ELSEVIER

Fluid queues and regular variation

O.J. Boxma *

CWI, P.O. Box 94079, 1090 GB Amsterdam, Netherlands;
Tilburg University, Faculty of Economics, P.O. Box 90153, 5000 LE Tilburg, Netherlands

Abstract

This paper considers a fluid queueing system, fed by N independent sources that alternate between silence and activity periods. We assume that the distribution of the activity periods of one or more sources is a regularly varying function of index ζ. We show that its fat tail gives rise to an even fatter tail of the buffer content distribution, viz., one that is regularly varying of index $\zeta + 1$. In the special case that $\zeta \in (-2, -1)$, which implies long-range dependence of the input process, the buffer content does not even have a finite first moment.

As a queueing-theoretic by-product of the analysis of the case of N identical sources, with $N \to \infty$, we show that the busy period of an M/G/∞ queue is regularly varying of index ζ iff the service time distribution is regularly varying of index ζ.

Keywords: Fluid queue; Regular variation; Long-range dependence; Buffer content

1. Introduction

Recent measurements (see e.g. [11,15]) have revealed that in high-speed telecommunication networks, like the ATM-based Broadband ISDN, traffic conditions can occur that exhibit long-range dependence and burstiness over an extremely wide range of time scales. In terms of probability distributions, such conditions may be due to the occurrence of interarrival or service time distributions with a fat, non-exponential, tail. It is important to obtain qualitative insight into the effect of such fat tails on performance measures like workload and waiting time distributions. This paper aims to contribute to such insight.

We study a fluid queueing system, that is fed by N independent on/off sources. Each source alternates between activity periods and silence periods; in the former periods a source feeds work into the buffer at constant rate (larger than the output rate of the buffer). This fluid queueing system has been found to be particularly well matched to the ATM environment. Since the fundamental publications of Anick et al. [2] and Kosten [10], it has become a key model for the performance analysis of high-speed

---

* E-mail: onno@cwi.nl.

telecommunication networks. However, in [2,10] and subsequent publications the activity and silence periods are exponentially distributed or determined by some Markov process, leading to exponential behaviour of the tail of the buffer content distribution. We are interested in the case that the activity periods of one or more of the sources have a fat tail (this may, for example, arise in file transfers). More precisely, we consider the case that the tail of some activity period distributions is *regularly varying*.

A measurable positive function f is called regularly varying of index ζ if, for all $x > 0$,

$$f(xt)/f(t) \to x^{\zeta}, \quad t \to \infty, \tag{1.1}$$

(cf. [3], p. 18). One writes $f \in R_{\zeta}$. When $\zeta = 0$, one speaks of a *slowly varying function*; this could for instance be a constant, or a logarithmic function. We shall say that a distribution $F(.)$ is regularly varying (or "has a regularly varying tail") when $F^c(t) := 1 - F(t)$ is a regularly varying function. The class of regularly varying distributions is an important one, containing for example the Pareto distribution. Of particular interest to us is the case that an activity period distribution has a regularly varying tail of index $\zeta \in (-2, -1)$. In that case the first moment of the distribution exists, but the variance is infinite. This case is known to give rise to long-range dependence. We shall roughly indicate this. Long-range dependence of a continuous-time process $\{Z_t, t \geq 0\}$ is defined by the property that $\text{cov}(Z_0, Z_t)$ is not integrable over t. Let Y_t be 1 when the source is active, and 0 otherwise. Observe that, for large values of t, $\text{cov}(Y_0, Y_t)$ is linearly related to the probability that the activity period that covers 0 is still going on at t. The resulting distribution of the residual activity period is again regularly varying, but of index $\zeta + 1 \in (-1, 0)$ (cf. [3], and see also Section 4 of this paper). Hence the covariance function is not integrable over t, implying long-range dependence.

For the traditional GI/G/1 queue, Cohen [6] has studied the effect of regularly varying interarrival or service time distributions on waiting-time and workload distributions. We shall exploit his main result, which states: the waiting-time distribution in the GI/G/1 queue is regularly varying of index $1 - \nu$ (with $\nu > 1$) iff the service time distribution is regularly varying of index $-\nu$. See Abate et al. [1] for a recent discussion of waiting-time tail probabilities in the GI/G/1 queue with a fat-tailed service time distribution, with an emphasis on approximations based on waiting-time tail asymptotics.

Other papers that are particularly relevant for our study are those of Brichet et al. [4] and Norros [13]. Norros [13] studies a fluid queue with Fractional Brownian Motion as input process. This self-similar input process exhibits long-range dependence. Norros analyzes the buffer content process of this model. Brichet et al. [4] consider the same model as we, restricting themselves to N identical sources. In the limit of a large number of sources, in heavy traffic, they obtain the important result that the tail of the buffer content distribution is Weibullian. They also indicate how their model relates to that of Norros.

Our paper is organized as follows. The fluid queueing system under consideration is described in Section 2. Section 3 displays for this model some key results of [7,9] that form the starting-point of our approach. Section 4 summarizes the main ingredients of the theory of regular variation. Our results are gathered in Sections 5–7. Section 5 considers the case of a single on/off source. We show that, analogously to the ordinary GI/G/1 queue, the buffer content distribution is regularly varying with parameter $1 - \nu$ iff the activity period distribution is regularly varying with parameter $-\nu$. In Section 6 it is shown that the buffer content distribution is still regularly varying with parameter $1 - \nu$ if, in addition to that single source, there are one or more sources with *exponential* activity and silence periods. Section 7 is concerned with the case of an infinite number of identical on/off sources, all with regularly varying activity period distributions. We present results for the regularly varying behaviour of (i) the period in which at least one source is active, (ii) the net buffer increment during such a period, and (iii) the buffer

content. The result (i) implies that the busy period distribution of an M/G/∞ queue is regularly varying of index $-\nu$ iff the service time distribution is regularly varying of index $-\nu$.

2. The model

Consider a fluid queueing system with an infinite storage capacity and constant, unit, output rate. This system receives input from N independent on/off sources. Source j has mutually independent alternating silence periods S_{ij} and activity periods A_{ij}, $j = 1, \ldots, N$, $i = 1, 2, \ldots$. Source j constantly transmits at rate r_j when active, so source j feeds $r_j A_{ij}$ traffic into the buffer during its ith activity period. The silence periods S_{ij} have a negative exponential distribution with mean $1/\lambda_j$, and the activity periods A_{ij} have distribution $A_j(\cdot)$ with $A_j(0+) = 0$ and with mean α_j and Laplace–Stieltjes Transform (LST) $\alpha_j(\cdot)$. We assume that $r_j > 1$, $j = 1, \ldots, N$. This assumption is somewhat restrictive and can be relaxed (see Remark 6.2 and also Section 5 of [9]); we shall discuss this issue in more detail in a future study.

The total traffic load offered to the buffer per unit time is assumed to be less than one:

$$L := \sum_{j=1}^{N} \frac{r_j \alpha_j \lambda_j}{1 + \alpha_j \lambda_j} < 1. \tag{2.1}$$

This is the ergodicity condition, cf. [9].

3. The fluid queue

Let V_t denote the content of the buffer at time t. Assume that the buffer is empty at time zero. We call C_n, $n = 1, 2, \ldots$, the length of the nth cumulative activity period after zero; this is a period in which at least one source is active. Such a period is followed by a period, with length I_{n+1}, in which all sources are silent. Denote by B_n the net increment of the buffer content during C_n, and by W_n the buffer content at the beginning of the nth cumulative activity period. It is easily seen (cf. Fig. 1 and [9]) that

$$W_{n+1} = \max[0, W_n + B_n - I_{n+1}], \quad n = 1, 2, \ldots, \tag{3.1}$$
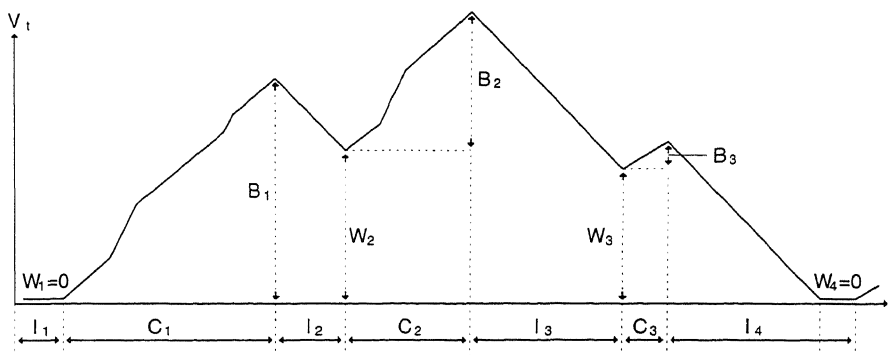$$W_1 = 0.$$

Fig. 1. The buffer content

The assumptions in the previous section about the silence periods imply that I_1, I_2, ... are independent and negative exponentially distributed with mean $1/\Lambda := 1/[\lambda_1 + \cdots + \lambda_N]$. Furthermore, B_1, B_2, ... are independent random variables; and W_n, B_n and I_{n+1} are mutually independent. Hence the sequence W_n, $n = 1, 2, \ldots$, has the same stochastic behaviour as the sequence of actual waiting times in an M/G/1 queue with arrival rate Λ and with service time distribution $B(\cdot)$, the distribution of B_n. So once we have determined $B(\cdot)$, M/G/1 theory tells us the steady-state distribution, $W(\cdot)$, of the buffer content at the start of cumulative activity periods. In this paper we shall concentrate on $W(\cdot)$. Knowledge about its behaviour will also enable us to draw conclusions about the buffer content distribution at arbitrary epochs; this will be done in Remarks 5.4 and 6.4.

Determination of $B(\cdot)$ is a far from trivial problem if $N > 1$, because the number of active sources may fluctuate during a cumulative activity period. Cohen tackles this problem in [7] for N identical sources with $r_1 = \cdots = r_N = 1$, and in [9] for N different sources. He observes that the activity and silence periods of the N sources give rise to N alternating renewal processes, and uses this to obtain the LST of the joint distribution of B and C (B and C denote generic random variables with joint distribution the joint limiting distribution of the random variables B_n and C_n). For the model with N different sources, with $\operatorname{Re}\omega \geq 0$, $\operatorname{Re} s > 0$, and some $y > 0$, according to [9],

$$\frac{1}{s + \Lambda(1 - \mathrm{E}[e^{-\omega(B+C)-sC}])} = \int_{t=0}^{\infty} e^{-st} \prod_{j=1}^{N} \left\{ \frac{1}{2\pi i} \int_{-i\infty+y}^{i\infty+y} \frac{e^{ut}}{u + \lambda_j[1 - \alpha_j\{\omega r_j + u\}]} du \right\} dt. \tag{3.2}$$

One can obtain the LST $\beta\{\omega\} := \mathrm{E}[e^{-\omega B}]$, if the validity of (3.2) can be extended to the domain $\operatorname{Re}\omega \geq 0$, $\operatorname{Re} s > -R - \operatorname{Re}\omega$ for some $R > 0$ so that $s = -\omega$ can be put; we shall return to this when the need arises, in Sections 5 and 6.

As a by-result of (3.2), Cohen [9] shows that

$$\Lambda\mathrm{E}[C] = -1 + \prod_{j=1}^{N}(1 + \alpha_j\lambda_j), \tag{3.3}$$

$$\Lambda\mathrm{E}[B] = 1 - (1 - L)\prod_{j=1}^{N}(1 + \alpha_j\lambda_j). \tag{3.4}$$

Note that the ergodicity condition $L < 1$ (cf. (2.1)) implies $\Lambda\mathrm{E}[B] < 1$.

In Sections 5–7 we shall use (3.1) and (3.2) to draw conclusions about the tail behaviour of the buffer contents W_n and V_t, given the tail behaviour of the activity periods A_{ij} of the sources. We shall assume that one or more of the sources have activity periods with regularly varying tails. Hence we need some key results concerning regularly varying distributions and their LST. These results are presented in the next section.

4. Regular variation

Regular variation is an important concept in probability theory and various other fields. The main reference text is the book [3], to which we refer for proofs and detailed discussions of the results gathered below. See (1.1) for the definition of regular variation. A slowly varying function (see below (1.1)) will in the sequel be denoted by $l(\cdot)$.

The next lemma links the behaviour of the tail of a probability distribution $F(t)$ on $[0, \infty)$ to the behaviour of its LST $\phi(s)$ for $s \to 0$. This is part of Theorem 8.1.6 on p. 333/334 of [3], originally due to Bingham and Doney. Assume that the first n moments $\mu_1, \ldots, \mu_n$ of $F(\cdot)$ are finite. Define

$$\phi_n(s) := (-1)^{n+1} \left[\phi(s) - \sum_{j=0}^{n} \mu_j \frac{(-s)^j}{j!} \right]. \tag{4.1}$$

Lemma 4.1. *Let* $v = n + \psi$ *with* $0 < \psi < 1$. *The following are equivalent:*

$$\phi_n(s) \sim s^v l(1/s), \quad s \to 0, \tag{4.2}$$

$$F^c(t) \sim \frac{(-1)^n}{\Gamma(1-v)} t^{-v} l(t), \quad t \to \infty. \tag{4.3}$$

In Theorem 8.1.6 of [3] the somewhat more complicated cases $\psi = 0, 1$ are also discussed.

The following two results, that will be useful in the sequel, follow quickly from Lemma 4.1 (but can also be obtained by more direct probabilistic and analytic arguments).

(i)

$$F^c(t) \in R_{-v} \Rightarrow \int_t^\infty F^c(x)dx \in R_{1-v}, \tag{4.4}$$

and in fact (cf. [3], p. 28, as part of the Karamata theorem):

$$\frac{t F^c(t)}{\int_{x=t}^\infty F^c(x)\, dx} \to v - 1, \quad t \to \infty. \tag{4.5}$$

This follows from Lemma 4.1 by observing that $\int_0^t F^c(x)dx$ has LST $[1 - \phi(s)]/s$.

(ii) The sum of two independent nonnegative random variables with regularly varying distributions of index v_1 respectively v_2 is again regularly varying, of index $\max(v_1, v_2)$.

5. N = 1

Consider the case of a single source ($N = 1$), transmitting at rate $r > 1$ when active (if $r \leq 1$ then the buffer is always empty). The following theorem states that the distribution $W(t) = \lim_{n \to \infty} P\{W_n < t\}$ of the buffer content at the beginning of activity periods is regularly varying at infinity of index $1 - v$ iff the activity period distribution of this source is regularly varying at infinity of index $-v$. Unlike in the remainder of the paper, we do not assume in the theorem below that the silence periods are negative exponentially distributed. We further suppress the index 1 in this section, so we write $A(t)$ instead of $A_1(t)$, etc.

Theorem 5.1. *For* $t \to \infty$, $l(t) \in R_0$, *and* $v > 1$,

$$A^c(t) = (v - 1)(\alpha/t)^v l(t) \iff W^c(t) = \frac{(r - 1)\alpha\lambda}{1 - (r - 1)\alpha\lambda} ((r - 1)\alpha/t)^{v-1} l(t). \tag{5.1}$$

Note. Actually in the "if" ("only if") part of (5.1) the second (first) equality sign should be replaced by "$\sim$"; for briefness sake we use the above notation in (5.1), and in subsequent relations of the same type.[1]

Proof. For $N = 1$, the nth cumulative activity period C_n coincides with the nth activity period A_n; and the net increment B_n of the buffer content during C_n equals $B_n = (r - 1)A_n$, with LST

$$\beta\{\omega\} = \alpha\{\omega(r - 1)\}, \quad \text{Re}\,\omega \geq 0. \tag{5.2}$$

Clearly, for $t \to \infty$ and with $l(t) \in R_0$,

$$A^c(t) = (v - 1)(\alpha/t)^v l(t) \iff B^c(t) = (v - 1)((r - 1)\alpha/t)^v l(t). \tag{5.3}$$

The crucial observation is (see (3.1)) that $W(t)$ behaves like the waiting-time distribution in the GI/G/1 queue with service times B_n and interarrival times I_n. We can now apply Theorem 1 of [6] for the ordinary GI/G/1 queue, which relates the tail behaviour of the waiting-time distribution $W_{GI/G/1}(t)$ and that of the service time distribution $B_{GI/G/1}(t)$. This theorem states that, for $t \to \infty$ and $v > 1$, and with β denoting mean service time, and ρ traffic load:

$$B^c_{GI/G/1}(t) = (v - 1)(\beta/t)^v l(t) \iff W^c_{GI/G/1}(t) = \frac{\rho}{1 - \rho}(\beta/t)^{v-1} l(t). \tag{5.4}$$

The theorem follows from (5.3) and (5.4). Actually, the result of [6] has been extended [14] to the larger class of subexponential distributions; accordingly, one can extend Theorem 5.1 to that class of activity period distributions. $\square$

We close the section with five remarks.

Remark 5.1. If $A^c(t) \in R_{-v}$ with $v \in (1, 2)$ (the case of long-range dependence, as observed in Section 1), then the first moment of $W(t)$ is infinite. This emphasizes the profound effect of long-range dependence on the buffer content, as has also been observed in [4] and [13].

Remark 5.2. Note that if the *silence periods* have a regularly varying tail, this does not imply that $W^c(t)$ is regularly varying. That makes sense: an extremely long silence period will most likely lead to an empty queue. See [4] for some further remarks about fat-tailed silence periods.

Remark 5.3. We return to the case of exponential silence periods. Formula (3.1) implies that the LST of $W(\cdot)$ is given by the M/G/1 Pollaczek–Khintchine formula: with $\beta := E[B_n]$,

$$\int_{t=0-}^{\infty} e^{-\omega t} dW(t) = \frac{1 - \beta\lambda}{1 - \lambda(1 - \beta\{\omega\})/\omega}, \quad \text{Re}\,\omega \geq 0. \tag{5.5}$$

By inversion,

$$W(t) = \sum_{n=0}^{\infty} (1 - \beta\lambda)(\beta\lambda)^n \left\{ \int_{u=0}^{t} \frac{1 - B(u)}{\beta} du \right\}^{n*}, \quad t \geq 0. \tag{5.6}$$

[1] *Note added in proof:* The reader is referred to [5] for a different discussion of the case $N = 1$.

It is interesting to observe that Theorem 5.1 easily follows from (5.5) and Lemma 4.1; Brichet et al. [4] make a similar observation for the ordinary M/G/1 queue. An insightful way to understand the relation between the regularly varying behaviour of the service-time and waiting-time distributions in an M/G/1 queue is to consider (5.6). Firstly, as remarked in (i) below Lemma 4.1, the residual service time distribution $\int_0^t [1 - B(u)]/\beta du$ has a regularly varying tail of index $1 - \nu$: the integration increases the index by one. Probabilistically this also makes sense: the residual part of a random variable with a regularly varying distribution will have a distribution with an even heavier tail. Secondly, the n-fold convolution of the residual service time distribution also has a regularly varying tail of index $1 - \nu$ (see (ii) below Lemma 4.1), but the coefficient is multiplied by n. In probabilistic terms: when the sum of n independent identically distributed random variables with regularly varying distribution exceeds a large value t, this is probably due to one of these n random variables being large. Finally, the multiplicative factor $((r - 1)\alpha\lambda)/(1 - (r - 1)\alpha\lambda) = \beta\lambda/(1 - \beta\lambda)$ arises in the theorem because a weighted sum is taken in (5.6): this only affects the coefficient of the regularly varying function, multiplying it by $\sum_{n=0}^{\infty} n(1 - \beta\lambda)(\beta\lambda)^n = \beta\lambda/(1 - \beta\lambda)$.

Remark 5.4. So far for W_n; what about the buffer content V_t at some arbitrary time t? In an ordinary M/G/1 queue these two quantities have the same limiting distribution. That is not true in the present model, but they are clearly quite closely related. Observe that the distribution of the sum of the independent random variables W_n and B_n has a regularly varying tail of index $1 - \nu$ iff the same holds for W_n (cf. the end of Section 4). Relating V_t to the immediately preceding end or beginning of an activity period (whichever has come last) makes it clear that the tail of its distribution is regularly varying of index $1 - \nu$ iff the same holds for W_n. We omit the details.

Remark 5.5. With a view towards the analysis of the case $N = 2$ in the next section, it is instructive to derive (5.2) from (3.2) in the case $N = 1$; remember that the latter formula yields the joint distribution of a cumulative activity period C and the net increment B during C. Observe that the second integral in (3.2) is the inversion formula for the Laplace Transform $[u + \lambda(1 - \alpha\{\omega r + u\})]^{-1}$, whereas the first integral in (3.2) is the Laplace transform of the second integral. Consequently the successive application of the two integral operators yields the integrand back again, with its argument replaced by s. Hence, from (3.2), for $\operatorname{Re}\omega \geq 0$, $\operatorname{Re}s > 0$,

$$\frac{1}{s + \Lambda(1 - \operatorname{E}[e^{-\omega(B+C)-sC}])} = \frac{1}{s + \lambda(1 - \alpha\{\omega r + s\})}.\tag{5.7}$$

So $\operatorname{E}[\exp(-\omega(B + C) - sC)] = \alpha\{\omega r + s\}$ for $\operatorname{Re}\omega \geq 0$, $\operatorname{Re}s > 0$ (note that $\Lambda = \lambda$ for $N = 1$). It is allowed to take $s = -\omega$ (see below (3.2)), since $r > 1$: $\alpha\{\omega(r - 1)\}$ is analytic for $\operatorname{Re}\omega \geq 0$, and the "M/G/1-like" denominator $\omega - \lambda(1 - \alpha\{\omega(r - 1)\})$ has no zeros in $\operatorname{Re}\omega \geq 0$, except for $\omega = 0$. We thus have rederived (5.2).

6. $N = 2$

Consider the case of two sources. The first source has activity period distribution $A_1(t)$ that is regularly varying at infinity of index $-\nu$, whereas the second source has a negative exponential activity period distribution; the silence periods of both sources are negative exponentially distributed. From the previous section we know that, in the absence of the second source, $W^c(t) \in R_{1-\nu}$. Below we derive the

LST of $W(t)$; the obtained expression will allow us to show that $W^c(t) \in R_{1-\nu}$ even in the presence of the second source. Moreover, it will also lead to the reverse statement that $A_1^c(t) \in R_{-\nu}$ if $W^c(t) \in R_{1-\nu}$.

Theorem 6.1. *For the case of two sources with exponential silence periods, and with exponential activity periods for source 2, and for $t \to \infty$, $l(t) \in R_0$, and $\nu > 1$,*

$$A_1^c(t) = (\nu - 1)(\alpha_1/t)^\nu l(t) \quad \Leftrightarrow \quad W^c(t) = \frac{\lambda_1 M(1 + \alpha_2\lambda_2)}{1 - \Lambda E[B]}(M/t)^{\nu-1}l(t), \tag{6.1}$$

with

$$M := \alpha_1\left(r_1 - 1 + \frac{\alpha_2\lambda_2 r_2}{1 + \alpha_2\lambda_2}\right). \tag{6.2}$$

Proof. Starting-point is formula (3.2) for $N = 2$. It will be used to obtain $E[e^{-\omega B}]$, and then $\int_{t=0-}^\infty e^{-\omega t}dW(t)$. We first evaluate the second term, X_2, of the product in the right-hand side (RHS) of (3.2): for $t \geq 0$, $\mathrm{Re}\,\omega \geq 0$, and some $y > 0$,

$$X_2 := \frac{1}{2\pi i}\int_{-i\infty+y}^{i\infty+y}\frac{e^{ut}}{u + \lambda_2(1 - \alpha_2\{\omega r_2 + u\})}du \tag{6.3}$$

$$= \frac{1}{2\pi i}\int_{-i\infty+y}^{i\infty+y}\frac{e^{ut}(1 + \alpha_2\omega r_2 + \alpha_2 u)}{(u + \lambda_2)(1 + \alpha_2\,\omega r_2 + \alpha_2 u) - \lambda_2}du$$

$$= \frac{1}{2\pi i\alpha_2}\int_{-i\infty+y}^{i\infty+y}\frac{e^{ut}(1 + \alpha_2\omega r_2 + \alpha_2 u)}{(u - u_+)(u - u_-)}du,$$

with poles

$$u_{+,-} := \frac{-[1 + \alpha_2\omega r_2 + \alpha_2\lambda_2] \pm [(1 + \alpha_2\omega r_2 + \alpha_2\lambda_2)^2 - 4\alpha_2^2\lambda_2\omega r_2]^{1/2}}{2\alpha_2}. \tag{6.4}$$

Note that both poles $u_{+,-}$ of the integrand in (6.3) are in the left half-plane. This is obvious for real ω. To see it for complex ω, apply Rouché's theorem. Consider the region G inside the closed contour $\hat{G}$ that consists of the line from $-iR$ to iR and a semi-circle to its left with radius R, and then let $R \to \infty$. The functions $f_1(u) := (u + \lambda_2)(1 + \alpha_2\omega r_2 + \alpha_2 u)$ and $f_2(u) := \lambda_2$ (appearing in the second line of (6.3)) are analytic inside G and continuous up to the boundary, while $|f_1(u)| > |f_2(u)|$ on $\hat{G}$; the statement now follows because $f_1(u)$ has two zeros inside G.

Subsequently take the contour integral of the integrand in (6.3) along $\hat{G}$. The contribution along the semi-circle disappears for $R \to \infty$, and what remains is the sum of the residues at the poles u_+ and u_-:

$$X_2 = \frac{e^{u_+t}}{(u_+ - u_-)\alpha_2}[1 + \alpha_2\omega r_2 + \alpha_2 u_+] \tag{6.5}$$

$$- \frac{e^{u_-t}}{(u_+ - u_-)\alpha_2}[1 + \alpha_2\omega r_2 + \alpha_2 u_-], \quad t \geq 0, \ \mathrm{Re}\,\omega \geq 0.$$

Substitution in (3.2) gives: for $\mathrm{Re}\,\omega \geq 0$, $\mathrm{Re}\,s > 0$, and some $y > 0$,

$$
\frac{1}{s + \Lambda(1 - \mathrm{E}[e^{-\omega(B+C)-sC}])}
$$

$$
= \frac{1 + \alpha_2\omega r_2 + \alpha_2 u_+}{(u_+ - u_-)\alpha_2} \int_{t=0}^{\infty} e^{-(s-u_+)t} \frac{1}{2\pi i} \int_{-i\infty+y}^{i\infty+y} \frac{e^{ut}}{u + \lambda_1(1 - \alpha_1\{\omega r_1 + u\})}\,du\,dt
$$

$$
- \frac{1 + \alpha_2\omega r_2 + \alpha_2 u_-}{(u_+ - u_-)\alpha_2} \int_{t=0}^{\infty} e^{-(s-u_-)t} \frac{1}{2\pi i} \int_{-i\infty+y}^{i\infty+y} \frac{e^{ut}}{u + \lambda_1(1 - \alpha_1\{\omega r_1 + u\})}\,du\,dt.
$$

As in Remark 5.5 we make the following observation. The two terms in the RHS of (6.6) are Laplace transforms of inversions of Laplace transforms. Successive application of the two integral operators yields the integrand with u replaced by $s - u_\pm$ (note that $\mathrm{Re}\,(s - u_\pm) > 0$): for $\mathrm{Re}\,\omega \geq 0$, $\mathrm{Re}\,s > 0$,

$$
\frac{1}{s + \Lambda(1 - \mathrm{E}[e^{-\omega(B+C)-sC}])} = \frac{1 + \alpha_2\omega r_2 + \alpha_2 u_+}{(u_+ - u_-)\alpha_2} \frac{1}{s - u_+ + \lambda_1(1 - \alpha_1\{\omega r_1 + s - u_+\})}
$$

$$
- \frac{1 + \alpha_2\omega r_2 + \alpha_2 u_-}{(u_+ - u_-)\alpha_2} \frac{1}{s - u_- + \lambda_1(1 - \alpha_1\{\omega r_1 + s - u_-\})}.
$$

$$\tag{6.7}$$

Multiply both sides of (6.7) by $(\Lambda\mathrm{E}B - 1)\omega$, and then take $s = -\omega$. Using the principle of permanence, one can show that it is allowed to continue (6.7) analytically to such values of s. The LST's $\alpha_1\{\cdot\}$ in the RHS of (6.7) are analytic for these values of s and ω. The LHS of (6.7) (after multiplication by ω) is analytic at least upto $s = -\omega$. In fact, multiplying (6.7) by $(\Lambda\mathrm{E}B - 1)\omega$ and then taking $s = -\omega$ gives exactly the LST of the waiting-time distribution $W(t)$ (cf. (3.1) and (5.5)). For $\mathrm{Re}\,\omega \geq 0$:

$$
\int_{t=0-}^{\infty} e^{-\omega t}\,dW(t) = \frac{(1 - \Lambda\mathrm{E}[B])\omega}{\omega - \Lambda(1 - \mathrm{E}[e^{-\omega B}])}
$$

$$
= \frac{(1 - \Lambda\mathrm{E}[B])\omega}{(u_+ - u_-)\alpha_2}\Big[\frac{1 + \alpha_2\omega r_2 + \alpha_2 u_+}{\omega + u_+ - \lambda_1(1 - \alpha_1\{\omega(r_1 - 1) - u_+\})}
$$

$$
- \frac{1 + \alpha_2\omega r_2 + \alpha_2 u_-}{\omega + u_- - \lambda_1(1 - \alpha_1\{\omega(r_1 - 1) - u_-\})}\Big],
$$

$$\tag{6.8}$$

with $\Lambda\mathrm{E}[B]$ specified in (3.4).

We now restrict ourself to the case $\nu \in (1, 2)$ (the case of long-range dependence). Hence, see Lemma 4.1,

$$
\alpha_1\{\omega\} = 1 - \alpha_1\omega + \Gamma(2 - \nu)(\alpha_1\omega)^\nu, \quad \omega \to 0. \tag{6.9}
$$

It easily follows from (6.4) that, for $\omega \to 0$,

$$
u_+ = -\omega\frac{\alpha_2\lambda_2 r_2}{1 + \alpha_2\lambda_2} + O(\omega^2), \tag{6.10}
$$

$$
u_- = -\frac{1 + \alpha_2\lambda_2}{\alpha_2} - \omega\frac{r_2}{1 + \alpha_2\lambda_2} + O(\omega^2). \tag{6.11}
$$

Substitution of these formulas in (6.8) gives, using (3.4) and after tedious but straightforward calcula-

tions:

$$\int_{t=0-}^{\infty} e^{-\omega t} dW(t) = 1 - \lambda_1 \Gamma(2 - \nu)\alpha_1^\nu \left(r_1 - 1 + \frac{\alpha_2\lambda_2 r_2}{1 + \alpha_2\lambda_2}\right)^\nu \frac{1 + \alpha_2\lambda_2}{1 - \Lambda E[B]}\omega^{\nu-1}, \quad \omega \to 0. \tag{6.12}$$

The $\Rightarrow$ part of the theorem now follows by application of Lemma 4.1. The $\Leftarrow$ part is obtained similarly, again exploiting the relation (6.8) between the LST's of $A_1(t)$ and $W(t)$. For $\nu \geq 2$ a similar result can be obtained from (6.8), at the expense of more lengthy calculations.   □

Note that a less direct way to prove the theorem would have been to apply Lemma 4.1 to $\beta\{\omega\}$ (as specified by (6.7) with $s = -\omega$) to prove that $B^c(t) \in R_{-\nu}$ if $A_1^c(t) \in R_{-\nu}$. Subsequently $W^c(t) \in R_{1-\nu}$ follows by exploiting the M/G/1-like relation (3.1) between B_n and W_{n+1}, and the GI/G/1 result (5.4).

We close the section with four remarks.

Remark 6.1. Take $\alpha_2 = 0$ (i.e., the second source is never active) to get the $N = 1$ case of Section 5 back again. Formula (6.12) reduces to

$$\int_{t=0-}^{\infty} e^{-\omega t} dW(t) = 1 - \lambda_1 \Gamma(2 - \nu)\alpha_1^\nu(r_1 - 1)^\nu \frac{1}{1 - \lambda_1\alpha_1(r_1 - 1)}\omega^{\nu-1}, \quad \omega \to 0. \tag{6.13}$$

The presence of the second (exponential) source apparently does not change the fact that $W^c(t) \in R_{1-\nu}$, but the coefficient is increased by a multiplicative factor

$$\left(1 + \frac{\alpha_2\lambda_2 r_2}{1 + \alpha_2\lambda_2}\frac{1}{r_1 - 1}\right)^\nu \frac{1 + \alpha_2\lambda_2}{1 - \Lambda E[B]}(1 - \lambda_1\alpha_1(r_1 - 1)) \geq 1. \tag{6.14}$$

Remark 6.2. So far we have assumed that $r_1 > 1$ and $r_2 > 1$. It is intuitively clear that taking $r_2 \leq 1$ cannot influence the conclusions of this section. But if $r_1 \leq 1$, it is possible that the buffer content decreases most of the time during which source 1 is active. Clearly the average increment of the buffer content per time unit that source 1 is active equals $r_1 - 1 + \alpha_2\lambda_2 r_2/(1 + \alpha_2\lambda_2)$ (cf. the factor in (6.12)!). Our conjecture is that (i) if the latter quantity exceeds 0, then $A_1^c(t) \in R_{-\nu}$ implies $W^c(t) \in R_{1-\nu}$ even if $r_1 < 1$, and (ii) if that quantity is less than zero, then $W^c(t)$ is *not* regularly varying even if $A_1^c(t) \in R_{-\nu}$. So while the presence of exponential sources in the case $r_1 > 1$ only influences the coefficient of the regularly varying function $W^c(t)$ (see the factor in (6.12)), in the case $r_1 \leq 1$ these exponential sources may make the difference between $W^c(t)$ being regularly varying or not.

Remark 6.3. Let us briefly consider the case of N sources, with $A_1^c(t) \in R_{-\nu}$ and all other sources having exponentially distributed activity periods. The last $N - 1$ terms of the product in (3.2) give a product of $N - 1$ terms that are completely similar to X_2 in (6.5). Accordingly, the RHS of (6.7) is replaced by 2^{N-1} terms similar to those in (6.7). The LST of $W(t)$ is obtained just as above, taking $s = -\omega$; all this eventually leads to the important conclusion that the one source with regularly varying activity periods dominates the behaviour of the buffer. The analysis and conclusion can even be extended to the case in which the activity periods of sources $2, \ldots, N$ have an exponential tail. The details will be presented in a future paper. The results suggest that the conjecture of Remark 6.2 may be extended in the following sense: if $r_1 < 1$, then the sign of $r_1 - 1 + \sum_{j=2}^{N} \alpha_j\lambda_j r_j/(1 + \alpha_j\lambda_j)$ determines whether

$W^c(t)$ is regularly varying or not. Conditions like this, which are reminiscent of effective bandwidth, could be most valuable in admission control.

Remark 6.4. So far we have concentrated on the distribution of W_n. Now we turn to the buffer content V_t at an arbitrary time t and to V_{max}, the maximum of the buffer content during a busy period (a period in which the buffer is never empty). For V_t the argument presented in Remark 5.4 again implies that the tail of its distribution is regularly varying of index $1 - \nu$ iff the same holds for W_n. Regarding V_{max}, note that it is the maximum of $W_n + B_n$ over a busy period; since W_n and B_n have the same distributions as the corresponding quantities in the ordinary M/G/1 queue, the same holds for V_{max}. Cohen ([8], Formula (III.7.67)) shows that, in the M/G/1 queue,

$$P\{V_{max} \geq t\} = \frac{1}{\Lambda} \frac{\frac{dW(t)}{dt}}{W(t)}, \quad t > 0. \tag{6.15}$$

The monotone density theorem ([3], p. 39) tells us that, for $t \to \infty$, $W^c(t) \sim t^{1-\nu}l(t) \Rightarrow dW^c(t)/dt \sim (1 - \nu)t^{-\nu}l(t)$. Hence $P\{V_{max} \geq t\} \in R_{-\nu}$ when $W^c(t) \in R_{1-\nu}$; the tail of the distribution of V_{max} is regularly varying *of the same index* as the tail of the activity period distribution of source 1.

It may be counterintuitive that the tail of the buffer content distribution is heavier than that of its maximum. However, note that V_{max} is sampled only once per busy period, whereas a busy period in which some W_n exceeds a large value x will probably contain many such large values. In the same vein, note that if the second moment of the service (activity) time distribution is not finite (as is the case if $1 < \nu < 2$), then the first moment of $W(t)$ is not finite, whereas $E[V_{max}] < \infty$.

7. $N = \infty$

In this section we consider the case of N identical sources, each with activity period distribution $A(\cdot)$ regularly varying of index $-\nu$ and with exponentially distributed silence periods with mean N/Λ, and we let $N \to \infty$. Activity periods hence begin according to a Poisson process of rate Λ. Cohen ([7], Formula (2.2.7)) derives the following expression for the joint LST of B and C (he takes $r = 1$): for $\text{Re}\,\omega \geq 0$, $\text{Re}\,s > 0$, and some $y > 0$,

$$\frac{1}{s + \Lambda(1 - E[e^{-\omega(B+C)-sC}])} = \int_{t=0}^{\infty} e^{-st} e^{-\Lambda I(\omega r, t)} dt, \tag{7.1}$$

with (cf. [7], Formula (2.2.9)): for $t \geq 0$, $\text{Re}\,\omega \geq 0$, and some $y > 0$,

$$I(\omega, t) := \frac{1}{2\pi i} \int_{-i\infty+y}^{i\infty+y} e^{ut} \frac{1 - \alpha\{\omega + u\}}{u^2} du$$
$$= t(1 - \alpha\{\omega\}) + E[Ae^{-\omega A}] - E[(A - t)e^{-\omega A}(A \geq t)], \tag{7.2}$$

with $(\cdot)$ an indicator function. Formula (7.1), that can be seen as a limiting form of (3.2) with $\lambda_j = \Lambda/N$, is our starting-point. It can be used to prove that $A^c(t) \in R_{-\nu}$ implies that $C^c(t) \in R_{-\nu}$ and $B^c(t) \in R_{-\nu}$. The latter result immediately yields (as in Section 5) that $W^c(t) \in R_{1-\nu}$. The derivations are rather technical. For lack of space we shall therefore refrain from detailed proofs, and just sketch our approach. We start with $C^c(t)$.

The cumulative activity period distribution

Taking $\omega = 0$ in (7.1) and multiplying by s yields (use (7.2) and apply partial integration in the last step):

$$\frac{s}{s + \Lambda(1 - \mathrm{E}[e^{-sC}])} = \int_{t=0}^{\infty} se^{-st}\exp\left\{-\Lambda\alpha + \Lambda \int_{x=t}^{\infty} A^c(x)dx\right\} dt$$

$$= 1 + e^{-\Lambda\alpha} \int_{t=0}^{\infty} e^{-st}d\left\{e^{\Lambda \int_{x=t}^{\infty} A^c(x)dx}\right\}. \tag{7.3}$$

Now exploit the fact that – after some scaling – an LST of a proper probability distribution appears in the RHS of (7.3), so that we can apply Lemma 4.1. If $A^c(x) = x^{-\nu}l(x)$ for $x \to \infty$, with $\nu \in (1, 2)$, so that (see also (4.5)) $\int_{x=t}^{\infty} A^c(x)dx = \frac{1}{\nu-1}t^{1-\nu}l(t)$ for $t \to \infty$, then

$$\frac{s}{s + \Lambda(1 - \mathrm{E}[e^{-sC}])} = e^{-\Lambda\alpha} - e^{-\Lambda\alpha}\Lambda\Gamma(1 - \nu)s^{\nu-1}l(1/s), \quad s \to 0. \tag{7.4}$$

Hence

$$\mathrm{E}[e^{-sC}] = 1 - \frac{e^{\Lambda\alpha} - 1}{\Lambda}s - e^{\Lambda\alpha}\Gamma(1 - \nu)s^{\nu}l(1/s), \quad s \to 0. \tag{7.5}$$

Observe that, indeed, $\mathrm{E}[C] = (e^{\Lambda\alpha} - 1)/\Lambda$ (cf. Formula (2.3.2) of [7]). Yet another application of Lemma 4.1 leads to the conclusion that, for $t \to \infty$ and $\nu \in (1, 2)$,

$$A^c(t) = t^{-\nu}l(t) \Rightarrow C^c(t) = e^{\Lambda\alpha}t^{-\nu}l(t). \tag{7.6}$$

One can easily check that the reverse statement also holds. Furthermore, at the expense of additional calculations one can also handle the case $\nu \geq 2$.

Remark 7.1. The cumulative activity period distribution, $C(\cdot)$, in this case equals the busy period distribution of an M/G/∞ queue, as has been observed in Remark 2.4 of [7]. Indeed, it is the distribution of an uninterrupted period in which at least one source is active, active sources "emerging" according to a Poisson process with rate Λ. As a by-product of our investigations we have hence found the interesting result that the busy period of an M/G/∞ queue has a regularly varying tail of index $-\nu$ iff the tail of the service time distribution is regularly varying of index $-\nu$. For the M/G/1 queue a similar statement is known [12].

The net increment distribution

Cohen [7] presents LST expressions for the distributions of B (Formula (2.2.19)) and of $B + C$ (Formula (A.5.11)) that both are suitable for proving that $B^c(t) \in R_{-\nu}$ if $A^c(t) \in R_{-\nu}$. The M/G/1 relations (3.1) and (5.4) subsequently show that $W^c(t) \in R_{1-\nu}$ (and also yield the reverse statement). We restrict ourself to mentioning Cohen's Formula (2.2.19), that follows from (7.1) and (7.2) after some manipulations (take $r = 1$ for simplicity):

$$\Psi(\omega) := \frac{\omega - \Lambda(1 - \alpha\{\omega\})}{\omega - \Lambda(1 - \mathrm{E}[e^{-\omega B}])}$$

$$= 1 - \Lambda \int_{t=0}^{\infty} \left\{\int_{x=t}^{\infty} e^{-\omega(x-t)}dA(x)\right\} e^{-\Lambda I(\omega,t)} dt, \quad \mathrm{Re}\,\omega \geq 0. \tag{7.7}$$

One can prove that, with K_1 some constant,

$$\Psi(\omega) = e^{-\Lambda\alpha} - K_1\omega^{\nu-1}l(1/\omega), \quad \omega \to 0. \tag{7.8}$$

The $e^{-\Lambda\alpha}$ factor is readily checked by taking $\omega = 0$ in (7.7), but the second factor requires a lengthy calculation which we omit. It follows from (7.7) and (7.8) that

$$E[e^{-\omega B}] = 1 - \alpha(e^{\Lambda\alpha} - \frac{1}{\Lambda\alpha}(e^{\Lambda\alpha} - 1))\omega + K_2\omega^\nu l(1/\omega), \quad \omega \to 0, \tag{7.9}$$

yielding the correct result (cf. [7]) $E[B] = \alpha e^{\Lambda\alpha} - (e^{\Lambda\alpha} - 1)/\Lambda$ and implying that $B^c(t) \in R_{-\nu}$.

Remark 7.2. Let $\{X_1, X_2, \dots\}$ be a sequence of independent random variables with common distribution F. If there exist constants $a_n > 0$ and b_n such that the distribution of $(X_1 + \cdots + X_n - b_n)/a_n$ tends to some distribution U, then F is said to belong to the *domain of attraction* of U. It is well known (cf. [3], pp. 408, 409) that $F^c(t) \in R_{1-\nu}$, $\nu > 1$, iff F lies in the domain of attraction of the extremal law $\Phi_{\nu-1}(t) := \exp(-t^{1-\nu})$, $t \geq 0$. Interestingly, this Weibull distribution also appears in [4] (or rather $1 - \exp(-t^{\nu-1})$, belonging to the inverse random variable). As mentioned in the introduction, Brichet et al. [4] study the same model as we. They show that, *in heavy traffic*, the tail of the buffer content distribution is Weibullian with parameter $\nu - 1$ when $A^c(t) \in R_{-\nu}$ (according to the present study, $W^c(t) \in R_{1-\nu}$ so $W(t)$ lies in the domain of attraction of $\Phi_{\nu-1}$).

8. Conclusion

We have studied the buffer content distribution in a fluid queueing system fed by N independent sources. Our main results are:

- for $N = 1$, the tail of the buffer content distribution is regularly varying of index $1 - \nu$ iff the tail of the activity period distribution is regularly varying of index $-\nu$.
- for $N = 2$, with one exponential source, the tail of the buffer content distribution is regularly varying of index $1 - \nu$ iff the tail of the activity period distribution of the other source is regularly varying of index $-\nu$.
- for $N = \infty$, all sources being identical, we have shown (with much less rigour and detail) that the tail of the buffer content distribution is regularly varying of index $1 - \nu$ iff the tail of the activity period distribution is regularly varying of index $-\nu$. Moreover, the tail of the cumulative activity period distribution – which coincides with the busy period distribution in an M/G/∞ queue – is regularly varying of index $-\nu$ iff the tail of the activity period distribution is regularly varying of index $-\nu$.

We believe that the above results yield valuable insight into the effect of fat-tailed activity period distributions on buffer content behaviour, in a model that is a key model for the performance analysis of high-speed communication networks.

The results have been obtained under the assumption that each source transmits at a rate which is at least equal to the buffer output rate. These results can be relaxed; a first step in this direction is taken in Remark 6.2.

Several other avenues for research open themselves; we mention a few.

(i) The case of $N - 1$ sources with exponentially tailed activity periods and one source with regularly varying tail of its activity period distribution (see Remark 6.3).

(ii) The case of two sources, *both* with regularly varying activity periods (a paper is in preparation).

(iii) The relation between our results in Section 7 and the heavy-traffic results in [4] should be further investigated.

(iv) As pointed out in [1] in another context, it is far from certain that the asymptotic results for $W^c(t)$ lead to accurate approximations. For this purpose more terms in the asymptotic expansion may be required. Or perhaps the obtained LST expressions may be numerically inverted; in that respect the explicit LST for a "Pareto mixture of exponentials", introduced in [1], may be useful for us.

Acknowledgement

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Waiting-time tail probabilities in queues with long-tail service-time distributions, *Queueing Systems* **16** (1994) 311–338.

[2] D. Anick, D. Mitra and M.M. Sondhi, Stochastic theory of a data-handling system with multiple sources, *Bell Syst. Techn. J.* **61** (1982) 1871–1894.

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

[4] F. Brichet, J. Roberts, A. Simonian and D. Veitch, Heavy traffic analysis of a storage model with long range dependent on/off sources, Report FT-CNET (1995).

[5] G.L. Choudhury and W. Whitt, Long-tail buffer-content distributions in broadband networks, Report AT&T Bell Laboratories (1995).

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

[7] J.W. Cohen, Superimposed renewal processes and storage with gradual input, *Stoch. Proc. and their Appl.* **2** (1974) 31–58.

[8] J.W. Cohen, *The Single Server Queue* (2nd ed.), North-Holland, Amsterdam (1982).

[9] J.W. Cohen, On the effective bandwidth in buffer design for the multi-server channels, CWI Report BS-R9406 (1994).

[10] L. Kosten, Stochastic theory of data handling systems with groups of multiple sources, In: *Performance of Computer-Communication Systems*, W. Bux and H. Rudin (Eds.), North-Holland, Amsterdam (1984) pp. 321–331.

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

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

[13] I. Norros, A storage model with self-similar input, *Queueing Systems* **16** (1994) 387–396.

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

[15] W. Willinger, M.S. Taqqu, W.E. Leland and D.V. Wilson, Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements, *Statistical Science* **10** (1995) 67–85.