Coupled Processors with Regularly Varying Service Times

Sem Borst*,†,‡, Onno Boxma†,‡, Predrag Jelenković**

†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

**Department of Electrical Engineering
Columbia University
New York, NY 10027, USA

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

Abstract—**Consider two $M/G/1$ queues that are coupled in the following way. Whenever both queues are non-empty, each server serves its own queue at unit speed. However, if server 2 has no work in its own queue, then it assists server 1, resulting in an increased service speed $r_1^* \geq 1$ in the first queue. This kind of coupling is related to generalized processor sharing. We assume that the service request distributions at both queues are regularly varying at infinity of index $-\nu_1$ and $-\nu_2$, viz., they are heavy-tailed. Under this assumption, we present a detailed analysis of the tail behaviour of the workload distribution at each queue. If the guaranteed unit speed of server 1 is already sufficient to handle its offered traffic, then the workload distribution at the first queue is shown to be regularly varying at infinity of index $1-\nu_1$. But if it is not sufficient, then the workload distribution at the first queue is shown to be regularly varying at infinity of index $1-\min(\nu_1,\nu_2)$. In particular, traffic at server 1 is then no longer protected from worse behaving (heavier-tailed) traffic at server 2.**

Keywords—**Coupled processors, Generalized Processor Sharing, workload, tail behaviour, regular variation.**

I. Introduction

Consider the following model of two coupled $M/G/1$ queues, Q_1 and Q_2. Q_i, $i = 1, 2$, receives a Poisson arrival stream of customers of type i with arrival rate λ_i and required amounts of service that are i.i.d. random variables with distribution $B_i(\cdot)$, with mean β_i and Laplace-Stieltjes transform (LST) $\beta_i\{s\}$. B_i denotes a random variable with distribution $B_i(\cdot)$, $i = 1, 2$. Denote the average amount of traffic offered per unit of time at Q_i by $\rho_i := \lambda_i\beta_i$. The arrival processes at the two queues, and the families of required service amounts in both streams, are independent of each other. Whenever there is work of each type, each server serves its own queue with speed 1. However, if server 2 is idle then the speed of server 1 is $r_1^* \geq 1$, and if server 1 is idle then the speed of server 2 is $r_2^* \geq 1$. In a sense, the servers are coupled, and a server with no work at its own queue is able to assist the other server.

This coupled-processors model has been analysed by Fayolle and Iasnogorodski [11] and by Konheim, Meilijson and Melkman [13] in the case of negative exponentially distributed service requests, and by Cohen and Boxma [9] in the case of generally distributed service requests. Konheim et al. apply the uniformisation technique; Fayolle and Iasnogorodski determine the joint queue length distribution by formulating and solving a Riemann-Hilbert boundary value problem; and Cohen and Boxma obtain the joint distribution of the workloads in both queues by formulating and solving a Wiener-Hopf boundary value problem.

The coupled-processors model is highly relevant for Generalized Processor Sharing (GPS). GPS-based scheduling algorithms, such as *Weighted Fair Queueing*, have emerged as an important mechanism for achieving differentiated quality-of-service in integrated-services networks. The GPS discipline operates as follows. Consider $N \geq 2$ sources sharing a link of unit rate. There is a nonnegative weight ϕ_i associated with source i, with $\sum_{i=1}^{N} \phi_i = 1$. If the buffer content of each source is positive, then source i is served at rate ϕ_i. But if some of the sources have an empty buffer, then the excess service capacity is redistributed among the sources with non-empty buffers in proportion to their respective weights. See [10] for a formal description of the evolution of the buffer content process.

The queueing analysis of GPS is extremely difficult. Interesting partial results were obtained in [2], [10], [14], [17]. If $N = 2$, then the above coupled-processors model with $r_1^* = r_2^* = 2$ coincides with the GPS model with equal weights; hence the exact queue length analysis in [11], [13], for the case of exponentially distributed service requests, applies to this special GPS case. Furthermore, the exact analysis of the joint workload process in [9], which holds for *generally distributed* service requests, is also applicable. The latter study forms the starting-point of the present paper.

Our goal is to investigate the influence of heavy-tailed service request distributions on the tail behaviour of the workload distributions at the two coupled processors. The motivation for this investigation is the following. Statistical data analysis has provided convincing evidence of heavy-tailed traffic characteristics in high-speed communication networks (see, e.g., the forthcoming book [16]). This has stimulated much research into the effect of heavy-tailed traffic on key performance measures like waiting times and workloads. An important question is: To which extent are performance measures for one type of input traffic affected by worse (i.e., heavier-tailed) input traffic of another type? In two recent studies [4], [5], we have partially answered this question for GPS. Using a sample-path analysis to determine lower and upper bounds for buffer content (workload) tails, we have identified conditions under which the buffer content of an individual source with long-tailed traffic characteristics behaves similarly as when served at a constant rate which is equal to the maximum feasible average rate for that source to be stable – regardless of the possibility that other sources have heavier-tailed input traffic. Under those conditions, GPS-based scheduling mechanisms apparently are able to protect individual connections.

In the present paper we identify a situation in which such a protection is *not* given. The exact joint steady-state distribution of the two workloads, which has been obtained in [9], subsequently allows us to exactly quantify the workload tail behaviour, and to determine to what extent the protection fails. We perform this tail behaviour analysis under the assumption of regularly varying service request distributions. Regularly varying distributions form an important class of heavy-tailed distributions, with well-studied properties [3].

While the results in [9] allow us, in principle, to study the workload tail behaviour for all (r_1^*, r_2^*) combinations with $r_1^* \geq 1$, $r_2^* \geq 1$, we have decided to restrict ourselves in this paper to $r_1^* \geq 1$, $r_2^* = 1$; analysis of the general case is the subject of a forthcoming study. The reason for the restriction is, that the case $r_2^* = 1$ is relatively simple and transparent: Q_2 is not affected by Q_1, and the influence of the service request tail at Q_2 on Q_1 can be sharply identified *and* interpreted. This yields much insight into more complicated cases, for which there is little hope of an exact analysis.

The paper is organised in the following way. Section II contains those results from the exact coupled-processors analysis of [9] that will be used in the sequel. We subsequently distinguish two cases: $\rho_1 < 1$ and $\rho_1 > 1$. In the former case, server 1 is able to handle its offered traffic, even if Q_2 were never empty. In the latter case, server 1 needs the assistance of server 2; this is the case where 'the protection fails'. The workload asymptotics for $\rho_1 < 1$ are analysed in Section III, and those for $\rho_1 > 1$ in Section IV. The latter section contains our main result: The tail of the workload distribution at Q_1 is shown to be regularly varying of index $1 - \min(\nu_1, \nu_2)$, i.e., the heaviest-tailed service request distribution determines the tail behaviour of the workload distribution. Section V contains conclusions and suggestions for future work. Some definitions and results regarding regularly varying and long-tailed distributions are gathered in the appendix.

II. Preliminaries

In this section we summarise those results of Section III.3.7 of [9] that will be used in the analysis of the tail behaviour of the workloads in the coupled-processors model. We refer to Section III.3.7 of [9] for a discussion of the ergodicity conditions; for the moment it suffices to observe that at least one of the conditions $\rho_1 < 1, \rho_2 < 1$ should be satisfied, but not necessarily both. For example, if $\rho_1 > 1$ then it is still possible that the server at Q_2 sufficiently often faces no work at its own queue and is able to serve the other queue. We restrict ourselves in the sequel to the steady-state situation.

V_i denotes the steady-state workload at Q_i; $(\cdot)$ is used to denote an indicator function. For Re $s_1 \geq 0$, Re $s_2 \geq 0$, let

$$\psi(s_1, s_2) := \mathrm{E}[\mathrm{e}^{-s_1 V_1 - s_2 V_2}], \tag{1}$$

$$\psi_1(s_2) := \mathrm{E}[\mathrm{e}^{-s_2 V_2}(V_1 = 0)], \tag{2}$$

$$\psi_2(s_1) := \mathrm{E}[\mathrm{e}^{-s_1 V_1}(V_2 = 0)], \tag{3}$$

$$\psi_0 := \mathrm{P}(V_1 = 0, V_2 = 0). \tag{4}$$

Formula (2.16) of Chapter III.3 of [9] (in the sequel we omit mentioning Chapter III.3 when referring to formulas from that chapter) expresses $\psi(s_1, s_2)$ into $\psi_1(s_2)$, $\psi_2(s_1)$, and ψ_0. For our purposes it is sufficient to study the LST's of the marginal workload distributions. In particular, we concentrate on the workload at Q_1. From (2.16) of [9] it follows that, for Re $s \geq 0$,

$$\psi(s, 0) = \mathrm{E}[\mathrm{e}^{-sV_1}] = \frac{(1-\rho_1)s}{s - \lambda_1(1 - \beta_1\{s\})} \left[\frac{\psi_1(0)}{1-\rho_1} + \frac{r_1^* - 1}{1-\rho_1}(\psi_0 - \psi_2(s))\right]. \tag{5}$$

Note that the first term in the righthand side is the Pollaczek-Khintchine LST of the workload in $M/G/1$ queue Q_1 in isolation (with service speed 1). We now discuss $\psi_2(s)$. In [9] a distinction is made between the special case $1/r_1^* + 1/r_2^* = 1$ (which corresponds directly to generalized processor sharing) and the case $1/r_1^* + 1/r_2^* \neq 1$. Let us concentrate on the latter more general case, which is of more interest for our purposes (in the next two sections, we take $r_2^* = 1$). According to (6.22) of [9],

$$\frac{1}{r_2^*}[\psi_2(\delta_1(w)) - \psi_0] = \frac{1}{r_1^* r_2^*} \frac{\mathrm{e}^{-P_1(0) - R_2(0)}}{1 - 1/r_1^* - 1/r_2^*} [1 - \mathrm{e}^{-R_1(w) + R_2(w)}], \quad \mathrm{Re}\, w \geq 0. \tag{6}$$

We still have to specify the functions $P_i(w)$, $R_i(w)$ and $\delta_1(w)$. For $i = 1, 2$,

$$P_i(w) := \sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathrm{E}[\mathrm{e}^{-w\sigma_n^{(i)}}(\sigma_n^{(i)} < 0)], \quad \mathrm{Re}\, w \leq 0, \tag{7}$$

$$R_i(w) := \sum_{n=1}^{\infty} \frac{b_i^n}{n} \mathrm{E}[\mathrm{e}^{-w\sigma_n^{(i)}}(\sigma_n^{(i)} > 0)], \quad \mathrm{Re}\, w \geq 0. \tag{8}$$

Here

$$b_1 := \rho_1(1 - \frac{1}{r_2^*}) + \frac{\rho_2}{r_2^*}, \tag{9}$$

$$b_2 := \rho_2(1 - \frac{1}{r_1^*}) + \frac{\rho_1}{r_1^*}, \quad (10)$$

and for $i = 1, 2$,

$$\sigma_n^{(i)} := X_{i1} + \ldots + X_{in}, \quad (11)$$

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

$$\begin{aligned} X_{11} &= \hat{P}_1 \quad w.p. \ \frac{\rho_1}{b_1}(1 - \frac{1}{r_2^*}), \\ &\quad -\hat{P}_2 \quad w.p. \ \frac{\rho_2}{b_1 r_2^*}, \end{aligned} \quad (12)$$

$$\begin{aligned} X_{21} &= \hat{P}_1 \quad w.p. \ \frac{\rho_1}{b_2 r_1^*}, \\ &\quad -\hat{P}_2 \quad w.p. \ \frac{\rho_2}{b_2}(1 - \frac{1}{r_1^*}). \end{aligned} \quad (13)$$

$\hat{P}_i$ denotes a busy period in an $M/G/1$ queue that has exactly the same traffic characteristics as Q_i and has service speed 1, and that *starts with an exceptional first service* that has distribution $\int_0^x \frac{1-B_i(u)}{\beta_i} du$ (a residual service time; we denote such a random variable by B_i^{res}).

We also have to specify $\delta_1(w)$, which plays a key role in the analysis of this coupled-processors model. The function

$$f_1(s, w) := \lambda_1(1 - \beta_1\{s\}) - s + w \quad (14)$$

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 zeros $s = \delta_1(0) > 0$ 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

$$f_2(s, w) := \lambda_2(1 - \beta_2\{s\}) - s - w. \quad (15)$$

The different behaviour of $\delta_1(w)$ for w near 0 for $\rho_1 < 1$ and $\rho_1 > 1$ will be reflected in different tail behaviour of the workload distribution at Q_1 for these two cases. In Section III we consider the case $\rho_1 < 1$, and in Section IV the case $\rho_1 > 1$.

III. Workloads for the case $\rho_1 < 1$

Firstly, remember that $r_2^* = 1$. Hence, Q_2 is not influenced by Q_1; it is an ordinary $M/G/1$ queue. It follows from Cohen [7] that $P(V_2 > t)$ is regularly varying of index $1 - \nu_2$ at infinity iff the tail of the service request distribution $P(B_2 > t)$ is regularly varying of index $-\nu_2$ at infinity (see the appendix for the definition of regularly, and slowly, varying functions), and more precisely:

$$P(B_2 > t) \sim \frac{C_2}{-\Gamma(1 - \nu_2)} t^{-\nu_2} l_2(t), \quad t \to \infty, \quad (16)$$

iff

$$P(V_2 > t) \sim \frac{C_2}{\beta_2 \Gamma(2 - \nu_2)} \frac{\rho_2}{1 - \rho_2} t^{1-\nu_2} l_2(t), \quad t \to \infty. \quad (17)$$

Here and in the sequel, $f(t) \sim g(t)$ denotes: $\lim_{t\to\infty} f(t)/g(t) = 1$; and $l(\cdot)$ and $l_i(\cdot)$ will be used to denote slowly varying functions.

Having established the tail behaviour of the workload at Q_2, we concentrate on Q_1 in the remainder of this section. Assume that $P(B_1 > t)$ is regularly varying at infinity of index $-\nu_1$:

$$P(B_1 > t) = \frac{C_1}{-\Gamma(1 - \nu_1)} t^{-\nu_1} l_1(t), \quad t \to \infty. \quad (18)$$

Let us assume that $1 < \nu_1 < 2$; higher values of ν_1 can be handled with minor adaptations. According to Lemma A.1, (18) with $1 < \nu_1 < 2$ is equivalent with

$$\frac{1 - \beta_1\{s\}}{\beta_1 s} = 1 - \frac{C_1}{\beta_1} s^{\nu_1 - 1} l_1(\frac{1}{s}), \quad s \downarrow 0. \quad (19)$$

We are interested in the tail behaviour of the workload distribution at Q_1. We intend to show that it is similar to that of V_2 as given above, but with index $1 - \nu_1$ instead of $1 - \nu_2$; i.e., in the case $\rho_1 < 1$, the index of regular variation of the tail of $P(V_1 > t)$ is not influenced by Q_2. Our approach is as follows. If we can determine the behaviour of $E[e^{-sV_1}]$ for $s \downarrow 0$, then we can invoke Lemma A.1 to determine the behaviour of $P(V_1 > t)$ for $t \to \infty$. Formula (5) expresses $E[e^{-sV_1}]$ into $\psi_2(s)$. Formula (6) expresses $\psi_2(s)$, or rather $\psi_2(\delta_1(w))$, into $R_1(w)$ and $R_2(w)$. Therefore we now concentrate on the behaviour of the latter functions for $w \downarrow 0$. Note that, since $r_2^* = 1$, we have $b_1 = \rho_2$ and $X_{11} = -\hat{P}_2 < 0$ $w.p.$ 1, which implies that $R_1(w) \equiv 0$. According to (6.21) of [9],

$$\psi(0) = e^{-R_1(0) - R_2(0)}. \quad (20)$$

In combination with the above, (6) reduces to

$$\psi_2(\delta_1(w)) = \psi_0 e^{R_2(w)}, \quad \text{Re } w \geq 0. \quad (21)$$

Before focusing on $R_2(w)$, we study the behaviour of $\delta_1(w)$ for $w \downarrow 0$. Let P_1 denote a random variable with distribution the steady-state distribution of a busy period in the $M/G/1$ queue with arrival rate λ_1 and service time distribution $B_1(\cdot)$, viz., Q_1 in isolation. Comparing (14) with the Takács equation for the busy period LST $E[e^{-wP_1}]$, cf. p. 250 of Cohen [8], it is seen that

$$\delta_1(w) = w + \lambda_1(1 - E[e^{-wP_1}]). \quad (22)$$

De Meyer and Teugels [15] have proven that $P(P_1 > t)$ is regularly varying at infinity of index $-\nu_1$ iff $P(B_1 > t)$ is regularly varying at infinity of index $-\nu_1$, and if either holds then, for $t \to \infty$,

$$\begin{aligned} P(P_1 > t) &\sim \frac{1}{1 - \rho_1} P(\frac{B_1}{1 - \rho_1} > t) \\ &\sim \frac{C_1}{-\Gamma(1 - \nu_1)} (\frac{1}{1 - \rho_1})^{\nu_1 + 1} t^{-\nu_1} l_1(t). \end{aligned} \quad (23)$$

Lemma A.1 then gives the behaviour of $E[e^{-wP_1}] - 1$ for $w \downarrow 0$. We conclude that, if (18) holds, then

$$\delta_1(w) = \frac{w}{1 - \rho_1} - \lambda_1 C_1 \frac{w^{\nu_1}}{(1 - \rho_1)^{\nu_1 + 1}} l_1(\frac{1}{w}), \quad w \downarrow 0. \quad (24)$$

In addition, using (14):

$$w = \delta_1^{-1}(s) = s - \lambda_1(1-\beta_1\{s\}) = (1-\rho_1)s + \lambda_1 C_1 s^{\nu_1} l_1(\frac{1}{s}), \quad s \downarrow 0. \tag{25}$$

In the study of $R_2(w)$, a key role is played by the LST of $\hat{P}_1$, a busy period in Q_1 in isolation that is started with a *residual* service time. From (6.4) of [9],

$$\mathrm{E}[\mathrm{e}^{-w\hat{P}_1}] = \frac{1-\beta_1\{\delta_1(w)\}}{\beta_1\delta_1(w)}, \ \mathrm{Re}\, w \geq 0. \tag{26}$$

It is now readily verified, using (19), (24) and (26), that

$$1 - \mathrm{E}[\mathrm{e}^{-w\hat{P}_1}] \sim \frac{C_1}{\beta_1}(\frac{w}{1-\rho_1})^{\nu_1-1} l_1(\frac{1}{\delta_1(w)}), \ w \downarrow 0. \tag{27}$$

Hence, using Lemma A.1, $\mathrm{P}(\hat{P}_1 > t)$ is seen to be regularly varying at infinity of index $1-\nu_1$:

$$\mathrm{P}(\hat{P}_1 > t) \sim \frac{C_1}{\beta_1\Gamma(2-\nu_1)}((1-\rho_1)t)^{1-\nu_1} l_1(t), \quad t \to \infty. \tag{28}$$

The difference with (23) 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 behaviour of $R_2(w)$. Observe that $R_2(w)$ is the LST of

$$r_2(t) := \sum_{n=1}^{\infty} \frac{b_2^n}{n} \mathrm{P}(0 < X_{21} + \ldots + X_{2n} < t), \quad t > 0. \tag{29}$$

Consider, for $t > 0$,

$$R_2(0) - r_2(t) = \sum_{n=1}^{\infty} \frac{b_2^n}{n} \mathrm{P}(X_{21} + \ldots + X_{2n} > t). \tag{30}$$

Introduce Bernoulli random variables Y_i, $i = 1, 2, \ldots$, with $\mathrm{P}(Y_i = 1) = p$, $\mathrm{P}(Y_i = 0) = 1-p$, with $p := \frac{\rho_1}{b_2 r_1^*}$. Using (13) and introducing the i.i.d. random variables $\hat{P}_{1i}$ respectively $\hat{P}_{2i}$ that have the same distribution as $\hat{P}_1$ respectively $\hat{P}_2$, we can write for $i \geq 1$:

$$X_{2i} = Y_i\hat{P}_{1i} - (1-Y_i)\hat{P}_{2i}.$$

Hence, for $t > 0$,

$$\mathrm{P}(X_{21} + \ldots + X_{2n} > t) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \mathrm{P}(\sum_{i=1}^{k} \hat{P}_{1i} - \sum_{i=k+1}^{n} \hat{P}_{2i} > t). \tag{31}$$

Since $\hat{P}_{1i} \in \mathcal{R}(1-\nu_1)$, the class of regularly varying functions of index $1-\nu_1$, we also have (see [3]): $\sum_{i=1}^{k} \hat{P}_{1i} \in \mathcal{R}(1-\nu_1)$. The class $\mathcal{L}$ of long-tailed distributions (see the appendix) contains $\mathcal{R}(1-\nu_1)$, and therefore $\sum_{i=1}^{k} \hat{P}_{1i} \in \mathcal{L}$. Since $\sum_{i=k+1}^{n} \hat{P}_{2i} > 0$ *w.p.* 1, we can apply the following well-known property of $\mathcal{L}$ (cf. [3]):

$$\mathrm{P}(\sum_{i=1}^{k} \hat{P}_{1i} - \sum_{i=k+1}^{n} \hat{P}_{2i} > t) \sim \mathrm{P}(\sum_{i=1}^{k} \hat{P}_{1i} > t), \quad t \to \infty. \tag{32}$$

Hence, for $t \to \infty$,

$$\begin{aligned} &\mathrm{P}(X_{21} + \ldots + X_{2n} > t) \\ \sim\ & \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \mathrm{P}(\sum_{i=1}^{k} \hat{P}_{1i} > t) \\ \sim\ & \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} k\mathrm{P}(\hat{P}_{1i} > t) \\ =\ & np\mathrm{P}(\hat{P}_{1i} > t). \end{aligned} \tag{33}$$

The second $\sim$ sign follows from a well-known property of the class of regularly varying distributions (again, cf. [3]). We conclude from (30), (33) and (28) that, for $t \to \infty$,

$$\begin{aligned} R_2(0) - r_2(t) &\sim \frac{b_2 p}{1-b_2} \mathrm{P}(\hat{P}_1 > t) \\ &\sim \frac{1}{(1-b_2)r_1^*} \frac{\lambda_1 C_1}{\Gamma(2-\nu_1)} ((1-\rho_1)t)^{1-\nu_1} l_1(t). \end{aligned} \tag{34}$$

Note that $b_2 < 1$ if $\rho_1 < 1$, $\rho_2 < 1$, $r_1^* \geq 1$, which is the case under consideration in this section. Again applying Lemma A.1, for $w \downarrow 0$,

$$R_2(w) - R_2(0) \sim -\frac{\lambda_1 C_1}{(1-b_2)r_1^*}(\frac{w}{1-\rho_1})^{\nu_1-1} l_1(\frac{1}{w}). \tag{35}$$

It follows from (35), (20) and (21) that, for $w \downarrow 0$,

$$\begin{aligned} &\psi_2(\delta_1(w)) - \mathrm{e}^{-P_1(0)} \\ \sim\ & -\mathrm{e}^{-P_1(0)} \frac{\lambda_1 C_1}{(1-b_2)r_1^*}(\frac{w}{1-\rho_1})^{\nu_1-1} l_1(\frac{1}{w}). \end{aligned} \tag{36}$$

From (7), (11) and the fact that $X_{1i} < 0$ *w.p.* 1 (cf. (12)):

$$P_1(0) = \sum_{n=1}^{\infty} \frac{b_1^n}{n} = -\ln(1-b_1).$$

Using this formula and the fact that $b_1 = \rho_2$ if $r_2^* = 1$ (cf. (9)), we get from (36): For $w \downarrow 0$,

$$\begin{aligned} &\psi_2(\delta_1(w)) - (1-\rho_2) \\ \sim\ & -(1-\rho_2)\frac{\lambda_1 C_1}{(1-b_2)r_1^*}(\frac{w}{1-\rho_1})^{\nu_1-1} l_1(\frac{1}{w}). \end{aligned} \tag{37}$$

Finally, see (24), for $s \downarrow 0$,

$$\psi_2(s) - (1-\rho_2) \sim -(1-\rho_2)\frac{\lambda_1 C_1}{(1-b_2)r_1^*} s^{\nu_1-1} l_1(\frac{1}{s}). \tag{38}$$

Using (5), (19) and (38), and the fact that (cf. (2.23) of [9], or (5))

$$\frac{\psi_1(0)}{1-\rho_1} + \frac{r_1^*-1}{1-\rho_1}[\psi_0 - \psi_2(0)] = 1, \tag{39}$$

it follows that, for $s \downarrow 0$,

$$\begin{aligned} &\mathrm{E}[\mathrm{e}^{-sV_1}] - 1 \sim -[\frac{1}{1-\rho_1} \\ -\ & \frac{(r_1^*-1)(1-\rho_2)}{1-\rho_1}\frac{1}{(1-b_2)r_1^*}]\lambda_1 C_1 s^{\nu_1-1} l_1(\frac{1}{s}). \end{aligned} \tag{40}$$

Using (10), we can rewrite this into

$$\mathrm{E}[\mathrm{e}^{-sV_1}] - 1 \sim -\frac{\lambda_1 C_1}{K-\rho_1} s^{\nu_1 - 1} l_1(\frac{1}{s}), \quad s \downarrow 0, \tag{41}$$

with $K := \rho_2 + (1-\rho_2) r_1^*$. Applying Lemma A.1 once more, we have proven the main result of this section:

Theorem III.1: If $\mathrm{P}(B_1 > t)$ is regularly varying at infinity of index $-\nu_1 \in (-2,-1)$, as given in (18), and if $\rho_1 < 1$, then $\mathrm{P}(V_1 > t)$ is regularly varying at infinity of index $1-\nu_1$, as given below:

$$\begin{aligned} \mathrm{P}(V_1 > t) &\sim \frac{1}{K-\rho_1} \frac{\lambda_1 C_1}{\Gamma(2-\nu_1)} t^{1-\nu_1} l_1(t) \\ &\sim \frac{\rho_1}{K-\rho_1} \mathrm{P}(B_1^{res} > t), \quad t \to \infty. \end{aligned} \tag{42}$$

Remark III.1: The above theorem implies that (cf. [7]) $\mathrm{P}(V_1 > t)$ behaves exactly as if Q_1 is an $M/G/1$ queue in isolation, with server speed K. Indeed K can be interpreted as the average available service speed for Q_1; $K = 1$ if $r_1^* = 1$. Note that $K - \rho_1 = r_1^*(1-b_2)$. The distribution of B_2 only plays a role via its mean. The theorem has a similar flavor as the 'reduced load equivalence' results of Agrawal et al. [1] for fluid queues, w.r.t. taking the influence of type-2 traffic into account.

Remark III.2: In [4], [5] similar results have been obtained for a related model with generalized processor sharing. The method employed in [4], [5] is to derive lower and upper bounds for the workload tail, which asymptotically coincide.

Remark III.3: In the case $r_2^* = 1$, which was studied in this section, Q_1 behaves as an $M/G/1$ queue with two service speeds. During $\exp(\lambda_2)$ periods the service speed is r_1^*, and during busy periods of Q_2 the service speed is 1. All those periods are independent. As far as we know, there are no exact results known for the workload distribution in an $M/G/1$ queue with speeds that change according to an alternating renewal process (except for various studies regarding the case of an alternation between positive speed and *zero* speed). The exact analysis of the above-mentioned case is implicitly contained in the analysis in Chapter III.3 of [9]. It should be noted that an $M/G/1$ busy period cannot represent any arbitrary distribution of a non-negative low-speed period.

In the present section we have assumed that $\rho_1 < 1$, i.e., the server in Q_1 would have been able to handle all the work in its queue without any assistance of the server at Q_2 (without periods of high speed r_1^*). It makes sense that in this case the tail behaviour of V_1 is not really influenced by Q_2, except for the factor K. One may expect this to be different when $\rho_1 \geq 1$. The case $\rho_1 > 1$ will be investigated in the next section. The boundary case $\rho_1 = 1$ is the topic of a later study.

IV. Workloads for the case $\rho_1 > 1$

As in the previous section, $r_2^* = 1$ so that Q_2 is not influenced by Q_1. We assume that both (18) and (16) hold, i.e., both $\mathrm{P}(B_1 > t)$ and $\mathrm{P}(B_2 > t)$ are regularly varying at infinity, with indices $1 < \nu_1, \nu_2 < 2$. Higher values of ν_i can be handled with minor adaptations. Starting-point for studying the tail behaviour of workload V_1 is again Relation (5) for its LST, but we can no longer use (21) for the term $\psi_2(s)$ which is contained in it. 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 zeros of $f_1(s,w)$, cf. (14). In [9] it is observed that $\frac{d}{ds} f_1(s,w)$ 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 $\delta_1(w)$, $\mathrm{Re}\, w \geq 0$, into $\mathrm{Re}\, w < 0$. For $\rho_2 < 1$, $\rho_1 \geq 1$, and $w \in [w_0, 0]$, the *two* zeros of $f_1(s,w)$ in $[0, \delta_1(0)]$ will be indicated by $\epsilon_1(w)$ and $\delta_1(w)$, and such that

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

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

Next to (21), there exists the following relation if $\rho_1 \geq 1$ (cf. (6.24) of [9], and choose $r_2^* = 1$; also remember the definition of $\delta_2(w)$ above (15)):

$$\begin{aligned} &[(1-\frac{1}{r_1^*})\frac{w}{\delta_2(w)} - \frac{1}{r_1^*}\frac{w}{\epsilon_1(w)}]\psi_2(\epsilon_1(w)) \\ -\; &\frac{w}{\delta_2(w)}[\frac{1}{r_1^*}(\psi_1(\delta_2(w)) - \psi_0) + \psi_0] = 0. \end{aligned} \tag{43}$$

To determine the behaviour of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$ (which eventually will give us the behaviour of $\mathrm{E}[\mathrm{e}^{-sV_1}]$ for $s \downarrow 0$, hence that of $\mathrm{P}(V_1 > t)$ for $t \to \infty$), we need to determine the behaviour, for $w \uparrow 0$, of $\epsilon_1(w)$, $\delta_2(w)$ and $\psi_1(\delta_2(w))$ – the terms that appear in (43). Take $w < 0$, $w \uparrow 0$. Then (cf. (24)):

$$\epsilon_1(w) = \frac{-w}{\rho_1 - 1} + \frac{\lambda_1 C_1}{\rho_1 - 1}(\frac{-w}{\rho_1 - 1})^{\nu_1} l_1(\frac{-1}{w}), \quad w \uparrow 0. \tag{44}$$

In view of the symmetry between the regularly-varying-tail assumptions (18) and (16) and between the definitions of $\delta_1(w)$ and $\delta_2(w)$, it is readily seen from (24) that

$$\delta_2(w) = \frac{-w}{1-\rho_2} - \frac{\lambda_2 C_2}{1-\rho_2}(\frac{-w}{1-\rho_2})^{\nu_2} l_2(\frac{-1}{w}), \quad w \uparrow 0. \tag{45}$$

For $\rho_2 < 1$, $\psi_1(\delta_2(w))$ is specified by Formula (6.23) of [9]: For $\mathrm{Re}\, w \leq 0$,

$$\begin{aligned} \psi_1(\delta_2(w)) &= \psi_0 - r_1^* \mathrm{e}^{-P_1(0) - R_2(0)} \\ &\quad (1 - \mathrm{e}^{P_1(w) - P_2(w)}) \\ &= \psi_0(1 - r_1^*) + r_1^* \psi_0 \mathrm{e}^{P_1(w) - P_2(w)}. \end{aligned} \tag{46}$$

The last equality sign is verified by using (20).
It follows from (7) that

$$\begin{aligned} P_1(w) &= \sum_{n=1}^{\infty} \frac{b_1^n}{n} (\mathrm{E}[\mathrm{e}^{w\hat{P}_2}])^n \\ &= -\ln(1 - b_1 \mathrm{E}[\mathrm{e}^{w\hat{P}_2}]), \quad \mathrm{Re}\, w \leq 0. \end{aligned} \tag{47}$$

Hence, cf. (26) or Formula (6.5) of [9], for $\mathrm{Re}\, w \leq 0$,

$$\mathrm{e}^{P_1(w)} = \frac{1}{1 - b_1 \mathrm{E}[\mathrm{e}^{w\hat{P}_2}]} = \frac{1}{1 - \rho_2 \frac{1-\beta_2\{\delta_2(w)\}}{\beta_2 \delta_2(w)}}. \tag{48}$$

(Remember that $b_1 = \rho_2$ when $r_2^* = 1$.) Using (16), Lemma A.1 and (45), we obtain for $w \uparrow 0$:

$$e^{P_1(w)} = \frac{1}{1-\rho_2}[1 - \frac{\lambda_2 C_2}{1-\rho_2}(\frac{-w}{1-\rho_2})^{\nu_2-1} l_2(\frac{1}{\delta_2(w)})]. \quad (49)$$

We now turn to $e^{-P_2(w)}$. The analysis is similar to that of $e^{R_2(w)}$ in the previous section. Observe that $P_2(w)$ is the LST of, for $t > 0$,

$$p_2(t) := \sum_{n=1}^{\infty} \frac{b_2^n}{n} \mathrm{P}(-t < X_{21} + \ldots + X_{2n} < 0). \quad (50)$$

Consider, for $t > 0$,

$$P_2(0) - p_2(t) = \sum_{n=1}^{\infty} \frac{b_2^n}{n} \mathrm{P}(X_{21} + \ldots + X_{2n} < -t). \quad (51)$$

The calculations in (31)-(34) for $R_2(0) - r_2(t)$ require a slight adaptation because if $\rho_1 > 1$ then the busy period $\hat{P}_1$ is defective. It follows from (14) that

$$\rho_1 \frac{1 - \beta_1\{\delta_1(w)\}}{\beta_1 \delta_1(w)} = 1 - \frac{w}{\delta_1(w)},$$

so that, using (26),

$$\mathrm{P}(\hat{P}_1 < \infty) = \frac{1}{\rho_1}. \quad (52)$$

We can now mimic the calculations in (31)-(34): For $t \to \infty$,

$$\begin{aligned}
& P_2(0) - p_2(t) \\
\sim\ & \sum_{n=1}^{\infty} \frac{b_2^n}{n} \sum_{k=0}^{n} \binom{n}{k} (\frac{p}{\rho_1})^k (1-p)^{n-k} \\
& \mathrm{P}(\sum_{i=k+1}^{n} \hat{P}_{2i} > t) \\
\sim\ & \sum_{n=1}^{\infty} \frac{b_2^n}{n} \sum_{k=0}^{n} \binom{n}{k} (\frac{p}{\rho_1})^k (1-p)^{n-k} \\
& (n-k)\mathrm{P}(\hat{P}_2 > t) \\
=\ & \sum_{n=1}^{\infty} \frac{b_2^n}{n} n(1-p)(\frac{p}{\rho_1} + 1 - p)^{n-1} \mathrm{P}(\hat{P}_2 > t) \\
=\ & \frac{b_2(1-p)}{1 - b_2(\frac{p}{\rho_1} + 1 - p)} \mathrm{P}(\hat{P}_2 > t) \\
=\ & \frac{\rho_2}{1-\rho_2} \mathrm{P}(\hat{P}_2 > t). \quad (53)
\end{aligned}$$

In the first step we have used a property of the class $\mathcal{L}$ of long-tailed distributions which allowed us to omit the finite sum $\sum_{i=1}^{k} \hat{P}_{1i}$; in the second step, an elementary property of regularly varying functions is used, and in the last step we have used (10). Using the counterpart of (28) for $\hat{P}_2$, it finally follows that, for $t \to \infty$,

$$P_2(0) - p_2(t) \sim \frac{1}{1-\rho_2} \frac{\lambda_2 C_2}{\Gamma(2-\nu_2)} ((1-\rho_2)t)^{1-\nu_2} l_2(t), \quad (54)$$

yielding, for $w \uparrow 0$,

$$P_2(w) - P_2(0) \sim -\frac{\lambda_2 C_2}{1-\rho_2} (\frac{-w}{1-\rho_2})^{\nu_2-1} l_2(\frac{-1}{w}). \quad (55)$$

Using (20), $R_1(0) = 0$ and $P_i(0) + R_i(0) = -\ln(1-b_i)$, $i = 1, 2$, it follows that

$$\psi_0 \frac{e^{-P_2(0)}}{1-b_1} = 1 - b_2.$$

Combining this result with (46), (49) and (55) yields: For $w \uparrow 0$,

$$\begin{aligned}
& \psi_1(\delta_2(w)) - \psi_0(1 - r_1^*) + r_1^*(1-b_2) \\
=\ & \psi_1(\delta_2(w)) - \psi_1(0) \\
=\ & o(w^{1-\nu_2} l_2(\frac{-1}{w})). \quad (56)
\end{aligned}$$

The first equality follows from Formula (2.23) of [9], or indirectly from (39). The second equality follows from the interesting fact that the $w^{1-\nu_2}$ factors in $e^{P_1(w)}$ and $e^{-P_2(w)}$ are multiplied by the same constant, with different signs.

Remark IV.1: Notice that, with $r_2^* = 1$, Q_2 is an $M/G/1$ queue in isolation. According to [7], the tail of its workload distribution, $\mathrm{P}(V_2 > t)$, is regularly varying at infinity of index $1 - \nu_2$. However, it follows from (56) that $\mathrm{P}(V_1 = 0, V_2 > t) = o(t^{1-\nu_2} l_2(t))$, $t \to \infty$. The explanation is the following. The workload in Q_1 has a positive drift $\rho_1 - 1$ when $V_2 > 0$. Therefore $\mathrm{P}(V_1 = 0 | V_2 > t) = o(1)$ 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 behaviour of $\psi_1(\delta_2(w))$ for $w \uparrow 0$ allows us to determine the behaviour of $\psi_2(\epsilon_1(w))$ for $w \uparrow 0$. Using Relation (43) between $\psi_2(\epsilon_1(w))$ and $\psi_1(\delta_2(w))$, along with the asymptotic results (44) and (45) for $\epsilon_1(w)$ and $\delta_2(w)$, it follows after some calculations that, for $w \uparrow 0$,

$$\begin{aligned}
& \psi_2(\epsilon_1(w)) - (1-\rho_2) \quad (57) \\
\sim\ & -\frac{\rho_1 - 1}{r_1^*(1-b_2)} \lambda_2 C_2 (\frac{-w}{1-\rho_2})^{\nu_2-1} l_2(\frac{-1}{w}).
\end{aligned}$$

Using (44) once more, we have for $s \downarrow 0$:

$$\begin{aligned}
& \psi_2(s) - (1-\rho_2) \quad (58) \\
\sim\ & -\frac{\rho_1 - 1}{r_1^*(1-b_2)} \lambda_2 C_2 (s\frac{\rho_1 - 1}{1-\rho_2})^{\nu_2-1} l_2(\frac{1}{s}).
\end{aligned}$$

Finally we are ready to determine the tail behaviour of the workload V_1 at Q_1. The LST of V_1 is given by (5). The first factor in its righthand side is the LST of the workload distribution in Q_1 in isolation, with a server that always has speed 1 (the Pollaczek-Khintchine workload LST in the $M/G/1$ queue); this factor would give a $t^{1-\nu_1}$ tail behaviour, cf. (16) and (17) where the relevant $M/G/1$ theory (but for Q_2) is given. Using (39) and (58), the second factor in the righthand side of (5) is seen to yield a $t^{1-\nu_2}$ tail behaviour. To see which term 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 B_1 dominates, and (41) still holds when $\rho_1 > 1$:

$$\mathrm{E}[e^{-sV_1}] - 1 \sim -\frac{\lambda_1 C_1}{K - \rho_1} s^{\nu_1 - 1} l_1(\frac{1}{s}), \quad s \downarrow 0, \quad (59)$$

with $K = \rho_2 + (1-\rho_2) r_1^*$. Remember that $K - \rho_1 = r_1^*(1-b_2)$.
Case 2: $\nu_1 > \nu_2$. In this case the heavier tail of B_2 dominates, resulting in: For $s \downarrow 0$,

$$\begin{aligned} & \mathrm{E}[e^{-sV_1}] - 1 \\ \sim & -\frac{1-\frac{1}{r_1^*}}{1-b_2} \lambda_2 C_2 \left(s\frac{\rho_1-1}{1-\rho_2}\right)^{\nu_2-1} l_2\left(\frac{1}{s}\right) \\ = & -\frac{r_1^*-1}{K-\rho_1} \lambda_2 C_2 \left(s\frac{\rho_1-1}{1-\rho_2}\right)^{\nu_2-1} l_2\left(\frac{1}{s}\right). \end{aligned} \quad (60)$$

Case 3: $\nu_1 = \nu_2$. In this case, addition of the righthand sides of (59) and (60) gives the right asymptotic behaviour of $\mathrm{E}[e^{-sV_1}] - 1$.

Applying Lemma A.1 again, we have proven the main theorem of this section:

Theorem IV.1: If $\mathrm{P}(B_i > t)$, $i = 1, 2$, is regularly varying at infinity of index $-\nu_i \in (-2, -1)$, as given in (18), (16), and if $\rho_1 > 1$, then $\mathrm{P}(V_1 > t)$ is regularly varying at infinity of index $1 - \min(\nu_1, \nu_2)$:
If $\nu_1 < \nu_2$, then

$$\begin{aligned} \mathrm{P}(V_1 > t) & \sim \frac{1}{K-\rho_1} \frac{\lambda_1 C_1}{\Gamma(2-\nu_1)} t^{1-\nu_1} l_1(t) \\ & \sim \frac{\rho_1}{K-\rho_1} \mathrm{P}(B_1^{res} > t), \quad t \to \infty; \end{aligned} \quad (61)$$

If $\nu_1 > \nu_2$, then for $t \to \infty$:

$$\mathrm{P}(V_1 > t) \sim \frac{r_1^*-1}{K-\rho_1} \frac{\lambda_2 C_2}{\Gamma(2-\nu_2)} \left(\frac{\rho_1-1}{1-\rho_2}\right)^{\nu_2-1} t^{1-\nu_2} l_2(t). \quad (62)$$

If $\nu_1 = \nu_2$, then for $t \to \infty$:

$$\begin{aligned} & \mathrm{P}(V_1 > t) \sim \frac{1}{K-\rho_1} \frac{\lambda_1 C_1}{\Gamma(2-\nu_1)} t^{1-\nu_1} l_1(t) \\ + & \frac{r_1^*-1}{K-\rho_1} \frac{\lambda_2 C_2}{\Gamma(2-\nu_1)} \left(\frac{\rho_1-1}{1-\rho_2}\right)^{\nu_1-1} t^{1-\nu_1} l_2(t). \end{aligned} \quad (63)$$

The above result implies the following. If the tail of B_1 is heavier than that of B_2, then $\mathrm{P}(V_1 > t)$ behaves exactly as if Q_1 is an $M/G/1$ queue in isolation, with server speed K (which is the average available speed for Q_1). But if the tail of B_2 is heavier than that of B_1 and $\rho_1 > 1$ (server 1 needs the help of server 2), then the former tail behaviour determines that of $\mathrm{P}(V_1 > t)$.

Remark IV.2: Formula (62) has the following interesting interpretation. First notice that the workload of Q_1 has a positive drift $\rho_1 - 1$ during the busy periods P_2 of Q_2, and a negative drift $\rho_1 - r_1^*$ during the $(\exp(\lambda_2)$ distributed) idle periods of Q_2. Now consider a fluid queue fed by one on/off source. The off-periods are $\exp(\lambda_2)$ distributed, and the on-periods are distributed like the busy periods of Q_2 (which is an $M/G/1$ queue in isolation, since $r_2^* = 1$). During off-periods, the buffer content V of the fluid queue decreases at rate $r_1^* - \rho_1$. During on-periods, the buffer content V increases at rate $\rho_1 - 1$. Jelenković and Lazar [12] have proven for this model that, with P_2^{res} denoting a residual busy period, for $t \to \infty$:

$$\mathrm{P}(V > t) \sim \frac{(1-\rho_2)\rho_2(r_1^*-1)}{K-\rho_1} \mathrm{P}\left(P_2^{res} > \frac{t}{\rho_1 - 1}\right). \quad (64)$$

To handle the latter tail probability, use the result of De Meyer and Teugels for the relation between the tail of the regularly varying service time distribution in an $M/G/1$ queue and the tail of its busy period (change indices 1 into 2 in (23)). The interesting conclusion then is, that the tail behaviour of the workload in this fluid queue is *equivalent* to the tail behaviour of V_1. This gives very useful insight into the workload tail behaviour under more complicated GPS disciplines, in cases where the guaranteed rate of a source is not sufficient to handle all its work.

V. Conclusions

In this paper we have studied a model of two coupled $M/G/1$ queues. The service speed at the first queue is increased during periods in which the second queue is empty. Under the assumption that the service request distributions at both queues are regularly varying at infinity of index $-\nu_1$ and $-\nu_2$, we have presented a detailed analysis of the tail behaviour of the workload distribution at each queue. If the guaranteed unit speed of server 1 is already sufficient to handle its offered traffic, then the workload distribution at the first queue is regularly varying at infinity of index $1 - \nu_1$. But if it is not sufficient, then the workload at Q_1 has a positive drift during regularly varying busy periods of Q_2, and the workload distribution at the first queue is regularly varying at infinity of index $1 - \min(\nu_1, \nu_2)$. In particular, traffic at server 1 is then no longer protected from worse behaving (heavier-tailed) traffic at server 2.

We believe that these results form a useful step towards determining the extent to which GPS-based scheduling algorithms are able to protect individual connections. Several extensions are possible, and we intend to study these in a following paper: (i) the special case $\rho_1 = 1$; (ii) the special case $\frac{1}{r_1^*} + \frac{1}{r_2^*} = 1$; (iii) the general case $r_1^* \geq 1$, $r_2^* \geq 1$; (iv) one of the two service request distributions has an exponential tail.

The thus obtained results, along with the results obtained in [4], [5], should give insight into the performance of a wide range of GPS-based scheduling disciplines, and into the effect of heavy-tailed input characteristics. This might be useful in various respects, e.g., in making appropriate choices for the weight factors ϕ_i in GPS.

References

[1] Agrawal, R., Makowski, A.M., Nain, P. (1998). On a reduced load equivalence for a fluid model under subexponential assumptions. Report INRIA Sophia-Antipolis.

[2] D. Bertsimas, I.Ch. Paschalidis and J.N. Tsitsiklis (1997). *Large deviations analysis of the generalized processor sharing policy.* Report Boston University.

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

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

[5] S.C. Borst, O.J. Boxma and P.R. Jelenković (1999). *Asymptotic behaviour of generalized processor sharing with long-tailed traffic sources.* Report CWI, June 1999; in these Proceedings.

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

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

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

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

[10] P. Dupuis and K. Ramanan (1998). *A Skorokhod Problem formulation and large deviation analysis of a processor sharing model.* Queueing Systems **28**, 109-124.

[11] G. Fayolle and R. Iasnogorodski (1979). *Two coupled processors: the reduction to a Riemann-Hilbert problem.* Z. Wahrsch. Verw. Gebiete **47**, 325-351.

[12] P.R. Jelenković and A.A. Lazar (1999). *Multiplexing on-off sources with subexponential on periods.* Adv. Appl. Probab. **31**, 394-421.

[13] A.G. Konheim, I. Meilijson and A. Melkman (1981). *Processor sharing of two parallel lines.* J. Appl. Probab. **18**, 952-956.

[14] L. Massoulié (1998). *Large deviations for polling and weighted fair queueing service systems.* Report France Télécom-CNET.

[15] A. de Meyer and J.L. Teugels (1980). *On the asymptotic behaviour of the distributions of the busy period and service time in $M/G/1$.* J. Appl. Probab. **17**, 802-813.

[16] K. Park and W. Willinger (1999). *Self-similar Network Traffic and Performance Evaluation* (Wiley, New York).

[17] Z.-L. Zhang, D. Towsley and J. Kurose (1995). *Statistical analysis of the generalized processor sharing discipline.* IEEE J. Sel. Areas Commun. **13**, 1071-1080.

APPENDIX

I. HEAVY TAILS

Definition A.1: A distribution function $F(\cdot)$ on $[0,\infty)$ is called *long-tailed* ($F(\cdot) \in \mathcal{L}$) if

$$\lim_{x\to\infty} \frac{1-F(x-y)}{1-F(x)} = 1, \quad \text{for all real } y.$$

A well-known subclass of the class of long-tailed distributions is the class of regularly varying distributions $\mathcal{R}$ (this class contains the Pareto distribution):

Definition A.2: A distribution function $F(\cdot)$ on $[0,\infty)$ is called *regularly varying of index* $-\nu$ ($F(\cdot) \in \mathcal{R}(-\nu)$) if

$$F(x) = 1 - \frac{l(x)}{x^{\nu}}, \quad \nu \geq 0,$$

where $l(x) : R_+ \to R_+$ is a function of slow variation, i.e., $\lim_{x\to\infty} l(\eta x)/l(x) = 1, \eta > 1$.

A key reference is [3]. The following lemma (cf. Lemma 2.2 in [6], which is an extension of Theorem 8.1.6 in [3]), links the regularly varying tail behaviour of $\mathrm{P}(Z > t)$ for $t \to \infty$ to the behaviour of its LST $f(s)$. It plays a key role in the proofs of our main results.

Lemma A.1: Let Z be a non-negative random variable with LST $f(s)$, $l(t)$ a slowly varying function, $\nu \in (n, n+1)$ ($n \in N$) and $C \geq 0$. Then the following are equivalent:

(i) $\mathrm{P}(Z > t) = [C + o(1)]t^{-\nu}l(t), \quad t \to \infty$;

(ii) $\mathrm{E}[Z^n] < \infty$ and $f(s) - \sum_{j=0}^{n} \frac{\mathrm{E}[Z^j](-s)^j}{j!} = (-1)^n\Gamma(1-\nu)[C + o(1)]s^{\nu}l(1/s), \quad s \downarrow 0$.