widehat{\varepsilon}(x+ak)\}$ customer μ is declared red and all the customers in the set $\{\mu+1, \dots, k\}$ stay green. Indeed for $j : \mu+1 \le j \le k$,

$$B_j \le \widehat{\varepsilon}(x+ak) < \delta(x+ak) < B_\mu^{1-\beta}.$$

Clearly, in the event $\{\mu \le k, M_{[\mu+1,k]} \le \widehat{\varepsilon}(x+ak)\}$, by the definition of μ,

$$\max(B_0, \dots, B_{j-1}) = B_\mu.$$

Thus,

$$B_j < B_\mu^{1-\beta} = \max(B_0, \dots, B_{j-1})^{1-\beta}$$

implies that customer $j : \mu + 1 \le j \le k$ is declared green.

That implies that in this event the sojourn time W_k of customer k satisfies

$$W_k \le S_{\mu-1} + B_\mu^{1-\beta} + S_{[\mu+1,k]} + B_k,$$

where $S_{[\mu+1,k]} = \sum_{l=\mu+1}^{k}(B_{l-1} - A_l)$. Therefore, the second term in the RHS of (11) is bounded from above by

$$\begin{aligned}
&\mathbf{P}\{\nu > k, \mu \le k, B_\mu^{1-\beta} \le (1-\varepsilon)(x+ak), \max(M_{[0,\mu-1]}M_{[\mu+1,k]}) \le \widehat{\varepsilon}(x+ak), W_k > x, G_k\} \\
&\le \sum_{i=0}^{k} \mathbf{P}\Big\{\nu > i, \mu = i, B_i^{1-\beta} \le (1-\varepsilon)(x+ak), \\
&\max(M_{[0,i-1]}, M_{[i+1,k]}) \le \widehat{\varepsilon}(x+ak), S_{i-1} + B_i^{1-\beta} + S_{[i+1,k]} + B_k > x\Big\} \\
&\le \sum_{i=0}^{k} \mathbf{P}\Big\{\nu > i, \mu = i, \max(M_{[0,i-1]}, M_{[i+1,k]}) \le \widehat{\varepsilon}(x+ak), \\
&S_{i-1} + (1-\varepsilon)(x+ak) + S_{[i+1,k]} + B_k > x\Big\} \\
&\le \sum_{i=0}^{k} \mathbf{P}\{\max(M_{[0,i-1]}, M_{[i+1,k]}) \le \widehat{\varepsilon}(x+ak), S_{i-1}+S_{[i+1,k]}+(1-\varepsilon)(x+ak) > x-\varepsilon x/2\} \\
&+ \sum_{i=0}^{k} \mathbf{P}\Big\{\nu > i, \mu = i, S_{i-1} + S_{[i+1,k]} + (1-\varepsilon)(x+ak) \le x - \varepsilon x/2, \\
&\qquad\qquad S_{i-1} + S_{[i+1,k]} + (1-\varepsilon)(x+ak) + B_k > x\Big\} \equiv P_1 + P_2. \quad (13)
\end{aligned}$$

First,

$$P_1 \le k\mathbf{P}\{S_{k-1} + (1-\varepsilon)ak > \varepsilon x/2, M_{[0,k-2)} \le \widehat{\varepsilon}(x+ak)\}.$$

Let $\widehat{\xi}_i = \xi_i + a$ and $\widehat{S}_k = S_k + ak$. We can now apply the Fuk-Nagaev inequality (see Lemma 5.1 below) to obtain

$$\begin{aligned}
&\mathbf{P}\{S_{k-1} + (1-\varepsilon)ak > \varepsilon x/2, M_{k-2} \le \widehat{\varepsilon}(x+ak)\} \\
&\qquad \le \mathbf{P}\{\widehat{S}_{k-1} > \varepsilon(x/2+ak), \max(\widehat{\xi}_1, \dots, \widehat{\xi}_{k-1}) \le \widehat{\varepsilon}(x+ak)\} \le \frac{C}{(x+ak)^{\alpha+2}},
\end{aligned}$$

when $\widehat{\varepsilon} > 0$ is taken sufficiently small. Therefore,

$$P_1 \le \frac{Ck}{(x+ak)^{\alpha+2}} \le \frac{C/a}{(x+ak)^{\alpha+1}} \le \varepsilon_0/4\mathbf{P}\{B^{1-\beta} > x+ak\}, \quad (14)$$

for sufficiently large x since $\mathbf{P}\{B^{1-\beta} > x\}$ is regularly varying with the parameter $-(\gamma+1) > -(\alpha+1)$.

Second,

$$P_2 \le \sum_{i=0}^{k} \mathbf{P}\{\nu > i, \mu = i, B_k > \varepsilon x/2\}$$
$$\le \mathbf{E}\nu \mathbf{P}\{B^{1-\beta} > \delta(x+ak)\}\mathbf{P}\{B > \varepsilon x/2\} \le \varepsilon_0/4 \mathbf{P}\{B^{1-\beta} > x + ak\}, \quad (15)$$

where the latter inequality holds for large x since $\mathbf{P}\{B^{1-\beta} > x\}$ is a regularly varying function.

Summing the Equations (10), (12) (14) and (15) we arrive at the conclusion. The proof of Lemma 4.1 is complete.

Lemma 4.2. *Consider a GI/GI/1 queue with the service discipline defined above. Assume that $\alpha > 1$ and $\beta \in (0, 1/(\alpha+1))$. Then the following asymptotic upper bound holds*

$$\mathbf{P}^{green}(x) \le \frac{1+o(1)}{a} \int_x^\infty P\{B > y^{1/(1-\beta)}\} dy, \quad x \to \infty.$$

Proof of Lemma 4.2. Let $\varepsilon > 0$ be a constant which we define later. Let K be an integer such that $K(\alpha-1) > (\alpha+1)$. Let $\delta = \min(1/(2K), \varepsilon)$ and let

$$\mu = \min\{i \ge 0 : B_i^{1-\beta} > \delta(x+ak)\}.$$

Then,

$$\begin{aligned} \mathbf{P}\{\nu > k, W_k > x, G_k\} &= \mathbf{P}\{\nu > k, \mu \le k, W_k > x, G_k\} \\ &+ \mathbf{P}\{\nu > k, \mu > k, W_k > x, G_k\} \equiv P_3 + P_4. \end{aligned}$$

By Lemma 4.1 there exist $\varepsilon_0 > 0$ and $x_0 > 0$ such that for $x > x_0$ the following inequality holds:

$$P_3 \le (E\nu + \varepsilon_0)\mathbf{P}(B^{1-\beta} > x + ak). \quad (16)$$

Further let N_k be the number of big customers

$$N_k = \#\{0 \le i \le k : B_i > \varepsilon(x+ak)\}.$$

Then the second probability is bounded as follows:

$$P_4 \le \mathbf{P}\{\nu > k, \mu > k, N_k \le K, W_k > x, G_k\} + \mathbf{P}\{\nu > k, N_k > K\} \equiv P_{41} + P_{42}.$$

To bound P_{41} consider the event $\{N_k \le K\}$. In this event there are $r \le K$ customers whose service demand exceeds $\varepsilon(x+ak)$. Let their numbers be $i_1, \dots, i_r$:

$$B_{i_1} > \varepsilon(x+ak), \dots, B_{i_r} > \varepsilon(x+ak).$$

In the event $\{\mu > k\}$ all these customers $i_1, \ldots, i_r$ will be declared red, since the maximal service demand of the first k customers is less than $(\delta(x+ak))^{1/(1-\beta)}$ and $\delta \leq \varepsilon$. Therefore, customer k will interrupt the service of any of the customers $i_1, \ldots, i_r$. Thus the total contribution of the customers $i_1, \ldots, i_r$ to the waiting time of customer k is at most

$$B_{i_1}^{1-\beta} + \cdots + B_{i_r}^{1-\beta}.$$

In the event $\{\mu > k\}$, the latter contribution is less than

$$B_{i_1}^{1-\beta} + \cdots + B_{i_r}^{1-\beta} \leq r\delta(x+ak) \leq K\delta(x+ak) \leq 0.5(x+ak).$$

Therefore the waiting time V_k of customer k satisfies

$$\begin{aligned} V_k &\leq S_k - (\xi_{i_1+1} + \xi_{i_2+1} \cdots + \xi_{i_r+1}) + B_{i_1}^{1-\beta} + \cdots + B_{i_r}^{1-\beta} \\ &\leq S_k - (\xi_{i_1+1} + \xi_{i_2+1} \cdots + \xi_{i_r+1}) + 0.5(x+ak). \end{aligned}$$

Note that

$$\begin{aligned} \mathbf{P}\{V_k > 0.75x\} &\leq \sum_{r=0}^{K} \binom{k}{r} \mathbf{P}\{S_{k-r} + 0.5(x+ak) > 0.75x, \max(B_0, \ldots, B_{k-r}) \leq \varepsilon(x+ak)\} \\ &\leq Kk^K \max_{0 \leq r \leq K} \mathbf{P}\{S_{k-r} + 0.5ak > 0.25x, \max(B_0, \ldots, B_{k-r}) \leq \varepsilon(x+ak)\} \\ &\leq \frac{CK}{(x+ak)^{\alpha+2}}, \end{aligned}$$

for sufficiently small ε. To pick such $\varepsilon > 0$ we use the Fuk-Nagaev inequality again, see Lemma 5.1 below. Thus, since in the event G_k, the sojourn time $W_k = V_k + B_k$, we have,

$$\begin{aligned} P_{41} &\leq \frac{CK}{(x+ak)^{\alpha+2}} + \mathbf{P}\{\nu > k, \mu > k, N_k \leq K, W_k > x, V_k \leq 0.75x, G_k\} \\ &\leq \frac{CK}{(x+ak)^{\alpha+2}} + \mathbf{P}\{\nu > k, B_k > 0.25x\} \\ &\leq \varepsilon_0 \mathbf{P}\{B^{1-\beta} > x+ak\} + \mathbf{P}\{\nu > k\}\mathbf{P}\{B_k > 0.25x\}, \end{aligned} \tag{17}$$

for sufficiently large x. Here, we have used the fact that $\mathbf{P}(B^{1-\beta} > x)$ is regularly varying with the parameter $-(\gamma+1) > -\alpha - 1$. Also,

$$P_{42} \leq (k\mathbf{P}(B > \varepsilon(x+ak))^K \leq (x+ak)^{-\alpha-1} \leq \varepsilon_0 \mathbf{P}(B^{1-\beta} > x+ak), \tag{18}$$

for sufficiently large x. Here, we have used the fact that $K(\alpha-1) > \alpha+1$ and that $\mathbf{P}(B^{1-\beta} > x)$ is regularly varying with the parameter $-(\gamma+1) > -\alpha - 1$.

We are now in position to estimate $P^{green}(x)$. Using the equations (16), (17) and (18) we have for sufficiently large x,

$$\begin{aligned} P^{green}(x) &\leq \sum_{k=0}^{\infty}\Big((\mathbf{E}\nu+3\varepsilon_0)\mathbf{P}\{B^{1-\beta}>x+ak\}+\mathbf{P}\{\nu>k\}\mathbf{P}\{B_k>0.25x\}\Big) \\ &\leq (\mathbf{E}\nu+3\varepsilon_0)\sum_{k=0}^{\infty}\mathbf{P}\{B^{1-\beta}>x+ak\}+\mathbf{E}\nu\mathbf{P}\{B>0.25x\} \\ &\leq \frac{\mathbf{E}\nu+4\varepsilon_0}{a}\int_x^{\infty}\mathbf{P}\{B^{1-\beta}>z\}dz. \end{aligned}$$

Since the constant $\varepsilon_0>0$ is arbitrary the latter equation implies the statement of the Lemma.

The proof of Lemma 4.2 is complete.

4.2 Upper bound for $P^{red}(x)$

In this subsection we give bounds for $P^{red}(x)$.

Lemma 4.3. *Consider a GI/GI/1 queue with the service discipline defined above. Assume that $\alpha>1$ and $\beta\in(0,1/(\alpha+1))$. Then the following asymptotic upper bound holds*

$$\mathbf{P}^{red}(x)=o\left(\int_x^{\infty}P\{B>z^{1/(1-\beta)}\}dz\right),\quad x\to\infty.$$

Proof of Lemma 4.3. Fix a constant $\varepsilon<\alpha-\gamma$. Let bp be the length of the first busy period and, by Theorem 5.1 from the appendix, $\mathbf{P}(bp>x)\leq C^*\mathbf{P}\{B>x\}$ for some constant C^*. Since $\alpha-\varepsilon>\gamma$ we have,

$$\sum_{k=0}^{x^{\varepsilon}}\mathbf{P}\{\nu>k,W_k>x\}\leq x^{\varepsilon}\mathbf{P}\{bp>x\}=o\left(\int_x^{\infty}\mathbf{P}\{B^{1-\beta}>z\}dz\right).$$

Thus we need to give bounds for

$$\sum_{k=x^{\varepsilon}}^{\infty}\mathbf{P}\{\nu>k,W_k>x,\overline{G}_k\}.$$

Note also that

$$\begin{aligned} &\sum_{k=x^{\varepsilon}}^{\infty}\mathbf{P}\{\nu>k,W_k>x,B_k>x^{\frac{\gamma+\varepsilon}{\alpha}},\overline{G}_k\}\leq\sum_{k=x^{\varepsilon}}^{\infty}\mathbf{P}\{\nu>k,B_k>x^{\frac{\gamma+\varepsilon}{\alpha}}\} \\ &=\sum_{k=x^{\varepsilon}}^{\infty}\mathbf{P}\{\nu>k\}\mathbf{P}\{B_k>x^{\frac{\gamma+\varepsilon}{\alpha}}\}\leq\mathbf{E}\nu\mathbf{P}\{B>x^{\frac{\gamma+\varepsilon}{\alpha}}\}=o\left(\int_x^{\infty}\mathbf{P}\{B^{1-\beta}>y\}dy\right), \end{aligned}$$

since $\mathbf{P}\{B > x^{\frac{\gamma+\varepsilon}{\alpha}}\}$ is regularly varying with the parameter $-(\gamma+\varepsilon)$ and $\int_x^\infty \mathbf{P}\{B^{1-\beta} > y\}dy$ is regularly varying with the parameter $-\gamma$. Thus, we have just shown that

$$P^{red}(x) = \sum_{k=x^\varepsilon}^{\infty} \mathbf{P}\{\nu > k, W_k > x, B_k \le x^{\frac{\gamma+\varepsilon}{\alpha}}, \overline{G}_k\} + o\left(\int_x^\infty \mathbf{P}\{B^{1-\beta} > y\}dy\right). \quad (19)$$

Now the kth customer, with $\{B_k \in dy\}$, is red if and only if all preceding customers in the same busy period satisfy the inequality $M_{k-1} = \max(B_0, \ldots, B_{k-1}) \le y^{1/(1-\beta)}$. Consequently,

$$\mathbf{P}\{\nu > k, W_k > x, B_k \le x^{\frac{\gamma+\varepsilon}{\alpha}}, \overline{G}_k\} = \int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}} \mathbf{P}\{\nu > k, W_k(y) > x, M_{k-1} \le y^{1/(1-\beta)}, B_k \in dy\},$$

where $W_k(y)$ is the sojourn time of customer k in the event $\{B_k \in dy\}$. Fix a constant $C > 0$ which we define later. Let $N_k(y) = \#\{0 \le i < k : B_k > y + ka/C\}$ and let K be an integer such that $K > \frac{\gamma/\varepsilon+1}{\alpha-1}$. Uniformly in y we have,

$$\mathbf{P}\{N_k(y) > K\} \le \binom{k}{K} \mathbf{P}\{B > y + ka/C\}^K \le (k\mathbf{P}\{B > ka/C\})^K = k^{(1-\alpha)K} L(k),$$

for some regularly varying function $L(k)$. Then,

$$\begin{aligned}
&\sum_{k=x^\varepsilon}^{\infty} \int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}} \mathbf{P}\{\nu > k, W_k(y) > x, M_{k-1} \le y^{1/(1-\beta)}, N_k(y) > K, B_k \in dy\} \\
&\le \sum_{k=x^\varepsilon}^{\infty} \int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}} \mathbf{P}\{N_k(y) > K\} \mathbf{P}\{B_k \in dy\} \\
&\qquad \le \sum_{k=x^\varepsilon}^{\infty} k^{(1-\alpha)K} L(k) \sim x^{(1+(1-\alpha)K)\varepsilon} L(x^\varepsilon) = o\left(\int_x^\infty \mathbf{P}\{B^{1-\beta} > z\}dz\right),
\end{aligned}$$

since $(1 + (1-\alpha)K)\varepsilon < -\gamma$ due to our choice of K. Therefore, we can continue (19) and obtain

$$\begin{aligned}
P^{red}(x) = \sum_{k=x^\varepsilon}^{\infty} \int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}} &\mathbf{P}\{\nu > k, W_k > x, M_{k-1} \le y^{1/(1-\beta)}, N_k(y) \le K, B_k \in dy\} \\
&+ o\left(\int_x^\infty \mathbf{P}\{B^{1-\beta} > z\}dz\right). \quad (20)
\end{aligned}$$

To proceed further, we need to consider the situation when the service demands of at most K customers in the set $\{0, 1, 2, \ldots, k-1\}$ can exceed $y + ka/C$ on the event $\{B_k \in dy, M_{k-1} \le y^{1/(1-\beta)}\}$. Let the service demands of customer $i_1, \ldots, i_r$ for $r \le K$ be greater than $y + ka/C$. Then, since $M_{k-1} \le y^{1/(1-\beta)}$ and by the definition of the service discipline, all of them will be declared red after receiving some initial amount of service. Thus the total contribution of the customers $i_1, \ldots, i_r$ to the waiting time of customer k is at most

$$B_{i_1}^{1-\beta} + \cdots + B_{i_r}^{1-\beta} \le Ky.$$

Therefore the waiting time of customer k is at most

$$\begin{aligned} V_k(y) &\equiv S_k - (\xi_{i_1+1} + \xi_{i_2+1} \cdots + \xi_{i_r+1}) + B_{i_1}^{1-\beta} + \cdots + B_{i_r}^{1-\beta} \\ &\leq S_k - (\xi_{i_1+1} + \xi_{i_2+1} \cdots + \xi_{i_r+1}) + Ky. \end{aligned}$$

Note that

$$\begin{aligned} \mathbf{P}\{V_k(y) > (K+C)y\} &\leq \sum_{r=0}^{K} \binom{k}{r} \mathbf{P}\{S_{k-r} + Ky > (K+C)y, M_{k-r} \leq y + ka/C\} \\ &\leq Kk^K \max_{0\leq r\leq K} \mathbf{P}\{S_{k-r} > Cy, M_{k-r} \leq y + ka/C\}. \end{aligned}$$

We can now apply the Fuk-Nagaev inequality, see Lemma 5.1 below. For that let $\widetilde{\xi} = \xi + a$ and $\widetilde{S}_n = \sum_{i=1}^n \widetilde{\xi}_i$. Let $t > 1$ be such that $\mathbf{E}|\widetilde{\xi}|^t < \infty$. Then, for any $\delta > 0$ there exists k_0 such that for $k > k_0$ and $y \geq 0$,

$$|\mu(y + ka)| \leq \delta, \quad A(t, y + ka)/y^{t-1} \leq \delta,$$

where $\mu(y) = \mathbf{E}\{\widetilde{\xi}, |\widetilde{\xi}| \leq y\}$ and $A(y) = \mathbf{E}\{|\widetilde{\xi}|^t, |\widetilde{\xi}| \leq y\}$. Then,

$$\begin{aligned} \max_{0\leq r\leq K} \mathbf{P}\{S_{k-r} > Cy, M_{k-r} \leq y+ka/C\} &= \max_{0\leq r\leq K} \mathbf{P}\{\widetilde{S}_{k-r} > Cy+(k-r)a, M_{k-r} \leq y+ka/C\} \\ &\leq \exp\left\{ \frac{Cy+ka}{y+ka/C} - \left(\frac{Cy+ka}{y+ka/C}\right) \ln(1 + \frac{Cy+ka}{\delta}) \right\} = \frac{e^C}{(1+\frac{Cy+ka}{\delta})^C}. \end{aligned}$$

Therefore,

$$\mathbf{P}\{V_k(y) > (K+C)y\} \leq e^C \frac{Kk^K}{(1+\frac{Cy+ka}{\delta})^C}.$$

This implies that we can choose C sufficiently large to ensure that

$$\sum_{k=x^\varepsilon}^{\infty} \int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}} \mathbf{P}\{N_K(y) \leq K, V_k(y) > (K+C)y, B_k \in dy\} = o\left(\int_x^\infty \mathbf{P}\{B^{1-\beta} > z\} dz \right).$$

Therefore, we can rewrite (20) to obtain

$$\begin{aligned} P^{red}(x) = \sum_{k=x^\varepsilon}^{\infty} \int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}} \mathbf{P}\Big\{ &\nu > k, W_k(y) > x, M_{k-1} \leq y^{1/(1-\beta)}, \\ &N_k(y) \leq K, V_k(y) \leq (K+C)y, B_k \in dy \Big\} + o\left(\int_x^\infty \mathbf{P}\{B^{1-\beta} > z\} dz \right). \quad (21) \end{aligned}$$

We are now in position to analyze $W_k(y)$. Clearly, $W_k(y)$ is bounded from above by the sum

$$W_k(y) \leq V_k(y) + bp(y + V_k(y)),$$

where $bp(y+V_k(y))$ is the length of the busy period of the system with initial workload $y+V_k(y)$. Thus,

$$\sum_{k=x^\varepsilon}^{\infty}\int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}}\mathbf{P}\{\nu>k,W_k>x,M_{k-1}\le y^{1/(1-\beta)},N_k(y)\le K,V_k(y)\le (K+C)y,B_k\in dy\}$$
$$\le\sum_{k=x^\varepsilon}^{\infty}\int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}}\mathbf{P}\{\nu>k\}\mathbf{P}(B_k\in dy)\mathbf{P}(bp(y+(K+C)y)>x). \quad (22)$$

Now, for $C_1=1+K+C$, using integration by parts twice, we obtain:

$$\int_0^{x^{\frac{\gamma+\varepsilon}{\alpha}}}\mathbf{P}(B\in dy)\mathbf{P}(bp(y+(K+C)y)>x)\le\int_0^{\infty}\mathbf{P}(C_1B\in dy)\mathbf{P}(bp(y)>x)$$
$$=\int_0^{\infty}d_y(\mathbf{P}(bp(y)>x))\mathbf{P}\{C_1B>y\}\le C_2\int_0^{\infty}d_y(\mathbf{P}(bp(y)>x))\mathbf{P}\{B>y\}$$
$$=C_2\int_0^{\infty}\mathbf{P}(B\in dy)\mathbf{P}(bp(y)>x)=C_2\mathbf{P}\{bp>x\},$$

for some constant $C_2>0$. Using the latter inequality implies that we can continue (21) and (22) to obtain that

$$P_{red}(x)\le C_2\sum_{k=x^\varepsilon}^{\infty}\mathbf{P}\{\nu>k\}\mathbf{P}\{bp>x\}+o\left(\int_x^{\infty}\mathbf{P}\{B^{1-\beta}>z\}dz\right)$$
$$\le C_2\mathbf{E}\nu\mathbf{P}\{bp>x\}+o\left(\int_x^{\infty}\mathbf{P}\{B^{1-\beta}>z\}dz\right).$$

It is sufficient now to apply Theorem 5.1 and use the fact that $\mathbf{P}\{B>x\}=o\left(\int_x^{\infty}\mathbf{P}\{B^{1-\beta}>z\}dz\right)$ to obtain the statement of the Lemma. The proof of Lemma 4.3 is complete.

5 Appendix: Asymptotics for the maximum workload and the length of the busy period

5.1 Estimates for the busy period of the single-server queue

Let $\xi=B_{i-1}-A_i$ and $S_n=\sum_{i=1}^n\xi_i$. Let $\nu=\min\{n\ge 1:S_n\le 0\}$ and $bp=B_1+\cdots+B_\nu$ be the length of the busy period. Then the following theorem holds, see [10].

Theorem 5.1. *Assume that* $\mathbf{E}\xi<0$ *and that* $\mathbf{P}\{B>x\}$ *is regularly varying with the parameter* α. *Then,*

$$\mathbf{P}\{bp>x\}\sim\mathbf{E}\nu\mathbf{P}\{B>(1-\rho)x\}, \quad (23)$$

where $\rho=\mathbf{E}B/\mathbf{E}A<1$. *In particular,* $\mathbf{P}\{bp>x\}\le C^*\mathbf{P}\{B>x\}$ *for some constant* C^*.

5.2 Fuk-Nagaev inequality

The following lemma is the Fuk-Nagaev inequality, see [3, Theorem 2, eq. (7)].

Lemma 5.1. *Let* $1 \leq t \leq 2$. *Let* $\mu(y) = \mathbf{E}\{\xi, |\xi| \leq y\}$ *and* $A(t,y) = \mathbf{E}\{|\xi|^t, |\xi| \leq y\}$. *Then,*

$$\mathbf{P}\{S_n > x, M_n \leq y\} \leq \exp\left\{\frac{x}{y} - \left(\frac{x - n\mu(y)}{y} + \frac{A(t,y)}{y^t}\right) \ln\left(\frac{xy^{t-1}}{A(t,y)} + 1\right)\right\}, \qquad (24)$$

where $M_n = \max(\xi_1, \ldots, \xi_n)$.

Acknowledgment
We gratefully acknowledge fruitful discussions with Professor Resnick, and with Professor Kherani who referred us to [5, 6] and outlined that the alternative discipline suggested at the end of Section 1 is closely related to the AWP (Adaptive Window Protocol) discussed by Kherani and Kumar in their closed loop analysis of an Internet link carrying http-like traffic. The research of Onno Boxma was supported by the European Network of Excellence Euro-NF and by the BRICKS project.

References

[1] Asmussen, S. (2003). *Applied Probability and Queues,* 2nd edn. Springer-Verlag, New York.

[2] Borst, S.C., Boxma, O.J., Núñez-Queija, R. and Zwart, A.P. (2003). The impact of the service discipline on delay asymptotics. *Performance Evaluation* **54**, 175-206.

[3] Fuk, D. Kh. and Nagaev, S.V. (1971). Probability inequalities of sums of independent random variables. *Theory Probab. Appl.* **16** (4), 643–660.

[4] Harchol-Balter, M. (ed.) (2007). Special Issue on New Perspectives in Scheduling. *Perf. Eval. Review* **34** (4), 1-70.

[5] Kherani, A.A. and Kumar, A. (2003). The lightening effect of adaptive window control. *IEEE Comm. Lett.* **7** (5), 284–286.

[6] Kherani, A.A. and Kumar, A. (2003). Closed loop analysis of the bottleneck buffer under adaptive window controlled transfer of HTTP-like traffic. *Proc. INFOCOM 2003.*

[7] Kingman, J.F.C. (1962). The effect of queue discipline on waiting time variance. *Proc. Camb. Phil. Soc.* **58**, 163-164.

[8] Stolyar, A. and Ramanan, K. (2001). Largest weighted delay first scheduling: large deviations and optimality. *Ann. Appl. Probab.* **11**, 1-48.

[9] Tambouratzis, D.G. (1968). On a property of the variance of the waiting time in a queue. *J. Appl. Probab.* **5**, 702-703.

[10] Zwart, B. (2001). Tail asymptotics for the busy period in the GI/G/1 queue. *Math. Oper. Res.* **26**, 485-493.