Centrum voor Wiskunde en Informatica

Two Coupled Queues with Heterogeneous Traffic

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

Probability, Networks and Algorithms (PNA)

PNA-R0107 June 30, 2001

Report PNA-R0107
ISSN 1386-3711

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

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

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

Two Coupled Queues with Heterogeneous Traffic

Sem Borst[*,**,†], Onno Boxma[*,**], Miranda van Uitert[*]
email: sem@cwi.nl, boxma@win.tue.nl, miranda@cwi.nl

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

[**] *Department of Mathematics & Computing 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 with two heterogeneous traffic classes, one having light-tailed characteristics, the other one exhibiting heavy-tailed properties. When both classes are backlogged, the two corresponding queues are each served at a certain nominal rate. However, when one queue empties, the service rate for the other class increases. This dynamic sharing of surplus service capacity is reminiscent of the Generalized Processor Sharing (GPS) discipline. GPS-based scheduling algorithms, such as Weighted Fair Queueing, provide a candidate implementation mechanism for achieving differentiated Quality-of-Service in a DiffServ architecture.

We characterize the asymptotic workload behavior of both traffic classes. The tail of the workload distribution of the *heavy-tailed* class is asymptotically equivalent to that of the heavy-tailed class in isolation – but with its nominal service rate inflated by the slack capacity of the light-tailed class. For the *light-tailed* class, we show a sharp dichotomy in the qualitative behavior, depending on whether its load exceeds its nominal service rate or not. In underload scenarios, the tail of its workload distribution is equivalent to that of the light-tailed class in isolation, multiplied with a certain pre-factor. The pre-factor represents the probability that the heavy-tailed class is backlogged long enough for the light-tailed class to build up a large workload. This provides a measure for the extent to which the light-tailed class benefits from sharing surplus capacity with the heavy-tailed class. In contrast, in overload situations, the light-tailed class is adversely affected by the heavy-tailed class, and inherits its traffic characteristics.

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

Keywords and Phrases: coupled processors, Generalised Processor Sharing (GPS), heavy-tailed traffic, queue-length asymptotics, regular variation.

Note: Work carried out under the project PNA2.1 "Communication and Computer Networks".

1 Introduction

The next-generation Internet is expected to support a wide variety of services, such as voice, video, and data applications. Voice and video communications induce far more stringent Quality-of-Service (QoS) requirements than the typical sort of data applications which currently account for the bulk of the Internet traffic. The integration of heterogeneous services thus raises the need for differentiated QoS, catering to the specific requirements of the various traffic flows.

One potential approach to achieve service differentiation is through the use of discriminatory scheduling algorithms, which distinguish between packets of various traffic streams. Because of scalability issues, it is practically infeasible though to manipulate packets at the granularity level of individual traffic flows in the core of any large-scale high-speed network. To avoid these complexity problems, traffic flows may instead be aggregated into a small number of classes with roughly similar features, with scheduling mechanisms acting at the coarser level of aggregate streams. With a little simplification, the majority of applications may for example be broadly categorized into just two classes, one containing *streaming* traffic (e.g. audio and video communications), the other one comprising *elastic* traffic (e.g. file transfers). This is a crucial element of the DiffServ proposal [16], which defines the EF class (Expedited Forwarding) for real-time traffic, and the AF class (Assured Forwarding) for best-effort type of traffic.

In view of the real-time requirements, it is desirable that streaming applications receive some sort of priority over elastic traffic, at least over short time scales. Strict priority scheduling may however not be ideal, since it may lead to starvation of the best-effort traffic. Even temporary starvation effects may cause end-to-end flow control mechanisms such as TCP to suffer a severe degradation in throughput performance. The Generalized Processor Sharing (GPS) discipline provides a potential mechanism for implementing priority scheduling in a tunable way, with strict priority scheduling as an extreme option [22], [23]. In GPS-based scheduling algorithms, such as Weighted Fair Queueing, the link capacity is shared in proportion to certain class-defined weight factors. By setting the weight factor for the best-effort class relatively low, one can still provide some degree of priority to the streaming applications, while avoiding starvation of the elastic traffic.

As a pre-requisite for achieving differentiated QoS, scheduling algorithms must be able to cope with heavy-tailed traffic phenomena. Extensive traffic measurements in high-speed communication networks have indicated that bursty traffic behavior may extend over a wide range of time scales, and may manifest itself in long-range dependence and self-similarity, see [19], [25]. The occurrence of these phenomena is commonly attributed to extreme variability and heavy-tailed characteristics in the traffic patterns, see [3], [15], [29]. These observations have triggered a strong interest in queueing models with heavy-tailed traffic processes, as reflected in the survey paper [11] as well as the recent publications [24], [27].

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 scheduling discipline, and the effect of flow control mechanisms such as TCP, see for instance [2].

In the present paper, we specifically examine the potential role of GPS-related scheduling mechanisms in protecting light-tailed traffic flows from the impact of heavy-tailed traffic processes. We consider a queueing model with two traffic classes, one having light-tailed (exponential) service requests, the other one exhibiting heavy-tailed characteristics. 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 when

the queue of class 2 empties and vice versa. We investigate the extent to which the workload behavior of each class is affected by the interaction with the other class.
The paper fits into two strands of research: (i) GPS queues with heavy-tailed traffic flows; (ii) the interplay of light-tailed and heavy-tailed traffic processes. (i) References [5]-[8] and [18] consider the queueing behavior of heavy-tailed traffic flows under the GPS discipline. The results show a stark contrast in qualitative behavior, depending on the relative values of the weight parameters. For certain weight combinations, an individual heavy-tailed flow is effectively served at a constant rate, and essentially immune from excessive activity of 'heavier-tailed' flows. For other weight parameters, however, a flow may be strongly affected by the activity of heavier-tailed flows, and inherit their traffic characteristics.
Most of these results were obtained by deriving probabilistic lower and upper bounds for the workload of an individual flow, and showing that these bounds asymptotically coincide. In [7] and in Section 6 of [8], a different approach is followed, adopting transform techniques. This approach builds on the detailed analysis of the joint workload distribution in a related coupled-processors model presented in [14] using the boundary value method. In the present paper we follow a similar approach, with the difference that we now focus on a light-tailed traffic flow. (The former, probabilistic approach is used in a companion paper [9], yielding qualitatively similar results.)
(ii) Several recent studies have revealed an interesting dichotomy in the interplay of exponential and heavy-tailed traffic processes. E.g., [13] considers an M/M/1 queue which alternates between exponentially distributed periods of high service speed and heavy-tailed periods of low service speed. If the offered traffic load is smaller than the low speed, then the workload tail is *semi-exponential*; otherwise it inherits the heavy-tailed behavior of the (residual) low-speed periods. A related phenomenon is observed in [10] 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 [30].
The main results of the present paper are as follows. First of all, we show that the workload tail of the exponential class 1 is also semi-exponential in case the traffic loads ρ_1 and ρ_2 of classes 1 and 2 are below the nominal service rate 1. In addition, the workload tail is found to be purely exponential if $\rho_1 < 1$ and $\rho_2 > 1$, and heavy-tailed if $\rho_1 > 1$ and $\rho_2 < 1$. The workload tail of the heavy-tailed class 2 on the other hand is shown to be asymptotically equivalent to that of the heavy-tailed class in isolation – but with its nominal service rate inflated by the slack capacity of the exponential class.
The paper is organized in the following way. In Section 2, we present a detailed model description. Section 3 contains some known results that will be used in the subsequent analysis. In Sections 4 and 5, we obtain the workload asymptotics for the exponential class; Section 4 focuses on the case $\rho_1 < 1$, and Section 5 is devoted to the case $\rho_1 > 1$. The boundary case $\rho_1 = 1$ is rather subtle, and will not be considered here. The workload asymptotics for the heavy-tailed class are derived in Section 6.

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. We assume that class-1 traffic has light-tailed characteristics, whereas class-2 traffic shows heavy-tailed features. Specifically, the service requirement of class-1 customers

is exponentially distributed with mean β_1. Class-2 customers require an amount of service whose distribution $B_2(\cdot)$ is regularly varying of index $-\nu_2$, with $1 < \nu_2 < 2$. With minor modifications, the analysis may be extended to values $\nu_2 > 2$.
There are separate queues maintained for each class. When both classes are backlogged, the queues are each served at unit rate. However, the service rate for class i increases to $r_i \geq 1$ when the queue of the other class is empty. Thus, the two queues are coupled through the mechanism for sharing surplus service capacity. For $r_i = 1/\phi_i$, with $\phi_1 + \phi_2 = 1$, the system may equivalently be viewed as a two-class GPS queue 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 however we concentrate on the more general case $1/r_1 + 1/r_2 \neq 1$; see also Remark 4.2 below.
Throughout the paper, we assume that the ergodicity conditions are satisfied. These conditions are discussed in Section III.3.7 of [14]. 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.
We conclude the section with introducing some notation that will be used throughout the paper. 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 denote that $\lim_{t\to\infty} f(t)/g(t) = 1$, or equivalently, $f(t) = g(t)(1 + \mathrm{o}(1))$ as $t \to \infty$. Similarly, in considering two Laplace transforms $h(s)$ and $k(s)$ for $s \to 0$ (or for $s \downarrow s_1$), we write $h(s) \sim k(s)$ when their ratio tends to 1 in the limit.
The fact that $B_2(\cdot)$ is regularly varying of index $-\nu_2$ means that [4]

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

with $l_2(\cdot)$ some slowly varying function, i.e., $\lim_{t\to\infty} l_2(\eta t)/l_2(t) = 1$, $\eta > 1$. 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

$$\mathbb{P}\{\mathbf{B}_2^r > t\} \sim \frac{C_2}{\beta_2 \Gamma(2-\nu_2)} t^{1-\nu_2} l_2(t), \quad t \to \infty. \tag{2}$$

3 Preliminary results

In this section we review some preliminary results which will play a crucial role in the analysis. Denote by $\mathbf{V}_i$ a random variable representing the class-i workload in steady state. For $c > 0$, denote by $\mathbf{V}_i^c$ a random variable representing the steady-state workload of class 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 [14] (in the sequel we omit Chapter III.3 when referring to formulas from [14]), for Re $s \geq 0$,

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

with, since we have assumed the class-1 service times to be exponentially distributed,

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}] = \frac{(1-\rho_1)(1+\beta_1 s)}{1-\rho_1+\beta_1 s}. \tag{4}$$

It should be noted that the denominator has a pole $s = s_1 := \lambda_1 - 1/\beta_1 < 0$; this pole s_1 will play an essential role in the analysis of the next section.
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

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}] = \mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}][1 - \frac{r_1-1}{1-\rho_1}(\psi_2(s) - \psi_2(0))]. \tag{5}$$

We now focus on the function $\psi_2(s)$. According to Formulas (6.21), (6.22), and (6.23) of [14],

$$\psi_2(\delta_1(w)) - \psi_0 = \frac{1}{r_1}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1 - \mathrm{e}^{-R_1(w)+R_2(w)}], \quad \text{Re } w \geq 0, \tag{6}$$

and

$$\psi_1(\delta_2(w)) - \psi_0 = \frac{1}{r_2}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1 - \mathrm{e}^{P_1(w)-P_2(w)}], \quad \text{Re } w \leq 0, \tag{7}$$

with

$$\psi_0 = \mathrm{e}^{-P_1(0)-R_2(0)}. \tag{8}$$

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

$$R_i(w) := \sum_{n=1}^{\infty} \frac{b_i^n}{n}\mathbb{E}[\mathrm{e}^{-w\sigma_n^{(i)}}\mathrm{I}_{\{\sigma_n^{(i)}>0\}}], \quad \text{Re } w \geq 0, \tag{9}$$

and

$$P_i(w) := \sum_{n=1}^{\infty} \frac{b_i^n}{n}\mathbb{E}[\mathrm{e}^{-w\sigma_n^{(i)}}\mathrm{I}_{\{\sigma_n^{(i)}<0\}}], \quad \text{Re } w \leq 0, \tag{10}$$

with

$$b_1 := \rho_1(1-\frac{1}{r_2}) + \frac{\rho_2}{r_2}, \tag{11}$$

$$b_2 := \rho_2(1-\frac{1}{r_1}) + \frac{\rho_1}{r_1}, \tag{12}$$

and for $i = 1, 2$,

$$\sigma_n^{(i)} := \mathbf{X}_{i1} + \ldots + \mathbf{X}_{in}, \tag{13}$$

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

$$\mathbf{X}_{11} = \begin{cases} \hat{\mathbf{P}}_1 & w.p.\ \pi_1 := \frac{\rho_1}{b_1}(1-\frac{1}{r_2}), \\ -\hat{\mathbf{P}}_2 & w.p.\ 1-\pi_1 = \frac{\rho_2}{b_1 r_2}, \end{cases} \tag{14}$$

$$\mathbf{X}_{21} = \begin{cases} \hat{\mathbf{P}}_1 & w.p.\ \pi_2 := \frac{\rho_1}{b_2 r_1}, \\ -\hat{\mathbf{P}}_2 & w.p.\ 1-\pi_2 = \frac{\rho_2}{b_2}(1-\frac{1}{r_1}). \end{cases} \tag{15}$$

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

$$f_1(s, w) := \lambda_1(1 - \beta_1\{s\}) - s + w = \frac{\rho_1 s}{1 + \beta_1 s} - s + w, \tag{16}$$

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

$$f_2(s, w) := \lambda_2(1 - \beta_2\{s\}) - s - w. \tag{17}$$

The following two lemmas 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 $s \downarrow s_1 = \lambda_1 - 1/\beta_1$, respectively (remember that s_1 is the pole of the LST of $\mathbf{V}_1^1$, cf. (4)).

The next result is formulated in Lemma 2.2 in [12] as an extension of Theorem 8.1.6 in [4].

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* $s \downarrow 0$*.*

The next lemma is established in [28]. It relates the asymptotic behavior of a function $f(t)$ for $t \to \infty$ and that of its Laplace transform $\phi(s)$ for s near its poles.

Lemma 3.2 *Define for* $d > 0$*,*

$$f(t) = \frac{1}{2\pi i} \int_{d-i\infty}^{d+i\infty} \mathrm{e}^{st} \phi(s) \mathrm{d}s,$$

where $s = x + iy$ *and the path of integration is the straight line* $x = d$*, chosen so that* $\phi(s)$ *is analytic for* $x \geq d$*. If*

(i) $\phi(s)$ *is analytic for* $x \geq a - \delta$ *(*$\delta > 0$*), except at* k *points* $s_1, \ldots, s_p, \ldots, s_k$ *on* $x = a$*;*

(ii) *near each such point* s_p*, we have*

$$(s - s_p)\phi(s) = \sum_{n=0}^{\infty} a_{np}(s - s_p)^n + (s - s_p)^{\gamma_p} \sum_{n=0}^{\infty} b_{np}(s - s_p)^n,$$

where $0 < \gamma_p < 1$*, and the series converge for* $|s - s_p| < l$ *(*$l > 0$*);*

(iii) *$\phi(s) \to 0$ as $y \to \pm\infty$, uniformly in x for $a - \delta \leq x \leq d$ $(d > a)$, and in such a manner that $\int |\phi(s)| \mathrm{d}y$ converges at $y = \pm\infty$,*

then,

$$f(t) \sim \sum_{p=1}^{k} \mathrm{e}^{s_p t} \left(a_{0p} + \frac{\sin(\pi\gamma_p)}{\pi} \sum_{n=0}^{\infty} (-1)^n b_{np} \Gamma(\gamma_p + n) t^{-\gamma_p - n} \right), \qquad t \to \infty.$$

4 Workload asymptotics for class 1 in underload

In this section we derive the asymptotic behavior of $\mathbb{P}\{\mathbf{V}_1 > t\}$ for the case $\rho_1 < 1$. We first assume that also $\rho_2 < 1$. Below we will consider the case $\rho_2 > 1$.

Theorem 4.1 *If $\rho_1 < 1$, $\rho_2 < 1$, then*

$$\begin{aligned} \mathbb{P}\{\mathbf{V}_1 > t\} &\sim \rho_1 \mathrm{e}^{(\lambda_1 - 1/\beta_1)t} \frac{1}{K_2 - \rho_2} \frac{\lambda_2 C_2}{\Gamma(2-\nu_2)} \left(\frac{\rho_1(1-\rho_2)}{1-\rho_1} t \right)^{1-\nu_2} l_2(t) \\ &\sim \mathbb{P}\{\mathbf{V}_1^1 > t\} \frac{\rho_2}{K_2 - \rho_2} \mathbb{P}\{\mathbf{B}_2^r > \frac{\rho_1(1-\rho_2)}{1-\rho_1} t\}, \qquad t \to \infty, \end{aligned} \tag{18}$$

with $K_2 := \rho_1 + (1-\rho_1) r_2$ representing the average service rate for class 2 when continuously backlogged.

Before giving the proof of the above theorem, we first provide an intuitive interpretation. Invoking a result of Pakes [21], Formula (18) may be rewritten as

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim \mathbb{P}\{\mathbf{V}_1^1 > t\} \mathbb{P}\{\mathbf{V}_2^{K_2} > \frac{1-\rho_2}{\hat{\rho}_1 - 1} t\}, \qquad t \to \infty, \tag{19}$$

with $\hat{\rho}_1 := 1/\rho_1 > 1$. To understand the above formula, it is useful to draw a comparison with the workload $\mathbf{V}_1^1$ of class 1 when 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' activity. Specifically, class 1 must essentially behave 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 class 2 is continuously backlogged while class 1 shows deviant behavior. In fact, given that the class-1 workload reaches a large level, class 2 is constantly backlogged with overwhelming probability, since otherwise class 1 should show even greater anomalous activity. The probability of that happening is negligibly small compared to that of class 2 being continuously backlogged, because of the highly bursty nature of class-2 traffic.
Note that the normal drift in the workload of class 2 is $\rho_2 - 1 < 0$. Thus, in order for class 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 class-2 workload is at least $\frac{1-\rho_2}{\hat{\rho}_1 - 1} x$. During that prior period, class 1 shows average behavior, and does not receive any surplus capacity from class 2, so that queue 1 is empty a fraction $1 - \rho_1$ of the time. Thus, the average service rate for class 2 during that period is $K_2 = \rho_1 + (1-\rho_1) r_2$. Hence, the probability that class 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 (19).

We now give the formal proof of Theorem 4.1. The approach may be outlined as follows. We are going to apply Lemma 3.2 with $f(t) = \mathbb{P}\{\mathbf{V}_1 > t\}$, and hence with Laplace transform $\phi(s) = (1-\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}])/s$. From Equations (3) and (4), we see that $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]$ has a singularity in $s = s_1 = \lambda_1 - 1/\beta_1 < 0$. The fact that the workload of class 1 cannot exceed that when served in isolation implies $\mathbb{P}\{\mathbf{V}_1 > t\} \leq \mathbb{P}\{\mathbf{V}_1^1 > t\}$, so that $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]$ does not have any singularities right of the singularity $s = s_1$ of $\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1^1}]$. In applying Lemma 3.2 we thus need to consider

$$(s-s_1)\phi(s) = (s-s_1)\frac{1-\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]}{s}, \tag{20}$$

and determine the series expansion for s near s_1.
Substituting (5) into (20), we obtain

$$\begin{aligned}(s-s_1)\frac{1-\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]}{s} &= \rho_1 + (r_1-1)\frac{1+\beta_1 s}{\beta_1 s}(\psi_2(s)-\psi_2(0)) \\ &= \rho_1 + (r_1-1)\frac{1+\beta_1 s}{\beta_1 s}(\psi_2(s_1)-\psi_2(0)) \\ &+ (r_1-1)\frac{1+\beta_1 s}{\beta_1 s}(\psi_2(s)-\psi_2(s_1)).\end{aligned} \tag{21}$$

Below we will show that $\psi_2(s_1) - \psi_2(0) = (1-\rho_1)/(r_1-1)$. But first we turn to the most involved part of the proof: determining the series expansion of the term $\psi_2(s) - \psi_2(s_1)$ in (21) for s near s_1. Note that we cannot use (6), since

$$w = \delta_1^{-1}(s) = s - \frac{\rho_1 s}{1+\beta_1 s} = \frac{\beta_1 s}{1+\beta_1 s}(s-s_1) \sim \frac{\rho_1 - 1}{\rho_1}(s-s_1) \uparrow 0 \tag{22}$$

for $s \downarrow s_1$, while (6) is only valid for $\mathrm{Re}\, w \geq 0$.
In the appendix we prove that for $s \downarrow s_1$,

$$\psi_2(s) - \psi_0 \sim \frac{1}{r_1}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1 - \mathrm{e}^{P_1(\delta_1^{-1}(s))-P_2(\delta_1^{-1}(s))}\frac{r_1}{(r_1-1)r_2}]. \tag{23}$$

Using (8), (22) and (23), we obtain, for $s \downarrow s_1$,

$$\begin{aligned}\psi_2(s) - \psi_2(s_1) &= \psi_2(s) - \psi_0 - [\psi_2(s_1) - \psi_0] \\ &\sim \frac{1}{(r_1-1)r_2}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[\mathrm{e}^{P_1(0)-P_2(0)} - \mathrm{e}^{P_1(\delta_1^{-1}(s))-P_2(\delta_1^{-1}(s))}] \\ &= \frac{1}{(r_1-1)r_2}\frac{\mathrm{e}^{-P_2(0)-R_2(0)}}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1 - \mathrm{e}^{P_1(\delta_1^{-1}(s))-P_1(0)-(P_2(\delta_1^{-1}(s))-P_2(0))}].\end{aligned} \tag{24}$$

Note that

$$P_i(0) + R_i(0) = \sum_{n=1}^{\infty}\frac{b_i^n}{n} = -\ln(1-b_i), \tag{25}$$

and $\mathrm{e}^x = 1 + x + \mathrm{O}(x^2)$ for x small. Hence, for $s \downarrow s_1$,

$$\psi_2(s) - \psi_2(s_1) \sim \frac{1}{(r_1-1)r_2}\frac{1-b_2}{1-\frac{1}{r_1}-\frac{1}{r_2}}[P_2(\delta_1^{-1}(s)) - P_2(0) - (P_1(\delta_1^{-1}(s)) - P_1(0))]. \tag{26}$$

We now determine the behavior of $P_2(w)-P_2(0)-(P_1(w)-P_1(0))$ for $w \uparrow 0$, which corresponds to $s \downarrow s_1$. It is easily seen that $P_i(w)$ is the LST of

$$p_i(t) := \sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathbb{P}\{-t < \sigma_n^{(i)} < 0\}.$$

Using Equations (13), (14), (15), we have

$$P_i(0) - p_i(t) = \sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathbb{P}\{-\sigma_n^{(i)} > t\} =$$

$$\sum_{n=1}^{\infty} \frac{b_i^n}{n} \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \pi_i^k (1-\pi_i)^{n-k} \mathbb{P}\{\sum_{j=1}^{n-k} \hat{\mathbf{P}}_{2,j} - \sum_{j=1}^{k} \hat{\mathbf{P}}_{1,j} > t\}. \tag{27}$$

From Formula (6.5) of [14], it follows that

$$\mathbb{E}[\mathrm{e}^{w\hat{\mathbf{P}}_2}] = \frac{1-\beta_2\{\delta_2(w)\}}{\beta_2 \delta_2(w)}, \quad \text{Re } w \leq 0. \tag{28}$$

Since $\mathbb{P}\{\mathbf{B}_2 > t\}$ is regularly varying of index $-\nu_2$, we have, using Lemma 3.1,

$$1 - \frac{1-\beta_2\{s\}}{\beta_2 s} \sim \frac{C_2}{\beta_2} s^{\nu_2 - 1} l_2(\frac{1}{s}), \quad s \downarrow 0. \tag{29}$$

As derived in Formula (45) of [7],

$$\delta_2(w) \sim \frac{-w}{1-\rho_2} \left[1 - \frac{\lambda_2 C_2}{1-\rho_2} \left(\frac{-w}{1-\rho_2} \right)^{\nu_2 - 1} l_2(\frac{-1}{w}) \right], \quad w \uparrow 0. \tag{30}$$

Using Equations (28), (29), and (30),

$$1 - \mathbb{E}[\mathrm{e}^{w\hat{\mathbf{P}}_2}] \sim \frac{C_2}{\beta_2} \left(\frac{-w}{1-\rho_2} \right)^{\nu_2 - 1} l_2(\frac{-1}{w}), \quad w \uparrow 0. \tag{31}$$

Applying Lemma 3.1, we find that Relation (31) gives

$$\mathbb{P}\{\hat{\mathbf{P}}_2 > t\} \sim \frac{C_2}{\beta_2 \Gamma(2-\nu_2)} \left((1-\rho_2)t\right)^{1-\nu_2} l_2(t), \quad t \to \infty, \tag{32}$$

indicating that $\mathbb{P}\{\hat{\mathbf{P}}_2 > t\}$ is regularly varying of index $1-\nu_2$.
The latter fact, in combination with $\mathbb{P}\{\hat{\mathbf{P}}_1 < \infty\} = 1$, implies (cf. [4]):

$$\mathbb{P}\{\sum_{j=1}^{n-k} \hat{\mathbf{P}}_{2,j} - \sum_{j=1}^{k} \hat{\mathbf{P}}_{1,j} > t\} \sim \mathbb{P}\{\sum_{j=1}^{n-k} \hat{\mathbf{P}}_{2,j} > t\} \sim (n-k)\mathbb{P}\{\hat{\mathbf{P}}_2 > t\}, \quad t \to \infty.$$

Using the above relation in (27), we obtain, for $t \to \infty$,

$$\begin{aligned} P_i(0) - p_i(t) &\sim \sum_{n=1}^{\infty} \frac{b_i^n}{n} \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \pi_i^k (1-\pi_i)^{n-k} (n-k) \mathbb{P}\{\hat{\mathbf{P}}_2 > t\} \\ &= (1-\pi_i) \sum_{n=1}^{\infty} b_i^n \sum_{k=0}^{n-1} \left(\begin{array}{c} n-1 \\ k \end{array} \right) \pi_i^k (1-\pi_i)^{n-k-1} \mathbb{P}\{\hat{\mathbf{P}}_2 > t\} \\ &= (1-\pi_i) \sum_{n=1}^{\infty} b_i^n \mathbb{P}\{\hat{\mathbf{P}}_2 > t\} \\ &= \frac{(1-\pi_i)b_i}{1-b_i} \mathbb{P}\{\hat{\mathbf{P}}_2 > t\}. \end{aligned} \tag{33}$$

Applying Lemma 3.1 (or rather a minor adaptation, since that lemma is formulated in terms of non-negative random variables), we deduce from (32) and (33):

$$P_i(w) - P_i(0) \sim -\frac{(1-\pi_i)b_i}{1-b_i}\frac{C_2}{\beta_2}\left(\frac{-w}{1-\rho_2}\right)^{\nu_2-1} l_2(\frac{-1}{w}), \qquad w \uparrow 0. \tag{34}$$

Using Equations (11), (12),

$$\frac{(1-\pi_2)b_2}{1-b_2} - \frac{(1-\pi_1)b_1}{1-b_1} = \frac{(1-\rho_1)\rho_2}{(1-b_1)(1-b_2)}(1-\frac{1}{r_1}-\frac{1}{r_2}).$$

Combining the above two relations, we obtain, as $w \uparrow 0$,

$$P_2(w) - P_2(0) - (P_1(w) - P_1(0)) \sim -\frac{(1-\rho_1)\rho_2}{(1-b_1)(1-b_2)}(1-\frac{1}{r_1}-\frac{1}{r_2})\frac{C_2}{\beta_2}\left(\frac{-w}{1-\rho_2}\right)^{\nu_2-1} l_2(\frac{-1}{w}).$$

Substituting this into (26), we have, for $s \downarrow s_1$,

$$\psi_2(s) - \psi_2(s_1) \sim -\frac{1}{(r_1-1)r_2}\frac{(1-\rho_1)\rho_2}{1-b_1}\frac{C_2}{\beta_2}\left(\frac{-\delta_1^{-1}(s)}{1-\rho_2}\right)^{\nu_2-1} l_2(\frac{-1}{\delta_1^{-1}(s)}). \tag{35}$$

It remains to calculate the term $\psi_2(s_1) - \psi_2(0)$ in (21). From (6), noting that $\delta_1^{-1}(0) = 0$,

$$\psi_2(0) - \psi_0 = \frac{1}{r_1}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1 - \mathrm{e}^{-R_1(0)+R_2(0)}]. \tag{36}$$

Combining (23) with (36) and using (8) and (25),

$$\begin{aligned}
\psi_2(s_1) - \psi_2(0) &= \psi_2(s_1) - \psi_0 - [\psi_2(0) - \psi_0] \\
&= \frac{1}{r_1}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[\mathrm{e}^{-R_1(0)+R_2(0)} - \mathrm{e}^{P_1(0)-P_2(0)}\frac{r_1}{(r_1-1)r_2}] \\
&= \frac{1}{r_1}\frac{1}{1-\frac{1}{r_1}-\frac{1}{r_2}}[\mathrm{e}^{-P_1(0)-R_1(0)} - \mathrm{e}^{-P_2(0)-R_2(0)}\frac{r_1}{(r_1-1)r_2}] \\
&= \frac{1}{r_1}\frac{1}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1-b_1-\frac{(1-b_2)r_1}{(r_1-1)r_2}] \\
&= \frac{1-\rho_1}{r_1-1}.
\end{aligned} \tag{37}$$

Substituting (35) and (37) into (21) with $\delta_1^{-1}(s) \sim \frac{\rho_1-1}{\rho_1}(s-s_1)$ as in (22), and observing that $\frac{1+\beta_1 s}{\beta_1 s} \to \frac{\rho_1}{\rho_1-1}$ for $s \downarrow s_1$, we obtain

$$(s-s_1)\frac{1-\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_1}]}{s} = \frac{\rho_1\rho_2}{(1-b_1)r_2}\frac{C_2}{\beta_2}\left(\frac{1-\rho_1}{\rho_1(1-\rho_2)}(s-s_1)\right)^{\nu_2-1} l_2(\frac{1}{s-s_1}) + \mathrm{O}(s-s_1).$$

In terms of Lemma 3.2, we have now determined a series expansion of $(s-s_1)\phi(s)$ with $a_{01} = 0$, $\gamma_1 = \nu_2 - 1$, and

$$b_{01} = \frac{\rho_1\rho_2}{(1-b_1)r_2}\frac{C_2}{\beta_2}\left(\frac{1-\rho_1}{\rho_1(1-\rho_2)}\right)^{\nu_2-1}.$$

Applying Lemma 3.2, using that $\Gamma(\nu_2-1)\sin(\pi(\nu_2-1))/\pi = 1/\Gamma(2-\nu_2)$ and $(1-b_1)r_2 = K_2 - \rho_2$, then gives the statement of the theorem. There is one problem left in applying Lemma 3.2: verifying Condition (iii). It follows from (21) and (37) that

$$\phi(s) = \frac{1}{s} + (r_1-1)\frac{1+\beta_1 s}{\beta_1 s}\frac{\psi_2(s)-\psi_2(s_1)}{s-s_1}. \tag{38}$$

The condition that $\int |\phi(s)|\mathrm{d}y$ converges at $y = \pm\infty$ is not fulfilled. However, as observed in [13] in the study of an M/M/1 queue in a heavy-tailed environment, a detailed analysis of Sutton's proof of Lemma 3.2 reveals that this convergence condition is not strictly necessary; it is sufficient that $\phi(s) = \phi(x + \mathrm{i}y)$ and its derivative w.r.t. y are in absolute value bounded by L_1/y and L_2/y^2, respectively, for x near s_1 and for y sufficiently large, for some positive constants L_1 and L_2. It easily follows from (38) that these conditions are indeed satisfied.

Remark 4.1 *In determining the coefficient b_{01}, we have disregarded the slowly varying function $l_2(\cdot)$. Formally, this causes a problem with the application of Lemma 3.2. The treatment of that – technical – problem is outside the scope of the present paper. The problem does not arise in the slightly less general case of regular variation with $l_2(\cdot) \equiv 1$.*

Remark 4.2 *In Section 2, we have excluded the case $1/r_1 + 1/r_2 = 1$. However, this case does not seem to represent a boundary case in the above theorem for $1/r_1 + 1/r_2 \to 1$, suggesting that the theorem then remains valid.*

We now turn to the case $\rho_2 > 1$. In this case, it is relatively likely for class 2 to be continuously backlogged during the period in which class 1 shows deviant behavior. This suggests that the pre-factor of $\mathbb{P}\{\mathbf{V}_1^1 > t\}$ should be O(1), as is confirmed by the next theorem.

Theorem 4.2 *If $\rho_1 < 1$, $\rho_2 > 1$, then*

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim H_2 \rho_1 \mathrm{e}^{(\lambda_1 - 1/\beta_1)t} \sim H_2 \mathbb{P}\{\mathbf{V}_1^1 > t\}, \qquad t \to \infty, \tag{39}$$

with

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

Proof
The proof of the above theorem is largely similar to that of Theorem 4.1. The main difference is that the distribution of $\hat{\mathbf{P}}_2$ is now defective with $\mathbb{P}\{\hat{\mathbf{P}}_2 < \infty\} = 1/\rho_2$ (cf. [7]). Instead of (34), we now have, using Equations (10), (13), (14), (15),

$$P_i(w) - P_i(0) \to -\sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathbb{P}\{\sigma_n^{(i)} = -\infty\}, \qquad w \uparrow 0.$$

with

$$\mathbb{P}\{\sigma_n^{(i)} = -\infty\} = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \pi_i^k (1 - \pi_i)^{n-k} \mathbb{P}\{\sum_{j=1}^{n-k} \hat{\mathbf{P}}_{2,j} - \sum_{j=1}^{k} \hat{\mathbf{P}}_{1,j} = \infty\}. \tag{40}$$

Since $\mathbb{P}\{\hat{\mathbf{P}}_1 < \infty\} = 1$, we may write

$$\mathbb{P}\{\sum_{j=1}^{n-k} \hat{\mathbf{P}}_{2,j} - \sum_{j=1}^{k} \hat{\mathbf{P}}_{1,j} = \infty\} = 1 - \Big(\mathbb{P}\{\hat{\mathbf{P}}_2 < \infty\}\Big)^{n-k} = 1 - \left(\frac{1}{\rho_2}\right)^{n-k}.$$

Using the above relation in (40), we obtain

$$P_i(w) - P_i(0) \to -\ln\left[\frac{\rho_2 - \rho_2 b_i \pi_i - b_i + b_i \pi_i}{\rho_2 (1 - b_i)}\right], \qquad w \uparrow 0,$$

yielding

$$e^{P_2(w)-P_2(0)-(P_1(w)-P_1(0))} \to \frac{-r_2+\rho_1 r_2-\rho_1+\rho_2}{(-r_1+\rho_2 r_1-\rho_2+\rho_1)(r_2-1)}, \qquad w \uparrow 0.$$

Substituting this into (24),

$$\psi_2(s)-\psi_2(s_1) \to -\frac{\rho_2-1}{(r_1-1)(r_2-1)}, \qquad s \downarrow s_1. \tag{41}$$

Substituting (41) and (37) into (21), and observing that $\frac{1+\beta_1 s}{\beta_1 s} \to \frac{\rho_1}{\rho_1-1}$ for $s \downarrow s_1$, we obtain

$$(s-s_1)\frac{1-\mathbb{E}[e^{-s\mathbf{V}_1}]}{s} = \rho_1 H_2 + \mathrm{O}((s-s_1)^{\nu_2-1}).$$

In terms of Lemma 3.2, we have now determined a series expansion of $(s-s_1)\phi(s)$ with $a_{01}=\rho_1 H_2$ and $\gamma_1=\nu_2-1$. Applying Lemma 3.2 then completes the proof.

5 Workload asymptotics for class 1 in overload

In this section we derive the asymptotic behavior of $\mathbb{P}\{\mathbf{V}_1 > t\}$ for the case $\rho_1 > 1$ (this forces $\rho_2 < 1$ to ensure ergodicity).

Theorem 5.1 *If $\rho_1 > 1$, $\rho_2 < 1$, then*

$$\begin{aligned}\mathbb{P}\{\mathbf{V}_1 > t\} &\sim \frac{r_1-1}{K_1-\rho_1}\frac{\lambda_2 C_2}{\Gamma(2-\nu_2)}\left(\frac{1-\rho_2}{\rho_1-1}t\right)^{1-\nu_2} l_2(t) \\ &\sim \frac{(r_1-1)\rho_2}{K_1-\rho_1}\mathbb{P}\{\mathbf{B}_2^r > \frac{1-\rho_2}{\rho_1-1}t\},\end{aligned} \tag{42}$$

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

The above result is similar to that derived in Section 6.3 of [8] for the case where the service time distribution of class-1 customers is not exponential but regularly varying of index $-\nu_1$, with $\nu_1 > \nu_2$ (class-2 traffic is heavier-tailed). The proof of Theorem 5.1 largely mimics the proof of the latter result, and in fact reveals that the same asymptotic behavior holds for any service time distribution of class-1 customers such that $\mathbb{P}\{\mathbf{B}_1 > t\} = \mathrm{o}(t^{-\nu_2})$ as $t \to \infty$.

Proof
As mentioned above, the proof follows the derivation in Section 6.3 of [8]. It uses Formula (6.24) of [14], where now the *second* zero $\epsilon_1(w)$ of the function $f_1(s,w)$ as defined in (16) plays a key role. In particular, it may be shown that

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

Substituting in (5), we obtain

$$1-\mathbb{E}[e^{-s\mathbf{V}_1}] \sim \frac{r_1-1}{(1-b_2)r_1}\lambda_2 C_2\left(\frac{\rho_1-1}{1-\rho_2}s\right)^{\nu_2-1} l_2(\frac{1}{s}), \qquad s \downarrow 0.$$

Applying Lemma 3.1 then completes the proof.

□

Using a result of De Meyer & Teugels [20], Formula (42) may be rewritten as

$$\mathbb{P}\{\mathbf{V}_1 > t\} \sim (1-\rho_2)\frac{(r_1-1)\rho_2}{K_1-\rho_1}\mathbb{P}\{\mathbf{P}_2^r > \frac{t}{\rho_1-1}\}, \tag{43}$$

with $\mathbf{P}_2$ a random variable representing the busy period of class 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 of class 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 class 2 generates a large burst, so that it becomes backlogged for a long period of time. Class 1 is then only served at rate 1, while it generates traffic at average rate $\rho_1 > 1$. Thus, the workload of class 1 has positive drift, and class 1 will soon become backlogged too (if it not already is), and remain so for as long as class 2 is backlogged. Consequently, class 2 will experience a busy period as if it were served at a constant rate 1. During that period, queue 1 will roughly grow at rate $\rho_1 - 1 > 0$. Only after queue 2 empties, queue 1 will start to drain again at approximately rate $r_1 - \rho_1$.
Of course, queue 1 may also build up a large workload when class 1 itself generates a large amount of traffic. However, these effects are insignificant compared to the build-up that occurs during a busy period of class 2, because of the relatively smooth nature of class-1 traffic.
Thus, the workload of class 1 should asymptotically behave 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 of class 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 (43).

6 Workload asymptotics for class 2

In this section we characterize the asymptotic behavior of $\mathbb{P}\{\mathbf{V}_2 > t\}$. It turns out that – in contrast to the light-tailed class – the qualitative behavior of $\mathbb{P}\{\mathbf{V}_1 > t\}$ does not strongly depend on whether $\rho_2 < 1$ or $\rho_2 > 1$ (in the latter case assuming $\rho_1 < 1$ to ensure ergodicity).

Theorem 6.1

$$\mathbb{P}\{\mathbf{V}_2 > t\} \sim \frac{1}{M_2-\rho_2}\frac{\lambda_2 C_2}{\Gamma(2-\nu_2)}t^{1-\nu_2}l_2(t) \sim \frac{\rho_2}{M_2-\rho_2}\mathbb{P}\{\mathbf{B}_2^r > t\}, \tag{44}$$

with $M_2 := \max\{1, K_2\}$.

The above result is similar to that obtained in Sections 6.2 and 6.3 of [8] for the case where the service time distribution of class-1 customers is not exponential but regularly varying of index $-\nu_1$, under the additional condition that $\nu_1 > \nu_2$ (class-2 traffic is heavier-tailed) in case $\rho_2 > 1$. The proof of Theorem 6.1 closely follows the proof of the latter result, and in fact shows that the same asymptotic behavior applies for any service time distribution of class-1 customers, provided $\mathbb{P}\{\mathbf{B}_1 > t\} = \mathrm{o}(t^{-\nu_2})$ in case $\rho_2 > 1$.

Proof
Interchanging the class indices in (5) yields

$$\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_2}] = \frac{(1-\rho_2)s}{s-\lambda_2(1-\beta_2\{s\})}[1-\frac{r_2-1}{1-\rho_2}(\psi_1(s)-\psi_1(0))]. \tag{45}$$

Applying Lemma 3.1, we find that (1) implies

$$1-\frac{1-\beta_2\{s\}}{\beta_2 s} \sim \frac{C_2}{\beta_2}s^{\nu_2-1}l_2(\frac{1}{s}), \qquad s \downarrow 0.$$

Mimicking the derivations in Sections 6.2 and 6.3 of [8] (where it is assumed that *both* traffic classes have regularly varying service time distributions), it may be checked that

$$\psi_1(s)-\psi_1(0) \sim -\frac{1-\rho_1}{M_2-\rho_2}\lambda_2 C_2 s^{\nu_2-1}l_2(\frac{1}{s}), \qquad s \downarrow 0.$$

Substituting into (45), we obtain

$$1-\mathbb{E}[\mathrm{e}^{-s\mathbf{V}_2}] \sim \frac{1}{M_2-\rho_2}\lambda_2 C_2 s^{\nu_2-1}l_2(\frac{1}{s}), \qquad s \downarrow 0.$$

Applying Lemma 3.1 then completes the proof.

□

Invoking a result of Pakes [21], Formula (44) may be rewritten as

$$\mathbb{P}\{\mathbf{V}_2 > t\} \sim \mathbb{P}\{\mathbf{V}_2^{M_2} > t\}. \tag{46}$$

Thus, the workload of class 2 is asymptotically equivalent to that in an isolated system, where class-2 traffic is served at a constant rate M_2. This may be intuitively interpreted as follows. Large-deviations arguments suggest that the most likely scenario for the workload of class 2 to reach a large level is that class 2 itself generates a large amount of traffic, while class 1 shows average behavior. Specifically, suppose that class 2 generates a large burst, so that it becomes backlogged for a long period of time. During that period, queue 1 will not receive any surplus capacity, and thus be empty only a fraction $1-\rho_1$ of the time, assuming $\rho_1 < 1$. In case $\rho_1 > 1$, queue 1 will be constantly backlogged too during that period. Thus, the average service rate for class 2 while backlogged is either $\rho_1 + (1-\rho_1)r_2$ in case $\rho_1 < 1$, or just 1 in case $\rho_1 > 1$, which may be compactly denoted as $M_2 := \max\{1, K_2\}$. As a result, class 2 is effectively served at a constant rate M_2, as confirmed by Formula (46). A related reduced-load equivalence result is established in [1].

Acknowledgment The authors gratefully acknowledge some useful comments of Qing Deng.

A Proof of Formula (23)

For $s \downarrow s_1$,

$$\psi_2(s)-\psi_0 \sim \frac{1}{r_1}\frac{\psi_0}{1-\frac{1}{r_1}-\frac{1}{r_2}}[1-\mathrm{e}^{P_1(\delta_1^{-1}(s))-P_2(\delta_1^{-1}(s))}\frac{r_1}{(r_1-1)r_2}].$$

Proof
Again using the fact that $\mathbb{P}\{\mathbf{V}_1 > t\} \leq \mathbb{P}\{\mathbf{V}_1^1 > t\}$, it may be verified that $\psi_2(s)$ is analytic for Re $s > s_1$. Analytic continuation of Formula (6.2) in [14], taking $w = \delta_1^{-1}(s)$, then shows that for $s \downarrow s_1$,

$$\psi_2(s) - \psi_0 = \frac{1}{r_1} \frac{\psi_0}{1 - \frac{1}{r_1} - \frac{1}{r_2}} + \left[\psi_1(\delta_2(\delta_1^{-1}(s))) - \psi_0 - \frac{\psi_0}{r_2} \frac{1}{1 - \frac{1}{r_1} - \frac{1}{r_2}} \right] \frac{s + (r_2 - 1)\delta_2(\delta_1^{-1}(s))}{s(r_1 - 1) + \delta_2(\delta_1^{-1}(s))}. \quad (47)$$

Using Relations (22), (30), for $s \downarrow s_1$,

$$\delta_2(\delta_1^{-1}(s)) \sim \delta_2(\frac{\rho_1 - 1}{\rho_1}(s - s_1)) \sim -\frac{\rho_1 - 1}{\rho_1(1 - \rho_2)}(s - s_1). \quad (48)$$

Taking $w = \delta_1^{-1}(s)$ in (7), we obtain, for $s \downarrow s_1$,

$$\psi_1(\delta_2(\delta_1^{-1}(s))) - \psi_0 - \frac{\psi_0}{r_2} \frac{1}{1 - \frac{1}{r_1} - \frac{1}{r_2}} = -\frac{\psi_0}{r_2} \frac{1}{1 - \frac{1}{r_1} - \frac{1}{r_2}} \mathrm{e}^{P_1(\delta_1^{-1}(s)) - P_2(\delta_1^{-1}(s))}. \quad (49)$$

Substituting (48) and (49) into (47) completes the proof.

□

References

[1] Agrawal, R., Makowski, A.M., Nain, Ph. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems* **33**, 5–41.

[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] Beran, J., Sherman, R., Taqqu, M.S., Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. IEEE *Trans. Commun.* **43**, 1566–1579.

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

[5] Borst, S.C., Boxma, O.J., Jelenković, P.R. (1999). Generalized processor sharing with long-tailed traffic sources. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 345–354.

[6] Borst, S.C., Boxma, O.J., Jelenković, P.R. (2000). Asymptotic behavior of generalized processor sharing with long-tailed traffic sources. In: *Proc. Infocom 2000 Conference*, Tel-Aviv, Israel, 912–921.

[7] 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.

[8] 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. Submitted for publication.

[9] Borst, S.C., Mandjes, M., Van Uitert, M.J.G. (2001). Generalized Processor Sharing queues with heterogeneous traffic classes. CWI Report PNA-R0106.

[10] 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. Submitted for publication.

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

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

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

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

[15] Crovella, M., Bestavros, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In: *Proc. ACM Sigmetrics '96*, 160–169.

[16] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., Weiss, W. (1999). A framework for differentiated services. IETF RFC 2475.

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

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

[19] Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V. (1994). On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM *Trans. Netw.* **2**, 1–15.

[20] 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.

[21] Pakes, A.G. (1975). On the tails of waiting-time distributions. *J. Appl. Prob.* **12**, 555–564.

[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] Paxson, A., Floyd, S. (1995). Wide area traffic: the failure of Poisson modeling. IEEE/ACM *Trans. Netw.* **3**, 226-244.

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

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

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

[29] Willinger, W., Taqqu, M.S., Sherman, R., Wilson, D.V. (1997). Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM *Trans. Netw.* **5**, 71–86.

[30] Zwart, A.P., Borst, S.C., Mandjes, M. (2000). Exact asymptotics for fluid queues fed by multiple heavy-tailed On-Off flows. SPOR Report 2000-14, Eindhoven University of Technology. In: *Proc. INFOCOM 2001*, 279–288.