A fluid queue with a finite buffer and subexponential input

Memorandum COSOR 98-25, 1998, Eindhoven University of Technology

A Fluid Queue with a Finite Buffer and Subexponential Input

A.P. Zwart

Eindhoven University of Technology
Department of Mathematics and Computing Science
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
email: zwart@win.tue.nl

ABSTRACT

We consider a fluid model similar to that of Kella and Whitt [33], but with a buffer having finite capacity K. The connections between the infinite buffer fluid model and the G/G/1 queue established in [33] are extended to the finite buffer case: It is shown that the stationary distribution of the buffer content is related to the stationary distribution of the finite dam. We also derive a number of new results for the latter model. In particular, an asymptotic expansion for the loss fraction is given for the case of subexponential service times. The stationary buffer content distribution of the fluid model is also related to that of the corresponding model with infinite buffer size, by showing that the two corresponding probability measures are proportional on $[0,K)$ if the silence periods are exponentially distributed. These results are applied to obtain large buffer asymptotics for the loss fraction and the mean buffer content when the fluid queue is fed by N on-off sources with subexponential on-periods. The asymptotic results show a significant influence of heavy-tailed input characteristics on the performance of the fluid queue.

1991 Mathematics Subject Classification: 60K25, 90B22.
Keywords & Phrases: on-off processes, finite buffers, fluid queues, loss fractions, proportionality of probability measures, regenerative processes, regular variation, subexponentiality, Tauberian theorems.

1. Introduction

In this paper we study a fluid model with a buffer having finite capacity K. Fluid models, and more in particular fluid queues fed by a number of on-off sources, have received much attention in recent literature, see e.g. [2, 10, 11, 30] and references therein. Most of these studies are motivated by performance issues arising in modern communication systems, like the internet and ATM-networks. An important topic in current research is the fluid queue fed by on-off sources of which the activity and/or silence periods have heavy-tailed distributions. The reason for this is that recent traffic measurements have shown that traffic in e.g. Local Area Networks [45], Wide Area Networks [39], and VBR video [6], exhibit phenomena like self-similarity and long-range dependence - phenomena that can be explained by on-off sources with heavy-tailed activity periods and/or silence periods. However, it is generally assumed that buffer sizes are infinite, so that no fluid is lost.

The main purpose of this paper is twofold. Firstly, we analyse these fluid models in the case that buffers are finite, the case occurring in many practical situations. Secondly, we want to investigate the influence of heavy-tailed input characterisics on performance measures like the loss fraction and the mean buffer content.

A secondary purpose of this study is to establish simple relations between the fluid model with finite buffer size K and the infinite buffer model, and between the finite-buffered fluid model and the G/G/1 queue with finite capacity K, also known as the finite dam (see [17, 36]). The results available for the latter models can then be used to analyse the former. Although we are particularly interested in the heavy-tailed case as described above, the relationships established in this paper are also applicable to study fluid models under more classical assumptions, e.g. when all random variables involved have finite moment generating functions.

First results on fluid queues with finite buffers and heavy tails are given in [26, 27, 40] where (the asymptotic behaviour of) the expected time to buffer overflow is studied. In [31] a discrete time queue with finite buffer is studied and the results are then applied to obtain asymptotic expansions for loss rates in fluid queues with finite buffers. Unfortunately, the results in [31] are not valid for the long-range dependent case.

The fluid model considered in this paper is –apart from the finite buffer size– the same as the model studied in [33], where some nice relations between this fluid model and a G/G/1 queue are established. It will be shown that some of these relations can be extended to the finite buffer case. The relation between the fluid model with finite buffer and the infinite buffer model is shown to hold in the special case that the time during which the buffer content decreases is exponentially distributed and that a stationary distribution of the infinite buffer content exists. We show that the two stationary distributions of the respective models are *proportional*, see Theorem 5.2 below.

Special attention is paid to the analysis of the fluid model in the case that the input process can be decribed by on-off sources with heavy-tailed activity period distributions. All the techniques used can be found in the books [8, 21]. We also refer to the survey [10] on fluid queues with heavy-tailed activity period distributions.

The asymptotic expansions that are derived for the loss fraction and the mean buffer content show that the heavy-tailed input characteristics have a significant influence on the performance of the fluid queue. In particular, loss fractions decay less than exponentially fast to zero when the buffer size gets large. This implies that very large buffers are needed to guarantee a small loss fraction, which is contrasting with the case where Cramér-type conditions are satisfied. In the latter case, the loss fraction is known to behave negative exponentially as function of the buffer size. Another performance measure which is influenced by heavy-tailed input is the mean buffer content. When the activity periods of the on-off sources have infinite second moments (in this case, the input process is long-range dependent), the mean buffer content behaves like a (positive) power of the buffer size when the latter gets large.

The paper is organised as follows. In Section 2, we introduce the fluid model and indicate its relation to the finite dam model with instantaneous input. We present some new results for the latter model in Sections 3 and 4. The main results for the fluid model can be found in Section 5. The results obtained in Sections 3–5 are applied in Section 6, where the fluid queue fed by a number of on-off sources is discussed. Section 7 treats the case of overloaded queues. Concluding remarks are given in Section 8. An alternative proof of Theorem 5.2

can be found in the appendix.

2. Preliminaries

In this section we describe the dynamics of the fluid model introduced by Kella and Whitt [33] and we extend this description to a fluid queue with a finite buffer; we adopt the notation of [33] in the sequel of the paper. There are four elements determining the dynamics of the fluid model: Two collections of random variables $\{D_k : k \geq 1\}$ and $\{U_k : k \geq 1\}$, and two collections of stochastic processes $\{\{R_k(t) : t \geq 0\} : k \geq 1\}$, and $\{\{S_k(t) : t \geq 0\} : k \geq 1\}$, both classes having right-continuous sample paths with left limits. In the terminology of [33], D_k and U_k can be interpreted as successive down- and up-times respectively, a terminology motivated by queues with service interruptions, see [33].

Fluid in the buffer increases according to $\{R_k(t) : t \geq 0\}$ during the $k-$th downtime (of the server) and fluid in the buffer decreases by the stochastic process $\{S_k(t) : t \geq 0\}$ during the $k-$th uptime. Therefore we use another terminology, which is motivated by fluid queues: We shall call D_i an activity period (of a global fluid source) and U_i a silence period.

Define

$$T_k = D_1 + U_1 + \cdots + D_k + U_k, \qquad k \geq 1, \tag{2.1}$$

and $T_0 = 0$. The net input process $Y(t)$ and the (infinite) buffer content process $Z(t)$ are then given by, cf. [33],

$$\begin{aligned} Y(t) &= Y(T_k-) + R_{k+1}(t - T_k), && T_k \leq t < T_k + D_{k+1}, \\ Y(t) &= Y(T_k-) + R_{k+1}(D_{k+1}-) - S_{k+1}(t - T_k - D_{k+1}), && T_k + D_{k+1} \leq t < T_{k+1}, \end{aligned} \tag{2.2}$$

for $k \geq 0$ with $Y(0-) = 0$, and

$$Z(t) = Y(t) - \min\{0, \inf\{Y(s) : 0 \leq s \leq t\}\}, \qquad t \geq 0. \tag{2.3}$$

In this paper we assume that the *main independence assumption* stated in [33] holds, i.e. $\{(D_k, U_k, \{R_k(t) : t \geq 0\}, \{S_k(t) : t \geq 0\}) : k \geq 1\}$ is an i.i.d. sequence. Moreover, it is assumed that the moments $\mathbb{E}[D_1]$, $\mathbb{E}[U_1]$, $\mathbb{E}[R_1(D_1-)]$, and $\mathbb{E}[S_1(U_1-)]$ are finite. Then, under the condition $\mathbb{E}[R_1(D_1-)]/\mathbb{E}[S_1(U_1-)] < 1$, it is shown in [33] that $Z(T_k-)$ converges in distribution to a random variable W if $k \to \infty$. Moreover, when D_1, U_1, and $D_1 + U_1$ are non-lattice, the buffer content process $Z(t)$ converges in distribution to a random variable Z.

It is obvious that the distribution of W corresponds to the waiting time distribution of the G/G/1 queue with service times $R_1(D_1-)$ and interarrival times $S_1(U_1-)$. One of the main contributions of Kella and Whitt is that they relate the distribution of Z to the stationary waiting time distribution of a G/G/1 queue, see Theorems 4–6 in [33].

Next, we introduce the fluid model with finite buffer size $K > 0$. For each K, the buffer content $Z^K(t)$ at time t can be described by $Z^K(0) = Z^K(T_0) = 0$, and

$$Z^K(T_{k+1}-) = \max\{\min\{Z^K(T_k-) + R_{k+1}(D_{k+1}-), K\} - S_{k+1}(U_{k+1}-), 0\}, \tag{2.4}$$

$$Z^K(t) = \min\{Z^K(T_k-) + R_{k+1}(t - T_k-), K\}, \qquad T_k \leq t < T_k + D_{k+1},$$

$$Z^K(t) = \max\{0, \min\{Z^K(T_k-) + R_{k+1}(D_{k+1}-), K\} - S_{k+1}(t - T_k - D_{k+1})\},$$

$$T_k + D_{k+1} \leq t < T_{k+1}. \qquad (2.5)$$

The dynamics of the finite buffer model are the same as that of the infinite buffer model, except that when the buffer content reaches level K, all the fluid offered to the buffer during the remaining activity period will be lost.

It is easily shown that $Z^K(T_{k+1}-)$ can be identified with the waiting time of the $(k+1)-$st customer in the G/G/1 queue with finite capacity K, in which the interarrival times and service times are distributed as $S_1(U_1-)$ and $R_1(D_1-)$. Under the condition $\mathbb{P}(S_1(U_1-) = R_1(D_1-)) < 1$, it is shown in Section III.5.3 of [17] that $Z^K(T_{k+1}-)$ converges in distribution to a limiting random variable W^K if $k \to \infty$.

We wish to extend the results of Kella and Whitt to finite buffer queues; we will establish a relationship between the stationary distribution of the finite buffer model and the stationary distribution of the G/G/1 queue with a buffer having finite capacity K. The latter model is also known as the finite dam, cf. Chapter III.5 in [17].

First, we make some additional definitions. With $f(x) \sim g(x)$ we mean that $f(x)/g(x)$ converges to one. If the first moment ϕ of a non-negative random variable X exists, then the integrated tail distribution $\widetilde{F}$ of the (excess) random variable $\widetilde{X}$ is defined by

$$\widetilde{F}(x) = \frac{1}{\phi}\int_0^x (1 - F(u))\mathrm{d}u,$$

with F being the distribution function of a X. The $n-$fold convolution F^{n*} of F is defined by, for $x \geq 0$,

$$\begin{aligned} F^{0*}(x) &= 1, \\ F^{n*}(x) &= \int_0^x F^{(n-1)*}(x-u)\mathrm{d}F(u), \qquad n = 1, 2, \ldots . \end{aligned}$$

Define the environment indicator process by

$$I(t) = I_{\{T_k + D_{k+1} \leq t < T_{k+1} \text{ for some } k \geq 1\}},$$

so $I(t) = 1$ if the global fluid source is silent at time t. The amount of time the global fluid source is silent (resp. active) up to time t, $t \geq 0$, is defined by

$$C_s(t) = \int_0^t I(x)\mathrm{d}x, \qquad (2.6)$$

$$C_a(t) = t - C_s(t). \tag{2.7}$$

The inverse processes of C_s and C_a are defined by

$$C_s^{-1}(t) = \inf_{x \geq 0}\{C_s(x) > t\}, \tag{2.8}$$

$$C_a^{-1}(t) = \inf_{x \geq 0}\{C_a(x) > t\}. \tag{2.9}$$

Since the random variables D_i and U_i are finite a.s., we may assume that $C_s(t) \to \infty$ if $t \to \infty$ everywhere. So the following processes are well-defined,

$$Z_s^K(t) = Z^K(C_s^{-1}(t)), \qquad t \geq 0, \tag{2.10}$$

$$Z_a^K(t) = Z^K(C_a^{-1}(t)), \qquad t \geq 0. \tag{2.11}$$

Note that $Z_s^K(U_1) = Z^K(D_1 + U_1 + D_2)$. We similarly define $Z_a(t)$ and $Z_s(t)$ for the infinite buffer model. Cf. [33], we define the r.v. $R_1(\widetilde{D}_1)$ (which is non-trivial since R_1 and D_1 are dependent in general) by

$$\mathbb{P}(R_1(\widetilde{D}_1) > x) = \frac{1}{\mathbb{E}[D_1]}\mathbb{E}\left[\int_0^{D_1} 1_{\{R_1(t)>x\}}\mathrm{d}t\right] = \int_0^{\infty} \mathbb{P}(R_1(t) > x | D_1 > t)\mathrm{d}\mathbb{P}(\widetilde{D}_1 \leq t). \tag{2.12}$$

We are now ready to give the main result of this section, which can be viewed as an extension of Theorem 4 in [33]. Note that no assumptions on the traffic load are needed, since the state space is finite. Denote convergence in distribution by '$\Rightarrow$', and denote equality in distribution by '$\overset{d}{=}$'.

Theorem 2.1 *Suppose that the main independence assumption holds, that D_1, U_1, and $D_1 + U_1$ are non-lattice, and that $\mathbb{P}(S_1(U_1-) = R_1(D_1-)) < 1$. Then there exist r.v.'s Z_s^K, Z_a^K, Z^K, and I such that, when $t \to \infty$,*

1. $Z_a^K(t) \Rightarrow Z_a^K \overset{d}{=} \min\{W^K + R_1(\widetilde{D}_1), K\}$,
2. $Z_s^K(t) \Rightarrow Z_s^K$,
3. $[Z^K(t), I(t)] \Rightarrow [Z^K, I]$.

Here $R_1(\widetilde{D}_1)$ is independent of W^K, $Z_s^K \overset{d}{=} (Z^K | I = 0)$, $Z_a^K \overset{d}{=} (Z^K | I = 1)$, and

$$\mathbb{P}(Z^K > x) = p\mathbb{P}(Z_s^K > x) + (1-p)\mathbb{P}(Z_a^K > x), \tag{2.13}$$

where

$$p = \mathbb{P}(I = 1) = \frac{\mathbb{E}[U_1]}{\mathbb{E}[D_1] + \mathbb{E}[U_1]}. \tag{2.14}$$

Proof The proof is almost identical to the proof of Theorem 4 in [33]. The processes $[Z^K(t), I(t)]$, $Z_a^K(t)$ and $Z_s^K(t)$ are all regenerative with the exit times of state [0,1], resp. 0 (i.e. the end of idle periods) as regeneration points. The regeneration cycles are non-lattice when U_1, D_1, and $U_1 + D_1$ are non-lattice, due to the main independence assumption. Since all state spaces are finite, it is trivially seen that all regeneration cycles have finite means. The convergence of the processes $[Z^K(t), I(t)]$, $Z_a^K(t)$ and $Z_s^K(t)$ now follows by the results on pp. 125-127 of [3].

By a result of Green [25], we can study the sequences of activity periods and silence periods separately. This gives the relationship between the limiting distributions $[Z^K, I]$, Z_a^K, and Z_s^K, and the characterisation of the distribution of Z_a^K. ■

Remark 2.1 The non-lattice conditions can be omitted if U_1 is exponentially distributed and independent of D_1 and $\{R_1(t)\}$. In Theorem 2.1, the condition on $U_1 + D_1$ is imposed since U_1 and D_1 are allowed to be dependent. □

If the outflow in the buffer is constant during silence periods, then it is also possible to specify the limiting distribution Z_s^K.

Theorem 2.2 *Suppose the assumptions stated in Theorem 2.1 hold and that*

$$S_1(t) \equiv t.$$

Then $\{Z_s^K(t), t \geq 0\}$ is distributed as the workload process in the finite dam with capacity K, interarrival times U_1, and service times $R_1(D_1-)$.

Proof Similarly to the proof of Theorem 2 in [33]. Both processes have reflecting barriers in the origin and K, decrease linearly at rate 1, and have jumps of size $R_{k+1}(D_{k+1}-)$ at times $U_1 + \cdots + U_k$. ■

One can apply Theorems 2.1 and 2.2 to compute (characteristics of) the distribution of Z^K when the steady state distribution for the $G/G/1$ finite dam is tractable enough, which is the case for the M/G/1 and G/M/1 finite dams, see [17, 36]. In Section 5, we will further specify the distribution of Z^K by using Theorems 2.1 and 2.2. Both theorems indicate a clear relationship between the fluid model with gradual input and the G/G/1 finite dam with instantaneous input; we will study the latter model in the next two sections.

In the remainder of the paper, we assume that the buffer content declines linearly during silence periods, i.e. we assume that $S_1(t) \equiv t$. In this case, the fluid model can process one unit of fluid per unit of time. The amount of fluid offered to the system per unit of time, given by ρ, equals

$$\rho = \frac{\mathbb{E}[R_1(D_1-)] + \mathbb{E}[D_1]}{\lambda^{-1} + \mathbb{E}[D_1]}.$$

3. The stationary distribution of the finite dam

The random variable W^K in the previous section corresponds to the stationary waiting time distribution in the finite dam having capacity K. The relation between the models with gradual and instantaneous input will turn out to be useful in the rest of the paper. In this section, we give some new results for the finite dam. In particular, we give a relationship between the virtual and actual waiting time which is very similar to the relationship in the infinite buffer case. The latter is well known (see e.g. p. 189 in [3]).

First, we introduce some notation in the traditional queueing setting. Customers arrive at a single server queue (which is initially empty) with interarrival times $(A_n)_{n\geq 1}$. These customers have service times $(B_n)_{n\geq 1}$. It is assumed that the interarrival times and service times are all independent of each other and have the same distributions as random variables A and B. The means of A and B are denoted by λ^{-1} and β, respectively. The distribution function of the service time is denoted by $B(.)$. The traffic load $\hat{\rho}$ is given by $\hat{\rho} = \lambda\beta$ and is assumed to be strictly positive.

The waiting time of the $n-$th customer is given by W_n^K. When $W_n^K + B_n$ exceeds K, a quantity of $W_n^K + B_n - K$ is lost (so we consider partial overflow). Hence, W_n^K is given by $W_0^K = 0$ and (see e.g. Chapter III.5 in [17]),

$$W_{n+1}^K = \max\{\min\{W_n^K + B_n, K\} - A_{n+1}, 0\}. \tag{3.1}$$

Denote the stationary waiting time by W^K (cf. Section 2, with $B_n \equiv R_n(D_n-)$ and $A_n \equiv U_n$). We also consider the amount of work present in the system at time t, given by $V^K(t)$; its stationary distribution is denoted by V^K. Finally, the long-run fraction of work lost is defined by $L_{q,K}$.

3.1 *General results*

The loss fraction $L_{q,K}$ can be obtained by a simple renewal argument:

$$L_{q,K} = \frac{\mathbb{E}[\max\{W^K + B - K, 0\}]}{\mathbb{E}[B]} = \mathbb{P}(W^K + \widetilde{B} > K). \tag{3.2}$$

The second equality, which is quite useful for further analysis as will be shown below, can be obtained by partial integration. For the virtual waiting time V and the actual waiting time W in the GI/G/1 queue (with $\hat{\rho} < 1$), it is well known that (see e.g. [3, 15, 17])

$$V | V > 0 \overset{d}{=} W + \widetilde{B}. \tag{3.3}$$

The following result is very similar to (3.3) and appears to be new.

Theorem 3.1 *For all $\hat{\rho} > 0$ and $0 < K < \infty$,*

$$V^K | V^K > 0 \overset{d}{=} (W^K + \widetilde{B}) \mid W^K + \widetilde{B} \leq K, \tag{3.4}$$

$$\mathbb{P}(V^K > x) = \hat{\rho}\, \mathbb{P}(x < W^K + \widetilde{B} \leq K). \tag{3.5}$$

Proof The results can be obtained in a similar way as in the infinite buffer queue, namely by a level crossing argument, cf. [15]. Following the same lines as in [15] we obtain for almost every $0 < v < K$,

$$\frac{\mathrm{d}}{\mathrm{d}v}\mathbb{P}(V^K < v) = \hat{\rho}\int_0^v \frac{1 - \mathbb{P}(B < v - u)}{\beta}\mathrm{d}\mathbb{P}(W^K < u).$$

Hence, for $0 < x < K$,

$$\begin{aligned}\mathbb{P}(V^K < x) &= \mathbb{P}(V^K = 0) + \hat{\rho}\int_0^x \int_0^{x-u} \frac{\mathbb{P}(B > w)}{\beta} \mathrm{d}w \mathrm{d}\mathbb{P}(W^K < u) \\ &= \mathbb{P}(V^K = 0) + \hat{\rho}\mathbb{P}(W^K + \widetilde{B} < x).\end{aligned}$$

By Little's law for a busy server (see e.g. Example 4.3 in [46]) and (3.2) we have

$$\mathbb{P}(V^K = 0) = 1 - \hat{\rho}(1 - L_{q,K}) = 1 - \hat{\rho}\mathbb{P}(W^K + \widetilde{B} \leq K).$$

Hence, for $0 < x < K$,

$$\begin{aligned}\mathbb{P}(V^K < x) &= 1 - \hat{\rho}(\mathbb{P}(W^K + \widetilde{B} \leq K) - \mathbb{P}(W^K + \widetilde{B} < x)) \\ &= 1 - \hat{\rho}\mathbb{P}(x \leq W^K + \widetilde{B} \leq K).\end{aligned}$$

This expression is also valid for $x = 0$ and $x = K$, which yields (3.5) since $\widetilde{B}$ has a continuous distribution. It is easily shown from (3.5) that

$$\mathbb{P}(0 < V^K \leq x) = \hat{\rho}\mathbb{P}(W^K + \widetilde{B} \leq x).$$

Hence,

$$\begin{aligned}\mathbb{P}(V^K \leq x | V^K > 0) &= \frac{\mathbb{P}(0 < V^K \leq x)}{\mathbb{P}(V^K > 0)} = \frac{\hat{\rho}\mathbb{P}(W^K + \widetilde{B} \leq x)}{\hat{\rho}\mathbb{P}(W^K + \widetilde{B} \leq K)} \\ &= \mathbb{P}(W^K + \widetilde{B} \leq x | W^K + \widetilde{B} \leq K).\end{aligned}$$

This proves (3.4). ■

3.2 Exponentially distributed interarrival times

If the interarrival times are exponentially distributed, then the following *proportionality relation* holds, cf. [14, 17, 29]:

$$\mathbb{P}(W^K \leq x) = \frac{\mathbb{P}(W \leq x)}{\mathbb{P}(W \leq K)}, \tag{3.6}$$

for $0 \leq x \leq K$. Proportionality relations like (3.6) have been applied in a number of studies to determine loss probabilities, see e.g. [20, 41, 23, 24, 9] and references therein. The main idea applied in these studies is to combine the proportionality result with Little's formula for a busy server (see e.g. Example 4.3 in [46]). Applying the latter together with PASTA to the finite and infinite buffer queue we obtain for $0 < \hat{\rho} < 1$ and $\hat{\rho} > 0$ respectively,

$$\mathbb{P}(W = 0) = 1 - \hat{\rho}, \tag{3.7}$$

$$\mathbb{P}(W^K = 0) = 1 - \hat{\rho}(1 - L_{q,K}). \tag{3.8}$$

Using the proportionality relation,

$$\frac{\mathbb{P}(W^K = 0)}{\mathbb{P}(W = 0)} = \frac{\mathbb{P}(W^K \leq K)}{\mathbb{P}(W \leq K)} = \frac{1}{\mathbb{P}(W \leq K)},$$

we obtain from (3.7) and (3.8),

$$L_{q,K} = \frac{1-\hat{\rho}}{\hat{\rho}} \left(\frac{1}{\mathbb{P}(W \leq K)} - 1 \right) = \frac{1-\hat{\rho}}{\hat{\rho}} \frac{\mathbb{P}(W > K)}{\mathbb{P}(W \leq K)}. \tag{3.9}$$

Remark 3.1 By PASTA, we have that $V^K \overset{d}{=} W^K$. Using this and the proportionality relation, it is also possible to derive (3.9) from (3.2) and (3.5). □

Remark 3.2 Another performance measure is the probability that the work offered by a customer (entering the system in its stationary regime) cannot be accepted completely; denote this probability by $P_{q,K}$. For the GI/G/1 finite dam, we have

$$P_{q,K} = \mathbb{P}(W^K + B > K). \tag{3.10}$$

(Note that $P_{q,K} = L_{q,K}$ in the GI/M/1 finite dam.) When $\hat{\rho} < 1$, we have the following remarkable relation for the M/G/1 finite dam. It follows from the proportionality relation that

$$P_{q,K} = \frac{\mathbb{P}(W + B > K) - \mathbb{P}(W > K)}{\mathbb{P}(W \leq K)}. \tag{3.11}$$

But this quantity can be identified with $\mathbb{P}(V_{max} > K)$, where V_{max} is the maximal content in the infinite dam during a busy cycle, see e.g. Section 3.3 in [14] or [17], p. 297, and p. 618, so we conclude that

$$P_{q,K} = \mathbb{P}(V_{max} > K) = \frac{1}{\lambda} \frac{\frac{\mathrm{d}}{\mathrm{d}K}\mathbb{P}(W \leq K)}{\mathbb{P}(W \leq K)}. \tag{3.12}$$

□

4. Asymptotic results for the finite dam

For the case $\hat{\rho} < 1$, we are interested in the asymptotic behaviour of $L_{q,K}$ when $K \to \infty$, in particular when the service time distribution is subexponential. In the case of exponentially distributed silence periods, it is possible to apply a classical result of Pakes [37] (see also [44]) for the GI/G/1 queue with infinite buffer size.

Theorem 4.1 *If $\hat{\rho} < 1$, $\widetilde{B}$ is subexponential, and the interarrival times are exponentially distributed, then*

$$L_{q,K} \sim \mathbb{P}(\widetilde{B} > K), \qquad K \to \infty. \tag{4.1}$$

Proof From [37] we have, if $\widetilde{B}$ is subexponential,

$$\mathbb{P}(W > x) \sim \frac{\hat{\rho}}{1-\hat{\rho}} \mathbb{P}(\widetilde{B} > x), \qquad x \to \infty. \tag{4.2}$$

Theorem 4.1 now follows directly from (3.9) and (4.2). ■

When the interarrival times have a general distribution, the proportionality relation does not hold, so it is not possible to apply results for the infinite dam directly. However, it is still possible to extend Theorem 4.1 to the case of generally distributed interarrival times. This is established in the following theorem, under the additional assumption that the service time distribution is regularly varying (see [8]).

Theorem 4.2 *For generally distributed interarrival times and $\hat{\rho} < 1$, (4.1) holds if the service time distribution is regularly varying of index $-\nu$, $\nu > 1$.*

Proof Note that $L_{q,K} \geq \mathbb{P}(\widetilde{B} > K)$, so it suffices to show that

$$\limsup_{K\to\infty} \frac{\mathbb{P}(W^K + \widetilde{B} > K)}{\mathbb{P}(\widetilde{B} > K)} \leq 1. \tag{4.3}$$

Let $\phi(K)$ be a function such that $\phi(K) \to \infty$ and $\phi(K)/K \to 0$ if $K \to \infty$, and let $\varepsilon > 0$. Write

$$\mathbb{P}(W^K + \widetilde{B} > K) = P_{1,K} + P_{2,K} + P_{3,K}, \tag{4.4}$$

with

$$\begin{aligned} P_{1,K} &= \mathbb{P}(W^K + \widetilde{B} > K; W^K \leq \varepsilon K), & (4.5)\\ P_{2,K} &= \mathbb{P}(W^K + \widetilde{B} > K; \varepsilon K < W^K \leq K - \phi(K)), & (4.6)\\ P_{3,K} &= \mathbb{P}(W^K + \widetilde{B} > K; W^K > K - \phi(K)). & (4.7) \end{aligned}$$

Since $P_{1,K} \leq \mathbb{P}(\widetilde{B} > (1-\varepsilon)K)$ and since $\widetilde{B}$ is regularly varying of index $1-\nu$, we have

$$\limsup_{K\to\infty} \frac{P_{1,K}}{\mathbb{P}(\widetilde{B} > K)} \leq \left(\frac{1}{1-\varepsilon}\right)^{\nu-1}, \qquad \forall \varepsilon > 0. \tag{4.8}$$

We can bound $P_{2,K}$ using that W^K is stochastically dominated by W:

$$P_{2,K} \leq \mathbb{P}(\widetilde{B} \geq \phi(K))\mathbb{P}(W^K > \varepsilon K) \leq \mathbb{P}(\widetilde{B} \geq \phi(K))\mathbb{P}(W > \varepsilon K).$$

Using (4.2) for the GI/G/1 queue and the fact that $\widetilde{B}$ is regularly varying we obtain for each $\varepsilon > 0$,

$$\lim_{K\to\infty} \frac{\mathbb{P}(W > \varepsilon K)}{\mathbb{P}(\widetilde{B} > K)} = \frac{\hat{\rho}}{1-\hat{\rho}} \varepsilon^{1-\nu},$$

which implies, since $\phi(K) \to \infty$ if $K \to \infty$,

$$\limsup_{K\to\infty} \frac{P_{2,K}}{\mathbb{P}(\widetilde{B} > K)} = 0, \qquad \forall \varepsilon > 0. \tag{4.9}$$

Finally, we deal with the last term. Note that

$$P_{3,K} \leq \mathbb{P}(W^K \geq K - \phi(K)). \tag{4.10}$$

We make some additional definitions. Define the random walk $(S_n)_{n\geq 0}$ by $S_0 = 0$, and for $n \geq 1$,

$$S_n = \sum_{i=1}^{n} (B_i - A_i). \tag{4.11}$$

Note that this random walk has negative drift $\beta - \lambda^{-1}$. We also define the sequence of random variables $(\bar{W}_n^K)_{n\geq 0}$ by $\bar{W}_0^K = 0$, and $\bar{W}_{n+1}^K = \min\{\max\{\bar{W}_n^K + B_n - A_{n+1}, 0\}, K\}$. Denote the stationary solution of this recursion by $\bar{W}^K$. From the construction of both W^K and $\bar{W}^K$ it is clear that $\mathbb{P}(W^K \geq x) \leq \mathbb{P}(\bar{W}^K \geq x)$, $0 \leq x \leq K$. Hence,

$$P_{3,K} \leq \mathbb{P}(\bar{W}^K > K - \phi(K)). \tag{4.12}$$

We now use a representation of the distribution of $\bar{W}^K$ in terms of an absorption probability of the random walk (S_n), which seems to be due to Lindley [34], see also [35]. Define the stopping times

$$\tau(K) = \inf\{n : S_n \geq K - \phi(K)\}, \qquad \tau'(K) = \inf\{n : S_n \leq -\phi(K)\}.$$

Then, cf. [34, 35],

$$\mathbb{P}(\bar{W}^K > K - \phi(K)) = \mathbb{P}(\tau(K) < \tau'(K)). \tag{4.13}$$

Rewriting this yields

$$\mathbb{P}(\bar{W}^K > K - \phi(K)) = \mathbb{P}(S_1, ..., S_{\tau(K)-1} > -\phi(K) | \tau(K) < \infty)\mathbb{P}(\tau(K) < \infty).$$

Since $\sup_n S_n$ can be identified with W, and $\tau(K) < \infty$ iff $\sup_n S_n > K - \phi(K)$, this equals

$$\mathbb{P}(S_1, ..., S_{\tau(K)-1} > -\phi(K) | \tau(K) < \infty)\mathbb{P}(W > K - \phi(K)).$$

Using $\phi(K)/K \to 0$, we have by (4.2) that

$$\mathbb{P}(W > K - \phi(K))/\mathbb{P}(\widetilde{B} > K) \to \frac{\hat{\rho}}{1-\hat{\rho}}, \qquad K \to \infty.$$

So we can conclude that $P_{3,K} = \mathrm{o}(\mathbb{P}(\widetilde{B} > K))$ if

$$\mathbb{P}(S_{\tau(K)-1} > -\phi(K) | \tau(K) < \infty) \to 0, \qquad K \to \infty. \tag{4.14}$$

For this we use a theorem of Asmussen and Klüppelberg, see Theorem 1.1 in [4]. This result provides the following conditional limit theorem for $S_{\tau(K)-1}$ (which is the last value of S_n before making a jump to level $K-\phi(K)$). Define $a(u)=\int_u^\infty(1-B(z))\mathrm{d}z/(1-B(u))$. Then,

$$\lim_{K\to\infty}\mathbb{P}(-S_{\tau(K)-1}/a(K)>x|\tau(K)<\infty)=(1+x/(\nu-1))^{1-\nu},\qquad x\geq 0. \tag{4.15}$$

Note that $a(K)\sim K/(\nu-1)$ if $K\to\infty$, so $\phi(K)/a(K)\to 0$ if $K\to\infty$. Hence,

$$\mathbb{P}(S_{\tau(K)-1}>-\phi(K)|\tau(K)<\infty)=\mathbb{P}(-S_{\tau(K)-1}/a(K)<\phi(K)/a(K)|\tau(K)<\infty)\to 0.$$

This proves (4.14). Hence, we have for each $\varepsilon>0$ that

$$\limsup_{K\to\infty}L_{q,K}/\mathbb{P}(\widetilde{B}>K)\leq\left(\frac{1}{1-\varepsilon}\right)^{\nu-1}, \tag{4.16}$$

which implies (4.3) by letting $\varepsilon\to 0$. ■

Remark 4.1 Jelenković [31] analyses the loss fraction in a discrete time queue. The evolution of this queueing model (which is equivalent to the G/G/1 queue with uniformly bounded actual waiting time, see e.g. Chapter III.4 in [17]) can be described by the random variables $\bar{W}_n^K$ in the proof of Theorem 4.2. In this model, the loss fraction equals $\mathbb{P}(\bar{W}^K+\widetilde{B}-A>K)$. Using the proof of Theorem 4.2, it is not difficult to show that this expression is asymptotically equal to $\mathbb{P}(\widetilde{B}>K)$, if the service time distribution is regularly varying and $K\to\infty$. This generalizes the result in [31], where it is assumed that the service time distribution is regularly varying of non-integer index $-\nu$, $\nu>2$. □

5. The stationary distribution of the fluid queue

In this section, we study the distribution of the steady state buffer content Z^K in the fluid queue. Under certain assumptions, we express the distribution of Z^K completely in terms of W^K, thereby extending the results in [33] to the finite buffer case. In a special case, it is also possible to express the distribution of Z^K in terms of Z, by showing that the two probability measures are proportional.

Theorem 5.1 *For $\rho>0$ and $0\leq x<K$,*

$$\mathbb{P}(Z^K>x)=(1-p)\mathbb{P}(W^K+R_1(\widetilde{D}_1)>x)+p\hat{\rho}\mathbb{P}(K\geq W^K+\widetilde{R}_1(D_1)>x). \tag{5.1}$$

In particular, if the silence periods are exponentially distributed, then

$$\mathbb{P}(Z^K>x)=(1-p)\mathbb{P}(W^K+R_1(\widetilde{D}_1)>x)+p\mathbb{P}(W^K>x), \tag{5.2}$$

with p given by (2.14).

Proof In view of Theorem 2.1, we only need to specify the distribution of Z_s^K. By Theorem 2.2, we have $Z_s^K\overset{d}{=}V^K$. The first part of the theorem now follows from Theorem 3.1 and the second part can be obtained using PASTA. ■

In the case that U_1 has an exponential distribution and $\rho < 1$, one can establish the following relation between the distributions of Z^K, Z, and W. Recall that W can be identified with the waiting time distribution of the $GI/G/1$ queue with service time $R_1(D_1-)$ and interarrival time U_1.

Theorem 5.2 *If U_1 is exponentially distributed and if $\rho < 1$, then, for $0 \leq x < K$,*

$$\mathbb{P}(Z^K \leq x) = \frac{\mathbb{P}(Z \leq x)}{\mathbb{P}(W \leq K)}. \tag{5.3}$$

Proof Use the second part of the previous theorem, the proportionality relation (3.6), and

$$\mathbb{P}(Z \leq x) = (1-p)\mathbb{P}(W + R_1(\widetilde{D}_1) \leq x) + p\mathbb{P}(W \leq x), \tag{5.4}$$

cf. [33]. ■

Note that (5.3) is not valid for $x = K$ and note the appearance of the term $\mathbb{P}(W \leq K)$ (and not $\mathbb{P}(Z \leq K)$) in (5.3). An explanation for this is that the probability that the buffer is full ($\mathbb{P}(Z^K = K)$) is strictly positive. This is not the case when input is instantaneous, cf. (3.6).

In the proof of Theorem 5.2, we used the relation between the models with gradual and instantaneous input (Theorem 5.1 and (5.4)), and the proportionality relation (3.6) between the two models with instantaneous input. It is also possible to prove Theorem 5.2 directly (without using connections with models with instantaneous input) by a regenerative argument, for which we refer to the appendix.

In the remainder of this section, we will study two important performance measures: the long-run fraction of fluid lost, denoted by L_K, and the mean buffer content.

Theorem 5.3 *For all $\rho > 0$,*

$$L_K = \frac{\mathbb{E}[R_1(D_1-)]}{\mathbb{E}[D_1] + \mathbb{E}[R_1(D_1-)]} L_{q,K}, \tag{5.5}$$

where $L_{q,K} = \mathbb{P}(W^K + \widetilde{R}_1(D_1-) > K)$.

Proof We can establish a relation between the fluid model and the finite buffer queueing model in the following manner. Suppose that both models are fed by the same input process. The amount of fluid lost during the $k-$th activity (and silence) period in the fluid model is identical to the work lost of the $k-$th customer in the finite buffer queue. However, the amount of fluid offered during the $k-$th activity period is $R_k(D_k-) + D_k$, whereas the amount of work offered by the $k-$th customer equals $R_k(D_k-)$. The result now follows by the renewal reward theorem, see e.g. [43]. ■

Finally, we investigate the mean buffer content $\mathbb{E}[Z^K]$. We restrict ourself to the case $\rho < 1$ and exponentially distributed silence periods.

Theorem 5.4 *Under the conditions of Theorem 5.2,*

$$\mathbb{E}[Z^K] = \frac{1}{\mathbb{P}(W \leq K)} \int\limits_0^K \mathbb{P}(Z > x)dx - \frac{K\mathbb{P}(W > K)}{\mathbb{P}(W \leq K)}.$$

Proof Use the representation $\mathbb{E}[Z^K] = \int_0^{K-} \mathbb{P}(Z^K > x)\mathrm{d}x$, and the identity

$$\mathbb{P}(Z^K > x) = \frac{\mathbb{P}(Z > x) - \mathbb{P}(W > K)}{\mathbb{P}(W \leq K)},$$

which follows easily from Theorem 5.2. ■

Remark 5.1 Using the proportionality relations (3.6) and (5.3), it is possible to formulate heavy traffic limit theorems for W^K and Z^K, based on heavy traffic limits for the M/G/1 queue.

Suppose that silence periods are exponentially distributed, that $\rho < 1$ (hence $\hat{\rho} < 1$), and that a function $\Delta(\hat{\rho})$ exists such that $\Delta(\hat{\rho})W$ converges in distribution to a random variable W_{HT} if $\hat{\rho} \to 1$. This is the case if the second moment of the service time is finite (cf. the classical result of Kingman for the GI/G/1 queue, see e.g. [17]), but also if the service time distribution is regularly varying and has infinite variance, see e.g. [12, 19].

Under these assumptions, we can formulate a heavy traffic limit for W^K by letting $\hat{\rho} \to 1$ and $K \to \infty$ such that $K\Delta(\hat{\rho}) = c$ for some constant c. Using (3.6), it is not difficult to see that, if $\hat{\rho} \to 1$ and $\Delta(\hat{\rho})K \equiv c$,

$$\mathbb{P}(\Delta(\hat{\rho})W^K \leq x) \to \frac{\mathbb{P}(W_{HT} \leq x)}{\mathbb{P}(W_{HT} \leq c)}, \tag{5.6}$$

for $0 \leq x \leq c$. By (5.4), $\Delta(\hat{\rho})Z$ converges to the same heavy traffic limit as $\Delta(\hat{\rho})W$. Hence, $\Delta(\hat{\rho})Z^K$ has the same heavy traffic limit as $\Delta(\hat{\rho})W^K$ using (5.2) or (5.3).

These heavy traffic limits may provide good practical approximations for the distributions of W^K and Z^K, since ρ may be near to one for economic reasons and K may be large to ensure a small loss fraction. Our conjecture is that the heavy traffic limit theorem can be extended to generally distributed silence periods.

For a similar result for the G/G/1 queue with uniformly bounded actual waiting time (Chapter III.4 in [17]), we refer to [32] and references therein. □

6. Asymptotic results for the fluid queue

In this section, we use asymptotic results for $\mathbb{P}(Z > x)$ and $\mathbb{P}(W > x)$ (for large x) and the results derived in the previous sections to obtain asymptotic expansions for various performance measures of the finite-buffered fluid queue, like the loss probability and the mean buffer content, when the buffer size K gets large. We concentrate on the case where D_1 and $R_1(D_1-)$ have a subexponential tail. Furthermore, most of the asymptotic results will be derived in the important special case that the fluid queue is fed by N $(1 \leq N \leq \infty)$ on-off sources.

The general case will be treated in Subsection 6.1. In Subsection 6.2 we study the simplest possible fluid model, namely the case of a single on-off source. The last two subsections treat the case of multiple sources, where we distinguish the cases $N < \infty$ and $N = \infty$. Throughout this section, $L(.)$ is a slowly varying function.

6.1 General input

We start with the case of general input, where we assume that $R_1(D_1-)$ has a subexponential distribution. The following result follows immediately from Theorem 5.3 and the results in Section 4.

Theorem 6.1 *Under the conditions of Theorem 4.1 or 4.2,*

$$L_K \sim \frac{\mathbb{E}[R_1(D_1-)]}{\mathbb{E}[D_1] + \mathbb{E}[R_1(D_1-)]} \mathbb{P}(\widetilde{R}_1(D_1-) > K), \qquad K \to \infty. \tag{6.1}$$

Asymptotics for the mean buffer content are more difficult to obtain in general. Such a result would involve the tail behaviour of $R_1(\widetilde{D}_1)$ (cf. Theorem 3.15 in [10] and (2.12)), for which no results are available.

6.2 A single on-off source

Suppose that the fluid queue is driven by a single on-off source. When the source is active, it sends input with rate $r > 1$ during a period of D_1. Off-periods are exponentially distributed with parameter λ. In terms of the model in the previous sections, this implies that $R_1(t) \equiv (r-1)t$. In the terminology of [33], this is the linear fluid model with random disruptions, with the additional assumption that the idle periods are exponentially distributed.

In [33] the following relation is established between the distributions of Z and W:

$$\mathbb{P}(Z > x) = \rho \mathbb{P}(W + (r-1)\widetilde{D}_1 > x).$$

From Theorem 5.1 we get for the finite buffer case, if $0 \le x < K$,

$$\mathbb{P}(Z^K > x) = \rho \mathbb{P}(W^K + (r-1)\widetilde{D}_1 > x) - p\hat{\rho}\mathbb{P}(W^K + (r-1)\widetilde{D}_1 > K), \tag{6.2}$$

where we used the identity $\rho = \hat{\rho}p + 1 - p$. In this case, $\hat{\rho} = \lambda(r-1)\mathbb{E}[D_1]$, and $p = \frac{\lambda^{-1}}{\lambda^{-1} + \mathbb{E}[D_1]}$.

The asymptotic expansions for the loss probability and the mean buffer content given below are based on the following results. Suppose the distribution of $\widetilde{D}_1$ is subexponential. Then (see e.g. [10]):

$$\mathbb{P}(W > x) \sim \frac{\hat{\rho}}{1-\hat{\rho}} \mathbb{P}((r-1)\widetilde{D}_1 > x), \qquad x \to \infty, \tag{6.3}$$

$$\mathbb{P}(Z > x) \sim p\frac{\rho}{1-\rho} \mathbb{P}((r-1)\widetilde{D}_1 > x), \qquad x \to \infty. \tag{6.4}$$

Proposition 6.1 *If the distribution of $\widetilde{D}_1$ is subexponential and if the off-periods are exponentially distributed, then, for $\rho < 1$ and $K \to \infty$,*

$$L_K \sim \frac{r-1}{r}\mathbb{P}(\widetilde{D}_1 > \frac{K}{r-1}). \tag{6.5}$$

When the off-periods are generally distributed and $\mathbb{P}(D_1 > x) = L(x)x^{-\nu}$, $\nu > 1$:

$$L_K \sim \frac{(r-1)^\nu}{r(\nu-1)\mathbb{E}[D_1]}L(K)K^{1-\nu}, \qquad K \to \infty. \tag{6.6}$$

Proof Equation (6.5) follows immediately from Theorem 6.1 (or alternatively, use Theorem 5.3, (3.9), and (6.3)). Equation (6.6) follows from Theorem 6.1 and Karamata's theorem. For the latter, see Section 1.6 in [8]. ■

Remark 6.1 Awater ([5], p. 131) has suggested the following approximation for the fraction of fluid lost,

$$L_{K,app} = \frac{(1-\rho)\mathbb{P}(Z > K)}{1-\rho\mathbb{P}(Z > K)}.$$

It is shown numerically in [5] that $L_{K,app}$ can be a good approximation to L_K. Variants of $L_{K,app}$ have been shown to be exact in various other cases like the loss probability of a customer in the $M^X/G/1/B$ queue (see [23]) and the M/M/c queue with impatient customers (see [9]).

If we evaluate the performance of $L_{K,app}$ in the simplest possible case $R_1(t) \equiv (r-1)t$, then it is easily shown from Proposition 6.1 and (6.4) that the asymptotic behaviour of $L_{K,app}$ is not entirely correct: Under the conditions of Proposition 6.1, $L_K/L_{K,app}$ converges to a constant which is positive and finite, but not equal to one. The same conclusion can be drawn if the activity periods are exponentially distributed. □

We now turn to the mean buffer content, where we restrict ourself to the (important) special case of activity periods with infinite second moments (corresponding to long-range dependent input, see [10]). It is also assumed that the silence periods are exponentially distributed.

Proposition 6.2 *If $\mathbb{P}(D_1 > x) = L(x)x^{-\nu}$, $1 < \nu < 2$ and if the conditions in Theorem 5.2 hold, then*

$$\mathbb{E}[Z^K] \sim \frac{\rho}{1-\rho}\frac{(r-1)^{\nu-1}}{(\nu-1)\mathbb{E}[D_1]}\left[\frac{p}{2-\nu} - \frac{r-1}{r}\right]L(K)K^{2-\nu}, \tag{6.7}$$

if $K \to \infty$.

Proof We will obtain an asymptotic expansion for both terms in the formula for $\mathbb{E}[Z^K]$ given in Theorem 5.4:

$$\mathbb{E}[Z^K] = \frac{1}{\mathbb{P}(W \leq K)}\int_0^K \mathbb{P}(Z > x)\mathrm{d}x - \frac{K\mathbb{P}(W > K)}{\mathbb{P}(W \leq K)}. \tag{6.8}$$

For the second term we have, by (6.3) and the identity $\frac{\hat{\rho}}{1-\hat{\rho}} = \frac{\rho}{1-\rho}\frac{r-1}{r}$,

$$\frac{K\mathbb{P}(W > K)}{\mathbb{P}(W \leq K)} \sim \frac{\rho}{1-\rho}\frac{(r-1)^{\nu}}{r(\nu-1)\mathbb{E}[D_1]}L(K)K^{2-\nu}, \tag{6.9}$$

if $K \to \infty$. The tail behaviour for Z follows straightforwardly from that of $\widetilde{D}_1$, which follows by applying Karamata's theorem (see e.g. Section 1.6 in [8]). This gives for $x \to \infty$,

$$\mathbb{P}(Z > x) \sim p\frac{\rho}{1-\rho}\frac{(r-1)^{\nu-1}}{(\nu-1)\mathbb{E}[D_1]}L(x)x^{1-\nu}. \tag{6.10}$$

Applying Karamata's theorem once more to the first term in the right hand side of (6.8), we get

$$\int_0^K \mathbb{P}(Z > x)\mathrm{d}x \sim \frac{p(r-1)^{\nu-1}}{(\nu-1)\mathbb{E}[D_1]}\frac{\rho}{1-\rho}\frac{1}{2-\nu}K^{2-\nu}L(K). \tag{6.11}$$

The proof follows by combining (6.9) and (6.11), thereby noting that the constant in (6.11) is larger than the constant appearing in (6.9). This follows from $\rho = \lambda(r-1)\mathbb{E}[D_1] < 1$ and $1 < \nu < 2$, which implies $p/(2-\nu) > p > (r-1)/r$. ■

Loosely speaking, the mean buffer content behaves like a power of the buffer size in case of long-range dependent input. This shows once more that the impact of long-range dependence on the performance of fluid queues can be quite substantial – even if buffers are finite.

Remark 6.2 For the model with a single on-off source it is also possible to obtain multiterm asymptotic expansions or even explicit results for the loss fraction. The classes of (heavy-tailed) service time distributions introduced in [12, 1] lead to explicit results for the waiting time distribution in the M/G/1 queue. The results in [12, 1] may also be used to obtain more refined asymptotics and explicit results for the mean buffer content. □

6.3 A superposition of N on-off sources

The characteristics of this model are described as follows. When source i, $1 \leq i \leq N$, is on, it transmits fluid at rate $r_i \geq 1$ during an activity period A_i having mean α_i. The silence periods S_i are exponentially distributed with parameter λ_i. The stationary probability of silence p_i equals $1/(1+\alpha_i\lambda_i)$, the mean offered load per unit of time offered by source i is denoted by ρ_i and equals $r_i\frac{\lambda_i\alpha_i}{1+\lambda_i\alpha_i}$. Note that in our setting, $\rho = \rho_1 + \cdots + \rho_N$, $\lambda = \lambda_1 + \cdots + \lambda_N$, and $p = \prod_i p_i$. Using this, it is not difficult to calculate $\mathbb{E}[D_1]$ and $\mathbb{E}[R_1(D_1-)]$. The following result is part of Theorem 4.6 in [10].

Lemma 6.1 *Suppose that the activity periods of the sources 2,...,N are exponentially distributed and suppose that*

$$\mathbb{P}(A_1 > x) = L(x)x^{-\nu},$$

for $\nu > 1$. Suppose that $\rho < 1$ and define $c = 1 - \sum_{i=2}^{N} \rho_i$. Then, the following asymptotics hold for $x \to \infty$:

$$\mathbb{P}(W > x) \sim \frac{\lambda_1 (r_1 - c)\alpha_1}{c - \lambda_1 (r_1 - c)\alpha_1} \mathbb{P}((r_1 - c)\widetilde{A}_1 > x), \tag{6.12}$$

$$\mathbb{P}(Z > x) \sim p_1 \frac{\rho_1}{c - \rho_1} \mathbb{P}((r_1 - c)\widetilde{A}_1 > x). \tag{6.13}$$

Note that the asymptotics for $\mathbb{P}(W > x)$ and $\mathbb{P}(Z > x)$ are the same as in the case of a fluid queue fed by a single on-off source and output rate c. The on-off source with the heaviest tail dominates the asymptotic behaviour, whereas other sources only contribute to the asymptotics through their means. In [2], this notion is called a *reduced load balance* and is shown to hold for the tail of Z under much more general conditions. The conditions under which the tail behaviour for W in Lemma 6.1 is valid can also be weakened, see [10].

Lemma 6.1 leads to the following results for the loss fraction and the mean buffer content in the finite buffer case.

Theorem 6.2 *Suppose that the conditions stated in Lemma 6.1 are valid. Then, for $K \to \infty$,*

$$L_K \sim M \mathbb{P}((r_1 - c)\widetilde{A}_1 > K), \tag{6.14}$$

where M is given by

$$M = \frac{1 - \rho}{\rho} \frac{\lambda_1 (r_1 - c)\alpha_1}{c - \lambda_1 (r_1 - c)\alpha_1}.$$

If $1 < \nu < 2$, the mean buffer content satisfies for $K \to \infty$,

$$\mathbb{E}[Z^K] \sim \frac{(r - c)^{\nu - 1}}{(\nu - 1)\alpha_1} \left[p_1 \frac{\rho_1}{1 - \rho_1} \frac{1}{2 - \nu} - \frac{\rho}{1 - \rho} M \right] L(K) K^{2 - \nu}. \tag{6.15}$$

Proof The first part follows easily from Lemma 6.1, Theorem 3.1, and Theorem 5.3, or alternatively, from Theorem 6.1 and the tail behaviour of $R_1(D_1 -)$, which is given in Theorem 4.6 of [10]. The proof of the second part follows the same lines as the proof of Proposition 6.2 and is therefore omitted. ■

Theorem 6.2 can be generalised to more generally distributed activity periods (for all sources), as long as Lemma 6.1 remains valid under these assumptions. In particular, the results in [2] show that Lemma 6.1 and Theorem 6.2 remain true in the case of N regularly varying sources, as long as the tail activity period of the first source is heavier than is the case for the other sources.

6.4 Infinitely many sources

The fluid queue with an infinite number of on-off sources can be described as follows. Activations of sources occur according to a Poisson process with rate λ and have a duration of length A with mean α_1. Hence, the number of active sources has the same distribution as the number of customers present in an M/G/∞ queue. While a source is active, it transmits fluid at rate $r \geq 1$.

It is easily shown that

$$\mathbb{E}[D_1] = \frac{\mathrm{e}^{\lambda\alpha_1} - 1}{\lambda}, \quad \mathbb{E}[R_1(D_1-)] = r\alpha_1 \mathrm{e}^{\lambda\alpha_1} - \frac{\mathrm{e}^{\lambda\alpha_1} - 1}{\lambda}, \quad \rho = r\lambda\alpha_1.$$

We derive an asymptotic expression for the loss fraction. Instead of using results for the tail of W, we use an asymptotic result for the tail of $R_1(D_1-)$ and Theorem 6.1. The most general result for the former is given in Theorem 1 of [40] and only requires $r \geq 1$ and on-times with an intermediately regularly varying distribution function. Under these conditions we have that

$$\mathbb{P}(R_1(D_1-) > x) \sim \frac{1}{\rho}\mathbb{P}(A > x/(r-1+\rho)), \qquad x \to \infty, \tag{6.16}$$

so in case of regular variation we have

$$\mathbb{P}(R_1(D_1-) > x) \sim \frac{1}{\rho}(r-1+\rho)^{\nu} L(x) x^{-\nu}, \qquad x \to \infty.$$

Using this result together with Theorem 6.1 and Karamata's theorem gives

Theorem 6.3 *If $r \geq 1$, and if*

$$\mathbb{P}(A > x) = L(x)x^{-\nu}, \qquad \nu > 1,$$

then

$$L_K \sim \frac{e^{-\lambda\alpha_1}}{r\alpha_1}\frac{1}{\nu-1}\frac{1}{\rho}(r-1+\rho)^{\nu} L(K) K^{1-\nu}, \qquad K \to \infty.$$

For the tail of Z no results are available in this setting, although a conjecture is stated in [30]. This conjecture can be used to obtain an asymptotic expansion for the mean of Z^K.

Results for a discrete time version of this model can be found in [38]. For more results, like upper bounds for the tail of $R_1(D_1-)$, W and Z in this model, which may be used to obtain upper bounds for the loss fraction and the mean buffer content in the finite buffer case, we refer to Section 5.2 in [10] and to [27, 40].

7. Overloaded queues

In this section we look at the case when the traffic load is at least 1, i.e. when $\rho \geq 1$ (equivalently $\hat{\rho} \geq 1$). If the silence periods are exponentially distributed, it is possible to use the results for the M/G/1 queue with finite capacity K given in Section III.5 of [17]. For this model we develop asymptotic expansions for the loss fraction, which can easily be applied to the fluid model by means of Theorem 5.3. Starting point of our analysis is the following expression for $\mathbb{P}(W^K = 0)$, given on p. 535 of [17], which is, just as (3.3), valid for all $\hat{\rho} > 0$.

$$\mathbb{P}(W^K = 0) = \left[\frac{1}{2\pi \mathrm{i}} \int\limits_{s=-\mathrm{i}\infty+\varepsilon}^{\mathrm{i}\infty+\varepsilon} \frac{\mathrm{e}^{sK}}{s - \lambda + \lambda\beta(s)} \mathrm{d}s \right]^{-1}, \tag{7.1}$$

where $\beta(s)$ is the Laplace-Stieltjes transform of the service time B (which will equal $R_1(D_1-)$ when applied to the fluid model). ε must be chosen such that all zeroes of $s - \lambda + \lambda\beta(s)$ have real part smaller than ε. If $\hat{\rho} \leq 1$, any $\varepsilon > 0$ suffices. Note that the Laplace-Stieltjes transform of $\mathbb{P}(W^K = 0)^{-1}$ with respect to K is given by, for Re $s > \varepsilon$,

$$\int\limits_0^\infty \mathrm{e}^{-sK} \mathrm{d} \left[\mathbb{P}(W^K = 0)^{-1} \right] = \frac{s}{s - \lambda + \lambda\beta(s)}. \tag{7.2}$$

We now apply Equation (7.2) to derive asymptotic expressions for the loss probability when $\rho \geq 1$. Define β_i as the $i-$th moment of the service time B. We first consider the case $\rho = 1$ (and hence $\hat{\rho} = 1$).

Proposition 7.1 *Let $\hat{\rho} = 1$.*

1. *If $\beta_2 < \infty$, then*

$$L_{q,K} \sim \frac{\beta_2}{2\beta_1} \frac{1}{K}, \qquad K \to \infty.$$

2. *If $\mathbb{P}(B > x) = L(x)x^{-\nu}$, $1 < \nu < 2$, then*

$$L_{q,K} \sim \frac{1}{\beta_1} \frac{\pi}{\sin(\pi(\nu - 1))} L(K) K^{1-\nu}, \qquad K \to \infty.$$

Proof Both assertions will be proven by the use of Tauberian theorems. Let $\beta(s)$ be the LST of B. Since $\hat{\rho} = 1$, (3.8) reduces to

$$L_{q,K} = \mathbb{P}(W^K = 0). \tag{7.3}$$

First, we prove Part 1. Since $\beta_2 < \infty$, we have

$$\beta(s) = 1 - \beta_1 s + \frac{1}{2}\beta_2 s^2 + \mathrm{o}(s^2), \qquad s \downarrow 0. \tag{7.4}$$

Inserting (7.4) in (7.2) yields

$$\int\limits_0^\infty \mathrm{e}^{-sK} \mathrm{d} \left[\mathbb{P}(W^K = 0)^{-1} \right] = \frac{2\beta_1}{\beta_2 s} + \mathrm{o}(1/s), \qquad s \downarrow 0,$$

which gives Part 1 of Proposition 7.1 by using the (classical) Tauberian theorem, see e.g. [22], and Proposition 3.1.

We now turn to Part 2. By Theorem 8.1.6 in [8] (see also [7]), if $\mathbb{P}(B > x) = L(x)x^{-\nu}$, $1 < \nu < 2$, $\beta(s)$ satisfies

$$\beta(s) - 1 + \beta_1 s \sim -\Gamma(1-\nu)s^{\nu}L(1/s), \qquad s \downarrow 0. \tag{7.5}$$

This gives, since $\hat{\rho} = 1$,

$$\int_0^\infty \mathrm{e}^{-sK}\mathrm{d}\left[\mathbb{P}(W^K = 0)^{-1}\right] \sim \frac{\beta_1}{-\Gamma(1-\nu)}s^{1-\nu}/L(1/s), \qquad s \downarrow 0. \tag{7.6}$$

Applying the Tauberian theorem 1.7.1 in [8], we get for $K \to \infty$,

$$\mathbb{P}(W^K = 0)^{-1} \sim \frac{\beta_1}{-\Gamma(\nu)\Gamma(1-\nu)}K^{\nu-1}/L(K). \tag{7.7}$$

Part 2 now follows from $\Gamma(\nu)\Gamma(1-\nu) = \pi/\sin(\pi\nu)$ and $\sin(a) = -\sin(a-\pi)$. ■

Remark 7.1 We refrain from discussing the case in which $\mathbb{P}(B > x) = L(x)x^{-2}$ (and $\beta_2 = \infty$). The Tauberian theorems are now much more delicate, see e.g. Theorem 8.1.6 in [8]. □

Remark 7.2 Although the asymptotic formula for the loss probability in the case $\hat{\rho} < 1$ (given in Theorems 4.1 and 4.2) is independent of $\hat{\rho}$, it is not valid for $\hat{\rho} = 1$, as Proposition 7.1 shows. However, note that the asymptotic behaviour of the loss probability in the heavy-tailed (infinite variance) case is the same for $\hat{\rho} < 1$ and $\hat{\rho} = 1$, apart from a constant. Since $\sin x < x$ for $x > 0$, $\pi/\sin(\pi(\nu-1)) > 1/(\nu-1)$, so the constant in the asymptotic approximation for $L_{q,K}$ is strictly larger for $\hat{\rho} = 1$ than for $\hat{\rho} < 1$. □

When $\hat{\rho} > 1$, it follows immediately that $\mathbb{P}(W^K = 0) \to 0$ if $K \to \infty$, which gives

$$L_{q,K} \to 1 - \frac{1}{\hat{\rho}}. \tag{7.8}$$

Using a result of Cohen [16], it is easy to derive the rate of convergence.

Proposition 7.2 *If $\hat{\rho} > 1$, then we have for the M/G/1 queue with finite capacity K,*

$$L_{q,K} - \frac{\hat{\rho}-1}{\hat{\rho}} \sim -\delta\widetilde{\beta}'(\delta)e^{-\delta K}, \qquad K \to \infty, \tag{7.9}$$

where $\widetilde{\beta}(s) = \frac{1-\beta(s)}{\beta_1 s}$, $\widetilde{\beta}'(s)$ is the derivative of $\widetilde{\beta}(s)$, and δ is the unique positive real solution of

$$\hat{\rho}\widetilde{\beta}(s) = 1. \tag{7.10}$$

Proof Follows immediately from (3.8) and Part (iii) of Theorem 2.3 in [16]. ■

8. Concluding remarks

In this study, a fluid model with a finite buffer has been studied. It has been shown that the fluid model under consideration is related to the finite dam with instantaneous input, and to the fluid queue with infinite buffer. These relations are applied, along with some new results for the finite dam, to obtain asymptotic expansions for the loss fraction and the mean buffer content in case of heavy-tailed input. The results obtained show that the performance of the fluid queue is seriously affected by heavy-tailed input characteristics. In particular, the loss fraction decays to zero very slowly when the buffer size gets large.

There are several topics which may be interesting for further research. The asymptotic expansions developed in Section 6 may be of use to estimate the loss fraction and the mean buffer content of these models. It also might be possible to generalize these asymptotic results by generalising Theorem 4.2 from regularly varying to subexponential service time distributions.

A serious limitation of the fluid model discussed in this paper is that it does not cover the case in which on-off sources send at a peak rate which is smaller than the output rate. We note however, that the analysis of this model is already quite difficult in the infinite buffer case, see e.g. [2]. A challenging topic for future research is to obtain asymptotic expansions for the loss fraction for this practically important case.

9. Appendix

In this section we give a direct proof of the proportionality result in Section 5 (Theorem 5.2), because we believe it is interesting in itself. It is an extension of the proof of Hooghiemstra [29] for the proportionality result (3.6) for the M/G/1 finite dam. We start with two preliminary observations.

1. Let c and c^K be the length of a busy cycle for the infinite buffer model and the model with finite buffer K respectively. Then, the laws of Z and Z^K are given by, cf. [14, 3]

$$\mathbb{P}(Z \leq x) = \frac{1}{\mathbb{E}[c]}\mathbb{E}\left[\int_0^c 1_{[0,x]}(Z(t))\mathrm{d}t\right], \tag{9.1}$$

$$\mathbb{P}(Z^K \leq x) = \frac{1}{\mathbb{E}[c^K]}\mathbb{E}\left[\int_0^{c^K} 1_{[0,x]}(Z^K(t))\mathrm{d}t\right]. \tag{9.2}$$

2. Let $x < K < \infty$ and suppose that a downcrossing at level x occurs for the process $Z^K(t)$ for some t, so that the environment process $I(t) = 1$. Then, since U_1 is exponentially distributed, the time that elapses until I reaches zero is distributed as U_1, due to the memoryless property of the silence periods.

We now construct a stochastic process $\widehat{Z}^K(t)$ directly from $Z(t)$. Consider an arbitrary sample path of $Z(t)$, e.g. the sample path in Figure 1.

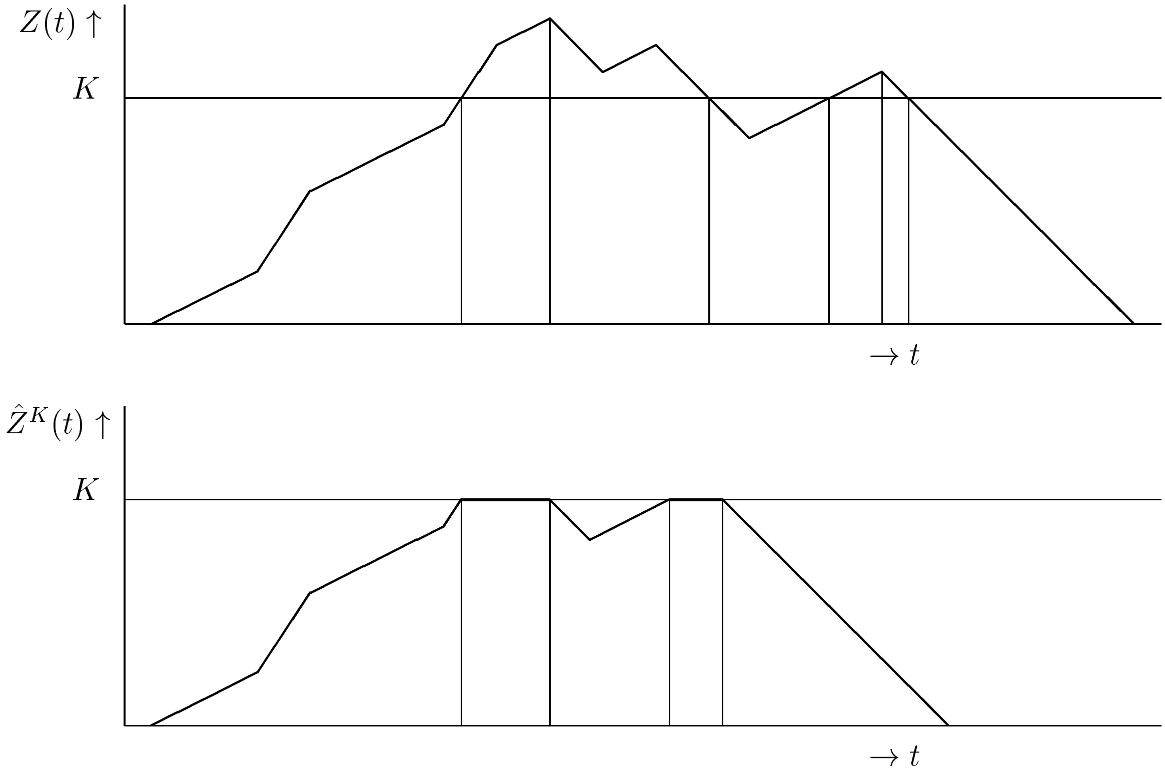

Figure 1: Construction of a sample path of $\widehat{Z}^K(t)$ from $Z(t)$.

The corresponding sample path for $\widehat{Z}^K(t)$ is constructed as follows. The parts of the sample path of $Z(t)$ below level K remain unchanged. Consider the parts of the sample path of $Z(t)$ between an upcrossing and a consecutive downcrossing of level K. Each of these parts can be divided into two sub-parts. The first sub-part is defined as the remaining activity period and the second sub-part as the remainder of the part. Now delete the second sub-part and truncate the first sub-part to level K, cf. Figure 1.

Since the silence periods in the infinite buffer model are exponentially distributed, the same holds for silence periods in the process $\{\widehat{Z}^K(t) : t \geq 0\}$, by Observation 2. It follows immediately from the construction of $\{\widehat{Z}^K(t)\}$ that the durations of the activity periods in $\{\widehat{Z}^K(t)\}$ have the same distribution as D_1, and are all independent and independent of the silence periods. Finally, the trajectories during activity periods kan be chosen identically (in distribution) and independent from each other according to the stochastic process $\{R_1(t) : t \geq 0\}$. Hence, the dynamics as $\{\widehat{Z}^K(t)\}$ satisfy the same dynamics as the process $\{Z^K(t)\}$, as defined by Equations (2.4) and (2.5). This proves that $\{\widehat{Z}^K(t) : t \geq 0\}$ has the same law as $\{Z^K(t) : t \geq 0\}$.

To simplify the notation, we now can define the process $Z^K(t)$ as

$$Z^K(t) := \widehat{Z}^K(t), \qquad\qquad t \geq 0.$$

It follows immediately from the construction of $Z^K(t)$ from $Z(t)$ that the number of downcrossings from level $x \leq K$ is the same for their respective sample paths, which implies that the number of downcrossings at level $x \leq K$ of the process $Z^K(t)$ during a busy cycle has the same distribution as that of $Z(t)$ for each $0 < K < \infty$.

A second implication of the construction carried out is that

$$\int_0^{c^K} 1_{[0,x]}(Z^K(t))\mathrm{d}t = \int_0^{c} 1_{[0,x]}(Z(t))\mathrm{d}t, \tag{9.3}$$

for $0 \leq x < K$. This implies the proportionality between the stationary distributions of $Z^K(t)$ and $Z(t)$, by (9.1) and (9.2): Define $\gamma = \frac{\mathbb{E}[c]}{\mathbb{E}[c^K]}$.

We relate γ to the loss fraction L_K by using variants of Little's formula, see also Section 3. The amount of work brought into the system per unit of time equals ρ in the infinite buffer model and $\rho(1 - L_K)$ in the finite buffer model. Hence, we have by Little's formula that

$$\mathbb{P}(Z = 0) = 1 - \rho, \tag{9.4}$$

and, for $K \geq 0$,

$$\mathbb{P}(Z^K = 0) = 1 - \rho(1 - L_K). \tag{9.5}$$

Consequently,

$$\gamma = \frac{1 - \rho(1 - L_K)}{1 - \rho}. \tag{9.6}$$

A straightforward computation (use (3.9) and Theorem 5.3) shows that $\gamma = 1/\mathbb{P}(W \leq K)$.

Acknowledgements

The author is indebted to Prof. O.J. Boxma and Prof. J.W. Cohen for stimulating discussions and useful remarks.

References

1. J. Abate and W. Whitt. Explicit M/G/1 waiting-time distributions for a class of long-tail service-time distributions, preprint, AT&T Labs, 1998.
2. R. Agrawal, A.M. Makowski and Ph. Nain. On a reduced load equivalence for fluid queues under subexponentiality. To appear in *Queueing Systems*, 1998.
3. S. Asmussen. *Applied Probability and Queues,* Wiley, Chicester, 1987.
4. S. Asmussen and C. Klüppelberg. Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic Processes and their Applications* 64:103-125, 1996.
5. G.A. Awater. *Broadband Communication – Modeling, Analysis and Synthesis of an ATM Switching Element,* Ph.D thesis, Delft University, 1994.
6. J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video. *IEEE Transactions on Communications*, 43:1566-1579, 1995.
7. N.H. Bingham and R.A. Doney. Asymptotic properties of supercritical branching processes I: The Galton-Watson process. *Advances in Applied Probability*, 6:711-731, 1974.
8. N.H. Bingham, C.M. Goldie, and J.L. Teugels. *Regular Variation,* Cambridge University Press, Cambridge, 1987.
9. N.K. Boots and H.C. Tijms. A multi-server queueing system with impatient customers. To appear in *Management Science,* 1998.
10. O.J. Boxma and V. Dumas. Fluid queues with heavy-tailed activity periods. Report PNA-R9705, CWI, Amsterdam, 1997. To appear in *Computer Communications.*
11. O.J. Boxma and V. Dumas. The busy period in the fluid queue. *Performance Evaluation Review,* 26:100-110, 1998.
12. O.J. Boxma and J.W. Cohen. The M/G/1 queue with heavy-tailed service time distribution. To appear in *Queueing Systems,* 1998.
13. J.W. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic*

Processes and their Applications, 2:31-58, 1974.

14. J.W. Cohen. *Regenerative Processes in Queueing Theory,* Springer Verlag, Berlin, 1976.
15. J.W. Cohen. On up- and downcrossings. *Journal of Applied Probability,* 14:405-410, 1977.
16. J.W. Cohen. On the maximal content of a dam and logarithmic concave renewal functions. *Stochastic Processes and their Applications*, 6:291-304, 1978.
17. J.W. Cohen. *The Single Server Queue,* Second edition, North Holland, Amsterdam, 1982.
18. J.W. Cohen. On the effective bandwidth in buffer design for the multi-server channels. Report BS-R9406, CWI, Amsterdam, 1994.
19. J.W. Cohen. Heavy-traffic limit theorems for the heavy-tailed GI/G/1 queue. Report PNA-R9719, CWI, Amsterdam, 1997.
20. D.J. Daley. General customer impatience in the queue GI/G/1. *Journal of Applied Probability,* 2:186-205, 1964.
21. P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events,* Springer Verlag, Berlin, 1997.
22. W. Feller. *An Introduction to Probability Theory and its Applications,* Volume II. Second edition, Wiley, New York, 1971.
23. F. Gouweleeuw. *A General Approach to Computing Loss Probabilities,* Ph.D Thesis, Free University, Amsterdam, 1996.
24. F. Gouweleeuw. Calculating the loss probability in a $BMAP/G/1/N+1$ queue. *Stochastic Models,* 12:473-492, 1996.
25. L. Green. A limit theorem on subintervals of interrenewal times. *Operations Research,* 30, 210-216, 1982.
26. D. Heath, S. Resnick, and G. Samorodnitsky. Patterns of buffer overflow in a class of queues with long memory in the input stream. *Annals of Applied Probability,* 7:1021-1057, 1997.
27. D. Heath, S. Resnick, and G. Samorodnitsky. How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. Technical report, School of OR & IE, Cornell University, 1997.
28. D. Heath, S. Resnick, and G. Samorodnitsky. Heavy tails and long range dependence in on/off processes and associated fluid models. *Mathematics of Operations Research,* 23, 145-165, 1998.
29. G. Hooghiemstra. A path construction for the virtual waiting time of an M/G/1 queue. *Statistica Neerlandica,* 41:175-181, 1987.
30. P. Jelenkovic and A. Lazar. Asymptotic results for multiplexing on-off sources with subexponential on-periods. To appear in *Advances in Applied Probability,* 1999.
31. P. Jelenkovic. Long-tailed loss rates in a GI/GI/1 queue with applications. Submitted to Queueing Systems, 1998.

32. D. Kennedy. Limit theorems for finite dams. *Stochastic Processes and their Applications,* 1:269-278, 1973.

33. O. Kella and W. Whitt. A storage model with a two-state random environment. *Operations Research,* 40, S257-S262, 1992.

34. D.V. Lindley. (Discussion of a paper by C.B. Winsten.) *Journal of the Royal Statistical Society, Series B,* 21:22-23, 1959.

35. R.M. Loynes. On a property of the random walks describing simple queues and dams. *Journal of the Royal Statistical Society, Series B,* 27:125-129, 1965.

36. Do Le Minh. The $GI/G/1$ queue with uniformly limited virtual waiting times; the finite dam. *Advances in Applied Probability,* 12:501-516, 1980.

37. A.G. Pakes. On the tails of waiting-time distributions. *Journal of Applied Probability,* 12:555-564, 1975.

38. M. Parulekar and A.M. Makowski. Tail probabilities for $\mathrm{M/G/}\infty$ input processes (I): Preliminary asymptotics. *Queueing Systems*, 27:271-296, 1997.

39. V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226-244, 1995.

40. S. Resnick and G. Samorodnitsky. Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues. Technical report, School of OR & IE, Cornell University, 1997.

41. R.E. Stanford. Reneging phenomena in single channel queues. *Mathematics of Operations Research*, 4:162-178, 1979.

42. L. Takács. *Combinatorial Methods in the Theory of Stochastic Processes,* Wiley, New York, 1967.

43. H.C. Tijms. *Stochastic Models – An Algorithmic Approach,* Wiley, Chicester, 1994.

44. N. Veraverbeke. Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stochastic Processes and their Applications,* 5:27-37, 1977.

45. 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*, 10:67-85, 1995.

46. W. Whitt. A review of $L = \lambda W$ and extensions. *Queueing Systems,* 9:235-268, 1991.