A stochastic mean-value method for the derivation of delay asymptotics in heavy-tailed processor-sharing systems

A Stochastic Mean-Value Method for the Derivation of Delay Asymptotics in Heavy-Tailed Processor-Sharing Systems*

Dennis van Ooteghem[†,⋆], Bert Zwart[†,⋆], Sem Borst[†,⋆,‡]

[†]Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[⋆]CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[‡]Bell Laboratories, Lucent Technologies
P.O. Box 636, Murray Hill, NJ 07974, USA

Abstract

We develop a stochastic mean-value method for the derivation of delay asymptotics in Processor-Sharing (PS) type systems with heavy-tailed service requirements. In order to demonstrate the strength of the approach, we apply the method to obtain the sojourn time asymptotics for a multi-class G/G/1 queue operating under the Discriminatory Processor-Sharing (DPS) discipline. Besides interesting from a queueing-theoretic perspective, DPS is also of practical relevance as it provides a useful paradigm for modelling the flow-level performance of differentiated resource-sharing mechanisms. Unlike for ordinary PS, however, the queue length for DPS does not have a simple distribution, and there are no manageable transform results available for the sojourn time. These circumstances seriously complicate the derivation of delay asymptotics using existing proof methods, and render DPS as a good 'test case' for judging the merits of alternative approaches. We use the stochastic mean-value method to show that under certain assumptions, the service requirement and sojourn time of a given class have similar tail behaviour, independent of the specific values of the DPS weights. The results suggest that DPS offers a useful instrument for effectuating preferential treatment to smaller service demands without inflicting excessive delays on larger requests. We also briefly discuss the potential applicability of the method for deriving the delay asymptotics under a broader class of resource-sharing strategies.

Keywords: differentiated services; (discriminatory) processor sharing; heavy-tailed traffic; regenerative processes; regular variation; sojourn time asymptotics; stochastic mean-value method.

*The work of the first and third author was in part financially supported through the EQUANET project. The second author was supported by an NWO-VENI grant.

1 Introduction

Over the past few years, the Processor-Sharing (PS) paradigm has been widely adopted as a natural abstraction for modelling the bandwidth sharing among dynamically interacting TCP flows. Independently, extensive measurement studies have indicated that the sizes of TCP transfers commonly exhibit heavy-tailed characteristics. These findings have triggered a strong interest in the delay asymptotics of PS queues with heavy-tailed service requirements. Unfortunately, the existing results invariably involve the assumption of Poisson arrivals and critically rely on the premise that the bandwidth is shared in a perfectly symmetric fashion. While the latter supposition is naturally restrictive by itself, it becomes prohibitively confining when one aims to explore structural extensions, such as the role of service differentiation, the impact of variations in the bandwidth, the integration with non-elastic traffic, or the behaviour in networks with multiple bottlenecks. Motivated by these observations, we develop in the present paper an alternative method for deriving sojourn time asymptotics in heavy-tailed PS-type systems which totally avoids the use of any specific properties of the M/G/1 PS queue. To illustrate the approach, we apply the method to a Discriminatory Processor-Sharing (DPS) G/G/1 queue, which has proved extremely difficult to analyze using existing techniques.

As noted above, various methods have been employed for establishing the sojourn time asymptotics in heavy-tailed M/G/1 PS queues. Zwart & Boxma [19] used transform techniques to obtain the delay asymptotics for regularly varying service requirements. They proved that the tail of the delay distribution is asymptotically equivalent to that of the service requirement distribution, up to a constant factor. Subsequently, Zwart [17] generalised the result to *multi-class* PS queues. Núñez-Queija [12, 13] constructed a proof based on conditional moments to extend the tail equivalence result to PS models with a time-varying service capacity and *intermediately* regularly varying service requirements. Jelenković & Momčilović [11] devised a probabilistic proof method to generalise the result to a larger subclass of subexponential distributions with a so-called square-root insensitivity property. The latter class includes Weibull distributions with an index parameter smaller than 1/2. They further showed that the result is sharp, in the sense that the tail equivalence does not hold for Weibull distributions with a larger index parameter. Recently, Guillemin, Robert & Zwart [8] established a general criterion for the validity of a tail equivalence, which they applied to PS queues with state-dependent service rates and blocking and/or reneging.

The above results crucially rely on the assumption of Poisson arrivals and the fact that the service capacity is shared in an egalitarian manner. The latter assumption is typically not valid for competing TCP flows that traverse heterogeneous routes and experience different loss rates and round-trip delays. Besides TCP-related effects, the heterogeneous bandwidth shares may also result from deliberate service differentiation. Unfortunately, the existing proof methods do not readily extend to asymmetric resource-sharing nor do they carry over to network scenarios or integration with non-elastic streams. Specifically, Tauberian techniques intrinsically rely on transform results which are only available for a limited class of models. Probabilistic approaches usually exploit the fact that the queue length in the M/G/1 PS queue has a simple geometric distribution, which no longer holds when the service capacity varies over time or is shared in an asymmetric fashion. These facts limit the applicability of existing proof methods to a restrictive class of models.

In the present paper we develop an alternative method which systematically circumvents the use of any particular properties of the M/G/1 PS queue, and can be brought to bear on a wide variety of resource-sharing models with heavy-tailed service requirements. The method that we construct is based on a theorem for regenerative processes, namely the *stochastic mean-value theorem*, see for instance Corollary 1.4, p. 171 of Asmussen [1]. Using this theorem, the stationary probability of an event can be expressed as the expected occurrence ratio of the event during a regenerative cycle. It then suffices to analyse this cycle in order to determine the probability of the event of interest, and in particular to obtain lower and upper bounds. Since the analysis of a regenerative cycle does not involve the stationary queue length distribution, the proposed method can also be applied to a wide range of extensions of the egalitarian PS model.

To illustrate the approach, we use the method to obtain the sojourn time asymptotics for a heavy-tailed G/G/1 queue operating under the DPS discipline. The DPS discipline provides a convenient approach for modelling the flow-level performance of asymmetric resource-sharing strategies. DPS is a multi-class extension of the ordinary egalitarian PS policy, where the various classes are assigned positive weight factors. The service capacity is shared among all users present in proportion to the respective class-dependent weights. The results for DPS in the literature are surprisingly sparse. In a seminal paper, Fayolle, Mitrani & Iasnogorodski [6] obtain the conditional mean sojourn times as the solution of a system of integro-differential equations. For the case of exponentially distributed service requirements, they derive closed-form expressions and also determine the unconditional mean sojourn times from a system of linear equations. Rege & Sengupta [14] prove a decomposition theorem for the conditional sojourn time. They specifically establish that the sojourn time of a customer which finds n customers upon arrival can be decomposed into $n+1$ independent components, which can be characterised as the solution of a system of non-linear integral equations. In a further paper, Rege & Sengupta [15] obtain the moments of the queue length distribution from a system of linear equations for the case of exponentially distributed service requirements. The latter results have recently been extended to the class of phase-type distributions [9, 16].

The fact that DPS has largely eluded existing approaches, renders it as a good 'test case' for assessing the strength of the proposed method. We use the method to show that, under certain assumptions, the same tail equivalence holds as for the ordinary egalitarian PS discipline. It is worth observing that the same result was obtained in previous work [5], and that the contribution of the current paper is primarily methodological. While the proof in [5] was rather lengthy and technically complicated, the method devised here allows a proof that is shorter and has a simpler structure. More importantly, the proposed method hardly uses any specific features of the DPS discipline at all, and hence is likely to extend to a wider range of disciplines and network scenarios. We thus expect that the presented method provides a promising tool for the derivation of sojourn time asymptotics in a broad class of resource-sharing models with heavy-tailed service requirements.

The remainder of the paper is organised as follows. In Section 2, we present a detailed model description. We state the main result and provide a heuristic interpretation in Section 3. The proof can be found in Section 4 and the paper is concluded in Section 5 with a further discussion of the method and some challenges that remain.

2 Model description

We consider a single server of unit rate which is offered traffic from K distinct customer classes. Arrivals occur according to a renewal process with total rate λ, i.e., the mean interarrival time is $1/\lambda$. With probability p_i an arriving customer is of class i. Define $\lambda_i := p_i\lambda$ as the arrival rate of class-i customers.
The service requirements of class-i customers are independent and identically distributed copies of some generic random variable B_i. The service requirement of an arbitrary customer will be denoted by the generic random variable B defined by $\mathbb{P}\{B > x\} := \sum_{i=1}^{K} p_i\mathbb{P}\{B_i > x\}$. Define $\rho_i := \lambda_i\mathbb{E}\{B_i\}$ as the offered traffic load from class i, and denote by $\rho := \sum_{i=1}^{K} \rho_i$ the total offered traffic load. We assume that $\rho < 1$ to ensure that the system is stable.
Throughout the paper, we will make the assumption that some service requirements are regularly varying as specified in the next definition. For further background on the class of regularly varying distributions, we refer to Bingham *et al.* [3].

Definition 2.1 *A non-negative random variable X is regularly varying of index ν if*

$$\mathbb{P}\{X > x\} = l(x)x^{-\nu}, \quad \nu \geq 0,$$

where $l(\cdot)$ is a slowly varying function, i.e., $\lim_{x\to\infty} l(\eta x)/l(x) = 1$, $\eta > 1$.

The customers of the various classes are served according to the Discriminatory Processor-Sharing (DPS) discipline. In the DPS discipline, there is a positive weight g_i associated with each class-i customer. When there are n_i class-i customers present in the system, $i = 1, \ldots, K$, each class-j customer receives service at rate $\frac{g_j}{\sum_{i=1}^{K} n_i g_i}$. Note that in case $g_i = g$ for all $i = 1, \ldots, K$, and in particular in case $K = 1$, the DPS discipline reduces to the familiar egalitarian PS discipline. Denote the generic sojourn time of an arbitrary customer and that of class-i customers by V and V_i, respectively.

3 Main result

The next theorem presents the main result of the paper which relates the asymptotic tail distribution of the sojourn time of class-i customers to that of their service requirement. The same result was obtained in [5], but the main goal of the current paper is the development of a new proof concept which extends to a wider class of models with PS-type service disciplines. Here and throughout the paper, we use the notational convention $f(x) \sim g(x)$, $x \to \infty$, for any two real functions $f(\cdot)$ and $g(\cdot)$ to indicate that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

Theorem 3.1 *If B_i is regularly varying of index $\nu > 2$ and $\mathbb{P}\{B_j > x\} = o(\mathbb{P}\{B_i > x\})$ as $x \to \infty$ for all $j \neq i$, then*

$$\mathbb{P}\{V_i > x\} \sim \mathbb{P}\{B_i > (1-\rho)x\}, \quad x \to \infty.$$

Loosely phrased, the above theorem indicates that a large sojourn time of a customer is most likely caused by a large service requirement of that same customer. A simple heuristic explanation of this relationship was presented in [5]. We will briefly sketch this

argument as it offers the intuitive guidance for the rigorous proof that we provide in the next section.
We focus on a customer with a large service requirement (hereafter called a 'large' customer). This large customer will stay in the system for a long time, and during that period other customers with typical service requirements (hereafter called 'small' customers) will arrive, be served, and depart from the system. For the system to remain stable, the small customers will together require approximately a service rate $\rho < 1$. The large customer will consume what is left over, and thus receive roughly a service rate $1 - \rho$. This results in a sojourn time that is about $1/(1-\rho)$ times as large as the service requirement of the large customer, and that is exactly what the above theorem indicates. Of course, there are alternative ways in which a large sojourn time may occur, and the above theorem thus indirectly shows that these are exceedingly unlikely compared to the dominant scenario described above.
Observe that the weight factors do not play any role in the above heuristic arguments, which immediately explains why the asymptotic behaviour does not depend on the specific weight values. When the large customer has a large weight, it will obviously take a larger number of small customers to obtain a service rate ρ. Until the small customers receive that rate ρ, however, their number will tend to grow, so that eventually the small customers will always claim a service rate ρ, regardless of the weights.

In the special case $K = 1$, the DPS discipline reduces to the familiar egalitarian PS discipline, yielding the following corollary.

Corollary 3.2 *If $K = 1$ and B is regularly varying of index $\nu > 2$, then*

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{B > (1-\rho)x\}, \quad x \to \infty.$$

The above corollary is actually in itself a noticeable extension of known results for the ordinary PS discipline, which further testifies to the strength of the new approach. These previous results all require Poisson arrivals, and entail proofs that exploit knowledge of the geometric queue length distribution in that case. Without Poisson arrivals, the queue length distribution is not known, which complicates the proof.

Theorem 3.1 involves the assumption that both B_i and B are regularly varying of index $\nu > 2$, which in particular means that the service requirements have finite variance. While the assumption that B_i is regularly varying is natural (though possibly not strictly necessary), the intuitive arguments described above suggest that the additional assumptions that $\nu > 2$ and that B is regularly varying of the same index, may not be essential for the result to hold: we conjecture that Theorem 3.1 remains valid without the latter two assumptions, i.e., applies for any class i with a regularly varying service requirement distribution.

4 Proofs

We now provide the proof of Theorem 3.1. As mentioned earlier, the proof strategy closely follows the intuitive insight described in the previous section. Technically speaking, the proof relies on lower and upper bounds which asymptotically coincide.

Proposition 4.1 *(Lower bound) If B_i is regularly varying of index $\nu > 1$, then*

$$\mathbb{P}\{V_i > x\} \geq \mathbb{P}\{B_i > (1-\rho)x\}(1 + \mathrm{o}(1)), \quad x \to \infty.$$

The proof of the above lower bound is similar to that of Proposition 4.1.1 in Borst, Van Ooteghem & Zwart [5]. The proof that is presented in the latter paper simply uses the Law of Large Numbers to make the heuristic derivation sketched in the previous section rigorous.

Proposition 4.2 *(Upper bound) If B is regularly varying of index $\nu > 2$ and $\mathbb{P}\{B_j > x\} = \mathrm{o}(\mathbb{P}\{B_i > x\})$ as $x \to \infty$ for all $j \neq i$, then*

$$\mathbb{P}\{V_i > x\} \leq \mathbb{P}\{B_i > (1-\rho)x\}(1 + \mathrm{o}(1)), \quad x \to \infty.$$

The proof of the above upper bound is far more difficult than that of the lower bound. We will develop a method based on a theorem for regenerative processes, namely the *stochastic mean-value theorem*, see for instance Corollary 1.4, p. 171 of Asmussen [1]. Specifically, we will use the stochastic mean-value theorem to rewrite the probability $\mathbb{P}\{V_i > x\}$ in terms of the expected fraction of class-i customers with sojourn time larger than x arriving during a busy period:

$$\mathbb{P}\{V_i > x\} = \frac{1}{\mathbb{E}\{N\}p_i}\mathbb{E}\{\sum_{j=1}^{N} \mathbb{1}_{V^{(j)}>x}\mathbb{1}_{j\in T_i}\},$$

where $\mathbb{1}_A$ denotes the indicator function of the event A, and the event $j \in T_i$ indicates whether or not the j-th customer arriving during the busy period belongs to class i. We then divide the probability into several terms by conditioning on the number of 'large' customers that arrive during the busy period. Here and throughout the paper a customer is considered 'large' when its service requirement exceeds ϵx, for some constant $\epsilon > 0$ independent of x. We finally show that the term corresponding to exactly one large customer arrival dominates all others, and behaves as $\mathbb{P}\{B_i > (1-\rho)x\}$.

Proof of Proposition 4.2
We first introduce some useful notation. Denote by N the number of customers that arrive during a busy period of length P. Let $B^{(j)}$ be the service requirement of the j-th customer arriving during the busy period and let $V^{(j)}$ be its sojourn time. The variable $A(y)$ denotes the set of customers with service requirement larger than y that arrive during the busy period, i.e.,

$$A(y) := \{j \in \{1, \ldots, N\} : B^{(j)} > y\}.$$

and $N_{>y}$ indicates the cardinality of this set.
Observing that $V^{(j)} > x$ for some $j \in \{1, \ldots, N\}$ implies that $P > x$, we may then write

$$\mathbb{P}\{V_i > x\} = \frac{1}{\mathbb{E}\{N\}p_i}\mathbb{E}\{\sum_{i=1}^{N} \mathbb{1}_{V^{(j)}>x}\mathbb{1}_{P>x}\mathbb{1}_{j\in T_i}\} = \frac{1}{\mathbb{E}\{N\}p_i}(\sum_{k=0}^{2}(I_k + II_k) + III), \quad (1)$$

where

$$\begin{aligned} I_k &= \mathbb{E}\{\sum_{\substack{j=1 \\ j\notin A(\epsilon x)}}^{N} \mathbb{1}_{V^{(j)}>x}\mathbb{1}_{j\in T_i}\mathbb{1}_{P>x}\mathbb{1}_{N_{>\epsilon x}=k}\}, \\ II_k &= \mathbb{E}\{\sum_{j\in A(\epsilon x)} \mathbb{1}_{V^{(j)}>x}\mathbb{1}_{j\in T_i}\mathbb{1}_{P>x}\mathbb{1}_{N_{>\epsilon x}=k}\}, \\ III &= \mathbb{E}\{\sum_{j=1}^{N} \mathbb{1}_{V^{(j)}>x}\mathbb{1}_{j\in T_i}\mathbb{1}_{P>x}\mathbb{1}_{N_{>\epsilon x}\geq 3}\}. \end{aligned}$$

We will now deal with each of the terms in (1) separately.
We start with the dominant term II_1, and then proceed to show that all others can be asymptotically neglected. Clearly, $II_1 \leq \mathbb{E}\{\mathbb{1}_{P>x}\} = \mathbb{P}\{P > x\}$, and Theorem 5.3.1 in Zwart [18] implies

$$\mathbb{P}\{P > x\} \sim \mathbb{E}\{N\}\mathbb{P}\{B > (1-\rho)x\} \sim p_i\mathbb{E}\{N\}\mathbb{P}\{B_i > (1-\rho)x\}, \quad x \to \infty.$$

Turning to the term II_2, it immediately follows from Proposition 4.3 below and the fact that B_i is regularly varying that

$$II_2 \leq \mathbb{E}\{2\mathbb{1}_{N_{>\epsilon x}=2}\} = 2\mathbb{P}\{N_{>\epsilon x} = 2\} = \mathrm{o}(\mathbb{P}\{B_i > (1-\rho)x\}), \quad x \to \infty.$$

By definition of $A(\epsilon x)$, we have $II_0 = 0$.
In conclusion, $\frac{1}{\mathbb{E}\{N\}p_i}\sum_{k=0}^{2} II_k \leq \mathbb{P}\{B_i > (1-\rho)x\}(1+\mathrm{o}(1))$.

Next, we show that the terms I_k, $k = 0, 1, 2$, asymptotically vanish as well. The sojourn time of customers that arrive to a system with at most k large customers is smaller than the sojourn time of these same customers entering a system with k permanent customers. Define $N_{\leq \epsilon x,k}$ as the number of customers arriving during a busy period in a system with k permanent customers and all service requirements truncated at level ϵx. Denote by $V^{(j)}_{\leq \epsilon x,k}$ and $V_{\leq \epsilon x,k}$ the sojourn times of the j-th customer and an arbitrary customer entering this system, respectively. With these random variables, the above assertion may be formalised as follows:

$$I_k \leq \mathbb{E}\{\sum_{j=1}^{N_{\leq \epsilon x,k}} \mathbb{1}_{V^{(j)}_{\leq \epsilon x,k}>x}\} = \mathbb{E}\{N_{\leq \epsilon x,k}\}\mathbb{P}\{V_{\leq \epsilon x,k} > x\}.$$

First observe that the term $\mathbb{E}\{N_{\leq \epsilon x,k}\}$ is finite. In order to control the term $\mathbb{P}\{V_{\leq \epsilon x,k} > x\}$, we use sample-path analysis to find an upper bound for the sojourn time of these 'small' customers. We first introduce some additional notation. Specifically, tag such a customer, and denote its (possibly truncated) service requirement by B_0 and its sojourn time by V_0. To keep the notation simple, we assume that the tagged customer arrives at time $t = 0$. Let $W_{\leq \epsilon x,k}(0^-)$ be the amount of work in the system just before the tagged customer arrives. Let $B_{j,m}$ be the service requirement of the m-th class-j customer arriving after the tagged customer, let $T_{j,m}$ be the arrival time of this customer, and let $R_0(t)$ be the remaining service requirement of the tagged customer at time t. Furthermore, $N_j(s,t)$

denotes the number of class-j customers arriving during the time interval (s,t), excluding the tagged customer. Let $A(s,t)$ be the amount of traffic generated during the time interval (s,t). For any $y \geq 0$, let $A_{\leq y}(s,t)$ be a modified version of the process $A(s,t)$ where all service requirements are truncated at level y. For any $y \geq 0$ and $c > \rho$, we define $W^c_{\leq y} := \sup_{t\geq 0}[A_{\leq y}(0,t) - ct]$. This random variable represents the stationary workload in a system where work arrives according to the process $A_{\leq y}(0,t)$ and is served at constant rate c. We also repeatedly use the constant f_{DPS} to bound the ratio between the weight of class-i customers and the weight of other customers:

$$f_{DPS} := \frac{\max_{j=1,\ldots,K} g_j}{g_i}.$$

As in [5], it may be shown that the DPS discipline implies that for every $0 < \delta < 1 - \rho$,

$$\begin{aligned}
V_0(1-\rho-\delta) &\leq (k+1)B_0 + W_{\leq \epsilon x,k}(0^-) + \sum_{j=1}^{K}\sum_{m=1}^{N_j(0,V_0)} \min\{B_{j,m}, \frac{g_j}{g_i}R_0(T_{j,m})\} - (\rho+\delta)V_0 \\
&\leq (k+1)B_0 + W_{\leq \epsilon x,k}(0^-) + \sum_{j=1}^{K}\sum_{m=1}^{N_j(0,V_0)} \min\{B_{j,m}, f_{DPS}B_0\} - (\rho+\delta)V_0 \\
&\leq (k+1)\epsilon x + W_{\leq \epsilon x,k}(0^-) + A_{\leq f_{DPS}B_0}(0,V_0) - (\rho+\delta)V_0 \\
&\leq (k+1)\epsilon x + W_{\leq \epsilon x,k}(0^-) + W^{\rho+\delta}_{\leq \epsilon f_{DPS}x}.
\end{aligned}$$

Using the above sample-path inequality, we obtain

$$\begin{aligned}
\mathbb{P}\{V_{\leq \epsilon x,k} > x\} &\leq \mathbb{P}\{W_{\leq \epsilon x,k}(0^-) + W^{\rho+\delta}_{\leq \epsilon f_{DPS}x} > (1-(k+1)\epsilon)x\} \\
&\leq \mathbb{P}\{W_{\leq \epsilon x,k}(0^-) > \frac{1-(k+1)\epsilon}{2}x\} + \mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon f_{DPS}x} > \frac{1-(k+1)\epsilon}{2}x\}.
\end{aligned}$$

Lemmas 4.6 and 4.7 presented below show that both of the above terms are $o(\mathbb{P}\{B > x\})$ and thus $o(\mathbb{P}\{B_i > (1-\rho)x\})$ since B_i is regularly varying. We conclude that $I_k = o(\mathbb{P}\{B_i > (1-\rho)x\})$ for all k.
It remains to prove that the term III is asymptotically negligible. Applying Hölder's inequality, we may write

$$\begin{aligned}
III &\leq \mathbb{E}\{N \mathbb{1}_{N_{>\epsilon x} \geq 3}\} \\
&\leq (\mathbb{E}\{N^2\})^{1/2}(\mathbb{E}\{(\mathbb{1}_{N_{>\epsilon x}} \geq 3)^2\})^{1/2} \\
&= (\mathbb{E}\{N^2\})^{1/2}(\mathbb{P}\{N_{>\epsilon x} \geq 3\})^{1/2}.
\end{aligned}$$

Since B is regularly varying, we have that $\mathbb{P}\{B > x\} = l(x)x^{-\nu}$ for some slowly varying function $l(\cdot)$. Theorem 43.3 in Borovkov [4] guarantees the existence of a constant C such that $\mathbb{P}\{N > x\} \leq Cl(x)x^{-\nu}$. As $\nu > 2$, this gives that $\mathbb{E}\{N^2\} < \infty$. By a direct application of Proposition 4.4 below, we deduce that

$$\mathbb{P}\{N_{>\epsilon x} \geq 3\}^{1/2} = o(\mathbb{P}\{B_i > (1-\rho)x\}),$$

which completes the proof. □

The next proposition provides an upper bound for the probability that two or more large customers arrive during a busy period. The probability that one large customer arrives during a busy period is approximately equal to the probability that one out of O(1) customers is large, i.e., $O(x^{-\nu})$. During the sojourn time of such a large customer there are $O(x)$ 'possibilities' for another large customer to arrive. So the probability that a second large customer arrives during the sojourn time of the first large customer is of the order $x \cdot x^{-\nu}$. These heuristic arguments suggest that the probability that two or more large customers arrive during a busy period roughly behaves as $x^{-\nu}(x \cdot x^{-\nu})$, which is asymptotically smaller than the probability $\mathbb{P}\{B_i > (1-\rho)x\}$ for $\nu > 1$.

Proposition 4.3 *If B is regularly varying of index $\nu > 1$, then*

$$\mathbb{P}\{N_{>x} \geq 2\} = O(x\mathbb{P}\{B > x\}^2), \quad x \to \infty.$$

Proof

We first introduce some useful notation. Denote by $B^{(n)}$ the service requirement of the n-th customer in a busy period and by $A^{(n)}$ the interarrival time between the n-th and $(n+1)$-st customer in a busy period. Let $S_n := S_{n-1} + X_n$ be a random walk with initial value $S_0 = 0$ and increments $X_n := B^{(n)} - A^{(n)}$. We define $\tau_1(x) := \inf\{n : X_n \geq x\}$ as the first time that a jump in the random walk S_n is larger than x. Likewise, $\sigma(0) := \inf\{n : S_n \leq 0\}$ is the first time that the random walk is smaller than 0.

We study the behaviour of the random variable $M_{>x}$, representing the number of times that a jump in the random walk is larger than x during a busy period. This random variable is slightly different but strongly related to the random variable $N_{>x}$ as will be shown later.

Conditioning on the amount of work in the system $S_{\tau_1(x)}$ just after the arrival of the first large customer, we obtain

$$\begin{aligned}\mathbb{P}\{M_{>x} \geq 2\} &\leq \int_{y=0}^{\infty} \mathbb{P}\{M_{>x} \geq 2 \mid S_{\tau_1(x)} = x+y; \tau_1(x) < N\} \mathrm{d}\mathbb{P}\{S_{\tau_1(x)} \leq x+y; \tau_1(x) < N\} \\ &= \int_{y=0}^{\infty} \mathbb{P}\{\exists X_i > x; i \leq \sigma(0) \mid S_0 = x+y\} \mathrm{d}\mathbb{P}\{S_{\tau_1(x)} \leq x+y; \tau_1(x) < N\}.\end{aligned}$$

First, we consider the probability $\mathbb{P}\{\exists X_i > x; i \leq \sigma(0) \mid S_0 = z\}$. We use a branching argument to derive an asymptotic upper bound for the probability that a large jump in the random walk with initial value z occurs before it becomes negative. The order in which the work that enters the system is processed does not influence this probability, so we may choose this order in any way we want. During the time period $(0, z)$ we devote the total system capacity to the initial value z of the random walk, and subsequently we allocate the total capacity to the $N(0, z)$ jumps that have occurred in the meantime, one at a time. Each of these $N(0, z)$ jumps will start its own busy period. Since we allow arbitrary renewal arrival processes, the residual interarrival time until the second customer of each of these $N(0, z)$ busy periods is generally not distributed as A. We can however find an upper bound for the probability that a large customer arrives during such a 'busy period' by assuming that the second customer arrives immediately. This way we obtain $N(0, z)$ busy periods which each start with a simultaneous arrival of two customers.

To study the probability of a large jump occurring in one of these $N(0, z)$ busy periods, let T_n be a random walk with initial value $T_0 = 0$, $T_1 := B^{(1)} + B^{(2)} - A^{(1)}$ and $T_{n+1} :=$

$T_n + B^{(n+1)} - A^{(n)}$. We use N_2 to denote the number of customers that arrive during such a busy period and $M_{2,>x}$ to denote the number of jumps which are larger than x that occur during such a busy period. Define the cycle maximum of this random walk as $C_{2,\max} := \max\{T_n : n \leq N_2\}$. Using the above line of reasoning, we obtain the following inequalities

$$\begin{aligned}
\mathbb{P}\{\exists X_i > x; i \leq \sigma(0) \mid S_0 = z\} &\leq \mathbb{P}\{M_{2,>x}^{(1)} + M_{2,>x}^{(2)} + \ldots + M_{2,>x}^{(N(0,z))} \geq 1\} \\
&= \mathbb{P}\{\exists i \in 1, \ldots, N(0,z) : M_{2,>x}^{(i)} \geq 1\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}\{\exists i \in 1, \ldots, n : M_{2,>x}^{(i)} \geq 1\} \mathbb{P}\{N(0,z) = n\} \\
&\leq \sum_{n=1}^{\infty} n \mathbb{P}\{N(0,z) = n\} \mathbb{P}\{M_{2,>x} \geq 1\} \\
&\leq \mathbb{E}\{N(0,z)\} \mathbb{P}\{C_{2,\max} > x\}.
\end{aligned}$$

The random variable $N_2 := \min\{n \geq 1 \mid T_n \leq 0\} \stackrel{\mathrm{d}}{=} \min\{n \geq 1 \mid S_n \leq -B^{(1)}\}$ is a stopping time (where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution). We now use Theorem 1 from Foss & Zachary [7] which implies $\mathbb{P}\{C_{2,\max} > x\} = \mathbb{E}\{N_2\}\mathbb{P}\{X > x\}$ as $x \to \infty$. So for any $\delta > 0$, we can choose x sufficiently large such that

$$\mathbb{P}\{C_{2,\max} > x\} \leq (\mathbb{E}\{N_2\} + \delta)\mathbb{P}\{X > x\}.$$

By the elementary renewal theorem, $\frac{1}{z}\mathbb{E}\{N(0,z)\} \to \lambda$ as $z \to \infty$, so that for any $\delta > 0$, there exists a $\bar{z}$ such that $\mathbb{E}\{N(0,z)\} \leq (\lambda + \delta)z$ for all $z \geq \bar{z}$. So for any $\delta > 0$, we can again choose x sufficiently large such that

$$\begin{aligned}
\mathbb{P}\{\exists X_i > x; i \leq \sigma(0) \mid S_0 = x + y\} &\leq \mathbb{E}\{N(0, x+y)\}(\mathbb{E}\{N_2\} + \delta)\mathbb{P}\{X > x\} \\
&\leq (\lambda + \delta)(\mathbb{E}\{N_2\} + \delta)(x + y)\mathbb{P}\{X > x\}.
\end{aligned}$$

Using the above derivation and applying partial integration, we obtain

$$\begin{aligned}
\mathbb{P}\{M_{>x} \geq 2\} &\leq \int_{y=0}^{\infty} \mathbb{P}\{\exists X_i > x; i \leq \sigma(0) \mid S_0 = x + y\} \mathrm{d}\mathbb{P}\{S_{\tau_1(x)} \leq x + y; \tau_1(x) < N\} \\
&\leq \int_{y=0}^{\infty} -(\lambda + \delta)(\mathbb{E}\{N_2\} + \delta)(x + y)\mathbb{P}\{X > x\} \mathrm{d}\mathbb{P}\{S_{\tau_1(x)} > x + y; \tau_1(x) < N\} \\
&= -\mathbb{P}\{X > x\}[(\lambda + \delta)(\mathbb{E}\{N_2\} + \delta)(x + y)\mathbb{P}\{S_{\tau_1(x)} > x + y; \tau_1(x) < N\}]_{y=0}^{\infty} + \\
&\quad \mathbb{P}\{X > x\} \int_{y=0}^{\infty} (\lambda + \delta)(\mathbb{E}\{N_2\} + \delta)\mathbb{P}\{S_{\tau_1(x)} > x + y; \tau_1(x) < N\} \mathrm{d}y \\
&\leq (\lambda + \delta)(\mathbb{E}\{N_2\} + \delta)\mathbb{P}\{X > x\} x \mathbb{P}\{\tau_1(x) < N\} + \\
&\quad (\lambda + \delta)(\mathbb{E}\{N_2\} + \delta)\mathbb{P}\{X > x\} \int_{y=0}^{\infty} \mathbb{P}\{S_{\tau_1(x)} > x + y; \tau_1(x) < N\} \mathrm{d}y. \qquad (2)
\end{aligned}$$

By again using a cycle maximum C_{max}, now defined by $C_{max} := \max\{S_n : n \leq N\}$, the expression in (2) may be bounded by

$$\begin{aligned}\int_{y=0}^{\infty} \mathbb{P}\{S_{\tau_1(x)} > x+y; \tau_1(x) < N\}\mathrm{d}y &\leq \int_{y=0}^{\infty} \mathbb{P}\{C_{max} > x+y\}\mathrm{d}y \\ &= \int_{y=x}^{\infty} \mathbb{P}\{C_{max} > y\}\mathrm{d}y \\ &= \mathbb{E}\{C_{max}\}\mathbb{P}\{C^r_{max} > x\}. \qquad (3)\end{aligned}$$

Clearly, $\mathbb{P}\{\tau_1(x) < N\} \leq \mathbb{P}\{C_{max} > x\}$ and Theorem 1 from Foss & Zachary [7] shows that

$$\mathbb{P}\{C_{max} > x\} = \mathrm{O}(\mathbb{P}\{X > x\}) = \mathrm{O}(\mathbb{P}\{B > x\}), \quad x \to \infty.$$

Using the fact that B is regularly varying, it may further be readily shown that

$$\mathbb{P}\{C^r_{max} > x\} = \mathrm{O}(x\mathbb{P}\{B > x\}), \quad x \to \infty.$$

Combining the above with (2) and (3), we conclude

$$\mathbb{P}\{M_{>x} \geq 2\} = \mathrm{O}(\mathbb{P}\{x\mathbb{P}\{B > x\}^2), \quad x \to \infty.$$

It remains to establish the link between the variables $M_{>x}$ and $N_{>x}$. If we truncate the interarrival times at some value $A_{\max}$, then that will only increase the probability that a given number of large customers arrive during a busy period. If we choose $A_{\max}$ large enough, then the system will also remain stable. So for the purpose of finding an upper bound, we may assume the interarrival times to be bounded by some value $A_{\max}$. Using the fact that B is regularly varying, it then follows

$$\mathbb{P}\{N_{>x} \geq 2\} \leq \mathbb{P}\{M_{>x-A_{\max}} \geq 2\} = \mathrm{O}(x\mathbb{P}\{B > x\}^2), \quad x \to \infty.$$

□

The next proposition presents an upper bound for the probability that three or more large customers arrive during a busy period. The intuitive arguments are largely similar to those for the case of two large customers. Specifically, the heuristic approximation of the probability that a second large customer arrives during the sojourn time of the first one, directly extends to the second and third customer. This suggests that the probability that three large customers arrive during a busy period roughly behaves as $x^{-\nu}(x \cdot x^{-\nu})^2$, which is asymptotically smaller than $x\mathbb{P}\{B > x\}^2$ for $\nu > 1$.
Observing the strong similarity between Propositions 4.3 and 4.4, it is tempting to conjecture that the probability of k large customers arriving in a busy period is of the order $\mathrm{O}(x^{k-1}\mathbb{P}\{B > x\}^k)$. This hypothesis is however not true without further assumptions as will be shown in Section 5.

Proposition 4.4 *If B is regularly varying of index $\nu > 2$, then*

$$\mathbb{P}\{N_{>x} \geq 3\} = \mathrm{O}(x^2\mathbb{P}\{B > x\}^3), \quad x \to \infty.$$

Proof
The proof is similar to that of Proposition 4.3 for the case of two large customers. Conditioning on the amount of work in the system $S_{\tau_1(x)}$ just after the arrival of the first large customer, we obtain

$$\mathbb{P}\{M_{>x} \geq 3\} \leq \int_{y=0}^{\infty} \mathbb{P}\{\exists X_i, X_j > x; i, j \leq \sigma(0), i \neq j \mid S_0 = x+y\} \mathrm{d}\mathbb{P}\{S_{\tau_1(x)} \leq x+y; \tau_1(x) < N\}.$$

As before, we use a branching argument to derive an asymptotic upper bound for the probability $P(z) = \mathbb{P}\{\exists X_i, X_j > x; i, j \leq \sigma(0), i \neq j \mid S_0 = z\}$ that two large jumps occur in the random walk with initial value z before it becomes negative.

$$\begin{aligned} P(z) &\leq \mathbb{P}\{M_{2,>x}^{(1)} + M_{2,>x}^{(2)} + \ldots + M_{2,>x}^{(N(0,z))} \geq 2\} \\ &= \mathbb{P}\{\max_{i,j \leq N(0,z)} \{M_{2,>x}^{(i)} + M_{2,>x}^{(j)}\} \geq 2\} \\ &= I + II, \end{aligned}$$

where

$$\begin{aligned} I &= \mathbb{P}\{\exists i \in 1, \ldots, N(0,z) : M_{2,>x}^{(i)} \geq 2\} \\ II &= \mathbb{P}\{\exists i, j \in 1, \ldots, N(0,z), i \neq j : M_{2,>x}^{(i)} \geq 1, M_{2,>x}^{(j)} \geq 1\}. \end{aligned}$$

Counting the combinations that lead to two large customers, we obtain the following inequalities:

$$\begin{aligned} I &= \sum_{n=1}^{\infty} \mathbb{P}\{\exists i \in 1, \ldots, n : M_{2,>x}^{(i)} \geq 2\} \mathbb{P}\{N(0,z) = n\} \\ &\leq \sum_{n=1}^{\infty} n \mathbb{P}\{N(0,z) = n\} \mathbb{P}\{M_{2,>x} \geq 2\} \\ &\leq \mathbb{E}\{N(0,z)\} \mathbb{P}\{M_{2,>x} \geq 2\}, \end{aligned}$$

and

$$\begin{aligned} II &= \sum_{n=1}^{\infty} \mathbb{P}\{\exists i, j \in 1, \ldots, n, i \neq j : M_{2,>x}^{(i)} \geq 1, M_{2,>x}^{(j)} \geq 1\} \mathbb{P}\{N(0,z) = n\} \\ &\leq \sum_{n=1}^{\infty} n^2 \mathbb{P}\{N(0,z) = n\} (\mathbb{P}\{M_{2,>x} \geq 1\})^2 \\ &= \mathbb{E}\{N(0,z)^2\} (\mathbb{P}\{M_{2,>x} \geq 1\})^2 \\ &\leq \mathbb{E}\{N(0,z)^2\} (\mathbb{P}\{C_{2,\max} > x\})^2. \end{aligned}$$

Similar to the proof of Proposition 4.3, with only a slight modification in the first step of the proof, it follows that

$$\mathbb{P}\{M_{2,>x} \geq 2\} = \mathrm{O}(x \mathbb{P}\{B > x\}^2), \quad x \to \infty.$$

Theorem 1 from Foss & Zachary [7] implies that

$$\mathbb{P}\{C_{2,\max} > x\} = \mathrm{O}(\mathbb{P}\{X > x\}) = \mathrm{O}(\mathbb{P}\{B > x\}), \quad x \to \infty.$$

Applying the elementary renewal theorem, we conclude that $\mathbb{E}\{N(0,z)\} = \mathrm{O}(z)$ as $z \to \infty$. An extension of this theorem shows that $\mathbb{E}\{N(0,z)^2\} = \mathrm{O}(z^2)$ as $z \to \infty$. This extension requires one minor condition, namely finite variance of the interarrival times. However, this condition is not restrictive because if we truncate the interarrival times, then the variance of the interarrival times will be finite, while the number of large customers in a busy period will only increase.
Together, these results imply the existence of a constant C such that for sufficiently large x

$$P(x+y) \leq C(x+y)^2 \mathbb{P}\{B > x\}^2.$$

Applying partial integration, we thus deduce (for existence of the integral $\nu > 2$ is needed)

$$\begin{aligned}
\mathbb{P}\{M_{>x} \geq 3\} &\leq \mathbb{P}\{B > x\}^2 \int_{y=0}^{\infty} C(x+y)^2 \mathrm{d}\mathbb{P}\{S_{\tau_1(x)} \leq x+y; \tau_1(x) < N\} \\
&= \mathbb{P}\{B > x\}^2 [-C(x+y)^2 \mathbb{P}\{S_{\tau_1(x)} > x+y; \tau_1(x) < N\}]_{y=0}^{\infty} \\
&+ \mathbb{P}\{B > x\}^2 \int_{y=0}^{\infty} 2C(x+y) \mathbb{P}\{S_{\tau_1(x)} > x+y; \tau_1(x) < N\} \mathrm{d}y \\
&\leq \tilde{I} + \widetilde{II},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{I} &= Cx^2 \mathbb{P}\{B > x\}^2 \mathbb{P}\{\tau_1(x) < N\} \\
\widetilde{II} &= \mathbb{P}\{B > x\}^2 \int_{y=x}^{\infty} 2Cy \mathbb{P}\{C_{\max} > y\} \mathrm{d}y.
\end{aligned}$$

Since $\mathbb{P}\{\tau_1(x) < N\} = \mathrm{O}(\mathbb{P}\{B > x\})$, we have $\tilde{I} = \mathrm{O}(x^2 \mathbb{P}\{B > x\}^3)$.
A straightforward application of Theorem 1 from Foss & Zachary [7] shows that $\mathbb{P}\{C_{\max} > x\} = \mathrm{O}(\mathbb{P}\{B > x\})$ as $x \to \infty$. Because

$$\int_{y=x}^{\infty} y \mathbb{P}\{B > y\} \mathrm{d}y \sim \frac{1}{\nu - 2} x^2 \mathbb{P}\{B > x\}, \quad x \to \infty,$$

it then follows

$$\widetilde{II} = \mathrm{O}(x^2 \mathbb{P}\{B > x\}^3), \quad x \to \infty.$$

Using the relationship between $N_{>x}$ and $M_{>x}$ given at the end of the proof of Proposition 4.3 completes the proof. □

The next lemma is needed to prove Lemmas 4.6 and 4.7. The random variable $W^c_{\leq y}(t)$ used in the lemma represents the workload at time t in a system where work arrives according to the process $A_{\leq y}(s,t)$ and is served at constant rate c. The random variable $A_{\leq y}(s,t)$ is defined in the proof of Proposition 4.2.

Lemma 4.5 *If B is regularly varying of index $\nu > 1$ and t_a an arbitrary arrival epoch, then for any $c > \rho$ and $n \in \mathbb{N}$,*

$$\mathbb{P}\{W^c_{\leq x}(t_a^-) > (n+1)x\} = \mathrm{O}(x^{-n(\nu-1)}), \quad x \to \infty.$$

Proof
We bound the interarrival times by M. If we choose M sufficiently large, then the system remains stable. For any $y \geq 0$, $M > 0$, let $A_{\leq y,M}(s,t)$ be a modified version of the process $A_{\leq y}(s,t)$ where the interarrival times are truncated at level M. For any $c > \rho$, define $W^c_{\leq y,M} := \sup_{t\geq 0}[A_{\leq y,M}(0,t) - ct]$. The random variable $W^c_{\leq y,M}$ represents the stationary workload in a system where work arrives according to the process $A_{\leq y,M}(0,t)$ and is served at constant rate c. For any $y > 0$ and $c > \rho$, there exists a sufficiently large M so that $W^c_{\leq y,M}$ is a proper random variable. Theorem 2 in Jelenković [10] implies

$$\mathbb{P}\{W^c_{\leq x,M}(t_a^-) > (n+1)x\} = \mathrm{O}(x^{-n(\nu-1)}), \quad x \to \infty.$$

Because $\mathbb{P}\{W^c_{\leq x,M}(t) > x\} \geq \mathbb{P}\{W^c_{\leq x}(t) > x\}$ for all t and x, the statement follows. □

The next lemma pertains to the workload associated with small customers in a system with k permanent customers and service requirements truncated at level ϵx. It shows that the workload just before an arrival instant is small compared to the untruncated service requirements of customers.

Lemma 4.6 *If B is regularly varying of index $\nu > 1$ and t_a an arbitrary arrival epoch, then*

$$\mathbb{P}\{W_{\leq \epsilon x,k}(t_a^-) > x\} = \mathrm{o}(\mathbb{P}\{B_i > x\}), \quad x \to \infty.$$

Proof
For conciseness, we call customers with a service requirement smaller than ϵx 'small' customers. The DPS discipline implies that, the larger the number of small customers in the system, the larger their (combined) service rate. Denote by $N_k := \lceil k \frac{1+\rho}{1-\rho} \frac{\max_{j=1,\ldots,K} g_j}{\min_{j=1,\ldots,K} g_i} \rceil$ the smallest number of customers needed to ensure a total combined service rate of at least $(1+\rho)/2$ for these customers in the presence of the k permanent customers, with $\lceil x \rceil$ denoting the smallest integer greater than or equal to x.
Let $N_{\leq u}(t)$ be the number of customers in the system at time t with service requirement smaller than u. Define $t^* := \sup\{t \in (-\infty, t_a] : N_{\leq \epsilon x}(t) \leq N_k - 1\}$. By definition, the number of small customers in the system is constantly larger than or equal to N_k during the time interval $[t^*, t_a]$. Thus, during $[t^*, t_a]$ the small customers will constantly receive service at a rate larger than or equal to $(1+\rho)/2$. We deduce

$$\begin{aligned} W_{\leq \epsilon x,k}(t_a^-) &\leq N_k \epsilon x + \int_{t^*}^{t_a^-} \mathrm{d}W_{\leq \epsilon x}(t) \\ &\leq N_k \epsilon x + \int_{t^*}^{t_a^-} \mathrm{d}W^{(1+\rho)/2}_{\leq \epsilon x}(t) \\ &\leq N_k \epsilon x + W^{(1+\rho)/2}_{\leq \epsilon x}(t_a^-). \end{aligned}$$

Since $t = t_a$ is an arbitrary arrival epoch, we can use Lemma 4.5 to choose $\epsilon > 0$ sufficiently small so that

$$\mathbb{P}\{W_{\leq \epsilon x,k}(t_a^-) > x\} \leq \mathbb{P}\{W^{(1+\rho)/2}_{\leq \epsilon x}(t_a^-) > (1 - N_k\epsilon)x\} = \mathrm{o}(\mathbb{P}\{B_i > x\}, \quad x \to \infty.$$

□

The next lemma presents the final element that is needed in the proof of the upper bound. It implies that the stationary workload in a system where the service requirements are truncated at level ϵx is small compared to an untruncated service requirement.

Lemma 4.7 *If B is regularly varying of index $\nu > 1$, then*

$$\mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x} > x\} = \mathrm{o}(\mathbb{P}\{B_i > x\}), \quad x \to \infty.$$

Proof
Let t_a be an arbitrary arrival epoch and let $B^r_{\leq y}$ be a remaining service requirement in a system where the service requirements are truncated at level y. Then,

$$\begin{aligned}
\mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x} > x\} &= \mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x} > 0\}\mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x}(t_a^-) + B^r_{\leq \epsilon x} > x\} \\
&< \mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x}(t_a^-) + \epsilon x > x\} \\
&= \mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x}(t_a^-) > (1-\epsilon)x\}.
\end{aligned}$$

Using Lemma 4.5, we can choose $\epsilon > 0$ sufficiently small so that

$$\mathbb{P}\{W^{\rho+\delta}_{\leq \epsilon x}(t_a^-) > (1-\epsilon)x\} = \mathrm{o}(\mathbb{P}\{B_i > x\}), \quad x \to \infty.$$

□

5 Discussion

In the present paper we have developed a stochastic mean-value method for deriving the delay asymptotics in PS type systems with heavy-tailed service requirements. In order to demonstrate the potential use, we have applied the method to the DPS discipline, which has largely defied analysis via existing approaches.
The derivation of the delay asymptotics involved the assumption that the variance of the service requirements is finite. However, we conjecture that the latter assumption is not essential for the main result to hold, since the heuristic arguments presented at the beginning of Section 3 do not rely on finite variance in any way. Note that the assumption is in fact not used in obtaining the lower bound at all, and is only used in proving the upper bound. Avoiding the finite-variance assumption in the latter proof appears extremely difficult though.
Recall that we needed the assumption $\nu > 2$ to establish the following inequality at the end of the proof of Proposition 4.2:

$$III \leq (\mathbb{E}\{N^2\})^{1/2}(\mathbb{P}\{N_{>\epsilon x} \geq 3\})^{1/2} = \mathrm{o}(\mathbb{P}\{B_i > (1-\rho)x\}), \quad x \to \infty.$$

In order to extend the main result to the case $1 < \nu \leq 2$, a natural attempt would be to generalise the above inequality to

$$III \leq (\mathbb{E}\{N^{\nu-\delta}\})^{\frac{1}{\nu-\delta}}\mathbb{P}\{N_{>\epsilon x} > K\}^{1-\frac{1}{\nu-\delta}}.$$

To prove the latter inequality, it would suffice to show that for large enough K

$$\mathbb{P}\{N_{>\epsilon x} > K\}^{1-\frac{1}{\nu-\delta}} = \mathrm{o}(\mathbb{P}\{B_i > (1-\rho)x\}), \quad x \to \infty.$$

which seems plausible upon inspection of Propositions 4.3 and 4.4. These results suggest that

$$\mathbb{P}\{N_{>\epsilon x} > K\} = \mathrm{O}(x^{K-1}\mathbb{P}\{B > x\}^K), \quad x \to \infty,$$

and for $\nu > 1$, it is clear that a large enough K exists such that

$$(x^{K-1}\mathbb{P}\{B > x\}^K)^{1-\frac{1}{\nu-\delta}} = \mathrm{o}(\mathbb{P}\{B_i > (1-\rho)x\}), \quad x \to \infty.$$

Remember that the proofs of Propositions 4.3 and 4.4 are based on the intuitive notion that during the sojourn time of a 'large' customer there are $\mathrm{O}(x)$ 'possibilities' for another large customer to arrive. So the probability that a second large customer arrives during the sojourn time of the first large customer behaves roughly as $x \cdot x^{-\nu}$. Similarly, the probability that a third large customer arrives during the sojourn time of the second large customer is approximately $x \cdot x^{-\nu}$. This line of argument would suggest that the probability that K large customers arrive during a busy period behaves roughly as $x^{-\nu}(x \cdot x^{-\nu})^{K-1}$. However, there exist alternative scenarios for several large customers to arrive during a busy period. One of these involves the event that the service requirement of the first large customer does not just exceed x, but even x^ν. During the sojourn time of such an extremely large customer, it is nearly certain that a large customer of size $\mathrm{O}(x)$ will arrive. Moreover, the amount of work in the system will not decrease by any significant amount during this interarrival time. So with near certainty yet another large customer of size $\mathrm{O}(x)$ will arrive, and so forth. Thus, the probability that K large customers arrive during a busy period, is at least of the order of the probability that a customer with service requirement larger than x^ν arrives, which is approximately $(x^\nu)^{-\nu} = x^{-\nu^2}$.
It is clear that the latter scenario for K large customers to arrive during a busy period becomes the dominant one when K grows large, which implies that for $1 < \nu \leq 2$, there is no hope of proving

$$\mathbb{P}\{N_{>\epsilon x} > K\}^{1-\frac{1}{\nu-\delta}} = \mathrm{o}(\mathbb{P}\{B_i > (1-\rho)x\}), \quad x \to \infty.$$

Thus, it remains as a challenge to extend the results to the case of service requirements with infinite variance. We refer to [2] for further insights on the significance of infinite variance in DPS queues.

References

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

[2] Ayesta, U., Avrachenkov, K.E., Brown, P., Núñez-Queija, R. (2004). Discriminatory Processor Sharing revisited. Submitted for publication.

[3] Bingham, N.H., Goldie, C., Teugels, J. (1987). *Regular Variation.* Cambridge University Press.

[4] Borovkov, A.A. (1984). *Ergodicity and Stability of Stochastic Processes.* Wiley, Chichester.

[5] Borst, S.C., Van Ooteghem, D.T.M.B., Zwart, A.P. (2003). Tail asymptotics for Discriminatory Processor-Sharing queues with heavy-tailed service requirements. SPOR-Report 2003-25, Eindhoven University of Technology. *Perf. Eval.*, to appear.

[6] Fayolle, G., Mitrani, I., Iasnogorodski, R. (1980). Sharing a processor among many job classes. *J. ACM* **27**, 519–532.

[7] Foss, S., Zachary, S. (2003). The maximum on a random time interval of a random walk with long-tailed increments and negative drift. *Ann. App. Prob.* **13**, 37–53.

[8] Guillemin, F., Robert, Ph., Zwart, A.P. (2004). Tail asymptotics for processor sharing queues. *Adv. Appl. Prob.* **36**, 525–543.

[9] Haviv, M., Van der Wal, J. (2004). Waiting times in queues and in processor sharing systems with relative priorities. Submitted for publication.

[10] Jelenković, P.R. (1999). Network multiplexer with truncated heavy-tailed arrival streams. In: *Proc. IEEE Infocom '99*, New York, NY, USA, 625–633.

[11] Jelenković, P.R., Momčilović, P. (2003). Large deviation analysis of subexponential waiting times in a processor sharing queue. *Math. Oper. Res.* **28**, 587–608.

[12] Núñez-Queija, R. (2000). *Processor-Sharing Models for Integrated-Services Networks.* PhD Thesis, Eindhoven University of Technology.

[13] Núñez-Queija, R. (2002). Queues with equally heavy sojourn time and service requirement distributions. *Ann. Oper. Res.* **113**, 101–117.

[14] Rege, K.M., Sengupta, B. (1994). A decomposition theorem and related results for the discriminatory processor-sharing queue. *Queueing Systems* **18**, 333–351.

[15] Rege, K.M., Sengupta, B. (1996). Queue length distribution for the discriminatory processor-sharing queue. *Oper. Res.* **44**, 653–657.

[16] Van Kessel, G., Núñez-Queija, R., Borst, S.C. (2004). Differentiated bandwidth sharing with disparate flow sizes. Submitted for publication.

[17] Zwart, A.P. (1999). Sojourn times in a multiclass processor sharing queue. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 335–344.

[18] Zwart, A.P. (2001). *Queueing Systems with Heavy Tails.* PhD Thesis, Eindhoven University of Technology.

[19] Zwart, A.P., Boxma, O.J. (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems* **35**, 141–166.