Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

The asymptotic workload behavior of two coupled queues

S.C. Borst, O.J. Boxma, M.J.G. van Uitert

REPORT PNA-R0202 JANUARY 31, 2002

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

The Asymptotic Workload Behavior of Two Coupled Queues

Sem Borst[*,**,†], Onno Boxma[*,**], Miranda van Uitert[*]

[*] *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 consider a system of two coupled queues, Q_1 and Q_2. When both queues are backlogged, they are each served at unit rate. However, when one queue empties, the service rate at the other queue increases. Thus, the two queues are coupled through the mechanism for dynamically sharing surplus service capacity.

We derive the asymptotic workload behavior at Q_1 for various scenarios where at least one of the two queues has a heavy-tailed service time distribution. First of all, we consider a situation where the traffic load at Q_1 is *below* the nominal unit service rate. We show that if the service time distribution at Q_1 is *heavy-tailed*, then the workload behaves exactly as if Q_1 is served in isolation at a *constant* rate, which only depends on the service time distribution at Q_2 through its mean. In addition, we establish that if the service time distribution at Q_1 is *exponential*, then the workload distribution is either exponential or semi-exponential, depending on whether the traffic load at Q_2 exceeds the nominal service rate or not. Next, we focus on a regime where the traffic load at Q_1 *exceeds* the nominal service rate, so that Q_1 relies on the surplus capacity from Q_2 to maintain stability. In that case, the workload distribution at Q_1 is determined by the *heaviest* of the two service time distributions, so that Q_1 may inherit potentially heavier-tailed characteristics from Q_2.

2000 Mathematics Subject Classification: 60K25 (primary), 68M20, 90B18, 90B22 (secondary).

Keywords and Phrases: coupled processors, Generalized Processor Sharing (GPS), heavy-tailed traffic, light-tailed traffic, regular variation, workload asymptotics.

Note: Work carried out under the project PNA2 "Advanced Communication Networks".

1 Introduction

Extensive measurements have shown that traffic in high-speed communication networks may exhibit extreme variability and heavy-tailed characteristics. The bursty traffic behavior may extend over a wide range of time scales, and manifest itself in long-range dependence and self-similarity. These findings have triggered a strong interest in queueing models with heavy-tailed traffic processes [24].

Although the presence of heavy-tailed traffic characteristics is widely acknowledged, the practical implications for network performance and traffic engineering remain controversial. Particularly relevant issues in assessing the performance impact include the buffer size, the role of flow control mechanisms, and the effect of scheduling algorithms [1, 2, 13, 20].

Discriminatory scheduling algorithms play a crucial role in achieving differentiated Quality-of-Service in integrated networks, which support a wide variety of services, like voice, video and data applications. Streaming applications such as voice and video induce far more stringent Quality-of-Service requirements than the data applications currently accounting for the bulk of the Internet traffic. In view of the real-time requirements, it is desirable that delay-sensitive services receive some sort of priority over data applications, at least over short time scales. Strict priority scheduling may however not be ideal, since it may lead to starvation of the low-priority traffic. Generalized Processor Sharing (GPS)-based disciplines provide a suitable alternative, with strict priority scheduling as an extreme option [22], [23]. In GPS-based disciplines, such as Weighted Fair Queueing (WFQ), the link capacity is shared in proportion to certain class-defined weight factors.

In the present paper we examine the potential role of GPS-based scheduling disciplines in isolating or protecting traffic classes, in particular in the presence of heavy-tailed characteristics. We consider a system with two heterogeneous traffic classes, each with their own queue. The service capacity is dynamically shared in a GPS fashion. When both classes are backlogged, the two corresponding queues are each served at unit rate. However, the service rate for class 1 increases to r_1 when the queue of class 2 empties, and that for class 2 increases to r_2 when the queue of class 1 empties.

The present paper fits into two strands of research: (i) GPS queues with heavy-tailed traffic characteristics; (ii) the interplay between light-tailed and heavy-tailed traffic processes. We now proceed to discuss both aspects.

(i) A special case of the above so-called coupled-processors model has been studied in [4]. There it is assumed that both classes have heavy-tailed characteristics and the value of r_2 is taken to be 1, so that Q_2 is not influenced by Q_1. The results show a stark contrast in the workload behavior at Q_1, depending on whether the offered traffic at Q_1 exceeds the nominal unit service rate or not. In case the traffic load is below the nominal service rate, the workload at Q_1 behaves exactly as if Q_1 is served in isolation at a *constant* rate, which only depends on the service time distribution at Q_2 through its mean. In the opposite

case, the workload distribution at Q_1 is determined by the *heaviest* of the two service time distributions, so that Q_1 may inherit potentially heavier-tailed characteristics from Q_2.
In the present paper, we generalize the results from [4] in several ways. First of all, we remove the assumption that $r_2 = 1$, so that Q_2 is now also influenced by Q_1. Second, we extend the results to the case where one of the two classes may have a general service time distribution. We use an approach based on transform methods, building on the detailed analysis of the joint workload distribution in a coupled-processors model presented in [16]. Qualitatively similar results have been obtained in [5, 18, 19] using asymptotic lower and upper bounds.
(ii) As in [6], we also consider the case where the service time distribution at Q_1 is exponential, while the service time distribution at Q_2 is heavy-tailed. In case the traffic load at Q_1 is below the nominal service rate, the workload distribution at Q_1 is either exponential or semi-exponential, depending on whether the traffic load at Q_2 exceeds its nominal service rate or not. Partially related results were obtained in [7] using probabilistic methods. In contrast, if the traffic load at Q_1 exceeds the nominal service rate, then the workload distribution at Q_1 inherits the heavy-tailed characteristics of Q_2 as described above.
Several recent studies have revealed a similar dichotomy in the interplay between exponential and heavy-tailed traffic processes. For example, [12] shows a similar contrast in an M/M/1 queue which alternates between exponentially distributed periods of high service speed and heavy-tailed periods of low service speed. A related phenomenon is observed in [9] for an M/G/2 queue with heterogeneous servers, one having exponential properties, the other one exhibiting heavy-tailed characteristics. A similar situation is also encountered in queues fed by a superposition of light-tailed traffic and heavy-tailed On-Off sources [8, 28].
The paper is organized in the following way. In Section 2, we present a detailed model description. Section 3 contains some preliminary results which provide the basis for the subsequent analysis. In Section 4, we obtain the workload asymptotics at Q_1 in case its service time distribution is heavy-tailed, while the service time distribution at Q_2 is allowed to be general. In Section 5, we derive the workload asymptotics at Q_1 in case its service time distribution is exponential, while the service time distribution at Q_2 is assumed to be heavy-tailed. In both Sections 4 and 5 we assume (with ρ_i denoting the offered traffic at Q_i) that $\rho_1 < 1$, and consider the case $\rho_2 < 1$ as well as $\rho_2 > 1$. In Section 6, we characterize the workload asymptotics at Q_1 for the case where $\rho_1 > 1$ (which forces $\rho_2 < 1$) and the service time distribution at both Q_1 and Q_2 is heavy-tailed. In Section 7, we extend the results of Section 6 to situations where the service time distribution at one of the queues is heavy-tailed and the service time distribution at the other queue has a lighter, general tail. The boundary case where either $\rho_1 = 1$ or $\rho_2 = 1$ is rather delicate, and will not be considered here.

2 Model description

We consider a system with two heterogeneous traffic classes. Class-i customers arrive as a Poisson process of rate λ_i, and require an amount of service $\mathbf{B}_i$ with mean $\beta_i < \infty$ and Laplace-Stieltjes Transform (LST) $\beta_i\{s\} := \mathbb{E}[\mathrm{e}^{-s\mathbf{B}_i}]$, Re $s \geq 0$. Define $\rho_i := \lambda_i\beta_i$ as the traffic intensity of class i.

In the next sections we will consider various scenarios for the service time distributions of the two classes. In all scenarios, the service time distribution of at least one of the two classes will be assumed to be regularly varying. If the service time distribution of class i is regularly varying of index $-\nu_i$, denoted as $\mathbb{P}\{\mathbf{B}_i > t\} \in \mathcal{R}_{-\nu_i}$, then cf. [3]

$$\mathbb{P}\{\mathbf{B}_i > t\} \sim \frac{C_i}{-\Gamma(1-\nu_i)} t^{-\nu_i} l_i(t), \quad t \to \infty; \tag{1}$$

here $l_i(\cdot)$ is some slowly varying function, i.e., $\lim_{t\to\infty} l_i(\eta t)/l_i(t) = 1$, $\eta > 1$; C_i denotes a constant; $\Gamma(\cdot)$ represents the Gamma function. For any two real functions $f(\cdot)$ and $g(\cdot)$, we use the notational convention $f(t) \sim g(t)$ for $t \to \infty$ to indicate that $\lim_{t\to\infty} f(t)/g(t) = 1$, or equivalently, $f(t) = g(t)(1 + \mathrm{o}(1))$ as $t \to \infty$. Throughout the present paper, we assume that $1 < \nu_i < 2$, which is an interesting case since the variance of $\mathbf{B}_i$ is then infinite. Larger values of ν_i may be handled with minor modifications.

There are separate queues maintained for each class, Q_1 and Q_2. When both classes are backlogged, each of the queues is served at unit rate. However, the service rate at Q_i increases to $r_i \geq 1$ when the other queue is empty. Thus, the two queues are coupled through the mechanism for dynamically sharing surplus service capacity.

For $r_i = 1/\phi_i$, with $\phi_1 + \phi_2 = 1$, the model may equivalently be viewed as a two-class GPS system of capacity C with relative weight factors ϕ_i and the service times of class-i customers scaled by a factor $\phi_i C$. In the present paper, we concentrate on the more general case $1/r_1 + 1/r_2 \neq 1$. A careful examination shows however that the case $1/r_1 + 1/r_2 = 1$ does not represent a boundary case, suggesting that the main results of the paper remain valid when $1/r_1 + 1/r_2 \to 1$.

Throughout the paper, we assume that the ergodicity conditions are satisfied. These conditions are discussed in Section III.3.7 of [16]. Here, it suffices to observe that $\max\{\rho_1, \rho_2\} < 1$ is sufficient but not necessary for ergodicity, while $\min\{\rho_1, \rho_2\} < 1$ is necessary.

On several occasions we will have to deal with residual lifetimes of random variables. For any non-negative random variable $\mathbf{X}$ with $\mathbb{E}\mathbf{X} < \infty$, denote by $\mathbf{X}^r$ a random variable representing the residual lifetime of $\mathbf{X}$, i.e., $\mathbb{P}\{\mathbf{X}^r < t\} = \frac{1}{\mathbb{E}\mathbf{X}} \int_0^t \mathbb{P}\{\mathbf{X} > u\}\mathrm{d}u$. In particular, note that if $\mathbb{P}\{\mathbf{B}_i > t\} \in \mathcal{R}_{-\nu_i}$, $\nu_i > 1$, then

$$\mathbb{P}\{\mathbf{B}_i^r > t\} \sim \frac{C_i}{\beta_i\Gamma(2-\nu_i)} t^{1-\nu_i} l_i(t), \quad t \to \infty; \tag{2}$$

it is again regularly varying, but of index $1 - \nu_i$.

3 Preliminary results

In this section we review some preparatory results which will provide the starting point for the analysis. Denote by $\mathbf{V}_i$ a random variable representing the workload at Q_i in steady state. For $c > 0$, denote by $\mathbf{V}_i^c$ a random variable representing the steady-state workload at Q_i when served in isolation at a constant rate c. For Re $s_1 \geq 0$, Re $s_2 \geq 0$, let

$$\begin{aligned}
\psi(s_1, s_2) &:= \mathbb{E}[\mathrm{e}^{-s_1\mathbf{V}_1 - s_2\mathbf{V}_2}], \\
\psi_1(s_2) &:= \mathbb{E}[\mathrm{e}^{-s_2\mathbf{V}_2}\mathrm{I}_{\{\mathbf{V}_1=0\}}], \\
\psi_2(s_1) &:= \mathbb{E}[\mathrm{e}^{-s_1\mathbf{V}_1}\mathrm{I}_{\{\mathbf{V}_2=0\}}], \\
\psi_0 &:= \mathbb{P}\{\mathbf{V}_1 = 0, \mathbf{V}_2 = 0\},
\end{aligned}$$

with $\mathrm{I}_{\{A\}}$ denoting the indicator function of the event A.
According to Formula (2.16) of Chapter III.3 of [16] (in the sequel we omit Chapter III.3 when referring to formulas from [16]), for Re $s \geq 0$,

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] = \mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}][\frac{\psi_1(0)}{1-\rho_1} + \frac{r_1 - 1}{1-\rho_1}(\psi_0 - \psi_2(s))]. \tag{3}$$

Taking $s = 0$ in (3), we obtain

$$\frac{\psi_1(0)}{1-\rho_1} + \frac{r_1 - 1}{1-\rho_1}(\psi_0 - \psi_2(0)) = 1,$$

so that (3) may be rewritten as

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] = \mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}][1 - \frac{r_1 - 1}{1-\rho_1}(\psi_2(s) - \psi_2(0))]. \tag{4}$$

We now focus on the function $\psi_2(s)$. According to Formulas (6.21), (6.22) and (6.23) of [16],

$$\psi_2(\delta_1(w)) - \psi_0 = \frac{1}{r_1}\frac{\psi_0}{1 - \frac{1}{r_1} - \frac{1}{r_2}}[1 - \mathrm{e}^{-R_1(w)+R_2(w)}], \quad \text{Re } w \geq 0, \tag{5}$$

and

$$\psi_1(\delta_2(w)) - \psi_0 = \frac{1}{r_2}\frac{\psi_0}{1 - \frac{1}{r_1} - \frac{1}{r_2}}[1 - \mathrm{e}^{P_1(w)-P_2(w)}], \quad \text{Re } w \leq 0, \tag{6}$$

with

$$\psi_0 = \mathrm{e}^{-P_1(0)-R_2(0)}.$$

It remains to specify the functions $R_i(w)$, $P_i(w)$, and $\delta_i(w)$, $i = 1, 2$:

$$R_i(w) := \sum_{n=1}^{\infty} \frac{b_i^n}{n}\mathbb{E}[\mathrm{e}^{-w\sigma_n^{(i)}}\mathrm{I}_{\{\sigma_n^{(i)}>0\}}], \quad \text{Re } w \geq 0,$$

and

$$P_i(w) := \sum_{n=1}^{\infty} \frac{b_i^n}{n}\mathbb{E}[\mathrm{e}^{-w\sigma_n^{(i)}}\mathrm{I}_{\{\sigma_n^{(i)}<0\}}], \quad \text{Re } w \leq 0,$$

with

$$b_1 := \rho_1(1-\frac{1}{r_2})+\frac{\rho_2}{r_2}, \tag{7}$$

$$b_2 := \rho_2(1-\frac{1}{r_1})+\frac{\rho_1}{r_1}, \tag{8}$$

and for $i=1,2$,

$$\sigma_n^{(i)} := \mathbf{X}_{i1}+\ldots+\mathbf{X}_{in},$$

with $\mathbf{X}_{11},\ldots,\mathbf{X}_{1n}$ i.i.d. and $\mathbf{X}_{21},\ldots,\mathbf{X}_{2n}$ i.i.d., and

$$\mathbf{X}_{11} = \begin{cases} \hat{\mathbf{P}}_1 & w.p.\ \pi_1 := \frac{\rho_1}{b_1}(1-\frac{1}{r_2}), \\ -\hat{\mathbf{P}}_2 & w.p.\ 1-\pi_1 = \frac{\rho_2}{b_1 r_2}, \end{cases} \tag{9}$$

$$\mathbf{X}_{21} = \begin{cases} \hat{\mathbf{P}}_1 & w.p.\ \pi_2 := \frac{\rho_1}{b_2 r_1}, \\ -\hat{\mathbf{P}}_2 & w.p.\ 1-\pi_2 = \frac{\rho_2}{b_2}(1-\frac{1}{r_1}). \end{cases} \tag{10}$$

Here $\hat{\mathbf{P}}_i$ is a random variable representing the busy period of class i when served in isolation at unit rate, *starting with an exceptional first service time* $\mathbf{B}_i^r$ (a residual service time as defined in Section 2).

The functions $\delta_i(w)$, $i=1,2$, play a crucial role in the analysis: $\delta_1(w)$ is defined for Re $w \geq 0$ as zero of the function

$$f_1(s,w) := \lambda_1(1-\beta_1\{s\})-s+w, \tag{11}$$

which has for Re $w \geq 0$, $w \neq 0$, exactly one zero $s=\delta_1(w)$ in Re $s \geq 0$, and this zero has multiplicity one.

$f_1(s,0)$ has for $\rho_1 < 1$ exactly one zero $s=\delta_1(0)=0$ in Re $s \geq 0$, with multiplicity one;

$f_1(s,0)$ has for $\rho_1 = 1$ exactly one zero $s=\delta_1(0)=0$ in Re $s \geq 0$, with multiplicity two;

$f_1(s,0)$ has for $\rho_1 > 1$ two zeroes $s=\delta_1(0)>0$ and $s=\epsilon_1(0)=0$ in Re $s \geq 0$, each with multiplicity one.

Similarly, $\delta_2(w)$ is defined for Re $w \leq 0$ as zero of the function

$$f_2(s,w) := \lambda_2(1-\beta_2\{s\})-s-w. \tag{12}$$

For the case where class-1 customers have heavy-tailed service times, the following lemma will be instrumental in deriving the asymptotic behavior of $\mathbb{P}\{\mathbf{V}_1 > t\}$ as $t \to \infty$ from the behavior of $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]$ as $s \downarrow 0$, and vice versa. It has been formulated in Lemma 2.2 in [11] as an extension of Theorem 8.1.6 in [3].

Lemma 3.1 *Let* $\mathbf{Y}$ *be a non-negative random variable,* $l(t)$ *a slowly varying function,* $\nu \in (n,n+1)$ *(*$n \in \mathbb{N}$*), and* $D \geq 0$*. Then the following two statements are equivalent:*

(i) $\mathbb{P}\{\mathbf{Y} > t\} = (D+\mathrm{o}(1))t^{-\nu}l(t)$ *as* $t \to \infty$*;*

(ii) $\mathbb{E}[\mathbf{Y}^n] < \infty$ *and* $\mathbb{E}[\mathrm{e}^{-s\mathbf{Y}}] - \sum_{j=0}^{n} \frac{\mathbb{E}[\mathbf{Y}^j](-s)^j}{j!} = (-1)^n\Gamma(1-\nu)(D+\mathrm{o}(1))s^\nu l(1/s)$ *as* $s \downarrow 0$*.*

In case the class-1 customers have exponential service times, we use a theorem of Sutton [27] (see also [6]), relating the asymptotic behavior of a function to its Laplace transform.

4 Q_1 heavy-tailed; Q_2 general; $\rho_1 < 1$

4.1 Introduction and main result

In this section we consider the case where the service time distribution at Q_1 is regularly varying of index $-\nu_1$, i.e.,

$$\mathbb{P}\{\mathbf{B}_1 > t\} \sim \frac{C_1}{-\Gamma(1-\nu_1)} t^{-\nu_1} l_1(t), \quad t \to \infty, \tag{13}$$

with $l_1(\cdot)$ some slowly varying function. According to Lemma 3.1, (13) with $1 < \nu_1 < 2$ is equivalent with (using $\sim$ according to the same notational convention for $s \downarrow 0$ as previously for $t \to \infty$)

$$1 - \frac{1-\beta_1\{s\}}{\beta_1 s} \sim \frac{C_1}{\beta_1} s^{\nu_1 - 1} l_1(\frac{1}{s}), \quad s \downarrow 0. \tag{14}$$

The service time distribution at Q_2 is allowed to be general. We assume that $\rho_1 < 1$, so that even the lower service speed at Q_1 is sufficient for stability; the case $\rho_1 > 1$ is treated in Section 6. We intend to show that the workload distribution at Q_1 is regularly varying of index $1 - \nu_1$, irrespective of the service time distribution at Q_2. We first assume that $\rho_2 < 1$. Below we will consider the case $\rho_2 > 1$.

Theorem 4.1 *If $\mathbb{P}\{\mathbf{B}_1 > t\} \in \mathcal{R}_{-\nu_1}$, $1 < \nu_1 < 2$, as given in (13), and if $\rho_1 < 1$, $\rho_2 < 1$, then $\mathbb{P}\{\mathbf{V}_1 > t\} \in \mathcal{R}_{1-\nu_1}$ as given below,*

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \frac{1}{K_1 - \rho_1} \frac{\lambda_1 C_1}{\Gamma(2-\nu_1)} t^{1-\nu_1} l_1(t) \sim \frac{\rho_1}{K_1 - \rho_1} \mathbb{P}\{\mathbf{B}_1^r > t\}, \quad t \to \infty, \tag{15}$$

with $K_1 := \rho_2 + (1-\rho_2) r_1 \geq 1$ representing the average service rate at Q_1 when continuously backlogged.

Remark 4.1 *Qualitatively, the result of Theorem 4.1 resembles Theorem 3.1 in [5] for a different but related GPS model, which is proved using asymptotic lower and upper bounds.*

4.2 Discussion and interpretation

Invoking a result of Cohen [14], Formula (15) may be rewritten as

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \mathbb{P}\{\mathbf{V}_1^{K_1} > t\}, \quad t \to \infty. \tag{16}$$

Thus, asymptotically, the workload at Q_1 behaves exactly as if Q_1 is served in isolation at a *constant* rate K_1. This may be intuitively interpreted as follows. Large-deviations arguments suggest that the most likely scenario for the workload at Q_1 to reach a large level is that class 1 generates a large amount of traffic, while class 2 shows average behavior. Specifically, suppose that a class-1 customer arrives with a large service time, so that Q_1 becomes backlogged for a long period of time. During that period, Q_2 will not receive any

surplus capacity, and hence be empty only a fraction $1-\rho_2$ of the time. As a result, the average service rate at Q_1 while backlogged is $K_1 = \rho_2 + (1-\rho_2)r_1$. Thus, Q_1 is effectively served at a constant rate $K_1 \geq 1$, as confirmed by Formula (16).
It is worth observing that the above result holds regardless of the service time distribution at Q_2 as long as both $\rho_1 < 1$ and $\rho_2 < 1$. In that sense, Q_1 is not significantly affected by the interaction with Q_2. In particular, Q_1 is virtually immune from 'heavier'-tailed service time characteristics at Q_2. That will no longer be the case when $\rho_1 > 1$, as we will see in Section 6.

4.3 Proof of Theorem 4.1

The approach may be outlined as follows. Formula (4) expresses $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]$ into $\psi_2(s)$. Formula (5) relates $\psi_2(s)$, or rather $\psi_2(\delta_1(w))$, to $R_1(w)$ and $R_2(w)$. We use these formulas to derive the behavior of $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]$ for $s \downarrow 0$. Lemma 3.1 then yields the behavior of $\mathbb{P}\{\mathbf{V}_1 > t\}$ for $t \to \infty$. Therefore, we now concentrate on the behavior of $R_1(w)$ and $R_2(w)$ and, first, $\delta_1(w)$ for $w \downarrow 0$.
Let $\mathbf{P}_1$ denote a random variable with distribution the steady-state distribution of a busy period at Q_1 in isolation, i.e., an M/G/1 queue with arrival rate λ_1 and service time distribution $B_1(\cdot)$. Comparing (11) with the Takács equation for the busy-period LST $\mathbb{E}[\mathrm{e}^{-w\mathbf{P}_1}]$, cf. p. 250 of Cohen [15], it is seen that

$$\delta_1(w) = w + \lambda_1(1 - \mathbb{E}[\mathrm{e}^{-w\mathbf{P}_1}]).$$

De Meyer & Teugels [21] have proven that $\mathbb{P}\{\mathbf{P}_1 > t\}$ is regularly varying of index $-\nu_1$ iff $\mathbb{P}\{\mathbf{B}_1 > t\}$ is regularly varying of index $-\nu_1$, and if either holds then

$$\mathbb{P}\{\mathbf{P}_1 > t\} \sim \frac{1}{1-\rho_1}\mathbb{P}\{\frac{\mathbf{B}_1}{1-\rho_1} > t\}, \quad t \to \infty. \tag{17}$$

Lemma 3.1 then gives the behavior of $\mathbb{E}[\mathrm{e}^{-w\mathbf{P}_1}] - 1$ for $w \downarrow 0$. We conclude that, if (13) holds, then

$$\delta_1(w) - \frac{w}{1-\rho_1} \sim -\frac{\lambda_1 C_1}{1-\rho_1}(\frac{w}{1-\rho_1})^{\nu_1} l_1(\frac{1}{w}), \quad w \downarrow 0. \tag{18}$$

In addition, using (11), we have for $\delta_1^{-1}(s) = s - \lambda_1(1-\beta_1\{s\})$:

$$\delta_1^{-1}(s) - (1-\rho_1)s \sim \lambda_1 C_1 s^{\nu_1} l_1(\frac{1}{s}), \quad s \downarrow 0.$$

In the study of $R_i(w)$, a key role is played by the LST of $\hat{\mathbf{P}}_1$, a busy period at Q_1 in isolation that starts with a *residual* service time. From (6.4) of [16],

$$\mathbb{E}[\mathrm{e}^{-w\hat{\mathbf{P}}_1}] = \frac{1-\beta_1\{\delta_1(w)\}}{\beta_1\delta_1(w)}, \quad \mathrm{Re}\ w \geq 0.$$

It is now readily verified that

$$1 - \mathbb{E}[\mathrm{e}^{-w\hat{\mathbf{P}}_1}] \sim \frac{C_1}{\beta_1}(\frac{w}{1-\rho_1})^{\nu_1 - 1} l_1(\frac{1}{\delta_1(w)}), \quad w \downarrow 0,$$

and hence, using Lemma 3.1, $\mathbb{P}\{\hat{\mathbf{P}}_1 > t\}$ is seen to be regularly varying of index $1-\nu_1$,

$$\mathbb{P}\{\hat{\mathbf{P}}_1 > t\} \sim \frac{C_1}{\beta_1 \Gamma(2-\nu_1)}((1-\rho_1)t)^{1-\nu_1} l_1(t), \quad t \to \infty. \tag{19}$$

The difference with (17) is caused by the *residual* service time with which the busy period starts; it is regularly varying of one index higher than an ordinary service time.
We are now ready to study the tail behavior of $R_i(w)$. Observe that $R_i(w)$ is the LST of

$$r_i(t) := \sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathbb{P}\{0 < \mathbf{X}_{i1} + \ldots + \mathbf{X}_{in} < t\}, \quad t > 0.$$

Consider

$$R_i(0) - r_i(t) = \sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathbb{P}\{\mathbf{X}_{i1} + \ldots + \mathbf{X}_{in} > t\}, \quad t > 0. \tag{20}$$

Using well-known properties of long-tailed random variables [25] (see Appendix, Proposition 1.1), it may be shown as in [4] that, irrespective of the distribution of $\mathbf{B}_2$,

$$\mathbb{P}\{\mathbf{X}_{21} + \ldots + \mathbf{X}_{2n} > t\} \sim n\pi_2 \mathbb{P}\{\hat{\mathbf{P}}_1 > t\}, \quad t \to \infty. \tag{21}$$

We conclude from (19), (20) and (21) that

$$R_2(0) - r_2(t) \sim \frac{b_2\pi_2}{1-b_2} \mathbb{P}\{\hat{\mathbf{P}}_1 > t\} \sim \frac{1}{(1-b_2)r_1} \frac{\lambda_1 C_1}{\Gamma(2-\nu_1)} ((1-\rho_1)t)^{1-\nu_1} l_1(t), \quad t \to \infty. \tag{22}$$

Again applying Lemma 3.1,

$$R_2(w) - R_2(0) \sim -\frac{\lambda_1 C_1}{(1-b_2)r_1} \left(\frac{w}{1-\rho_1}\right)^{\nu_1 - 1} l_1\left(\frac{1}{w}\right), \quad w \downarrow 0. \tag{23}$$

Similarly, it is seen that

$$\mathbb{P}\{\mathbf{X}_{11} + \ldots + \mathbf{X}_{1n} > t\} \sim n\pi_1 \mathbb{P}\{\hat{\mathbf{P}}_1 > t\}, \quad t \to \infty,$$

leading to

$$R_1(w) - R_1(0) \sim -\frac{\lambda_1 C_1 (1 - \frac{1}{r_2})}{1-b_1} \left(\frac{w}{1-\rho_1}\right)^{\nu_1 - 1} l_1\left(\frac{1}{w}\right), \quad w \downarrow 0. \tag{24}$$

It follows from (5), (23) and (24) after a lengthy calculation that

$$\psi_2(\delta_1(w)) - \psi_2(0) \sim -\frac{\lambda_1 C_1 (1-\rho_2)}{(1-b_2)r_1} \left(\frac{w}{1-\rho_1}\right)^{\nu_1 - 1} l_1\left(\frac{1}{w}\right), \quad w \downarrow 0. \tag{25}$$

Finally, see (18),

$$\psi_2(s) - \psi_2(0) \sim -\frac{\lambda_1 C_1 (1-\rho_2)}{(1-b_2)r_1} s^{\nu_1 - 1} l_1\left(\frac{1}{s}\right), \quad s \downarrow 0. \tag{26}$$

Using (4), (14) and (26), it follows that

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] - 1 \sim -\left[\frac{1}{1-\rho_1} - \frac{(r_1-1)(1-\rho_2)}{1-\rho_1} \frac{1}{(1-b_2)r_1}\right] \lambda_1 C_1 s^{\nu_1 - 1} l_1\left(\frac{1}{s}\right), \quad s \downarrow 0.$$

Inserting (8), we can rewrite this into

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] - 1 \sim -\frac{\lambda_1 C_1}{K_1 - \rho_1} s^{\nu_1 - 1} l_1\left(\frac{1}{s}\right), \quad s \downarrow 0, \tag{27}$$

with $K_1 = \rho_2 + (1-\rho_2)r_1$. Applying Lemma 3.1 once more completes the proof of the theorem.

4.4 The case $\rho_2 > 1$

We now turn to the case $\rho_2 > 1$. The characterization of the overflow scenario provided in Subsection 4.2 suggests that Theorem 4.1 should then continue to hold, except that the service rate K_1 should be reduced to 1. Specifically, the most likely scenario for the workload at Q_1 to reach a large level is again that a class-1 customer arrives with a large service time, so that Q_1 becomes backlogged for a long period of time. As before, Q_2 will not receive any surplus capacity, and hence soon become persistently backlogged too, since now $\rho_2 > 1$. Thus, Q_1 is only served at rate 1 (rather than $K_1 \geq 1$) while backlogged, as is confirmed by the next theorem.

Theorem 4.2 *If $\mathbb{P}\{\mathbf{B}_1 > t\} \in \mathcal{R}_{-\nu_1}$, $1 < \nu_1 < 2$, as given in (13), and if $\rho_1 < 1$, $\rho_2 > 1$, then $\mathbb{P}\{\mathbf{V}_1 > t\} \in \mathcal{R}_{1-\nu_1}$ as given below,*

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \frac{1}{1-\rho_1}\frac{\lambda_1 C_1}{\Gamma(2-\nu_1)}t^{1-\nu_1}l_1(t) \sim \frac{\rho_1}{1-\rho_1}\mathbb{P}\{\mathbf{B}_1^r > t\}, \quad t \to \infty. \tag{28}$$

Proof

The proof is largely similar to that of Theorem 4.1. The main difference is that the distribution of $\hat{\mathbf{P}}_2$ is now defective; $\mathbb{P}\{\hat{\mathbf{P}}_2 < \infty\} = \frac{1}{\rho_2}$. As in Section IV of [4], it may be shown that the first part of (22) changes into

$$R_2(0) - r_2(t) \sim \frac{b_2\pi_2}{1 - b_2(\pi_2 + \frac{1-\pi_2}{\rho_2})}\mathbb{P}\{\hat{\mathbf{P}}_1 > t\}, \quad t \to \infty. \tag{29}$$

Similarly,

$$R_1(0) - r_1(t) \sim \frac{b_1\pi_1}{1 - b_1(\pi_1 + \frac{1-\pi_1}{\rho_2})}\mathbb{P}\{\hat{\mathbf{P}}_1 > t\}, \quad t \to \infty. \tag{30}$$

From (7)-(10), it may be checked that the constants in the two right-hand sides both equal $\frac{\rho_1}{1-\rho_1}$. Using (19) and Lemma 3.1, we obtain

$$R_1(w) - R_1(0) \sim R_2(w) - R_2(0) \sim -\frac{\lambda_1 C_1}{1-\rho_1}(\frac{-w}{1-\rho_1})^{\nu_1-1}l_1(\frac{-1}{w}), \quad w \uparrow 0. \tag{31}$$

Substituting into (5), we have

$$\psi_2(\delta_1(w)) - \psi_2(0) = \mathrm{o}(w^{\nu_1-1}l_1(\frac{1}{w})), \quad w \downarrow 0. \tag{32}$$

Using (4), (14) and (32) (instead of (25)), we obtain

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] - 1 \sim -\frac{\lambda_1 C_1}{1-\rho_1}s^{\nu_1-1}l_1(\frac{1}{s}), \quad s \downarrow 0.$$

Applying Lemma 3.1 then completes the proof of the theorem.

5 Q_1 exponential; Q_2 heavy-tailed; $\rho_1 < 1$

5.1 Introduction and main result

In this section we consider the case where the service time distribution at Q_1 is exponential. The service time distribution at Q_2 is assumed to be regularly varying of index $-\nu_2$, $1 < \nu_2 < 2$, i.e.,

$$\mathbb{P}\{\mathbf{B}_2 > t\} \sim \frac{C_2}{-\Gamma(1-\nu_2)} t^{-\nu_2} l_2(t), \quad t \to \infty, \tag{33}$$

with $l_2(\cdot)$ some slowly varying function. We assume that $\rho_1 < 1$, so that even the lower service speed at Q_1 is sufficient for stability; the case $\rho_1 > 1$ is covered in Section 7. We will show that the workload distribution at Q_1 is now (semi-)exponential, depending on whether $\rho_2 < 1$ or $\rho_2 > 1$. We first assume that $\rho_2 < 1$. Below we will deal with the case $\rho_2 > 1$. The results of this section have been proved in [6]. We include them here for the sake of completeness.

Theorem 5.1 *If $\mathbb{P}\{\mathbf{B}_1 > t\}$ is exponential, and $\mathbb{P}\{\mathbf{B}_2 > t\} \in \mathcal{R}_{-\nu_2}$, $1 < \nu_2 < 2$, as given in (33), and if $\rho_1 < 1$, $\rho_2 < 1$, then*

$$\begin{aligned} \mathbb{P}\{\mathbf{V}_1 > t\} &\sim \rho_1 \mathrm{e}^{(\lambda_1 - 1/\beta_1)t} \frac{1}{K_2 - \rho_2} \frac{\lambda_2 C_2}{\Gamma(2-\nu_2)} \left(\frac{\rho_1(1-\rho_2)}{1-\rho_1} t \right)^{1-\nu_2} l_2(t) \\ &\sim \mathbb{P}\{\mathbf{V}_1^1 > t\} \frac{\rho_2}{K_2 - \rho_2} \mathbb{P}\{\mathbf{B}_2^r > \frac{\rho_1(1-\rho_2)}{1-\rho_1} t\}, \quad t \to \infty, \end{aligned} \tag{34}$$

with $K_2 := \rho_1 + (1-\rho_1) r_2 \geq 1$ representing the average service rate at Q_2 when continuously backlogged.

Proof

The proof is presented in [6]. The approach may be outlined as follows. The term $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}]$, appearing in (4), represents the LST of the waiting-time distribution in an M/M/1 queue. It is well-known that this LST has a pole at $s = s_1 := \lambda_1 - 1/\beta_1 < 0$. A lemma of Sutton [27] relates the asymptotic behavior of a function $f(t)$ for $t \to \infty$ to that of its Laplace transform $\phi(s)$ for s near its poles. Take $f(t) = \mathbb{P}\{\mathbf{V}_1 > t\}$, so that $\phi(s) = \frac{1 - \mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]}{s}$. In applying Sutton's lemma, we then need to consider $(s - s_1)\phi(s)$, and determine the series expansion for s near s_1. The pole at $s = s_1$ is responsible for the occurrence of the term $\exp[s_1 t] = \exp[(\lambda_1 - 1/\beta_1)t]$ in (34).

Remark 5.1 *Qualitatively, the result of Theorem 5.1 mirrors that of Theorem 3.1 in [7] for a diferent but related GPS model, which is established using probabilistic techniques.*

5.2 Discussion and interpretation

Invoking a result of Cohen [14], Formula (34) may be rewritten as

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \mathbb{P}\{\mathbf{V}_1^1 > t\}\mathbb{P}\{\mathbf{V}_2^{K_2} > \frac{1-\rho_2}{\hat{\rho}_1 - 1}t\}, \quad t \to \infty, \tag{35}$$

with $\hat{\rho}_1 := 1/\rho_1 > 1$. To understand the above formula, it is useful to draw a comparison with the workload $\mathbf{V}_1^1$ when Q_1 is served in isolation at constant rate 1. Large-deviations results for the M/M/1 queue [26] suggest that the most likely way for $\mathbf{V}_1^1$ to reach a large level x is that class 1 temporarily experiences 'abnormal' traffic activity. Specifically, class 1 essentially behaves as if its traffic intensity were increased from the normal value ρ_1 to the value $\hat{\rho}_1$, causing a positive drift $\hat{\rho}_1 - 1 > 0$ in the workload. In order for the workload to reach a large level x, the deviant behavior must persist for a period of time $\frac{x}{\hat{\rho}_1 - 1}$.
Now observe that the above scenario may occur in the shared system as well, provided Q_2 is continuously backlogged while class 1 shows deviant behavior. In fact, given that the workload at Q_1 reaches a large level, Q_2 is constantly backlogged with overwhelming probability, since otherwise class 1 must show even greater anomalous activity. The probability of that happening is negligibly small compared to that of Q_2 being continuously backlogged, because of the highly bursty nature of class-2 traffic.
Note that the normal drift in the workload at Q_2 is $\rho_2 - 1 < 0$. Thus, in order for Q_2 to remain backlogged during the period of deviant behavior of class 1, an additional amount of work of at least $\frac{1-\rho_2}{\hat{\rho}_1 - 1}x$ must be accounted for. The most likely scenario is that class 2 generates a large amount of traffic prior to the deviant behavior of class 1, so that when that starts, the workload at Q_2 is at least $\frac{1-\rho_2}{\hat{\rho}_1 - 1}x$. During that prior period, class 1 shows average behavior, and Q_1 does not receive any surplus capacity from Q_2, so that Q_1 is empty a fraction $1 - \rho_1$ of the time. Thus, the average service rate at Q_2 during that period is $K_2 = \rho_1 + (1 - \rho_1)r_2$. Hence, the probability that Q_2 is sufficiently long backlogged is approximately equal to $\mathbb{P}\{\mathbf{V}_2^{K_2} > \frac{1-\rho_2}{\hat{\rho}_1 - 1}x\}$. Combined, these considerations yield Formula (35).

5.3 The case $\rho_2 > 1$

We now turn to the case $\rho_2 > 1$. In this case, it is relatively likely for Q_2 to be continuously backlogged during the period in which class 1 shows deviant behavior. This suggests that the asymptotic behavior of $\mathbb{P}\{\mathbf{V}_1 > t\}$ should not fundamentally change, except that the pre-factor of $\mathbb{P}\{\mathbf{V}_1^1 > t\}$ should be O(1), as is confirmed by the next theorem. Its proof may be found in [6].

Theorem 5.2 *If $\mathbb{P}\{\mathbf{B}_1 > t\}$ is exponential, and $\mathbb{P}\{\mathbf{B}_2 > t\} \in \mathcal{R}_{-\nu_2}$, $1 < \nu_2 < 2$, as given in (33), and if $\rho_1 < 1$, $\rho_2 > 1$, then*

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim H_2\rho_1 \mathrm{e}^{(\lambda_1 - 1/\beta_1)t} \sim H_2\mathbb{P}\{\mathbf{V}_1^1 > t\}, \quad t \to \infty, \tag{36}$$

with

$$H_2 := \frac{\rho_2 - 1}{r_2 - 1} \frac{1}{1 - \rho_1} = \frac{\rho_2 - 1}{K_2 - 1}.$$

Remark 5.2 *It follows from the discussion on p. 316 of [16] that $K_2 > \rho_2$ if the system is ergodic, as may also be argued from the interpretation of K_2 given earlier, which implies that $H_2 < 1$. This is consistent with the observation that $\mathbb{P}\{\mathbf{V}_1 > t\} \leq \mathbb{P}\{\mathbf{V}_1^1 > t\}$ for all $t \geq 0$ if $\rho_1 < 1$, since Q_1 is guaranteed to receive a minimum service rate of 1 when backlogged. Note that the latter observation is also in agreement with Theorems 4.1, 4.2, and 5.1.*

6 Q_1 heavy-tailed; Q_2 heavy-tailed; $\rho_1 > 1$; $\rho_2 < 1$

6.1 Introduction and main result

In this section we consider the case where the service time distribution at both Q_1 and Q_2 is regularly varying of index $-\nu_1$ and $-\nu_2$, respectively. We assume that $\rho_1 > 1$ so that the lower service speed alone is not sufficient to ensure stability of Q_1. Thus, Q_1 relies on surplus capacity from Q_2, which forces $\rho_2 < 1$. The next theorem shows that the workload distribution at Q_1 is then determined by the heaviest of the service time distributions at Q_1 and Q_2.

Theorem 6.1 *If $\mathbb{P}\{\mathbf{B}_i > t\} \in \mathcal{R}_{-\nu_i}$ as given in (1), $1 < \nu_i < 2$, $i = 1, 2$, and if $\rho_1 > 1$, $\rho_2 < 1$, then $\mathbb{P}\{\mathbf{V}_1 > t\} \in \mathcal{R}_{1-\min\{\nu_1,\nu_2\}}$ as given below:*
If $\nu_1 < \nu_2$, then

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \frac{1}{K_1 - \rho_1} \frac{\lambda_1 C_1}{\Gamma(2 - \nu_1)} t^{1-\nu_1} l_1(t) \sim \frac{\rho_1}{K_1 - \rho_1} \mathbb{P}\{\mathbf{B}_1^r > t\}, \quad t \to \infty; \tag{37}$$

If $\nu_1 > \nu_2$, then

$$\begin{aligned} \mathbb{P}\{\mathbf{V}_1 > t\} &\sim \frac{r_1 - 1}{K_1 - \rho_1} \frac{\lambda_2 C_2}{\Gamma(2 - \nu_2)} \left(\frac{1 - \rho_2}{\rho_1 - 1} t\right)^{1-\nu_2} l_2(t) \\ &\sim \frac{(r_1 - 1)\rho_2}{K_1 - \rho_1} \mathbb{P}\{\mathbf{B}_2^r > \frac{1 - \rho_2}{\rho_1 - 1} t\}, \quad t \to \infty; \end{aligned} \tag{38}$$

If $\nu_1 = \nu_2 = \nu$, then

$$\begin{aligned} \mathbb{P}\{\mathbf{V}_1 > t\} &\sim \frac{1}{K_1 - \rho_1} \frac{\lambda_1 C_1}{\Gamma(2 - \nu)} t^{1-\nu} l_1(t) + \frac{r_1 - 1}{K_1 - \rho_1} \frac{\lambda_2 C_2}{\Gamma(2 - \nu)} \left(\frac{1 - \rho_2}{\rho_1 - 1} t\right)^{1-\nu} l_2(t), \\ & t \to \infty. \end{aligned} \tag{39}$$

Here $K_1 := \rho_2 + (1 - \rho_2) r_1$ represents the average service rate at Q_1 when continuously backlogged.

Remark 6.1 *Qualitatively, the results for $\nu_1 < \nu_2$ and $\nu_1 > \nu_2$ in Theorem 6.1 are similar to Theorems 4.1 and 5.1 in [5], respectively.*

6.2 Discussion and interpretation

Theorem 6.1 shows that the workload distribution at Q_1 is determined by the heaviest of the service time distributions at Q_1 and Q_2. If the tail of $\mathbf{B}_1$ is heavier than that of $\mathbf{B}_2$, then the workload at Q_1 behaves exactly as if Q_1 is served in isolation at a constant rate K_1. The explanation is largely similar to that of Theorem 4.1 provided in Subsection 4.2. Specifically, the most likely scenario for the workload at Q_1 to reach a large level is again that a class-1 customer arrives with a large service time, so that Q_1 becomes backlogged for a long period of time. As result, Q_2 will not receive any surplus capacity, and hence be empty only a fraction $1-\rho_2$ of the time. Thus, the average service rate at Q_1 while backlogged is $K_1 = \rho_2 + (1-\rho_2)r_1$ as before.

Since $\rho_1 > 1$, a large workload at Q_1 may also occur when Q_2 is backlogged for a long period of time. The latter scenario is however far less likely when the tail of $\mathbf{B}_2$ is lighter than that of $\mathbf{B}_1$.

In contrast, if the tail of $\mathbf{B}_2$ is heavier than that of $\mathbf{B}_1$, then the tail of $\mathbf{B}_2$ determines that of the workload distribution at Q_1. Using a result of De Meyer & Teugels [21] (see also Formula (17)), Formula (38) for $\nu_1 > \nu_2$ may be rewritten as

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim (1-\rho_2)\frac{(r_1-1)\rho_2}{K_1-\rho_1}\mathbb{P}\{\mathbf{P}_2^r > \frac{t}{\rho_1-1}\}, \tag{40}$$

with $\mathbf{P}_2$ a random variable representing the busy period at Q_2 when served in isolation at unit rate. This may be heuristically explained as follows. Large-deviations arguments suggest that the most likely scenario for the workload at Q_1 to reach a large level is that class 2 generates a large amount of traffic, while class 1 *itself* shows average behavior. Specifically, suppose that a class-2 customer arrives with a large service time, so that Q_2 becomes backlogged for a long period of time. Then Q_1 is only served at unit rate, while class 1 generates traffic at an average rate $\rho_1 > 1$. Thus, the workload at Q_1 has a positive drift, and Q_1 will soon become backlogged too (if it not already is), and remain so for as long as Q_2 is backlogged. Consequently, Q_2 will experience a busy period as if it were served at unit rate. During that period, the workload at Q_1 will roughly grow at rate $\rho_1 - 1 > 0$. Only after Q_2 empties, Q_1 will start to drain again at approximately rate $r_1 - \rho_1$.

Of course, the workload at Q_1 may also build up when class 1 *itself* generates a large amount of traffic. However, this scenario is highly implausible compared to the build-up that occurs during a busy period at Q_2, because of the relatively smooth nature of class-1 traffic.

Thus, the workload at Q_1 basically behaves as that in a queue of capacity $r_1 - \rho_1$ which is fed by an On-Off source with as On- and Off-periods the busy and idle periods at Q_2, respectively, and with inflow rate $r_1 - 1 > 0$ when On, and fraction Off time $1-\rho_2$. In particular, the traffic intensity of the On-Off source equals $(r_1-1)\rho_2$. Using a result of Jelenković & Lazar [17], it may be verified that the workload behavior in this queue is exactly as indicated by Formula (40).

Remark 6.2 *Interestingly, the high service speed r_2 does not show up in either Theorem 4.1, Theorem 4.2, or Theorem 6.1 at all. In particular, the results coincide with those in [4] for the case $r_2 = 1$. This may be explained by observing that if the workload at Q_1 is large, then Q_2 operates at the lower service speed all the time, so that the high service speed is irrelevant. Similarly, the high service speed r_1 does not show up in either Theorem 5.1 or Theorem 5.2. In order for the workload at Q_1 to be large, Q_1 must have operated at the lower service speed all the time.*

6.3 Proof of Theorem 6.1

Starting point for studying the tail distribution of the workload $\mathbf{V}_1$ is again Relation (4) for its LST, but we can no longer use (5) for the term $\psi_2(s)$. The reason for this is the following. We want to let $s \to 0$, but $\delta_1(w) \to \delta_1(0) \neq 0$ for $w \to 0$ if $\rho_1 > 1$. Let us therefore take a closer look at the zeroes of $f_1(s,w)$, cf. (11). In [16] it is observed that $\frac{\mathrm{d}}{\mathrm{d}s} f_1(s,w)$ has, for real $s \geq 0$, no zero if $\rho_1 < 1$, one zero $s_0 = 0$ if $\rho_1 = 1$, and one zero $s_0 > 0$ if $\rho_1 > 1$. If $\rho_1 \geq 1$, then the point $w_0 := s_0 - \lambda_1(1 - \beta_1\{s_0\})$ is a second-order branch-point of the analytic continuation of $\delta_1(w)$, Re $w \geq 0$, into Re $w < 0$. For $\rho_2 < 1$, $\rho_1 \geq 1$, and $w \in [w_0, 0]$, the *two* zeroes of $f_1(s,w)$ in $[0, \delta_1(0)]$ will be denoted by $\epsilon_1(w)$ and $\delta_1(w)$, and such that

$\epsilon_1(w)$ maps $[w_0, 0]$ one-to-one onto $[0, s_0]$,

$\delta_1(w)$ maps $[w_0, 0]$ one-to-one onto $[s_0, \delta_1(0)]$.

If $\rho_1 \geq 1$, then (6.24) of [16] yields

$$[(1-\frac{1}{r_1})\frac{w}{\delta_2(w)} - \frac{1}{r_1}\frac{w}{\epsilon_1(w)}][\frac{1}{r_2}(\psi_2(\epsilon_1(w)) - \psi_0) - \frac{\psi_0}{r_1 r_2}\frac{1}{1-\frac{1}{r_1}-\frac{1}{r_2}}] = \tag{41}$$
$$- \quad [(1-\frac{1}{r_2})\frac{w}{\epsilon_1(w)} - \frac{1}{r_2}\frac{w}{\delta_2(w)}][\frac{1}{r_1}(\psi_1(\delta_2(w)) - \psi_0) - \frac{\psi_0}{r_1 r_2}\frac{1}{1-\frac{1}{r_1}-\frac{1}{r_2}}].$$

To determine the behavior of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$ (which eventually will give us the behavior of $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]$ for $s \downarrow 0$, hence that of $\mathbb{P}\{\mathbf{V}_1 > t\}$ for $t \to \infty$), we need to determine the behavior, for $w \uparrow 0$, of $\epsilon_1(w)$, $\delta_2(w)$ and $\psi_1(\delta_2(w))$ – the terms that appear in (41). Take $w < 0$, $w \uparrow 0$, then (cf. (18)),

$$\epsilon_1(w) + \frac{w}{\rho_1 - 1} \sim \frac{\lambda_1 C_1}{\rho_1 - 1}(\frac{-w}{\rho_1 - 1})^{\nu_1} l_1(\frac{-1}{w}), \quad w \uparrow 0. \tag{42}$$

In view of the symmetry between the two regularly-varying-tail assumptions and between the definitions of $\delta_1(w)$ and $\delta_2(w)$, it is readily seen from (18) that

$$\delta_2(w) + \frac{w}{1 - \rho_2} \sim -\frac{\lambda_2 C_2}{1 - \rho_2}(\frac{-w}{1 - \rho_2})^{\nu_2} l_2(\frac{-1}{w}), \quad w \uparrow 0. \tag{43}$$

For $\rho_2 < 1$, $\psi_1(\delta_2(w))$ is specified by Formula (6.23) of [16] (notice the symmetry with (5)),

$$\frac{1}{r_1}[\psi_1(\delta_2(w)) - \psi_0] = \frac{1}{r_1 r_2}\frac{\psi_0}{1 - \frac{1}{r_1} - \frac{1}{r_2}}[1 - \mathrm{e}^{P_1(w) - P_2(w)}], \quad \text{Re } w \leq 0. \tag{44}$$

We now turn to the study of $P_i(w)$. Compared to the study of $R_i(w)$, there is a slight difference. Now that $\rho_1 > 1$, the distribution of the busy period $\hat{\mathbf{P}}_1$ is defective: $\mathbb{P}\{\hat{\mathbf{P}}_1 < \infty\} = \frac{1}{\rho_1}$. As in (29), (30), we obtain, with $p_i(t) := \sum_{n=1}^{\infty} \frac{b_i^n}{n}\mathbb{P}\{-t < \mathbf{X}_{i1} + \ldots + \mathbf{X}_{in} < 0\}$,

$$P_2(0) - p_2(t) \sim \frac{b_2(1-\pi_2)}{1 - b_2(\frac{\pi_2}{\rho_1} + 1 - \pi_2)}\mathbb{P}\{\hat{\mathbf{P}}_2 > t\} = \frac{\rho_2}{1-\rho_2}\mathbb{P}\{\hat{\mathbf{P}}_2 > t\}.$$

Similarly,

$$P_1(0) - p_1(t) \sim \frac{b_1(1-\pi_1)}{1 - b_1(\frac{\pi_1}{\rho_1} + 1 - \pi_1)}\mathbb{P}\{\hat{\mathbf{P}}_2 > t\} = \frac{\rho_2}{1-\rho_2}\mathbb{P}\{\hat{\mathbf{P}}_2 > t\}.$$

Using the counterpart of (19) for $\hat{\mathbf{P}}_2$, and Lemma 3.1, it finally follows that (cf. (31))

$$P_1(w) - P_1(0) \sim P_2(w) - P_2(0) \sim -\frac{\lambda_2 C_2}{1-\rho_2}(\frac{-w}{1-\rho_2})^{\nu_2 - 1} l_2(\frac{-1}{w}), \quad w \uparrow 0.$$

Substituting into (44), we have

$$\psi_1(\delta_2(w)) - \psi_1(0) = \mathrm{o}(w^{\nu_2 - 1} l_2(\frac{-1}{w})), \quad w \uparrow 0. \tag{45}$$

As a brief intermezzo, we make the following observation. It follows from (45) that $\mathbb{P}\{\mathbf{V}_1 = 0, \mathbf{V}_2 > t\} = \mathrm{o}(t^{1-\nu_2} l_2(t))$, $t \to \infty$. This may be surprising in view of the fact that if Q_2 were an M/G/1 queue in isolation, then $\mathbb{P}\{\mathbf{V}_2 > t\} \sim C t^{1-\nu_2} l_2(t)$, cf. [14]. The explanation is the following. The workload at Q_1 has a positive drift $\rho_1 - 1$ when $\mathbf{V}_2 > 0$. Therefore $\mathbb{P}\{\mathbf{V}_1 = 0 | \mathbf{V}_2 > t\} = \mathrm{o}(1)$ for $t \to \infty$: when the workload at Q_2 is very large, it is highly unlikely that Q_1 is empty.

The above result for the behavior of $\psi_1(\delta_2(w))$ for $w \uparrow 0$ allows us to determine the behavior of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$. Using Relation (41) between $\psi_2(\epsilon_1(w))$ and $\psi_1(\delta_2(w))$, along with the asymptotic results (42) and (43) for $\epsilon_1(w)$ and $\delta_2(w)$, it follows after some calculations that

$$\begin{aligned} \psi_2(\epsilon_1(w)) - \psi_2(0) \quad &\sim \quad -\frac{\rho_1 - 1}{(1-b_2)r_1}\lambda_2 C_2 (\frac{-w}{1-\rho_2})^{\nu_2 - 1} l_2(\frac{-1}{w}) \mathrm{I}_{\{\nu_1 > \nu_2\}} \\ &- \quad \frac{(1-\rho_2)}{(1-b_2)r_1}\lambda_1 C_1 (\frac{w}{1-\rho_1})^{\nu_1 - 1} l_1(\frac{-1}{w}) \mathrm{I}_{\{\nu_1 < \nu_2\}}, \quad w \uparrow 0. \end{aligned}$$

Using (42) once more,

$$\begin{aligned} \psi_2(s) - \psi_2(0) \quad &\sim \quad -\frac{\rho_1 - 1}{(1-b_2)r_1}\lambda_2 C_2 (s\frac{\rho_1 - 1}{1-\rho_2})^{\nu_2 - 1} l_2(\frac{1}{s}) \mathrm{I}_{\{\nu_1 > \nu_2\}} \\ &- \quad \frac{(1-\rho_2)}{(1-b_2)r_1}\lambda_1 C_1 s^{\nu_1 - 1} l_1(\frac{1}{s}) \mathrm{I}_{\{\nu_1 < \nu_2\}}, \quad s \downarrow 0. \end{aligned} \tag{46}$$

Finally we are ready to determine the tail behavior of the workload $\mathbf{V}_1$ at Q_1. The LST of $\mathbf{V}_1$ is given by (4). The first factor in its right-hand side is the LST of the workload distribution at Q_1 when served in isolation at unit speed (the Pollaczek-Khintchine workload LST in the M/G/1 queue); this factor would give a $t^{1-\nu_1}$ tail behavior, cf. [14]. Using (46), the second factor in the right-hand side of (4) is seen to yield either a $t^{1-\nu_1}$ or a $t^{1-\nu_2}$ tail behavior. To see which of these two terms dominates, we have to distinguish between three cases: $\nu_1 < \nu_2$, $\nu_1 > \nu_2$ and $\nu_1 = \nu_2$.

Case i: $\nu_1 < \nu_2$. In this case the heavier tail of $\mathbf{B}_1$ dominates, and Formula (27) still holds,

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] - 1 \sim -\frac{\lambda_1 C_1}{K_1 - \rho_1} s^{\nu_1 - 1} l_1(\frac{1}{s}), \quad s \downarrow 0, \tag{47}$$

with $K_1 = \rho_2 + (1 - \rho_2) r_1$.

Case ii: $\nu_1 > \nu_2$. In this case the heavier tail of $\mathbf{B}_2$ dominates, resulting in

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] - 1 \sim -\frac{r_1 - 1}{K_1 - \rho_1} \lambda_2 C_2 (s\frac{\rho_1 - 1}{1 - \rho_2})^{\nu_2 - 1} l_2(\frac{1}{s}), \quad s \downarrow 0. \tag{48}$$

Case iii: $\nu_1 = \nu_2 = \nu$. In this case, addition of the right-hand sides of (47) and (48) gives the right asymptotic behavior of $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] - 1$.

The proof of the theorem is completed by yet another application of Lemma 3.1.

7 One queue heavy-tailed; other queue lighter tail; $\rho_1 > 1$; $\rho_2 < 1$

7.1 Introduction and main result

In this section we consider an extension of the model analyzed in the previous section. We show that the results obtained there, in fact continue to hold when the service time distribution at one of the queues is regularly varying, and the service distribution time at the other queue has a lighter, general tail. More precisely, we consider the following two cases. For case (i) we assume the service times at Q_1 to be regularly varying of index $-\nu_1$, $1 < \nu_1 < 2$, see (13) for the distribution function and (14) for the behavior of its LST. The service times at Q_2 have a lighter tail than those at Q_1, more precisely, we assume that $\mathbb{P}\{\mathbf{B}_2 > t\} = \mathrm{o}(t^{-\nu_1} l_1(t))$ for $t \to \infty$. For case (ii) we assume the service times at Q_2 to be regularly varying of index $-\nu_2$, $1 < \nu_2 < 2$, similar to $\mathbf{B}_1$ in case (i), and the service times at Q_1 to have a lighter tail, i.e., $\mathbb{P}\{\mathbf{B}_1 > t\} = \mathrm{o}(t^{-\nu_2} l_2(t))$ for $t \to \infty$. As in the previous section, we assume $\rho_1 > 1$ and $\rho_2 < 1$, so the lower speed at Q_2 is sufficient for stability, whereas Q_1 needs the surplus capacity from Q_2 to remain stable. We show that for case (i) Formula (37) still holds and that Formula (38) remains valid for case (ii).

Theorem 7.1 *If $\rho_1 > 1$, $\rho_2 < 1$, then $\mathbb{P}\{\mathbf{V}_1 > t\}$ is as given below:*
If $\mathbb{P}\{\mathbf{B}_1 > t\} \in \mathcal{R}_{-\nu_1}$, $1 < \nu_1 < 2$, and $\mathbb{P}\{\mathbf{B}_2 > t\} = \mathrm{o}(t^{-\nu_1} l_1(t))$, then

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \frac{1}{K_1 - \rho_1} \frac{\lambda_1 C_1}{\Gamma(2 - \nu_1)} t^{1-\nu_1} l_1(t) \sim \frac{\rho_1}{K_1 - \rho_1} \mathbb{P}\{\mathbf{B}_1^r > t\}, \quad t \to \infty. \tag{49}$$

If $\mathbb{P}\{\mathbf{B}_2 > t\} \in \mathcal{R}_{-\nu_2}$, $1 < \nu_2 < 2$, *and* $\mathbb{P}\{\mathbf{B}_1 > t\} = \mathrm{o}(t^{-\nu_2} l_2(t))$, *then*

$$\begin{aligned}\mathbb{P}\{\mathbf{V}_1 > t\} &\sim \frac{r_1 - 1}{K_1 - \rho_1} \frac{\lambda_2 C_2}{\Gamma(2-\nu_2)} (\frac{1-\rho_2}{\rho_1 - 1} t)^{1-\nu_2} l_2(t) \\ &\sim \frac{(r_1-1)\rho_2}{K_1 - \rho_1} \mathbb{P}\{\mathbf{B}_2^r > \frac{1-\rho_2}{\rho_1 - 1} t\}, \quad t \to \infty. \end{aligned} \tag{50}$$

Just as in the previous section, the tail of the service time distribution of one queue dominates the other. Therefore, the interpretation and explanation of the above results are the same as given in Subsection 6.2.

7.2 Proof of Theorem 7.1

The proof is similar to that of Theorem 6.1. Again we use Relation (4) to determine the tail distribution of $\mathbf{V}_1$. Also, because $\rho_1 \geq 1$ we have to use Relation (6.24) of [16] (see (41)) to find the behavior of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$. This means that we have to determine for both cases the behavior of $\epsilon_1(w)$, $\delta_2(w)$ and $\psi_1(\delta_2(w))$ for $w \uparrow 0$.
Before going into detail, we derive some results which we need in the remainder of the proof. If we assume for the distribution of a generic service time $\mathbf{B}$ that $\mathbb{P}\{\mathbf{B} > t\} = \mathrm{o}(t^{-\nu} l(t))$ for $t \to \infty$, $1 < \nu < 2$, then applying Lemma 3.1 with $D = 0$ gives the following behavior of its LST,

$$\beta\{s\} = 1 - \beta s + \mathrm{o}(s^{\nu} l(\frac{1}{s})), \quad s \downarrow 0, \tag{51}$$

with β denoting the first moment of $\mathbf{B}$. Now consider an isolated stable M/G/1 queue with generic service time $\mathbf{B}$ and arrival rate λ. Rewriting the well-known formula for the LST of the workload $\mathbf{V}^1$, and using (51), we obtain

$$1 - \mathbb{E}[\mathrm{e}^{-s\mathbf{V}^1}] = \mathrm{o}(s^{\nu-1} l(\frac{1}{s})), \quad s \downarrow 0. \tag{52}$$

Case i:
Similar to the proof of Theorem 6.1 (see (42)), we have

$$\epsilon_1(w) + \frac{w}{\rho_1 - 1} \sim \frac{\lambda_1 C_1}{\rho_1 - 1} (\frac{-w}{\rho_1 - 1})^{\nu_1} l_1(\frac{-1}{w}), \quad w \uparrow 0.$$

The behavior of $\delta_2(w)$ is determined as follows. According to the definition of $f_2(s, w)$ in (12),

$$\delta_2(w) = \lambda_2(1 - \beta_2\{\delta_2(w)\}) - w.$$

Using (51) for $\mathbf{B}_2$, it follows that

$$\delta_2(w) = \frac{-w}{1-\rho_2} + \mathrm{o}((\delta_2(w))^{\nu_1} l_1(\frac{-1}{w})), \quad w \uparrow 0,$$

and hence

$$\delta_2(w) = \frac{-w}{1-\rho_2} + \mathrm{o}((-w)^{\nu_1} l_1(\frac{-1}{w})), \quad w \uparrow 0. \tag{53}$$

It remains to determine the behavior of $\psi_1(\delta_2(w))$. Observe that $\psi_1(s) - \psi_1(0)$ is the LST of $\mathbb{P}\{\mathbf{V}_1 = 0, \mathbf{V}_2 > t\}$. Obviously,

$$\mathbb{P}\{\mathbf{V}_1 = 0, \mathbf{V}_2 > t\} \leq \mathbb{P}\{\mathbf{V}_2 > t\}.$$

Recall that Q_2 is guaranteed to receive a minimum service rate of 1 when backlogged. Hence, the following inequality holds,

$$\mathbb{P}\{\mathbf{V}_2 > t\} \leq \mathbb{P}\{\mathbf{V}_2^1 > t\}.$$

Using (52) it is readily seen that

$$\psi_1(\delta_2(w)) - \psi_1(0) = \mathrm{o}(w^{\nu_1 - 1} l_1(\frac{-1}{w})), \quad w \uparrow 0. \tag{54}$$

Now we have to substitute (42), (53) and (54) into (41) to find the behavior of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$. After some calculations we obtain

$$\psi_2(\epsilon_1(w)) - \psi_2(0) \sim -\frac{1-\rho_2}{(1-b_2)r_1}\lambda_1 C_1 (\frac{-w}{\rho_1 - 1})^{\nu_1 - 1} l_1(\frac{-1}{w}), \quad w \uparrow 0.$$

Using (42) once again results in

$$\psi_2(s) - \psi_2(0) \sim -\frac{1-\rho_2}{(1-b_2)r_1}\lambda_1 C_1 s^{\nu_1 - 1} l_1(\frac{1}{s}), \quad s \downarrow 0.$$

Now we have gathered all the ingredients necessary in (4) to find the tail behavior of $\mathbf{V}_1$, except the term $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}]$. Observing that this is the same factor as in the proof of Theorem 6.1, we obtain Formula (47).

Case ii:
Now we can use (43) as given in the proof of Theorem 6.1,

$$\delta_2(w) + \frac{w}{1-\rho_2} \sim -\frac{\lambda_2 C_2}{1-\rho_2}(\frac{-w}{1-\rho_2})^{\nu_2} l_2(\frac{-1}{w}), \quad w \uparrow 0.$$

To find the behavior of $\epsilon_1(w)$, we can apply the same method as we used to determine the behavior of $\delta_2(w)$ in case (i). According to the definition of $f_1(s, w)$ (see (11)),

$$\epsilon_1(w) = \lambda_1(1 - \beta_1\{\epsilon_1(w)\}) + w. \tag{55}$$

Using a second-order Taylor approximation for $\epsilon_1(w)$ around $w = 0$ (note that $\epsilon_1(0) = 0$) together with (51) in (55), we obtain

$$\epsilon_1(w) = \frac{-w}{\rho_1 - 1} + \mathrm{o}(w^{\nu_2} l_2(-\frac{1}{w})), \quad w \uparrow 0. \tag{56}$$

The behavior of $\psi_1(\delta_2(w))$ is the same as in the proof of Theorem 6.1 (see (45)),

$$\psi_1(\delta_2(w)) - \psi_1(0) = \mathrm{o}(w^{\nu_2 - 1} l_2(-\frac{1}{w})), \quad w \uparrow 0.$$

Again we have to substitute (43), (45), and (56) into (41) to find the behavior of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$. After some calculations we obtain

$$\psi_2(\epsilon_1(w)) - \psi_2(0) \sim -\frac{\rho_1 - 1}{(1-b_2)r_1}\lambda_2 C_2 (\frac{-w}{1-\rho_2})^{\nu_2 - 1} l_2(\frac{-1}{w}), \quad w \uparrow 0.$$

Using (56) once again gives

$$\psi_2(s) - \psi_2(0) \sim -\frac{\rho_1 - 1}{(1-b_2)r_1}\lambda_2 C_2 \left(s\frac{\rho_1 - 1}{1-\rho_2}\right)^{\nu_2 - 1} l_2(\frac{1}{s}), \quad s \downarrow 0.$$

Observing that the term $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}]$ satisfies Formula (52), we obtain Formula (48).

Acknowledgment The authors gratefully acknowledge Predrag Jelenković for several useful discussions.

References

[1] Anantharam, V. (1999). Scheduling strategies and long-range dependence. *Queueing Systems* **33**, 73–89.

[2] Arvidsson, A., Karlsson, P. (1999). On traffic models for TCP/IP. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 457–466.

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

[4] Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Coupled processors with regularly varying service times. In: *Proc. Infocom 2000 Conference*, Tel-Aviv, Israel, 157–164.

[5] Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. CWI Report PNA-R0016. *Queueing Systems*, to appear.

[6] Borst, S.C., Boxma, O.J., Van Uitert, M.J.G. (2001). Two coupled queues with heterogeneous traffic. In: *Proc. ITC-17*, 1003–1014.

[7] Borst, S.C., Mandjes, M., Van Uitert, M.J.G. (2001). Generalized Processor Sharing queues with heterogeneous traffic classes. CWI Report PNA-R0106. Shortened version in: *Proc. Infocom 2002 Conference*, New York NY, USA, to appear.

[8] Borst, S.C., Zwart, A.P. (2000). A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows. SPOR-Report 2000-04, Eindhoven University of Technology. *J. Appl. Prob.*, to appear.

[9] Boxma, O.J., Deng, Q., Zwart, A.P. (1999). Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers. Technical Memorandum COSOR 99-20, Eindhoven University of Technology. *Queueing Systems*, to appear.

[10] Boxma, O.J., Dumas, V. (1998). Fluid queues with heavy-tailed activity period distributions. *Computer Communications* **21**, 1509–1529.

[11] Boxma, O.J., Dumas, V. (1998). The busy period in the fluid queue. *Perf. Eval. Review* **26**, 100–110.

[12] Boxma, O.J., Kurkova, I.A. (2000). The M/M/1 queue in a heavy-tailed random environment. *Statistica Neerlandica* **54**, 221–236.

[13] Cao, J., Ramanan, K. (2001). A Poisson limit for the unfinished work of superposed point processes. Technical Memorandum. Bell Laboratories, Lucent Technologies, Murray Hill, NJ.

[14] Cohen, J.W. (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Prob.* **10**, 343–353.

[15] Cohen, J.W. (1982). *The Single Server Queue* (North-Holland Publ. Cy., Amsterdam; revised edition).

[16] Cohen, J.W., Boxma, O.J. (1983). *Boundary Value Problems in Queueing System Analysis* (North-Holland Publ. Cy., Amsterdam).

[17] Jelenković, P.R., Lazar, A.A. (1999). Asymptotic results for multiplexing subexponential on-off processes. *Adv. Appl. Prob.* **31**, 394–421.

[18] Jelenković, P.R., Momčilović, P. (2001). Network multiplexer with Generalized Processor Sharing and heavy-tailed on-off flows. In: *Proc. ITC-17*, 719–730.

[19] Kotopoulos, C., Likhanov, N., Mazumdar, R.R. (2001). Asymptotic analysis of the GPS system fed by heterogeneous long-tailed sources. In: *Proc. Infocom 2001 Conference*, Anchorage AK, USA, 299–308.

[20] Mandjes, M., Kim, J.-H. (2001). Large deviations for small buffers: an insensitivity result. *Queueing Systems* **37**, 349–362.

[21] De Meyer, A., Teugels, J.L. (1980). On the asymptotic behaviour of the distribution of the busy period and service time in M/G/1. *J. Appl. Prob.* **17**, 802–813.

[22] Parekh, A.K., Gallager, R.G. (1993). A generalized processor sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM *Trans. Netw.* **1**, 344–357.

[23] Parekh, A.K., Gallager, R.G. (1994). A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. IEEE/ACM *Trans. Netw.* **2**, 137–150.

[24] Park, K., Willinger, W. (eds.) (2000). *Self-Similar Network Traffic and Performance Evaluation* (Wiley, New York).

[25] Sigman, K. (ed.) (1999). *Queueing Systems* **33**. Special Issue on Queues with Heavy-Tailed Distributions.

[26] Shwartz, A., Weiss, A. (1995). *Large Deviations for Performance Analysis* (Chapman & Hall, London).

[27] Sutton, W.G.L. (1934). The asymptotic expansion of a function whose operational equivalent is known. *J. London Math. Soc.* **9**, 131–137.

[28] Zwart, A.P., Borst, S.C., Mandjes, M. (2000). Exact asymptotics for fluid queues fed by multiple heavy-tailed On-Off flows. *Ann. Appl. Prob.*, to appear. Shortened version in: *Proc. Infocom 2001 Conference*, Anchorage AK, USA, 279–288.