Tail asymptotics for discriminatory processor-sharing queues with heavy-tailed service requirements

Citation for published version (APA):

Document status and date:
Published: 01/01/2003

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Tail Asymptotics for Discriminatory Processor-Sharing Queues with Heavy-Tailed Service Requirements

Sem Borst†,*, Dennis van Ooteghem†,*, Bert Zwart†,*

†CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

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

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

Abstract

We derive the sojourn time asymptotics for a multi-class G/G/1 queue with regularly varying service requirements operating under the Discriminatory Processor-Sharing (DPS) discipline. DPS provides a natural approach for modelling the flow-level performance of differentiated bandwidth-sharing mechanisms. Under certain assumptions, we prove that the service requirement and sojourn time of a given class have similar tail behaviour, independent of the specific values of the DPS weights. As a by-product, we obtain an extension of the tail equivalence for ordinary Processor-Sharing (PS) queues to non-Poisson arrivals. The results suggest that DPS offers a potential instrument for effectuating preferential treatment to high-priority classes, without inflicting excessive delays on low-priority classes. To obtain the asymptotics, we develop a novel method which only involves information of the workload process and does not require any knowledge of the steady-state queue length distribution. In particular, the proof method brings sufficient strength to extend the results to scenarios with a time-varying service capacity.

Keywords: differentiated services; (discriminatory) processor sharing; heavy-tailed traffic; regular variation; sojourn time asymptotics.

1 Introduction

Over the past few years, the Processor-Sharing (PS) discipline has been widely adopted as a convenient paradigm for modelling the bandwidth sharing among dynamically interacting TCP flows. Independently, extensive measurement studies have indicated that file sizes in the Internet, and hence the data volumes of TCP flows, commonly exhibit heavy-tailed distributions. Therefore, the analysis of the tail behaviour of the sojourn time becomes crucial for understanding the performance of differentiated services.

The work of the first two authors was in part financially supported through the EQUANET project. The third author was supported by an NWO-VENI grant.
tailed features, see for instance [6]. These observations have triggered a huge interest in the delay characteristics of PS queues with heavy-tailed service requirements. Zwart & Boxma [25] obtained the delay asymptotics for regularly varying service requirements using transform techniques. 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 [23] generalised the result to multi-class PS queues. Using a proof based on conditional moments, Núñez-Queija [17, 18] extended the tail equivalence result to PS models with a time-varying service capacity and intermediately regularly varying service requirements. Jelenković & Momčilović [13] used 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, Boyer et al. [5] obtained a tail equivalence result for heavy-tailed PS queues with impatience and admission control as well as queues operating under a state-dependent PS discipline.

While the above results provide valuable insights, they invariably rely on the assumption that the service capacity is shared in an egalitarian manner. The actual bandwidth shares may however show substantial variation among competing users with heterogeneous characteristics. For instance, TCP flows that share a common bottleneck link but traverse heterogeneous routes, may experience diverse packet loss rates and round-trip delays. Because of TCP mechanics, these differences result in a significant discrepancy in the bandwidth shares, see for instance Altman, Jimenez & Kofman [2].

Besides TCP-related effects, the heterogeneity in bandwidth shares may also be due to deliberate service differentiation among competing users. As the Internet evolves to support an ever increasing range of services, there is a growing need for some form of service differentiation to satisfy the diverse requirements of heterogeneous applications. The ability to achieve different bandwidth shares is arguably one of the most fundamental vehicles for service differentiation. Both equation-based and general Additive-Increase Multiplicative-Decrease (AIMD) rate control algorithms provide potential instruments for differentiated bandwidth sharing [9, 10, 19, 22]. A somewhat special mechanism of this sort is the algorithm proposed in [14] for supporting low-priority data transfer by utilising excess bandwidth only. This is a somewhat degenerate case with an extreme degree of asymmetry between the low-priority transfers and regular TCP flows.

The Discriminatory Processor-Sharing (DPS) discipline provides a natural approach for modelling the flow-level performance of such differentiated bandwidth-sharing mechanisms. 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 weight factors. In case all weight factors are equal, the DPS discipline reduces to the familiar egalitarian PS policy. Note that DPS shows some similarity with the Generalised Processor-Sharing (GPS) discipline (or Generalised Head-Of-the-Line (HOL) PS), where the service capacity is also shared in accordance with class-dependent weight factors. In GPS, the capacity is not divided however among all users present, but distributed across (non-empty) classes (e.g. the users at the head-of-the-line of the various classes), irrespective of the actual number of users present.
The results for DPS in the literature are surprisingly sparse. In a seminal paper, Fayolle, Mitrani & Iasnogorodski [8] 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 [20] prove a decomposition theorem for the conditional sojourn time. They specifically show 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 [21] obtain the moments of the queue length distribution from a system of linear equations for the case of exponentially distributed service requirements.

In the present paper, we focus on the exact delay asymptotics for DPS queues with heavy-tailed service requirements. As the relative paucity of results suggests, the analysis of the DPS discipline is extremely difficult compared to that of the ordinary egalitarian PS policy. Most notably, the simple geometric queue length distribution for the standard PS discipline does not have any counterpart for DPS. In addition, there do not seem to be manageable transform results available for the sojourn time distribution. This circumstance considerably complicates the derivation of tail asymptotics, since the existing proof methods for the ordinary PS discipline rely either on transform techniques or probabilistic approaches that exploit knowledge of the queue length distribution. The derivation of tail asymptotics for DPS thus requires a fundamentally different approach to circumvent these difficulties. In the present paper we develop an approach that partially fulfills these requirements, but involves additional distributional assumptions.

Under the latter assumptions, we show that a similar tail equivalence result holds as for the ordinary egalitarian PS policy, regardless of the specific values of the weight factors. At first sight, the asymptotic insensitivity might be perceived as a somewhat negative fact, since it indicates that DPS is not capable of reducing the likelihood of extremely long delays, no matter how the weights are set. Some reflection however shows that such long delays are basically inevitable as they are typically incurred by large customers. In fact, the merit of DPS is not so much in avoiding long delays for large customers, but rather in ensuring short delays to small customers. In that sense, the insensitivity property may be interpreted as a rather positive result, because it also implies that preferential treatment of classes with large weights does not carry the penalty of increasing the occurrence of long delays for classes with smaller weights.

As alluded to above, the probabilistic sample-path approach that we develop avoids the use of explicit queue length information. As a side-benefit, it thus allows us to obtain an extension of the tail equivalence for ordinary PS queues to non-Poisson arrivals. In addition, the proof method turns out to be sufficiently powerful to extend the results to scenarios with a time-varying service capacity. Models with a time-varying service capacity have been of crucial importance in analysing the performance of elastic traffic sharing the bandwidth with higher-priority streaming traffic, see for instance [4, 7, 17].

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 involves lower and upper bounds which may be found in Subsections 4.1 and 4.2, respectively. In Section 5, we extend the results to scenarios with a time-varying service capacity.
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 \( \lambda \), 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

\[
P\{ B > x \} = \sum_{i=1}^{K} p_i P\{ B_i > x \}.
\]

Define \( \rho_i := \lambda_i E\{ B_i \} \) 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 \( \nu \) if

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

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. It relates the asymptotic tail distribution of the sojourn time of class-\( i \) customers to that of their service requirement. 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 \( \lim_{x \to \infty} \frac{f(x)}{g(x)} = 1 \). Note that the assumption that \( B_i \) and \( B \) are both regularly varying of the same index, is slightly milder than \( P\{ B_j > x \} = o(P\{ B_i > x \}) \) as \( x \to \infty \) for all \( j \neq i \).

**Theorem 3.1** If \( B_i \) and \( B \) are both regularly varying of index \( \nu > 2 \), then

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

Despite the remarkably simple form, the proof of the above theorem is quite technical; and will be provided in the next section. However, the result does have a simple interpretation, which will in fact serve as guidance for the proof. Informally speaking, the result shows that a large sojourn time of a customer is most likely due to an exceptionally large service
requirement of that same customer. Specifically, let us focus on a customer with an 
extremely large service requirement (hereafter called a ‘large’ customer). The large customer 
will stay in the system for a long time, and during that time 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.

Note 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 values 
of the weights. When the large customer has a large weight, it will obviously take a 
larger number of small customers to obtain a service rate \( \rho \). Until the small customers 
receive that rate \( \rho \), however, their number will tend to grow, so that eventually the small 
customers will always claim a service rate \( \rho \), 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. 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. 
Thus a fundamentally different method is required.

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. It remains as a challenge to develop a proof method that avoids the latter two 
assumptions.

4 Proofs

We now provide the proof of Theorem 3.1. The proof relies on lower and upper bounds 
which asymptotically coincide.

4.1 Lower bound

We start by proving the lower bound. The next proposition gives a lower bound for 
\( \mathbb{P}\{V_i > x\} \) in terms of \( \mathbb{P}\{B_i > x\} \).
Proposition 4.1.1 If \( B_i \) is regularly varying, then
\[
\mathbb{P}\{V_i > x\} \geq \mathbb{P}\{B_i > (1 - \rho)x\}(1 + o(1)), \quad x \to \infty.
\]
Note that the above proposition in fact only requires \( B_i \) to be regularly varying, and does not involve the additional assumption made in Theorem 3.1 that \( B \) is regularly varying of the same index \( \nu > 2 \).

In the proof of Proposition 4.1.1, we use the notion of a permanent customer. Specifically, the permanent customer is a fictitious class-\( i \) customer which arrives at time 0 and stays in the system forever. Denote by \( B_{\text{perm}}(s, t) \) the amount of service that the permanent customer would receive during the time interval \((s, t)\).

Proof of Proposition 4.1.1
Consider an arbitrary class-\( i \) tagged customer. For convenience, the arrival time of the tagged customer is taken to be time 0. Denote by \( B_0 \) and \( V_0 \) the service requirement and the sojourn time of the tagged customer, respectively. Observe that the amount of service received by the tagged customer during the time interval \((0, x)\) equals \( B_{\text{perm}}(0, x) \) for any \( x < V_0 \) and that the events \( V_0 > x \) and \( B_{\text{perm}}(0, x) < B_0 \) are identical. Thus, for any \( \delta > 0, \epsilon > 0 \),
\[
\mathbb{P}\{V_0 > x\} = \mathbb{P}\{B_{\text{perm}}(0, x) < B_0\} \geq \mathbb{P}\{B_0 > (1 - \rho + \delta + \epsilon)x\} \mathbb{P}\{B_{\text{perm}}(0, x) \leq (1 - \rho + \delta + \epsilon)x\}.
\]
Using Lemma 4.1.2 below, we obtain that the latter probability converges to 1 when \( x \) tends to \( \infty \). Letting \( \delta, \epsilon \downarrow 0 \) and using the fact that \( B_0 \) is regularly varying then completes the proof.

The heuristic arguments in Section 3 suggest that the system should remain stable in the presence of the permanent customer, regardless of the specific values of the DPS weights. Thus, the amount of service that the permanent customer would receive during a time interval \((0, t)\) is unlikely to exceed \((1 - \rho + \delta)t\) by a significant margin. This insight is formalized in the next lemma.

Lemma 4.1.2 For any \( \delta > 0, \epsilon > 0 \),
\[
\lim_{x \to \infty} \mathbb{P}\{\sup_{t \geq 0}[B_{\text{perm}}(0, t) - (1 - \rho + \delta)t] \leq \epsilon x\} = 1.
\]

Proof
We first introduce some useful notation. Let \( A_j(s, t) \) be the amount of traffic generated by class-\( j \) customers during the time interval \((s, t)\). For any \( y \geq 0 \), let \( A_{j, \leq y}(s, t) \) be a modified version of the process \( A_j(s, t) \) where all service requirements are truncated at level \( y \). For any \( c < \rho_j \), define \( U^c_j := \sup_{t \geq 0}[ct - A_j(0, t)] \) and \( U^c_{j,y} := \sup_{t \geq 0}[ct - A_{j, \leq y}(0, t)] \). Furthermore, let \( W_j(t) \) be the workload at time \( t \) of class-\( j \) customers. Finally, denote by \( B_j(s, t) \) the amount of service received during the time interval \((s, t)\) by class-\( j \) customers in the presence of the permanent customer. Then the following obvious equality holds
\[
B_{\text{perm}}(s, t) + \sum_{j=1}^{K} B_j(s, t) = t - s \tag{1}
\]
for all \( t \geq s \geq 0 \).

Using the identity relation \( W_j(t) = W_j(s) + A_j(s,t) - B_j(s,t) \) and taking \( s = 0 \), we obtain

\[
B_{\text{perm}}(0,t) = t + \sum_{j=1}^{K} [W_j(t) - W_j(0) - A_j(0,t)]
\]

\[
\leq (1 - \rho + \delta)t + \sum_{j=1}^{K} [W_j(t) + (\rho_j - \tilde{\delta})t - A_j(0,t)]
\]

\[
\leq (1 - \rho + \delta)t + \sum_{j=1}^{K} W_j(t) + \sum_{j=1}^{K} U_{j}^{\rho_j - \tilde{\delta}}(t).
\]

The statement of the lemma then follows by observing that \( U_{j}^{\rho_j - \tilde{\delta}}(t) \) and \( W_j(t) \) both have finite distributions since \( \mathbb{E}\{A_j(0,t)\} = \rho_j t \) and the various classes remain stable in the presence of the permanent customer. A formal proof argument proceeds as follows.

We will show that \( U_{j}^{\rho_j - \tilde{\delta}}(t) < \tilde{\delta} \epsilon x \) for all \( j = 1, \ldots, K \), with \( \tilde{\delta} := \delta/K \), implies that \( \sup_{t \geq 0} [B_{\text{perm}}(0,t) - (1 - \rho + \delta)T] \leq \epsilon x \).

Suppose that were not the case, i.e., there exists a \( T \geq 0 \) such that \( B_{\text{perm}}(0,T) - (1 - \rho + \delta)T > \epsilon x \).

Define \( T^* := \inf\{t \geq 0 : B_{\text{perm}}(t,T) \leq (1 - \rho + \delta)(T-t)\} \). Note that \( T^* \) is the first time epoch \( t \) at which the remaining service to be received by the permanent customer during the time interval \( (t,T) \) is less than \((1 - \rho + \delta)\) times the remaining duration of the time interval.

Using (1) and the definition of \( T^* \), we obtain

\[
B_{\text{perm}}(0,T) - (1 - \rho + \delta)T \leq B_{\text{perm}}(0,T^*) - (1 - \rho + \delta)T^*
\]

\[
= T^* - \sum_{j=1}^{K} B_j(0,T^*) - (1 - \rho + \delta)T^*
\]

\[
= (\rho - \tilde{\delta})T^* - \sum_{j=1}^{K} B_j(0,T^*)
\]

\[
= \sum_{j=1}^{K} [(\rho_j - \tilde{\delta})T^* - B_j(0,T^*)]. \tag{2}
\]

We now proceed to derive a lower bound for each of the terms in square brackets. In order to do so, we split the customers present during (part of) the time interval \((0,T^*)\) into two groups, depending on whether they arrive before or after time 0. We first introduce some additional notation. Let \( B_{j,m} \) be the service requirement of the \( m \)-th class-\( j \) customer arriving after time 0. Let \( B_{j,l}^{r} \) be the remaining service requirement of the \( l \)-th class-\( j \) customer in the system at time 0. Let \( T_{j,m} \) be the arrival time of the \( m \)-th class-\( j \) customer arriving after time 0. Let \( N_j(t) \) be the number of class-\( j \) customers in the system at time \( t \), excluding the permanent customer. Let \( N_j(s,t) \) be the number of class-\( j \) customers arriving to the system during the time interval \((s,t)\), excluding the permanent customer.
The DPS discipline implies that for any $s \in [0, T^*],$

$$B_j(0, T^*) = \sum_{l=1}^{N_j(0)} \min\{B_{j,l}, \frac{g_j}{g_i} B_{\text{perm}}(0, T^*)\} + \sum_{m=1}^{N_j(0,T^*)} \min\{B_{j,m}, \frac{g_j}{g_i} B_{\text{perm}}(T_{j,m}, T^*)\}$$

$$\geq \sum_{m=1}^{N_j(0,s)} \min\{B_{j,m}, \frac{g_j}{g_i} B_{\text{perm}}(T_{j,m}, T^*)\}$$

$$\geq \sum_{m=1}^{N_j(0,s)} \min\{B_{j,m}, \frac{g_j}{g_i} B_{\text{perm}}(s, T^*)\}.$$  \hspace{1cm} (3)

Using (1) and the definition of $T^*$ once more, we find

$$T^* \geq B_{\text{perm}}(0, T^*)$$

$$= B_{\text{perm}}(0, T) - B_{\text{perm}}(T^*, T)$$

$$> (1 - \rho + \delta)T + \varepsilon x - (1 - \rho + \delta)(T - T^*)$$

$$= (1 - \rho + \delta)T^* + \varepsilon x,$$

which means $(\rho - \delta)T^* > \varepsilon x,$ so that $S^* := T^* - \varepsilon x > 0.$

By definition of $T^*$, we have that $B_{\text{perm}}(s, T) > (1 - \rho + \delta)(T - s)$ for all $s \in [0, T^*)$ and $B_{\text{perm}}(T^*, T) = (1 - \rho + \delta)(T - T^*),$ and thus

$$B_{\text{perm}}(s, T^*) = B_{\text{perm}}(s, T) - B_{\text{perm}}(T^*, T)$$

$$> (1 - \rho + \delta)(T - s) - (1 - \rho + \delta)(T - T^*)$$

$$= (1 - \rho + \delta)(T^* - s).$$  \hspace{1cm} (4)

Combining (3) with (4), taking $s = S^*,$ we derive

$$B_j(0, T^*) - (\rho_j - \tilde{\delta})T^* \geq \sum_{m=1}^{N_j(0,S^*)} \min\{B_{j,m}, \frac{g_j}{g_i} B_{\text{perm}}(S^*, T^*)\} - (\rho_j - \tilde{\delta})T^*$$

$$\geq \sum_{m=1}^{N_j(0,S^*)} \min\{B_{j,m}, \frac{g_j}{g_i} (1 - \rho + \delta)\varepsilon x\} - (\rho_j - \tilde{\delta})S^* - (\rho_j - \tilde{\delta})(T^* - S^*)$$

$$= A_{j \leq \frac{g_j}{g_i}(1-\rho+\delta)\varepsilon x}(0, S^*) - (\rho_j - \tilde{\delta})S^* - (\rho_j - \tilde{\delta})(T^* - S^*)$$

$$\geq \inf_{t \geq 0} [A_{j \leq \frac{g_j}{g_i}(1-\rho+\delta)\varepsilon x}(0, t) - (\rho_j - \tilde{\delta})t] - (\rho_j - \tilde{\delta})\varepsilon x$$

$$= - \sup_{t \geq 0} [(\rho_j - \tilde{\delta})t - A_{j \leq \frac{g_j}{g_i}(1-\rho+\delta)\varepsilon x}(0, t)] - (\rho_j - \tilde{\delta})\varepsilon x$$

$$= - U^{\rho_j - \tilde{\delta}}_{j \leq \frac{g_j}{g_i}(1-\rho+\delta)\varepsilon x} - (\rho_j - \tilde{\delta})\varepsilon x$$

$$\geq - \rho_j\varepsilon x.$$  \hspace{1cm} (5)

Substituting (5) into (2), we conclude

$$B_{\text{perm}}(0, T) - (1 - \rho + \delta)T \leq \rho\varepsilon x,$$
contradicting the initial assumption. Thus, \( \mathbb{P}\{\sup_{t \geq 0} |B_{perm}(0, t) - (1 - \rho + \delta) t| \leq \epsilon x\} \geq \prod_{j=1}^{K} \mathbb{P}\{U_{j}^{\rho_{j} - \delta} t_{j} \leq (1 - \rho + \delta) \epsilon x\} \).

Observing that \( U_{j}^{\rho_{j} - \delta} t_{j} \) is a proper random variable for \( x \) sufficiently large completes the proof.

\[ \square \]

### 4.2 Upper Bound

The next proposition gives an upper bound for \( \mathbb{P}\{V_{i} > x\} \) in terms of \( \mathbb{P}\{B_{i} > (1 - \rho)x\} \).

**Proposition 4.2.1** If \( B_{i} \) and \( B \) are regularly varying of index \( \nu > 2 \), then

\[
\mathbb{P}\{V_{i} > x\} \leq \mathbb{P}\{B_{i} > (1 - \rho)x\}(1 + o(1)), \quad x \to \infty.
\]

In the proof of Proposition 4.2.1 we consider again an arbitrary class\(-i\) customer with service requirement \( B_{0} \) and sojourn time \( V_{0} \). We will divide the probability \( \mathbb{P}\{V_{0} > x\} \) into several parts, and then use sample-path techniques to prove that one of these parts asymptotically dominates all others, namely the part \( \mathbb{P}\{B_{0} > (1 - \rho)x\} \). Specifically, the proof involves conditioning on the length of the busy period before the tagged customer arrives and conditioning on the number of large customers (customers with a large service requirement, see the heuristic arguments described in Section 3) that have arrived during this time period. The length of the busy period in which the tagged customer arrives is denoted by \( P \). The part of the busy period before the tagged customer arrives, is a past busy period and is denoted by \( P^{r} \). We use the random variable \( N_{>u}(s, t) \) to denote the number of customers arriving during the time interval \((s, t)\) with service requirement strictly larger than \( u \), excluding the tagged customer.

**Proof of Proposition 4.2.1**

For every \( L \in \mathbb{N} \) and \( \alpha > 0 \),

\[
\mathbb{P}\{V_{0} > x\} = \mathbb{P}\{V_{0} > x; P^{r} \leq x^{\alpha}\} + \mathbb{P}\{V_{0} > x; P^{r} > x^{\alpha}\}
\]

\[
= \sum_{k=0}^{L} \mathbb{P}\{V_{0} > x; P^{r} \leq x^{\alpha}; N_{>ex}(-x^{\alpha}, 0) = k\} + \mathbb{P}\{V_{0} > x; P^{r} > x^{\alpha}\}
\]

\[
\leq \sum_{k=0}^{L} \mathbb{P}\{V_{0} > x; P^{r} \leq x^{\alpha}; N_{>ex}(-x^{\alpha}, 0) \geq L + 1\} + \mathbb{P}\{V_{0} > x; P^{r} > x^{\alpha}\}
\]

\[
+ \mathbb{P}\{N_{>ex}(-x^{\alpha}, 0) \geq L + 1\} + \mathbb{P}\{P^{r} > x^{\alpha}\}. \tag{6}
\]

We now consider each of the three terms in the last part of (6) separately. We first examine the term \( \mathbb{P}\{P^{r} > x^{\alpha}\} \). Since \( B \) is regularly varying of index \( \nu \), \( P \) is regularly varying of index \( \nu \) as well according to Lemma 5.3.1 in Zwart [24]. Hence, \( P^{r} \) is regularly varying of index \( \nu - 1 \). So there exists a slowly varying function \( l(\cdot) \) so that...
\[ \mathbb{P}\{P^r > x^\alpha\} = l(x)x^{\alpha(1-\nu)}. \] Take \( \alpha := -(\nu+\delta)/(1-\nu) \) and choose \( \delta > 0 \) sufficiently small so that \( \alpha < \nu \), which is possible as \( \nu > 2 \). Then,
\[ \mathbb{P}\{P^r > x^\alpha\} = l(x)x^{\alpha(1-\nu)} = l(x)x^{-(\nu+\delta)} = o(\mathbb{P}\{B_0 > x\}), \quad x \to \infty. \]

Next, we turn to the term \( \mathbb{P}\{N_{>tx}(x^\alpha,0) \geq L + 1\} \). Take \( L = \left\lceil \frac{x^{\nu+\delta}}{x^\alpha} \right\rceil \), with \( \lfloor x \rfloor \) denoting the largest integer smaller than or equal to \( x \), so that \( L + 1 > \frac{x^{\nu+\delta}}{x^\alpha} \). There exists a slowly varying function \( m(\cdot) \) such that \( \mathbb{P}\{B > x\} = m(x)x^{-\nu} \). According to Lemma 4.2.2 (which is stated below), there exist \( C > 0 \) and \( X > 1 \) so that for all \( x > X \),
\[ \mathbb{P}\{N_{>tx}(x^\alpha,0) \geq L + 1\} \leq C(x^\alpha\mathbb{P}\{B > x\})^{L+1} = C(m(x))^{L+1}x^{-(\alpha-\nu)(L+1)} \leq C(m(x))^{L+1}x^{-(\nu+\delta)}. \]

So \( \mathbb{P}\{N_{>tx}(x^\alpha,0) \geq L + 1\} = o(\mathbb{P}\{B_0 > x\}) \) as \( x \to \infty \).

Finally, we study the term \( \mathbb{P}\{V_0 > x | (P^r \preceq x^\alpha; N_{>tx}(x^\alpha,0) = k)\} \mathbb{P}\{N_{>tx}(x^\alpha,0) = k\} \). According to Lemma 4.2.3 (which is also formulated below), for any \( 0 < \delta < 1 - \rho \) there exists an \( \epsilon > 0 \) so that for all \( k \in \{1, \ldots, L\} \),
\[ \mathbb{P}\{V_0 > x | (P^r \preceq x^\alpha; N_{>tx}(x^\alpha,0) = k)\} \leq \mathbb{P}\{B_0 > \frac{1 - \rho - 3\delta}{k+1}x\}(1 + o(1)), \quad x \to \infty, \]

and using Lemma 4.2.2 and the fact that \( \alpha < \nu \), it follows that for all \( k \in \{1, \ldots, L\} \),
\[ \mathbb{P}\{N_{>tx}(x^\alpha,0) \geq k\} = o(1), \quad x \to \infty. \]

So it is readily seen that for any \( 0 < \delta < 1 - \rho \) there exists an \( \epsilon > 0 \) so that
\[ \sum_{k=0}^{L} \mathbb{P}\{V_0 > x | (P^r \preceq x^\alpha; N_{>tx}(x^\alpha,0) = k)\} \mathbb{P}\{N_{>tx}(x^\alpha,0) = k\} = \mathbb{P}\{B_0 > (1 - \rho - 3\delta)x\}(1 + o(1)), \quad x \to \infty. \]

The proof is completed by using the fact that \( B_0 \) is regularly varying.

We now present the proofs of Lemmas 4.2.2 and 4.2.3.

The next lemma shows that the probability that \( k \) large customers arrive in a time interval is related to the probability that \( k \) independently arriving customers are all large customers times the length of the time interval. This lemma is used in the proof of Proposition 4.2.1 to bound the probability that \( k \) large customers arrive during a time interval of length \( x^\alpha \).

**Lemma 4.2.2** For any \( \epsilon > 0 \), \( \alpha > 0 \) and \( k \in \mathbb{N} \),
\[ \mathbb{P}\{N_{>tx}(x^\alpha,0) \geq k\} = O((x^\alpha\mathbb{P}\{B > x\})^k), \quad x \to \infty. \]

**Proof**
We condition on the first moment that a large customer arrives in the observed time interval. For compactness, define the events:
Arr\_r(t): a customer with service requirement larger than \( u \) arrives at time \( t \);

Arr\_r(s, t): the first customer with service requirement larger than \( u \) which arrives after time \( s \) arrives at time \( t \).

Further define the random variables:

\( T_u \): the interarrival time of an arbitrary customer with service requirement larger than \( u \);

\( T^r_u \): the residual interarrival time of an arbitrary customer with service requirement larger than \( u \).

Note that

\[
P\{T^r_u \leq t\} = \frac{1}{\mathbb{E}\{T_u\}} \int_0^t \mathbb{P}\{T_u > z\} dz \leq \frac{t}{\mathbb{E}\{T_u\}} = \lambda t \mathbb{P}\{B > u\}, \quad (7)
\]

\[
P\{T_u \leq t\} = \mathbb{P}\{N_u(0, t) \geq 1\} \leq \mathbb{E}\{N_u(0, t)\} = \lambda t \mathbb{P}\{B > u\}. \quad (8)
\]

Using (7), we now obtain

\[
P\{N_u(-t, 0) \geq k\} = \int_0^t \mathbb{P}\{N_u(-t, 0) \geq k \mid \text{Arr}_u(-t, y - t)\} d\mathbb{P}\{T^r_u \leq y\}
\]

\[
= \int_0^t \mathbb{P}\{N_u(y - t, 0) \geq k - 1 \mid \text{Arr}_u(y - t)\} d\mathbb{P}\{T^r_u \leq y\}
\]

\[
\leq \mathbb{P}\{N_u(-t, 0) \geq k - 1 \mid \text{Arr}_u(-t)\} \int_0^t d\mathbb{P}\{T^r_u \leq y\}
\]

\[
= \mathbb{P}\{N_u(-t, 0) \geq k - 1 \mid \text{Arr}_u(-t)\} \mathbb{P}\{T^r_u \leq t\}
\]

\[
\leq \mathbb{P}\{N_u(-t, 0) \geq k - 1 \mid \text{Arr}_u(-t)\} \lambda t \mathbb{P}\{B > u\}.
\]

Similarly, using (8), we derive

\[
P\{N_u(-t, 0) \geq k \mid \text{Arr}_u(-t)\} = \int_0^t \mathbb{P}\{N_u(-t, 0) \geq k \mid \text{Arr}_u(-t, y - t)\} d\mathbb{P}\{T_u \leq y\}
\]

\[
= \int_0^t \mathbb{P}\{N_u(y - t, 0) \geq k - 1 \mid \text{Arr}_u(y - t)\} d\mathbb{P}\{T_u \leq y\}
\]

\[
\leq \mathbb{P}\{N_u(-t, 0) \geq k - 1 \mid \text{Arr}_u(-t)\} \lambda t \mathbb{P}\{B > u\}.
\]

Using induction, we find

\[
P\{N_u(-t, 0) \geq k\} \leq (\lambda t \mathbb{P}\{B > u\})^k.
\]

Taking \( t = x^\alpha \) and \( u = \epsilon x \), we have

\[
P\{N_{\epsilon x}(-x^\alpha, 0) \geq k\} \leq (\lambda x^\alpha \mathbb{P}\{B > \epsilon x\})^k.
\]

Because \( B \) is regularly varying, we finally conclude

\[
P\{N_{\epsilon x}(-x^\alpha, 0) \geq k\} = O((x^\alpha \mathbb{P}\{B > x\})^k), \quad x \to \infty.
\]

The next lemma implies that the tail of the sojourn time of the tagged customer is asymptotically bounded by the tail of the service requirement, up to a constant factor. The
value of the constant is determined by the number of large customers that arrive in the part of the busy period before the tagged customer arrives. The constant factor includes a coefficient \( f_{\text{DPS}} \), which depends on the specific values of the DPS weights, and is defined by \( f_{\text{DPS}} := \max_{j=1, \ldots, K} g_j / g_i \).

**Lemma 4.2.3** For any \( 0 < \delta < 1 - \rho \) and \( k \in \mathbb{N} \), there exists an \( \epsilon > 0 \) such that

\[
\mathbb{P}\{ V_0 > x \mid (P^r \leq x^\alpha; N_{>ex}(-x^\alpha, 0) = k) \} \leq \mathbb{P}\{ B_0 > \frac{1 - \rho - 3\delta}{kf_{\text{DPS}} + 1} x \} (1 + o(1)), \quad x \to \infty.
\]

**Proof**

We first prove a sample-path inequality and later translate this into a probabilistic bound. To obtain the sample-path relation, we split the customers present during (a part of) the sojourn time of the tagged customer into two groups, depending on whether they arrive before or after the tagged customer.

We first introduce some useful notation. Let \( A(s, t) \) be the total 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 the service requirements are truncated at level \( y \). For any \( c > \rho \), define \( W_{\leq c}(0, t) := A_{\leq y}(0, t) - ct \) and \( W_{\leq c} := \sup_{t \geq 0} [A_{\leq y}(0, t) - ct] \).

The random variable \( W_{\leq c}(0, t) \) represents the stationary workload in a system where work arrives according to the process \( A_{\leq y}(0, t) \) and is served at a constant rate \( c \). For any \( y > 0 \) and \( c > \rho \), \( W_{\leq c} \) is a proper random variable. Define \( W_{\leq c}(t) \) as the workload at time \( t \) associated with customers with service requirements smaller than \( ex \), excluding the tagged customer. Denote by \( B_0(s, t) \) the amount of service received by the tagged customer during the time interval \( (s, t) \), and define \( R_0(t) := B_0 - B_0(0, t) \) as the remaining service requirement of the tagged customer at time \( t \). We will also use the random variables \( B_{j,m}, B_{j,l}, T_{j,m}, N_j(t), \) and \( N_j(s, t) \) again, which are defined in the proof of Lemma 4.1.2.

The DPS discipline implies that for any \( 0 < \delta < 1 - \rho \),

\[
V_0(1 - \rho - \delta) = B_0 + \sum_{j=1}^{K} \sum_{l=1}^{N_j(0)} \min\{B_{j,l}^r, f_{\text{DPS}} B_0\} + \sum_{j=1}^{K} \sum_{m=1}^{N_j(0)} \min\{B_{j,m}, g_j / g_i, R_0(T_{j,m})\} - (\rho + \delta)V_0
\]

\[
\leq B_0 + \sum_{j=1}^{K} \sum_{l=1}^{N_j(0)} \min\{B_{j,l}^r, f_{\text{DPS}} B_0\} + \sum_{j=1}^{K} \sum_{m=1}^{N_j(0)} \min\{B_{j,m}, f_{\text{DPS}} B_0\} - (\rho + \delta)V_0
\]

\[
= B_0 + \sum_{j=1}^{K} \sum_{l=1}^{N_j(0)} \min\{B_{j,l}^r, f_{\text{DPS}} B_0\} + A_{\leq f_{\text{DPS}} B_0}(0, V_0) - (\rho + \delta)V_0
\]

\[
\leq B_0 + \sum_{j=1}^{K} \sum_{l=1}^{N_j(0)} \min\{B_{j,l}^r, f_{\text{DPS}} B_0\} + \sup_{t \geq 0} [A_{\leq f_{\text{DPS}} B_0}(0, t) - (\rho + \delta)t]
\]

\[
\leq B_0 + \sum_{j=1}^{K} \sum_{l=1}^{N_j(0)} \min\{B_{j,l}^r, f_{\text{DPS}} B_0\} + W_{\leq f_{\text{DPS}} B_0}^{\rho + \delta}.
\]
If at most \( k \) of the \( \sum_{j=1}^{K} N_j(0) \) customers present just before time \( t = 0 \) have a service requirement larger than \( cx \), then

\[
B_0 + \sum_{j=1}^{K} \sum_{l=1}^{N_j(0)} \min\{B^{r}_{j,l}, f_{DPS}B_0\} + W^\rho_{\leq f_{DPS}B_0} \leq (kf_{DPS} + 1)B_0 + W_{\leq cx}(0) + W^\rho_{\leq f_{DPS}B_0}.
\]

For compactness, define

\[
R := \{P^r \leq x^a; N_{> cx}(-x^a, 0) = k\},
\]

\[
D(x) := (kf_{DPS} + 1)B_0 + (\rho + \delta)x + W_{\leq cx}(0) + W^\rho_{\leq f_{DPS}B_0}.
\]

The probability of interest may then be written as

\[
\mathbb{P}\{V_0 > x \mid (P^r \leq x^a; N_{> cx}(-x^a, 0) = k)\}
= \mathbb{P}\{V_0(1 - \rho - \delta) > x(1 - \rho - \delta) \mid R\}
\leq \mathbb{P}\{(kf_{DPS} + 1)B_0 + (\rho + \delta)x + W_{\leq cx}(0) + W^\rho_{\leq f_{DPS}B_0} > x \mid R\}
= \mathbb{P}\{D(x) > x \mid R\}
= \mathbb{P}\{D(x) > x; W_{\leq cx}(0) \leq \delta x; W^\rho_{\leq f_{DPS}B_0} \leq \delta x \mid R\}
+ \mathbb{P}\{D(x) > x; W_{\leq cx}(0) > \delta x; W^\rho_{\leq f_{DPS}B_0} \leq \delta x \mid R\}
+ \mathbb{P}\{D(x) > x; W^\rho_{\leq f_{DPS}B_0} > \delta x \mid R\}
\leq \mathbb{P}\{D(x) > x; W_{\leq cx}(0) \leq \delta x; W^\rho_{\leq f_{DPS}B_0} \leq \delta x \mid R\}
+ \mathbb{P}\{W_{\leq cx}(0) > \delta x \mid R\} + \mathbb{P}\{W^\rho_{\leq f_{DPS}B_0} > \delta x\}.
\]

Using Lemmas 4.2.5 and 4.2.6, we obtain that the last two terms are both \( o(\mathbb{P}\{B_0 > x\}) \) as \( x \to \infty \).

The first probability may be rewritten as

\[
\mathbb{P}\{(kf_{DPS} + 1)B_0 + (\rho + \delta)x > x - W_{\leq cx}(0) - W^\rho_{\leq f_{DPS}B_0}; W_{\leq cx}(0) \leq \delta x; W^\rho_{\leq f_{DPS}B_0} \leq \delta x \mid R\}
\leq \mathbb{P}\{(kf_{DPS} + 1)B_0 + (\rho + \delta)x > (1 - 2\delta)x \mid R\}
= \mathbb{P}\{B_0 > \frac{(1 - \rho - 3\delta)}{kf_{DPS} + 1}x \mid R\}
= \mathbb{P}\{B_0 > \frac{(1 - \rho - 3\delta)}{kf_{DPS} + 1}x\}.
\]

The proof is completed by observing that \( B_0 \) is regularly varying.

The next lemma is needed to prove Lemmas 4.2.5 and 4.2.6. The random variable \( W^\nu_{\leq y}(t) \) used in the lemma represents the workload at time \( t \) in a system where work arrives according to the process \( A_{\leq y}(0, t) \) and is served at a constant rate \( c \).

**Lemma 4.2.4** If \( B \) is regularly varying of index \( \nu > 1 \) and \( t_a \) an arbitrary arrival epoch, then for any \( c > \rho \) and \( n \in \mathbb{N} \),

\[
\mathbb{P}\{W^\nu_{\leq x}(t_a) > (n + 1)x\} = O\left(x^{-n(\nu - 1)}\right), \quad x \to \infty.
\]
Proof
We bound the interarrival times by $M$. If we choose $M$ sufficiently large, then the system is still 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 $W_{\leq y,M}^c := \sup_{t \geq 0} [A_{\leq y,M}(0,t) - ct]$. The random variable $W_{\leq y,M}^c$ represents the stationary workload in a system where work arrives according to the process $A_{\leq y,M}(0,t)$ and is served at a constant rate $c$. For any $y > 0$ and $c > \rho$, there exists a sufficiently large $M$ so that $W_{\leq y,M}^c$ is a proper random variable. Theorem 2 in Jelenković [11] implies

$$
P\{W_{\leq y,M}^c(t_a) > (n + 1)x\} = O\left(x^{-n(u-1)}\right), \quad x \to \infty.
$$

Because $W_{\leq y,M}^c(t)$ is stochastically larger than $W_{\leq x}^c(t)$, the lemma follows. \hfill $\Box$

The next lemma indicates that the workload at time 0 associated with small customers is small compared to the service requirement of the tagged customer, as long as the number of large customers that arrive in the part of the busy period before time 0 is limited.

Lemma 4.2.5 For any $0 < \delta < 1 - \rho$ and $k \in \mathbb{N}$, there exists an $\epsilon > 0$ such that

$$
P\{W_{\leq \epsilon x}(0) > \delta x \mid (P^r \leq x^\alpha; N_{> \epsilon x}(-x^\alpha,0) = k)\} = o(P\{B_0 > x\}, \quad x \to \infty.
$$

Proof
For conciseness, we call customers with a service requirement smaller than $\epsilon x$ ‘small’ customers, and customers with service requirement larger than $\epsilon x$ ‘large’ customers (see also the heuristic arguments described in Section 3). The DPS discipline implies that, the larger the number of small customers present in the system, the larger their (combined) service rate. Denote by $N_k := \lceil k^{1+\rho/2} \max_{i=1}^{\min} \sum_{i=1}^{\min} r_i \rceil$ the smallest number of customers needed to ensure a total service rate of at least $(1 + \rho)/2$ for these customers in the presence of $k$ arbitrary other customers, with $[x]$ 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 $u$, excluding the tagged customer. Define $t^* := \sup\{t \in [-P^r, 0) : N_{\leq \epsilon x}(t) \leq N_k - 1\}$. By definition, the number of small customers in the system is constantly larger than or equal to $N_k$ during the time interval $[t^*, 0]$. Thus, during $[t^*, 0]$ the small customers will constantly receive service at a rate larger than or equal to $(1 + \rho)/2$. We deduce

$$
W_{\leq \epsilon x}(0) = W_{\leq \epsilon x}(0^-) \leq N_k \epsilon x + \int_{t^*}^{0^-} dW_{\leq \epsilon x}(t)
\leq N_k \epsilon x + \int_{t^*}^{0^-} dW_{\leq \epsilon x}^{(1+\rho)/2}(t)
\leq N_k \epsilon x + W_{\leq \epsilon x}^{(1+\rho)/2}(0^-).
$$

Since $t = 0$ is an arbitrary arrival epoch, we can use Lemma 4.2.4 to choose $\epsilon > 0$ sufficiently small so that

$$
P\{W_{\leq \epsilon x}(0) > \delta x \mid (P^r \leq x^\alpha; N_{> \epsilon x}(-x^\alpha,0) = k)\} \leq P\{W_{\leq \epsilon x}^{(1+\rho)/2}(0^-) > (\delta - N_k \epsilon)x\}
= o(P\{B_0 > x\}, \quad x \to \infty.
$$

$\Box$
The next lemma pertains to a system where the service requirements are truncated by the service requirement of the tagged customer, and implies that the stationary workload in such a system is still small compared to the service requirement of the tagged customer.

**Lemma 4.2.6** For any $0 < \delta < 1 - \rho$ and $c > 0$,

$$P\{W_{\leq cB_0}^\rho + \delta > \delta x\} = o(P\{B_0 > x\}), \quad x \to \infty.$$  

**Proof**  
For any $d > \rho$, define $W_{\leq cB_0}^{\rho + \delta} := \sup_{t \geq 0}[A(0,t) - dt]$. Then,

$$P\{W_{\leq cB_0}^{\rho + \delta} > \delta x\} = P\{W_{\leq cB_0}^{\rho + \delta} > \delta x; B_0 \leq \gamma x\} + P\{W_{\leq cB_0}^{\rho + \delta} > \delta x; B_0 > \gamma x\}$$

$$\leq P\{W_{\leq cB_0}^{\rho + \delta} > \delta x\} + P\{W_{\leq cB_0}^{\rho + \delta} > \delta x\}P\{B_0 > \gamma x\}.$$ 

Let $t_a$ be an arbitrary arrival epoch and $B_{c,y}^c$, a remaining service requirement in a system where the service requirements are truncated at level $y$. Then,

$$P\{W_{\leq cB_0}^{\rho + \delta} > \delta x\} = P\{W_{\leq cB_0}^{\rho + \delta} > 0\}P\{W_{\leq cB_0}^{\rho + \delta}(t_a) + B_{c,y}^c > \delta x\}$$

$$< P\{W_{\leq cB_0}^{\rho + \delta}(t_a) + c\gamma x > \delta x\}$$

$$= P\{W_{\leq cB_0}^{\rho + \delta}(t_a) > (\delta - c\gamma)x\}$$

Using Lemma 4.2.4, we may choose $\gamma > 0$ sufficiently small so that

$$P\{W_{\leq cB_0}^{\rho + \delta}(t_a) > (\delta - c\gamma)x\} = o(P\{B_0 > x\}), \quad x \to \infty.$$ 

The proof is completed by using the fact that $B_0$ is regularly varying. 

\[\square\]

## 5 Time-varying service rate

In this section we extend the results to the case where the service rate is time-varying. As observed in the introduction, models with varying service rate play a crucial role in modelling systems with service integration. Let $C(s,t)$ be the available service capacity during the time interval $(s,t)$, and let $c > \rho$ be the mean available service rate.

**Assumption 5.1** We assume that for all $\psi > 0$ and $\phi > 0$:

(i) $\lim_{x \to \infty} P\{\sup_{t \geq 0}[C(0,t) - (c + \psi)t] > \phi x\} = 0$;

(ii) $P\{\sup_{t \geq 0}[(c - \psi)t - C(0,t)] > \phi x\} = o(P\{B_0 > x\})$ as $x \to \infty$;

(iii) $P\{\sup_{t \geq 0}[(c - \psi)(-t) - C(-t,0)] > \phi x\} = o(P\{B_0 > x\})$ as $x \to \infty$.

Since $B_0$ is regularly varying, observe that the above assumption holds for all values of $\phi > 0$ whenever it holds for one such value.

The next theorem is an extension of the main theorem to time-varying service rates.
Theorem 5.2 If $B_i$ and $B$ are both regularly varying of index $\nu > 2$, and the service rate process $C(s, t)$ satisfies Assumption 5.1, then
\[
P\{V_0^{var} > x\} \sim P\{B_0 > (c-\rho)x\}, \quad x \to \infty.
\]

Before describing the various modifications to the proofs, we first consider the important special case where the service rate alternates between two values, a ‘low’ speed $c_0$, and a ‘high’ speed $c_1 \geq c_0$. We specifically assume that the service rate is regulated by a stationary alternating 0-1 renewal process $I(u)$, $u \geq 0$, i.e., assume that
\[
C(s, t) = c_0 t + (c_1 - c_0) \int_s^t I(u) du.
\]
The lengths of the 0- and 1-intervals of $I(u)$ are denoted by the generic random variables $T_0$ and $T_1$. The mean available service rate is $c = (1 - p)c_0 + pc_1$, with $p = P\{I(u) = 1\} = E(T_1) / (E(T_0) + E(T_1))$.

The next corollary gives a condition on the distribution of $T_0$ under which Assumption 5.1 is satisfied.

Corollary 5.3 Assume that $C(0, t)$ is regulated by an alternating renewal process as described above with $c > \rho$, and assume that $T_0$ is regularly varying with
\[
P\{T_0 > x\} = o(x^{-1}P\{B_0 > x\}), \quad x \to \infty.
\]
Then Assumption 5.1 is satisfied.

Proof

The equation
\[
\sup_{t \geq 0} [C(0, t) - (c + \psi)t] = \sup_{t \geq 0} [A_1(0, t) - (c - c_0 + \psi)t],
\]
with $A_1(0, t) := C(0, t) - c_0 t$, indicates that the quantity $\sup_{t \geq 0} [C(0, t) - (c + \psi)t]$ represents the stationary workload in a fluid queue of drain rate $c - c_0 + \psi$ fed by an On-Off process with generic On-period $T_1$, generic Off-period $T_0$, and peak rate $c_1 - c_0$. Part (i) of Assumption 5.1 then follows by observing that the traffic intensity of the On-Off process is given by $p(c_1 - c_0) = c_1 - c$, which is smaller than the service rate $c - c_0 + \psi$.

The equation
\[
\sup_{t \geq 0} [(c - \psi)t - C(0, t)] = \sup_{t \geq 0} [A_2(0, t) - (c_1 - c + \psi)t],
\]
with $A_2(0, t) = c_1 t - C(0, t)$, shows that the variable $\sup_{t \geq 0} [(c - \psi)t - C(0, t)]$ may be interpreted as the stationary workload in a fluid queue of drain rate $c_1 - c + \psi$ fed by an On-Off process with generic On-period $T_0$, generic Off-period $T_1$, and peak rate $c_1 - c_0$. Observe that the traffic intensity of the On-Off process is given by $(1 - p)(c_1 - c_0) = c_1 - c$, so that the queue is stable. Part (ii) of Assumption 5.1 then follows from the asymptotic behaviour of the workload distribution for a single On-Off process in Jelenković & Lazar [12].
Part (iii) of Assumption 5.1 may be shown in a similar fashion using reversibility of the On-Off process.

We now move to an outline of the proof of Theorem 5.2. The proof of this theorem shows strong overlap with the proof of Theorem 3.1. Only a few minor modifications are needed, which are described below. The notation is also largely inherited from that used before, and we refer to the original proofs for the introduction of most of the notation. In order to distinguish the random variables in the model with varying service rate from those in the model with fixed service rate, we add the superscript \( \text{var} \) to the corresponding symbols. When a random variable is independent of the service rate, the superscript \( \text{var} \) is omitted.

### 5.1 Modifications to the proof of the lower bound

The next proposition is the analogue of Proposition 4.1.1.

**Proposition 5.1.1** If \( B_i \) is regularly varying and the service rate process \( C(s, t) \) satisfies part (i) of Assumption 5.1, then

\[
P\{V^\text{var}_i > x\} \geq P\{B_i > (c - \rho)x\}(1 + o(1)), \quad x \to \infty.
\]

**Proof**

The proof is similar to that of Proposition 4.1.1. Consider an arbitrary class-\( i \) customer with service requirement \( B_0 \) and sojourn time \( V_0 \). Then, for any \( \delta, \epsilon, \phi, \psi > 0 \),

\[
P\{V_0 > x\} = P\{B^\text{perm}_{\text{var}}(0, x) < B_0\} \geq P\{B_0 > (c - \rho + \delta + \psi + \epsilon + \phi)x\}P\{B^\text{perm}_{\text{var}}(0, x) \leq (c - \rho + \delta + \psi + \epsilon + \phi)x\}.
\]

Using Lemma 5.1.2 below, we obtain that the latter probability converges to 1 when \( x \) tends to \( \infty \). Letting \( \delta, \epsilon, \phi, \psi \downarrow 0 \) and using the fact that \( B_0 \) is regularly varying then completes the proof.

**Lemma 5.1.2** If the service rate process satisfies part (i) of Assumption 5.1, then for any \( \delta > 0, \epsilon > 0, \psi > 0, \phi > 0 \),

\[
\lim_{x \to \infty} P\{\sup_{t \geq 0} [B^\text{perm}_{\text{var}}(0, t) - (c - \rho + \delta + \psi)t] \leq (\epsilon + \phi)x\} = 1.
\]

**Proof**

The proof involves a few minor modifications to that of Lemma 4.1.2. We will show that \( \sup_{t \geq 0} [C(0, t) - (c + \psi)t] \leq \phi x \) and

\[
U_{j \geq 0}^{\rho_j - \delta_j, \frac{\psi_j}{\rho_j}} (c - \rho + \delta + \psi)x < \delta \epsilon x
\]

for all \( j = 1, \ldots, K \) implies that \( \sup_{t \geq 0} [B^\text{perm}_{\text{var}}(0, t) - (c - \rho + \delta + \psi)t] \leq (\epsilon + \phi)x \).

Suppose that were not the case, i.e., there exists a \( T \geq 0 \) such that \( B^\text{perm}_{\text{var}}(0, T) - (c - \rho + \delta + \psi)T > (\epsilon + \phi)x \).

Instead of (1), the following evident equality applies

\[
B^\text{perm}_{\text{var}}(s, t) + \sum_{j=1}^{K} B^\text{var}_j(s, t) = C(s, t)
\]
for all $t \geq s \geq 0$.

Redefine $T^* := \inf \{ t \geq 0 : B^\var_{\text{perm}}(t, T) \leq (c - \rho + \delta + \psi)(T - t) \}$.

Similar to (2), we obtain
\[
B^\var_{\text{perm}}(0, T) - (c - \rho + \delta + \psi)T^* \leq B^\var_{\text{perm}}(0, T^*) - (c - \rho + \delta + \psi)T^*
\]
\[
= C(0, T^*) - \sum_{j=1}^{K} B^\var_j(0, T^*) - (c - \rho + \delta + \psi)T^*
\]
\[
= C(0, T^*) - (c + \psi)T^* + (\rho - \delta)T^* - \sum_{j=1}^{K} B^\var_j(0, T^*)
\]
\[
\leq \sum_{j=1}^{K} [(\rho_j - \tilde{\delta})T^* - B^\var_j(0, T^*)] + \phi x. \tag{9}
\]

We now proceed to derive a lower bound for each of the terms in square brackets.

The inequality in (3) remains unchanged.

As before, we find
\[
(c + \psi)T^* + \phi x \geq C(0, T^*)
\]
\[
\geq B^\var_{\text{perm}}(0, T^*)
\]
\[
= B^\var_{\text{perm}}(0, T) - B^\var_{\text{perm}}(T^*, T)
\]
\[
> (c - \rho + \delta + \psi)T + (\epsilon + \phi)x - (c - \rho + \delta + \psi)(T - T^*)
\]
\[
= (c - \rho + \delta + \psi)T^* + (\epsilon + \phi)x,
\]

which means $(\rho - \delta)T^* > \epsilon x$, so that $S^* := T^* - \epsilon x > 0$.

By definition of $T^*$, we have that $B^\var_{\text{perm}}(s, T) > (c - \rho + \delta + \psi)(T - s)$ for all $s \in [0, T^*)$, and thus
\[
B^\var_{\text{perm}}(s, T^*) = B^\var_{\text{perm}}(s, T) - B^\var_{\text{perm}}(T^*, T)
\]
\[
> (c - \rho + \delta + \psi)(T - s) - (c - \rho + \delta + \psi)(T - T^*)
\]
\[
= (c - \rho + \delta + \psi)(T^* - s).
\]

Combining the above two inequalities, taking $s = S^*$, we derive as before in equality (5)
\[
B^\var_j(0, T^*) - (\rho_j - \tilde{\delta})T^* \geq \sum_{m=1}^{N_j(0, S^*)} \min \{ \beta_j, \frac{\alpha_j}{g_k} (c - \rho + \delta + \psi)\epsilon x \} - (\rho_j - \tilde{\delta})S^* - (\rho_j - \tilde{\delta})(T^* - S^*)
\]
\[
= A_j \leq \frac{\alpha_j}{g_k} (c - \rho + \delta + \psi)\epsilon x (0, S^*) - (\rho_j - \tilde{\delta})S^* - (\rho_j - \tilde{\delta})\epsilon x
\]
\[
> -\rho_j \epsilon x. \tag{10}
\]

Substituting (10) into (9), we conclude
\[
B^\var_{\text{perm}}(0, T) - (c - \rho + \delta + \psi)T \leq (\rho \epsilon + \phi)x,
\]
contradicting the initial assumption.
Thus,
\[
\mathbb{P}\{\sup_{t \geq 0} [B_{\text{perm}}^\text{var}(0, t) - (c - \rho + \delta + \psi)t] \leq (\epsilon + \phi)x\} \\
\geq \mathbb{P}\{\sup_{t \geq 0} [C(0, t) - (c + \psi)t] \leq \phi x\} \prod_{j=1}^{K} \mathbb{P}\left\{U^\rho_{\frac{2\epsilon}{\delta}} \frac{j}{e}(c - \rho + \psi)x \leq \frac{\tilde{\epsilon}x}{x}\right\}.
\]

Using part (i) of Assumption 5.1 and observing that \(U^\rho_{\frac{2\epsilon}{\delta}} \frac{j}{e}(c - \rho + \psi)x\) is a proper random variable for \(x\) sufficiently large completes the proof. 

\[\square\]

### 5.2 Modifications to the proof of the upper bound

The next proposition is the analogue of Proposition 4.2.1.

**Proposition 5.2.1** If \(B_i\) and \(B\) are regularly varying of index \(\nu > 2\) and the service rate process \(C(s, t)\) satisfies parts (ii) and (iii) of Assumption 5.1, then
\[
\mathbb{P}\{V_{0}^\text{var} > x\} \leq \mathbb{P}\{B_0 > (c - \rho)x\}(1 + o(1)), \quad x \to \infty.
\]

**Proof**

Similar to the proof of Proposition 4.2.1, we divide the probability \(\mathbb{P}\{V_0 > x\}\) into several parts to obtain
\[
\mathbb{P}\{V_{0}^\text{var} > x\} \leq \sum_{k=0}^{L} \mathbb{P}\{V_{0}^\text{var} > x \mid ((P_{\text{var}})^r \leq x^\alpha; N_{\text{ex}}(-x^\alpha, 0) = k)\}\mathbb{P}\{N_{\text{ex}}(-x^\alpha, 0) = k\} \\
+ \mathbb{P}\{N_{\text{ex}}(-x^\alpha, 0) \geq L + 1\} + \mathbb{P}\{(P_{\text{var}})^r > x^\alpha\}.
\]

However, the analysis of some of these terms is different. The first term that is handled differently is the probability \(\mathbb{P}\{(P_{\text{var}})^r > x^\alpha\}\). Write \(\mathbb{P}\{(P_{\text{var}})^r > x^\alpha\} = I + II\), with
\[
I = \mathbb{P}\{(P_{\text{var}})^r > x; \sup_{t \geq 0} [(c - \psi)t - C(-t, 0)] \leq \phi x\}, \\
II = \mathbb{P}\{(P_{\text{var}})^r > x; \sup_{t \geq 0} [(c - \psi)t - C(-t, 0)] > \phi x\}.
\]

Clearly, part (iii) of Assumption 5.1 shows that term II is \(o(\mathbb{P}\{B_0 > x\})\) as \(x \to \infty\).

In order to bound term I, recall that \(W^d\) represents the stationary workload in a system where work arrives according to the process \(A(s, t)\) and is served at a constant rate \(d\). Term I may be compared to \(\mathbb{P}\{W^d > x\}\) using the following sample-path derivation.

Assume \((P_{\text{var}})^r > x\) and \(\sup_{t \geq 0} [(c - \psi)t - C(-t, 0)] \leq \phi x\). Then there exists a \(S > x\) such that \(A(-S, 0) - C(-S, 0) > 0\) and \(C(-S, 0) \geq (c - \psi)S - \phi x\). Combining these two inequalities, we obtain \(A(-S, 0) - (c - \psi)S > -\phi x\). Since \(S > x\), we have \(A(-S, 0) - (c - 2\psi)S > (\psi - \phi)x\). Hence, for all \(\psi > 0\) and \(\phi > 0\),
\[
W^{c-2\psi} = \sup_{s \geq 0} [A(-s, 0) - (c - 2\psi)s] > (\psi - \phi)x.
\]
Taking \( \phi = \psi/2 \) and \( \psi = (c - \rho)/4 \), we obtain
\[
W^{(c+\rho)/2} > \frac{c - \rho}{8} x.
\]
So we arrive at the following probabilistic bound,
\[
\mathbb{P}\{(P^{\var})^r > x; \sup_{t \geq 0}[(c - \psi)t - C(-t, 0)] \leq \phi x\} \leq \mathbb{P}\{W^{(c+\rho)/2} > \frac{c - \rho}{8} x\}.
\]

According to Pakes [16], \( \mathbb{P}\{W^d > x\} \sim \frac{d}{x^d} \mathbb{P}\{B^r > x\} \) for all \( d > \rho \). Since \( B \) is regularly varying of index \( \nu \), it follows that \( W^{(c+\rho)/2} \) is regularly varying of index \( \nu - 1 \). Taking \( \alpha = -(\nu + \delta)/(1 - \nu) \), we deduce in a similar fashion as for \( P^r \) in Proposition 4.2.1 that
\[
I = \mathbb{P}\{(P^{\var})^r > x^\alpha; \sup_{t \geq 0}[(c - \psi)t - C(-t, 0)] \leq \phi x\} = o(\mathbb{P}\{B_0 > x\}), \quad x \to \infty.
\]
The only other term in (11) that is treated differently is the probability \( \mathbb{P}\{V^{\var} > x\} \) \( ((P^{\var})^r \leq x^\alpha; N_{>\varepsilon x}(-x^\alpha, 0) = k) \). The sample-path inequality that is derived in the proof of Lemma 4.2.3 is slightly different in case of a varying service rate, and changes as follows
\[
C(0, V_0^{\var}) - (\rho + \delta)V_0^{\var} \leq B_0 + \sum_{j=1}^K \sum_{l=1}^{N^{\var}(0)} \min\{(B_j^{\var})^r, \delta_{\var}B_0\} + W^{\rho+\delta}_{\var} \leq f_{\var}B_0. \tag{12}
\]
The conditions \( (P^{\var})^r \leq x^\alpha \) and \( N_{>\varepsilon x}(-x^\alpha, 0) = k \) imply that the number of customers present at time 0 with a service requirement larger than \( \varepsilon x \) is at most \( k \). Under these conditions, it follows that
\[
B_0 + \sum_{j=1}^K \sum_{l=1}^{N^{\var}(0)} \min\{(B_j^{\var})^r, \delta_{\var}B_0\} + W^{\rho+\delta}_{\var} \leq (f_{\var} + 1)B_0 + W^{\var}_{\varepsilon} + W^{\var}_{f_{\var}}. \tag{13}
\]
For compactness, define
\[
R^{\var} = \{(P^{\var})^r \leq x^\alpha; N_{>\varepsilon x}(-x^\alpha, 0) = k\},
\]
\[
F = [C(0, V_0^{\var}) - (\rho + \delta)V_0^{\var}] + [(c - \psi)V_0^{\var} - C(0, V_0^{\var})].
\]
We may then write
\[
\mathbb{P}\{V_0^{\var} > x \mid R^{\var}\} = \mathbb{P}\{V_0^{\var}(c - \psi - \rho - \delta) > x(c - \psi - \rho - \delta) \mid R^{\var}\} = \mathbb{P}\{F > x(c - \psi - \rho - \delta) \mid R^{\var}\} = \hat{I} + \hat{\Pi},
\]
with
\[
\hat{I} = \mathbb{P}\{F > x(c - \psi - \rho - \delta); \sup_{t \geq 0}[(c - \psi)t - C(0, t)] \leq \phi x \mid R^{\var}\},
\]
\[
\hat{\Pi} = \mathbb{P}\{F > x(c - \psi - \rho - \delta); \sup_{t \geq 0}[(c - \psi)t - C(0, t)] > \phi x \mid R^{\var}\}.
\]
Clearly, part (ii) of Assumption 5.1 gives that term $\hat{I}$ is $o(\mathbb{P}\{B_0 > x\})$ as $x \to \infty$. Applying the inequalities (12) and (13), we derive

$$\hat{I} \leq \mathbb{P}\{F > x(c - \psi - \rho - \delta); (c - \psi)V_{0}^{\text{var}} - C(0, V_{0}^{\text{var}}) \leq \phi x \mid R^{\text{var}}\}$$

$$\leq \mathbb{P}\{(kf_{DPS} + 1)B_0 + W_{\leq \varepsilon x}^{\text{var}}(0) + W_{\leq f_{DPS}B_0}^{\text{var}} + \phi x > x(c - \psi - \rho - \delta) \mid R^{\text{var}}\}.$$ 

Once it has been shown that $\mathbb{P}\{W_{\leq \varepsilon x}^{\text{var}}(0) > \delta x \mid R^{\text{var}}\} = o(\mathbb{P}\{B_0 > x\})$ as $x \to \infty$, the remainder of the proof of Lemma 4.2.3 stays the same, and we obtain

$$\hat{I} \leq \mathbb{P}\{B_0 > \frac{c - \rho - 3\delta - \phi - \psi}{kf_{DPS} + 1}x\} + o(\mathbb{P}\{B_0 > x\}), \quad x \to \infty.$$ 

So it only remains to be shown that $\mathbb{P}\{W_{\leq \varepsilon x}^{\text{var}}(0) > \delta x \mid R^{\text{var}}\} = o(\mathbb{P}\{B_0 > x\})$ as $x \to \infty$. Lemma 4.2.5 indicated that $\mathbb{P}\{W_{\leq \varepsilon x}(0) > \delta x \mid (P^r \leq x^\alpha; N_{> \varepsilon x}(-x^\alpha, 0) = k)\} = o(\mathbb{P}\{B_0 > x\})$ and is served at a rate $dC(0, t)$. In this system, customers receive a fraction $d$ of the varying service rate $C(0, t)$. Denote by $W_{\leq \varepsilon x}^{d, \text{var}}(t)$ the stationary workload in a system where work arrives according to the process $A_{\leq y}(0, t)$ and is served at a rate $dC(0, t).$ Using the same principle as in Lemma 4.2.5, we obtain for some large $N \in \mathbb{N}$ and some small $\omega > 0$ that $W_{\leq \varepsilon x}^{\text{var}}(0)$ is stochastically smaller than $N_{\leq x} + W_{\leq \varepsilon x}^{(1-\omega), \text{var}}$. We may then write

$$\mathbb{P}\{W_{\leq \varepsilon x}^{\text{var}}(0) > \delta x\} \leq \mathbb{P}\{W_{\leq \varepsilon x}^{(1-\omega), \text{var}} > (\delta - N\epsilon)x\} = \hat{I} + \tilde{I},$$

with

$$\hat{I} = \mathbb{P}\{W_{\leq \varepsilon x}^{(1-\omega), \text{var}} > (\delta - N\epsilon)x; \sup_{t \geq 0}[(c - \psi)t - C(-t, 0)] > \phi x\},$$

$$\tilde{I} = \mathbb{P}\{W_{\leq \varepsilon x}^{(1-\omega), \text{var}} > (\delta - N\epsilon)x; \sup_{t \geq 0}[(c - \psi)t - C(-t, 0)] \leq \phi x\}.$$ 

Clearly, part (iii) of Assumption 5.1 means that term $\tilde{I}$ is $o(\mathbb{P}\{B_0 > x\})$ as $x \to \infty$. Further observe that

$$\tilde{I} \leq \mathbb{P}\{W_{\leq \varepsilon x}^{(1-\omega), \text{var}} > (\delta - N\epsilon)x; C(-(P^{\text{var}})^r, 0) \geq (c - \psi)(P^{\text{var}})^r - \phi x\}$$

$$\leq \mathbb{P}\{W_{\leq \varepsilon x}^{(1-\omega)(c - \psi)} > (\delta - N\epsilon - \phi)x\}.$$ 

Since $\omega$ and $\psi$ may be chosen in such a way that $(1 - \omega)(c - \psi) > \rho$, a simple application of Lemma 4.2.4 completes the proof. \hfill \Box

References


