



Centrum voor Wiskunde en Informatica

REPORTRAPPORT

Fluid Queues with Long-tailed Activity Period Distributions

O.J. Boxma and V. Dumas

Probability, Networks and Algorithms (PNA)

PNA-R9705 April 30, 1997

Report PNA-R9705
ISSN 1386-3711

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

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.

SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

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

Fluid Queues with Long-tailed Activity Period Distributions

O.J. Boxma and V. Dumas

CWI

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

ABSTRACT

This is a survey paper on fluid queues, with a strong emphasis on recent attempts to represent phenomena like long-range dependence. The central model of the paper is a fluid queueing system fed by N independent sources that alternate between silence and activity periods. The distribution of the activity periods of at least one source is assumed to be long-tailed, which may give rise to long-range dependence. We consider the effect of this tail behaviour on the steady-state distributions of the buffer content at embedded points in time and at arbitrary time, and on the busy period distribution. Both exact results and bounds are discussed.

1991 Mathematics Subject Classification: 60K25, 68M20, 90B22.

Keywords & Phrases: Fluid queue, on/off sources, long-range dependence, long-tailed, subexponential, regularly varying, buffer content, busy period, *GI/G/1* queue.

Note: This is an invited survey, to appear in the special issue on 'Stochastic Analysis and Optimisation of Communication Systems' of the journal 'Computer Communications'. The work is carried out under the project LRD in PNA2.1.

1. INTRODUCTION

In modern communication networks, a hierarchy of time scales can be distinguished [49]: call scale, burst scale and cell scale. At the *call scale*, the largest time scale, the entities are the holding times of arriving calls or service demands. During such holding times, small units of traffic called *cells* are generated by a source. In ATM traffic, e.g., a cell is a 53-byte packet. This *cell scale* granularity is ignored at the intermediate *burst scale*, where one simply characterizes the input process by its rate. Hence *fluid* models are natural on the latter time scale. A burst of consecutive cells, generated by a source, becomes the amount of fluid produced by this source during one of its activity periods. As long as there is fluid in the buffer, there is outflow at a constant rate. Fluid models have, in the last 15 years, become firmly established as key models for capturing the behaviour of a wide range of, in particular ATM-based, communication networks at the burst scale.

Early studies of fluid models can be found in the applied probability literature under the heading of *dam models* (see for example [16] or [47] and references therein). In the

early seventies L. Kosten, modelling a class of data communication networks, analysed fluid queues fed by an infinite number of on/off sources, viz., sources that alternate between active (on) and silent (off) periods. His path-breaking studies were published in an internal publication series of Delft University of Technology ([34, 35, 37], successively discussing the cases of exponential, Erlang and hyperexponential on-period distributions), and first did not reach a large audience. A similar fate was met by other pioneering studies, of Rubinovitch [51] (busy period results for the case of N identical sources), Cohen [15] (a very deep and detailed study about the case of identical sources with generally distributed on-periods) and Kaspi & Rubinovitch [30] (N non-identical sources). The fundamental paper of Anick, Mitra and Sondhi [2] has generated new interest in fluid queues. It considers N on/off sources with exponentially distributed on- and off-periods. Similar to Kosten's work, the equilibrium buffer content distribution in [2] is described by a set of differential equations. All the eigenvalues are obtained explicitly, and simple expressions are obtained for moments of the distribution and the asymptotic behaviour of the buffer content. This paper, and a new paper of Kosten [36], have contributed much towards establishing the fluid queue fed by on/off sources as a key model for representing traffic behaviour in modern communication networks. However, in these studies, and in important successors like [41] and [54], the on- and off-periods are exponentially distributed or determined by some Markov process, leading to an input process that is short-range dependent (see the definition in Subsection 2.1), and to exponential behaviour of the tail of the buffer content distribution.

It came as a shock when traffic measurements at Bellcore for Ethernet Local Area Network traffic clearly demonstrated [55] that short-range dependence assumptions are violated. The traffic plots show a striking similarity when one considers a time period of hours, minutes or milliseconds: bursty subperiods are alternated by less bursty subperiods on each scale. This scale-invariant or *self-similar* feature of Ethernet traffic, and the related phenomenon of *long-range dependence*, was also convincingly demonstrated in [38] using a careful statistical analysis. The presence of long-term correlations has also been demonstrated for traffic measurements on communication systems different from Ethernet, like Wide Area Networks [46] and VBR video [6]; see [23, 55] for additional information.

As observed in [55], in many cases on- and/or off-periods of actual traffic sources exhibit a tail behaviour that is far from exponential. The observation of long-range dependence and *non-exponential tail behaviour* has spurred a strong interest in the modelling and analysis of traffic in modern communication networks. Fluid queues fed by on/off sources with non-exponential on- and/or off-periods appear to be among the most natural models for these purposes, and considerable success has recently been obtained in their analysis. The present paper is an invited survey paper on fluid queues, with a strong emphasis on recent attempts to represent long-range dependence via non-exponential tails.

While in the classical queueing literature a typical result is that waiting time, workload and busy period distributions have an exponential tail whenever the service time

distribution has an exponential tail, the focus of many of the results surveyed in this paper is: if the activity period of one or more sources has a certain long-tailed behaviour (for example subexponential or regularly varying) then the buffer content and busy period exhibit a similar long-tailed behaviour. In the consideration of buffer content distribution, throughout the paper a distinction is being made between steady-state buffer content and buffer content at the end of a silence period.

The remainder of the paper is organized in the following way. In Section 2 we discuss the concepts of long-range dependence and self-similarity. We study ways to incorporate these features in fluid queues, and we mention a physical explanation for their occurrence in communication networks; their impact on network performance is also discussed.

Section 3 is devoted to the case of a fluid queue with an infinite buffer fed by a single source that alternates between silence and activity periods. The relative simplicity of this system allows us to admit a rather general input process during the activity periods. The close relation between the fluid queue and an ordinary GI/G/1 queue is explained, and subsequently exploited. Detailed analytic results for (the tail behaviour of) the buffer content distribution and the busy period distribution are presented.

Section 4 considers a fluid queue with an infinite buffer, fed by an arbitrary finite number of independent on/off sources (i.e., the inflow rate of each source during an activity period is constant). We assume that the off-periods are negative exponentially distributed. We first discuss the - rather scarce - exact analytic results for buffer content and busy period distribution, and subsequently we discuss bounds for the buffer content distribution.

Section 5 presents results for the same model as Section 4, but with an *infinite* number of on/off sources. Under additional assumptions, this limiting case allows a detailed treatment. This holds in particular when all off-periods are exponentially distributed, so that the times at which a source becomes active occur according to a Poisson process.

Section 6 contains conclusions and mentions some of the most challenging open problems.

Some key results about long-tailed, subexponential and regularly varying distributions are gathered in the appendices.

We end this introduction by listing some

Notations and conventions: except otherwise stated:

- any random variable X is assumed positive with unbounded support:

$$\forall x \geq 0 : \quad 0 < \mathbb{P}[X > x] \leq \mathbb{P}[X > 0] = 1.$$

- in a sum of random variables, all are assumed independent from each other.

If $\mathbb{E}[X] < \infty$, then X^* denotes a random variable of distribution:

$$\mathbb{P}[X^* > x] = \frac{1}{\mathbb{E}[X]} \int_x^{+\infty} \mathbb{P}[X > y] dy, \quad x \geq 0.$$

Other notations:

- $X \leq_{st} Y$ (resp. $X \simeq Y$) means that $\mathbb{P}[X > x] \leq \mathbb{P}[Y > x]$ (resp. $\mathbb{P}[X > x] = \mathbb{P}[Y > x]$) for all $x \geq 0$.
- LST for Laplace-Stieltjes transform.
- $f(t) \stackrel{t \rightarrow \infty}{\sim} g(t)$, or simply $f(t) \sim g(t)$, means that $f(t)/g(t) \rightarrow 1$ as $t \rightarrow \infty$.
- $f(t) = o(g(t))$ (resp. $f(t) = O(g(t))$) means that $f(t)/g(t) \rightarrow 0$ as $t \rightarrow \infty$ (resp. $\limsup_{t \rightarrow \infty} f(t)/g(t) < \infty$).
- $[x]^+ := \max(0, x)$, $x \in \mathbb{R}$.

2. LONG-RANGE DEPENDENCE

2.1 The concepts of long-range dependence and self-similarity

Let $r(s)$ denote the arrival rate at time s of traffic offered to a system. $T(t) := \int_0^t r(s) ds$ denotes the total amount of traffic offered to that system in the time interval $[0, t]$. We assume that $T(t)$ has stationary increments and that $\mathbb{E}[T(t)^2] < \infty$. The stationarity of the increments implies that $\mathbb{E}[T(t)] = mt$ for some constant m . Furthermore, the autocovariance function: $c(t) := \text{Cov}(r(s), r(s+t))$ does not depend on s (weak stationarity), and: $\text{Var}(T(t)) = 2 \int_0^t \int_0^u c(v) dv du$.

If $\int_0^\infty c(t) dt < \infty$, then the process is called *short-range dependent*, and $\text{Var}(T(t))$ is asymptotically linear (e.g., if $T(t)$ is a process with independent increments like a compound Poisson process, then $\text{Var}(T(t))$ is linear in t). Of more interest to us is the case in which $c(t) \rightarrow 0$ so slowly for $t \rightarrow \infty$, that $\int_0^\infty c(t) dt = \infty$. The process $T(t)$ is now called *long-range dependent* (LRD). In particular, assume that $c(t) \sim l(t)t^{-\nu}$, $0 < \nu < 1$, for $t \rightarrow \infty$, with $l(t)$ a slowly varying function (cf. Appendix B). Then, cf. [49] p. 325 or Appendix B, $\text{Var}(T(t)) \sim Kl(t)t^{2-\nu} = Kl(t)t^{2H}$ for large t , some K and $H = (2 - \nu)/2 \in (1/2, 1)$. The number H is called the *Hurst* parameter.

It should be noted that, if $\text{Var}(T(t)) = t^{2H}$, then the time-scaled process $T(\gamma t)$ has

$$\text{Var}(T(\gamma t)) = (\gamma t)^{2H} = \gamma^{2H} \text{Var}(T(t)),$$

hence $T(\gamma t)$ and $\gamma^H T(t)$ have the same correlation structure, i.e., the centered process $T(t) - mt$ is *second-order self-similar*. We remind the reader of the recent observations of both long-range dependence and self-similarity, that were mentioned in the Introduction.

Definition 2.1 A square integrable process $G(t)$ is called second-order self-similar with self-similarity (or Hurst) parameter H if, for any $\gamma > 0$, the processes $G(\gamma t)$ and $\gamma^H G(t)$ have the same second-order characteristics. It is called asymptotically second-order self-similar if the second-order characteristics of $G(\gamma t)$, suitably normalized, converge to those of a second-order self-similar process when $\gamma \rightarrow \infty$.

Definition 2.2 A process $G(t)$ is called (strictly) self-similar with Hurst parameter H if, for any $\gamma > 0$, the processes $G(\gamma t)$ and $\gamma^H G(t)$ have the same finite-dimensional distributions.

It follows from Definition 2.1 and the text above it, that asymptotic second-order self-similarity with $H > \frac{1}{2}$ is essentially equivalent with LRD. For Gaussian processes (processes of which all finite-dimensional distributions are Gaussian), self-similar and second-order self-similar are equivalent, since their finite-dimensional distributions are fully determined by their first and second moments. A very interesting Gaussian process is Fractional Brownian Motion (FBM). A normalized FBM $\{U(t), t \in (-\infty, \infty)\}$ with Hurst parameter $H \in [\frac{1}{2}, 1)$ is a stochastic process characterized by the following properties.

- $U(t)$ has stationary increments;
- $U(0) = 0$, and $\mathbb{E}[U(t)] = 0$ for all t ;
- $\mathbb{E}[U(t)^2] = |t|^{2H}$ for all t ;
- $U(t)$ has continuous sample paths;
- $U(t)$ is Gaussian.

Note that FBM is self-similar and LRD. In the special case $H = 1/2$, $U(t)$ is the standard Brownian motion. FBM with $H > 1/2$ has a long memory; it is typically proceeding in the same direction for a long time.

2.2 Modelling long-range dependence

In this subsection we discuss three ways to model LRD in an input process.

A first possibility is to use deterministic non-linear (chaotic) maps. Consider a single on/off source in discrete time. Starting from an $x_0 \in (0, 1)$, sequences (x_n, y_n) are generated by the iteration rule

$$\begin{aligned} x_{n+1} &= f_1(x_n), \quad y_n = 0, & \text{if } 0 \leq x_n \leq d, \\ x_{n+1} &= f_2(x_n), \quad y_n = 1, & \text{if } d < x_n < 1. \end{aligned}$$

$y_n = 1$ (0) corresponds to the source being on (off) at time n . The linear maps $f_1(x_n) = x_n/d$, $f_2(x_n) = (x_n - d)/(1 - d)$ (“Bernoulli shift”) lead to geometrically distributed on- and off-periods. But now change $f_1(\cdot)$ into the non-linear map (see [24])

$$f_1(x) = \epsilon + x + \frac{1 - \epsilon - d}{d^m} x^m, \quad (2.1)$$

with $\epsilon \ll d$. For $m=2$, it is remarked in [24] that the distribution of the off-periods has a power tail ($\sim k^{-2}$) while the on-periods are still approximately geometrically distributed. In [48] Pruthi considers the case in which both $f_1(\cdot)$ and $f_2(\cdot)$ are non-linear:

$$f_1(x) = \frac{x}{(1 - c_1 x^{m_1-1})^{1/(m_1-1)}},$$

$$f_2(x) = 1 - \frac{1 - x}{(1 - c_2(1 - x)^{m_2-1})^{1/(m_2-1)}},$$

where $c_1 = d^{1-m_1} - 1$, $c_2 = (1 - d)^{1-m_2} - 1$, with $d \in (0, 1)$, and $m_1, m_2 \geq 1$. The values $m_i \in (3/2, 2)$ correspond to finite mean and infinite variance, thus resulting in an LRD input process.

Chaotic maps like the above allow a concise description of complex traffic phenomena. Much research is to be done concerning the choice of maps that accurately fit actually observed traffic, and concerning the performance analysis of the resulting source models. The latter problem seems very difficult; in [24] a first attempt is made towards a performance analysis of the model resulting from (2.1).

A second way to introduce LRD in an input process is to take a fluid queue with as input process FBM (which is also self-similar). This is proposed in a very original paper of Norros [42], see also [43]. Norros studies the distribution of the buffer content Z for an input process $T(t) = mt + \sqrt{am}U(t)$ and outflow rate $C > m$. His main result is:

$$\mathbb{P}(Z > x) \geq 1 - \Phi\left(\frac{(C - m)^H x^{1-H}}{H^H (1 - H)^{1-H} \sqrt{am}}\right), \quad (2.2)$$

where $\Phi(\cdot)$ is the standard Normal distribution. Using the approximation $1 - \Phi(y) \approx \exp(-y^2/2)$, he suggests as an approximation:

$$\mathbb{P}(Z > x) \approx \exp\left(-\frac{(C - m)^{2H}}{2H^{2H} (1 - H)^{2-2H} am} x^{2-2H}\right). \quad (2.3)$$

Using the theory of large deviations, Duffield and O'Connell [21] (see also [20]) have shown that the approximation (2.3) is logarithmically accurate for large x . For $H = 1/2$, FBM reduces to Brownian motion and the Weibull distribution in (2.3) reduces to an exponential distribution.

A third way to introduce LRD in an input process is to take a fluid queue fed by a single on/off source, and to assume that a typical on-period A has the following tail:

$$\mathbb{P}[A > t] \stackrel{t \rightarrow \infty}{\sim} h_a t^{-a}, \quad (2.4)$$

and/or that a typical off-period S has the following tail:

$$\mathbb{P}[S > t] \stackrel{t \rightarrow \infty}{\sim} h_s t^{-s}, \quad (2.5)$$

with $1 < a, s < 2$ and h_a, h_s positive constants. Indeed (cf. [49]), in those cases $\text{Var}(T(t)) \sim t^{2H}$, where $H = (3 - a)/2$ respectively $H = (3 - s)/2$ if (2.4) respectively (2.5) holds, or $H = \frac{3 - \min(a, s)}{2}$ if both (2.4) and (2.5) hold. Either way, $2H > 1$ and the input process is LRD. In this paper we mainly concentrate on the latter way to introduce LRD.

If the tail of a distribution satisfies (2.4) with $1 < a < 2$, then one speaks of a *heavy-tailed* distribution (although some authors use the term heavy-tailed more generally to denote a non-exponential tail, giving it a similar meaning as *long-tailed*, a term defined in Appendix A). Of course, we can allow more sources, some of them having a heavy-tailed on- and/or off-period distribution. In fact, *the occurrence of heavy-tailed on- and/or off-periods of sources seems to provide the most natural explanation of LRD (and of the related concept of asymptotic second-order self-similarity) in aggregated packet traffic*. In [56] the fluid model with N identical independent on/off sources is considered, and it is assumed that the on-period distribution satisfies (2.4) with $1 < a < 2$ and/or the off-period distribution satisfies (2.5) with $1 < s < 2$. As observed above, one such source already gives rise to an LRD input process. But in [56] $N \rightarrow \infty$ is taken, and it is shown that *the aggregate traffic, suitably normalized, is FBM*. Hence it is not only LRD but also strictly self-similar. Analysis of Ethernet traffic from various sources indeed shows extremely widespread presence of heavy-tailed on/off periods [56]. More detailed further explanations for such phenomena typically refer to application-level characteristics, and include: (i) an empirically observed hyperbolic tail behaviour for file sizes residing in file servers; (ii) a Pareto-like tail behaviour for measured CPU time used by a typical UNIX process; (iii) measurement studies of an ISDN office automation application, suggesting that human-computer interactions occur over a wide range of time scales and thus may require models based on distributions with infinite variance; (iv) in the case of more recent measurements in local area networks [19], the observations on the size of documents residing on present-day WWW servers also suggest an infinite variance.

In view of all these measurements, an important research issue of course is the study of the impact of LRD, self-similar traffic and heavy tails on system performance. In [23], Ethernet traffic measurements have been used to demonstrate the very considerable impact that LRD has on queuing behaviour. The central issue in our paper is the impact of long tails on system performance.

Remark 2.3 *We shall not restrict ourselves to the special class of heavy-tailed distributions. In fact we are interested in all non-negative variables X for which does not hold that $\mathbb{P}[X > x] = O(e^{-\eta x})$, $x \rightarrow \infty$, with η a positive constant; in particular, we are interested in the class \mathcal{L} of long-tailed distributions (see Appendix A), its subclass \mathcal{S} of subexponential distributions (also to be discussed in Appendix A), and the class \mathcal{R} of regularly varying distributions (Appendix B), which is a subclass of \mathcal{S} .*

3. A SINGLE FLUID SOURCE

The model of what we call here a “fluid source” is fairly more general than that of the “on/off sources” that will be considered in superposition in the next two sections. One reason for this is that the analysis of a fluid queue fed by a single source is easier, and deep results can be obtained in a more general setting, but another reason is that our general model covers the case of a source obtained as the superposition of such on/off sources. Hence the results of the current section find applications even in some situations when several sources coexist.

The next subsection contains the description of the model, and the characterization of three variables of primary interest for the analysis of the stationary regime: the buffer content, first at the beginning of an activity period of the source (it is then denoted W), then in continuous time (we denote it by Z); and the busy period P of the fluid queue. In many aspects the results are related to classical analogs for the GI/G/1 or M/G/1 queue.

In Subsection 3.2, we show how long-range dependence occurs in the on/off source, and give an equivalent for the integrated covariance without any assumption on the tails of the activity and silence period distributions of the source.

Subsection 3.3 collects (and marginally completes) the results in the literature about the conditions under which the distributions of the stationary variables W , Z and P exhibit subexponential or regularly varying tails. Comparison of these results with the former characterization of LRD clearly shows that LRD may result in a poor performance of the queue (and how it may propagate in a network); but also that LRD and long tails of the stationary distributions are quite distinct phenomena, which may occur independently.

The section is concluded with two remarks that provide key interpretations of the results presented in the paper.

3.1 Model - Stationary characteristics

Consider a fluid queue with an infinite buffer, and an output rate equal to 1. This queue is fed by a source that alternates between silence periods S_n , $n \geq 1$, during which it generates no input, and activity periods A_n , $n \geq 1$, during which it generates fluid according to the rate process $(r_n(t))_{t \geq 0}$, $n \geq 1$. A crucial assumption will be that:

$$\forall n \geq 1, \forall t \geq 0 : \quad 1 \leq r_n(t) \leq R \quad (\text{where } R \text{ is some given constant}). \quad (3.1)$$

In particular, the n th activity period results in a *net* input equal to:

$$B_n := \int_0^{A_n} [r_n(t) - 1] dt.$$

Now we assume that the sequences $(S_n)_{n \geq 1}$, and $(A_n, r_n(t), t \geq 0)_{n \geq 1}$, are respectively i.i.d., and that they are independent. Then the beginnings of activity and silence periods form a so-called alternating renewal process. This model is a specialization of the storage model introduced by Kella and Whitt [31]. If $r_1(t) \equiv r > 1$, we shall

mention it as an *on/off source* (the case $r = 1$ is trivial, since obviously the fluid queue is constantly empty after some time).

Remark 3.1 *A generalization of the on/off source is the fluid model in which the sources can be in several (not necessarily two) states, transitions between states occurring according to a Markov chain. Interesting references on this subject are the papers of Asmussen [4] and Jelenkovic and Lazar [28].*

From now on, except otherwise stated, we assume that *the first period is an activity period*. Then the sequence $T_n := \sum_{k=1}^n (A_k + S_k)$ ($T_0 = 0$) depicts the successive epochs when the source starts an activity period. The main parameters are:

$$\alpha := \mathbb{E}[A_1], \quad \beta := \mathbb{E}[B_1], \quad \lambda := 1/\mathbb{E}[S_1],$$

all of them assumed *finite and non-null*.

The content of the buffer at time $t \geq 0$ will be denoted by $Z(t)$, and we shall first concentrate on: $W_n := Z(T_n)$. Notice that $(W_n)_{n \in \mathbb{N}}$ satisfies the recurrence equation:

$$W_{n+1} = [W_n + B_n - S_n]^+, \quad (3.2)$$

which characterizes the successive waiting times of a GI/G/1 queue with service times B_n and interarrival times S_n , $n \geq 1$, of traffic intensity $\lambda\beta$. In consequence, the classical results about the GI/G/1 queue apply (see for example Feller [25]). For the notation B_1^* used below, we refer the reader to the end of Section 1.

Theorem 3.2 *If $\lambda\beta \geq 1$ then $W_n \rightarrow \infty$ in probability as $n \rightarrow \infty$. If $\lambda\beta < 1$, then $W_n \rightarrow W$ in distribution as $n \rightarrow \infty$, where the law of the random variable W may be characterized by:*

$$\mathbb{P}[W > x] = (1 - \theta) \sum_{n \in \mathbb{N}} \theta^n \mathbb{P}[H_1 + \dots + H_n > x]. \quad (3.3)$$

In the above equation, $(H_n)_{n \geq 1}$ is the i.i.d. sequence of “ascending ladder heights” associated with the increments $(B_n - S_n)_{n \geq 1}$, that is, if $\tau = \inf\{n \geq 1 / \sum_{k=1}^n (B_k - S_k) > 0\}$:

$$\theta = \mathbb{P}[\tau < \infty], \quad \mathbb{P}[H_1 > x] = \mathbb{P}\left[\sum_{k=1}^{\tau} (B_k - S_k) > x \mid \tau < \infty\right].$$

In particular, if the distribution of S_1 is exponential: $\theta = \lambda\beta$ and $H_1 \simeq B_1^$, so that:*

$$\mathbb{P}[W > x] = (1 - \lambda\beta) \sum_{n \in \mathbb{N}} (\lambda\beta)^n \mathbb{P}[B_1^* + \dots + B_n^* > x]. \quad (3.4)$$

Remark 3.3 *Since most of the subsequent results are based on the above characterization of W , it is worth emphasizing the importance of Assumption (3.1): if we allowed $r_n(t)$ to be smaller than 1 for some values of t , then we would lose the notion of net input produced during an activity period, and the subsequent representation (3.2).*

Remark 3.4 Equation (3.4) is the famous Pollaczek-Khintchine formula, which in the $M/G/1$ setting relates the stationary waiting time distribution (W here) to the integrated service time distribution (B_1^* here). Equation (3.3) provides a similar representation for the $GI/G/1$ queue, though B_1^* does not appear explicitly here. But the proof of Theorem 3.13 will show that in the “subexponential case”, H_1 has essentially the same tail behaviour as B_1^* , and W has the same tail behaviour as the right-hand side of Equation (3.4).

The asymptotic behaviour of $Z(t)$ has been investigated in [31], where the analysis relies on the arguments used for the virtual waiting time in the classical $GI/G/1$ queue (see Asmussen [3], Chapter VIII, 3). When dealing with the continuous time behaviour of the fluid queue, we shall often use auxiliary parameters that naturally appear in formulas: the *stationary probability of silence*

$$p := \frac{1}{1 + \lambda\alpha},$$

and the *traffic intensity*

$$\rho := \frac{\lambda(\alpha + \beta)}{1 + \lambda\alpha},$$

that is the long-run average amount of fluid sent per time unit. As might be expected: $\lambda\beta < 1 \Leftrightarrow \rho < 1$.

Theorem 3.5 [31]. *If $\lambda\beta < 1$ and A_1 and S_1 have non-lattice distributions, then $Z(t) \rightarrow Z$ in distribution as $t \rightarrow \infty$, where the random variable Z is related to W by:*

$$\mathbb{P}[Z > x] = p\lambda\beta\mathbb{P}[W + B_1^* > x] + (1 - p)\mathbb{P}[W + B_1(A_1^*) > x]. \quad (3.5)$$

Here B_1^* and $B_1(A_1^*)$ are independent of W , and the latter is defined by:

$$\mathbb{P}[B_1(A_1^*) > x] = \frac{1}{\mathbb{E}[A_1]} \mathbb{E} \left[\int_0^{A_1} \mathbb{1}_{\{B_1(t) > x\}} dt \right], \quad \text{with } B_1(t) = \int_0^t [r_1(u) - 1] du.$$

In particular, if $r_1(t) \equiv r > 1$ (on/off source), then: $B_1^* \simeq B_1(A_1^*) \simeq (r - 1)A_1^*$, and $\beta = (r - 1)\alpha$, so that:

$$\mathbb{P}[Z > x] = \rho\mathbb{P}[W + (r - 1)A_1^* > x].$$

Another variable of interest is the *busy period* of the fluid queue, namely the typical period during which the queue generates an output. Denote by P_n (resp. I_n) the n th busy period (resp. the n th idle period), assuming that they are all almost surely finite. Notice that the sequence $(P_n, I_n)_{n \geq 1}$ is i.i.d., and that during P_n (resp. during I_n), the queue generates an output at rate 1 (resp. no output), due to our assumption that the active source sends input at rate ≥ 1 (see (3.1)). Hence the traffic on the output line is as generated by an on/off source of constant input rate 1, except that P_n and I_n

are *not independent* in general. Independence may be obtained by assuming that the silence periods S_n are exponentially distributed, and then $I_1 \simeq S_1$. Moreover, it is then possible to characterize the distribution of P_1 . In the following theorem, we introduce the Laplace-Stieltjes transforms:

$$\gamma[\omega, \theta] := \mathbb{E}[e^{-\omega A_1 - \theta B_1}], \quad \pi[\omega] := \mathbb{E}[e^{-\omega P_1}], \quad \omega \geq 0, \quad \theta \geq 0.$$

Theorem 3.6 *If the distribution of S_1 is exponential, then:*

$$P_1 \simeq A_1 + B_1 + P_1 + \dots + P_K, \quad (3.6)$$

where $(P_n)_{n \geq 1}$ is independent of (A_1, B_1, K) , and given (A_1, B_1) , the conditional distribution of K is Poisson of parameter λB_1 . In terms of Laplace-Stieltjes transforms, for $\omega > 0$, $\pi[\omega]$ is the unique solution in $(-\infty, 1]$ of:

$$\pi[\omega] = \gamma[\omega, \omega + \lambda(1 - \pi[\omega])]. \quad (3.7)$$

Moreover, P_1 is a.s. finite if and only if $\rho \leq 1$, and then:

$$\mathbb{E}[P_1] = \frac{\rho}{\lambda(1 - \rho)} \quad (= \infty \text{ if } \rho = 1). \quad (3.8)$$

Remark 3.7 *Equation (3.7) has already been obtained by Cohen [15] in the special case when the fluid source is obtained as the superposition of several on/off sources (see Section 4).*

In Formula (3.8) we recognize the mean busy period of an M/G/1 queue of traffic intensity ρ and arrival rate λ (see [16], Section II.4.4).

Proof:

The arguments are similar to those used for the M/G/1 queue (cf. [16], Section II.4.4). Assume that the system is empty at time 0 and that it starts an active period A_1 . Given that $A_1 = x$ and $B_1 = y$, then at time x the buffer content is y . If no new activity period starts before time $x + y$, then $P_1 = x + y$. Otherwise, the source starts sending input again at some time $x + y_1$, with $y_1 < y$. Then we may assume that the processing of the residual input $y - y_1$ is interrupted, and the queue treats the new input until the first time $t > x + y_1$ such that $Z(t) = y - y_1$ and the source is silent again. Clearly the time we have to wait until this event occurs is distributed as a busy period P_1 . After that the processing of the residual input $y - y_1$ can be restarted, possible interruptions being handled as the first one. Obviously the successive interruptions form an i.i.d. sequence $(P_n)_{n \geq 1}$, and the number of interruptions is an independent Poisson variable K , of parameter λy . The busy period is hence equal to:

$$x + y_1 + P_1 + y_2 + \dots + P_K + (y - y_1 - \dots - y_K) = x + y + P_1 + \dots + P_K,$$

which is Equation (3.6). Moreover, integration of the above formula with respect to (x, y) according to the distribution of (A_1, B_1) yields the right-hand side of (3.7).

The problem of uniqueness may be solved as for the M/G/1 queue, see Cohen [16] Section II.4.4. Similarly, $P_1 < \infty$ a.s. if and only if $\pi[0] = 1$, which is equivalent to

$$-\lambda \frac{\partial \gamma}{\partial \theta}[0, 0] \leq 1, \text{ that is } \lambda \beta \leq 1.$$

Finally (3.8) is obtained either directly from (3.6) or via differentiation from (3.7). ♠

3.2 Long-range dependence

Here we restrict our attention to the case when $r_n(t) \equiv r > 1$, i.e. to the on/off source (in Section 4 our approach is extended to a superposition of on/off sources). The problem at hand is: when does such a source exhibit LRD, and how can we relate the asymptotic dependence to the primitives of the source, that is the activity and silence period distributions? The main result is Theorem 3.9, which is new to the best of our knowledge.

Denote by $I(t)$ the indicator function of {source silent at time t } (take it right-continuous), and by $(I^*(t))_{t \geq 0}$ the stationary version of this process. It is obtained by imposing an independent delay T^* on the alternating renewal process $(A_n, S_n, n \geq 1)$: with probability $p = 1/(1 + \lambda\alpha)$, the delay is a silence period S_1^* ; with probability $1 - p$, the delay is formed of the succession of an activity period A_1^* and an independent silence period S_0 (such that $(S_n)_{n \geq 0}$ is i.i.d.).

Remark 3.8 *This construction can be justified by introducing the Markov process $(I(t), F(t))$, where $F(t)$ is the residual time until the end of the current (activity or silence) period. By Theorem 3.2 of Asmussen [3], VI.3, its stationary distribution is characterized by: $\mathbb{P}[I^*(0) = 1] = p$, and:*

$$\mathbb{P}[F^*(0) \leq t | I^*(0) = 1] = \mathbb{P}[S_1^* \leq t], \quad \mathbb{P}[F^*(0) \leq t | I^*(0) = 0] = \mathbb{P}[A_1^* \leq t],$$

which is equivalent to our formulation.

The stationary input rate process is simply: $r^*(t) = r \mathbb{1}_{\{I^*(t) = 0\}}$. Long-range dependence is analyzed in the following theorem.

Theorem 3.9 *Assume that the distribution of A_1 or S_1 is non-lattice. If $\mathbb{E}[A_1^2] < \infty$ and $\mathbb{E}[S_1^2] < \infty$, then the stationary process $(r^*(t))_{t \geq 0}$ is short-range dependent, and:*

$$\int_0^\infty \text{Cov}(r^*(0), r^*(u)) du = \frac{r^2 p^2 (1-p)^2}{2} \mathbb{E}[A_1 + S_1] \mathbb{E} \left[\left(\frac{A_1}{\mathbb{E}[A_1]} - \frac{S_1}{\mathbb{E}[S_1]} \right)^2 \right].$$

If $\mathbb{E}[A_1^2] = \infty$ or $\mathbb{E}[S_1^2] = \infty$, then the process is long-range dependent, and:

$$\int_0^t \text{Cov}(r^*(0), r^*(u)) du \stackrel{t \rightarrow \infty}{\sim} r^2 p(1-p) \left[p \int_0^t \mathbb{P}[A_1^* > u] du + (1-p) \int_0^t \mathbb{P}[S_1^* > u] du \right]. \quad (3.9)$$

The proof relies on the following lemmas, which are proven in Appendix C.

Lemma 3.10

$$\begin{aligned} & \frac{1}{r^2 p(1-p)} \int_0^t \text{Cov}(r^*(0), r^*(u)) du \\ &= \int_0^t (\mathbb{P}[I^*(u) = 0 | I^*(0) = 0] - \mathbb{P}[I^*(u) = 0 | I^*(0) = 1]) du \end{aligned} \quad (3.10)$$

$$= \int_0^t \mathbb{P}[A_1^* > u] du + \int_0^t (\mathbb{P}[S_1^* > u] - \mathbb{P}[A_1^* + S_0 > u]) \mathbb{P}[I(t-u) = 0] du. \quad (3.11)$$

Remark 3.11 Expression (3.10) may be rewritten as:

$$\int_0^t (\mathbb{P}[I^*(u) = 1 | I^*(0) = 1] - \mathbb{P}[I^*(u) = 1 | I^*(0) = 0]) du,$$

which shows that the roles of A_1 and S_1 are strictly equivalent in terms of long-range dependence (a phenomenon already noticed by Brichet et al. [10] and Heath et al. [27]). This is the main difference with the problem of the long-tailed stationary buffer content distribution, which depends only on the tail of A_1 (see Subsection 3.3).

Lemma 3.12 Assume that the distribution of A_1 or S_1 is non-lattice. Then for any measurable, non-null function $f: \mathbb{R}^+ \rightarrow \mathbb{R}^+$, such that $f(t) \rightarrow 0$ as $t \rightarrow \infty$:

$$\int_0^t f(u) \mathbb{P}[I(t-u) = 0] du \stackrel{t \rightarrow \infty}{\sim} (1-p) \int_0^t f(u) du.$$

Proof of Theorem 3.9:

For any non-negative random variable X such that $0 < \mathbb{E}[X] < \infty$, we have: $\mathbb{E}[X^*] = \mathbb{E}[X^2]/(2\mathbb{E}[X])$. The first part of the theorem follows easily in view of Lemmas 3.10 and 3.12.

For the second part, write: $\mathbb{P}[A_1^* + S_0 > u] = \mathbb{P}[A_1^* > u] + \mathbb{P}[A_1^* + S_0 > u \geq A_1^*]$, and notice that:

$$\int_0^\infty \mathbb{P}[A_1^* + S_0 > u \geq A_1^*] dt = \mathbb{E}[S_0] < \infty.$$

Now apply Lemmas 3.10 and 3.12 once again, thus obtaining Formula (3.9), which proves long-range dependence. \spadesuit

Formula (3.9) suggests the stronger result:

$$\text{Cov}(r^*(0), r^*(u)) \stackrel{u \rightarrow \infty}{\sim} r^2 p(1-p)(p\mathbb{P}[A_1^* > u] + (1-p)\mathbb{P}[S_1^* > u]).$$

It has actually been proved by Heath *et al.* ([27], Theorem 4.3) in the special case when $\mathbb{P}[A_1 > t]$ is of the form $l(t)/t^{1+\epsilon}$, where $l(t)$ is a slowly varying function and $\epsilon \in (0, 1)$, and $\mathbb{P}[S_1 > t] = o(\mathbb{P}[A_1 > t])$. The classical results of Karamata on functions of regular variation (see Lemma 7.7 or [7]) then yield: $\text{Cov}(r^*(0), r^*(u)) \sim r^2 p^2(1-p)l(u)/(\alpha \epsilon u^\epsilon)$.

Repeated applications of Karamata's Theorem also yield explicit equivalents of $\int_0^t \text{Cov}(r^*(0), r^*(u)) du$ and $\int_0^t \int_0^u \text{Cov}(r^*(0), r^*(v)) dv du (= \frac{1}{2} \text{Var}(T(t)))$, as pointed out in Subsection 2.1), which have already been obtained by Brichet *et al.* ([10], Appendix 5) and Willinger *et al.* ([56], Appendix).

3.3 Buffer contents with subexponential tails

In order to get precise results on long tails in the context of queues, it seems appropriate to work inside the class \mathcal{S} of distributions with subexponential tails, or even inside the subclass \mathcal{R} of distributions with regularly varying tails (see Appendices A and B). We are especially interested in expressions for the tails of the stationary variables (buffer content, busy period) in terms of those of the primitive variables (activity period, net input).

By applying a beautiful result on the stationary waiting time of the GI/G/1 queue, mainly due to Cohen [14] and Pakes [44], we obtain the following theorem.

Theorem 3.13 *Assume that $\lambda\beta < 1$. Still denoting by W the stationary buffer content at the beginning of activity periods, we have $\mathbb{P}[W > x] \in \mathcal{S}$ if and only if $\mathbb{P}[B_1^* > x] \in \mathcal{S}$, and if either is the case then:*

$$\mathbb{P}[W > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda\beta}{1 - \lambda\beta} \mathbb{P}[B_1^* > x]. \quad (3.12)$$

Conversely, if the distribution of S_1 is exponential and relation (3.12) holds, then both $\mathbb{P}[W > x]$ and $\mathbb{P}[B_1^ > x]$ belong to \mathcal{S} .*

Proof:

This result is a pure copy of Theorem 1 in [44] (which partly relies on the results of [14], later improved by Smith [53]). We only sketch the proof of the “easy part”, that is $\mathbb{P}[B_1^* > x] \in \mathcal{S}$ implies Formula (3.12) (hence $\mathbb{P}[W > x] \in \mathcal{S}$, by Corollary 7.6). The starting point is a formula observed by Cohen [14], and related to the Wiener-Hopf factorization (see also XII.3, Formula (3.7a) in Feller [25]):

$$\frac{\theta}{1 - \theta} \mathbb{P}[H_1 > x + D_1^*] = \frac{\lambda\beta}{1 - \lambda\beta} \mathbb{P}[B_1^* > x + S_1],$$

where D_1^* is independent of H_1 , and D_1 is the first ascending ladder height associated with the increments $(S_n - B_n)_{n \geq 1}$ (see Theorem 3.2). By Lemma 7.2, $\mathbb{P}[B_1^* > x] \in \mathcal{S}$ implies that $\mathbb{P}[B_1^* > x] \in \mathcal{L}$, which from the above formula easily yields:

$$\frac{\theta}{1-\theta} \mathbb{P}[H_1 > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda\beta}{1-\lambda\beta} \mathbb{P}[B_1^* > x].$$

In particular, by Corollary 7.6: $\mathbb{P}[H_1 > x] \in \mathcal{S}$. Now consider Formula (3.3). In view of Lemma 7.4, applying the criterion of dominated convergence, we obtain:

$$\lim_{x \rightarrow \infty} \frac{\mathbb{P}[W > x]}{\mathbb{P}[H_1 > x]} = (1-\theta) \sum_{n \in \mathbb{N}} \theta^n n = \frac{\theta}{1-\theta}.$$

The proof of the “easy part” is complete. ♠

Remark 3.14 *Without any assumption on the distributions, it is not difficult to check that: $\liminf_{x \rightarrow \infty} \mathbb{P}[H_1 + \dots + H_n > x] / \mathbb{P}[H_1 > x] \geq n$ for all n . Consequently, by applying Fatou’s lemma, we obtain: $\liminf_{x \rightarrow \infty} \mathbb{P}[W > x] / \mathbb{P}[H_1 > x] \geq \theta / (1 - \theta)$. Hence the equivalent in (3.12) is, at least when S_1 is exponential (hence $H_1 \simeq B_1^*$), a natural bound for the tail of W , reached only in the subexponential case.*

Thanks to the results of Kella and Whitt ([31], also exposed in Theorem 3.5), it is not difficult to obtain similar results for the stationary buffer content.

Theorem 3.15 *Assume that $\lambda\beta < 1$, and A_1 and S_1 have non-lattice distributions. If $\mathbb{P}[B_1^* > x] \in \mathcal{S}$ and $\mathbb{P}[B_1(A_1^*) > x] \in \mathcal{S}$, then $\mathbb{P}[Z > x] \in \mathcal{S}$ and:*

$$\mathbb{P}[Z > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda\beta}{1-\lambda\beta} \mathbb{P}[B_1^* > x] + (1-p) \mathbb{P}[B_1(A_1^*) > x]. \quad (3.13)$$

In particular, if $r_1(t) \equiv r > 1$ and $\mathbb{P}[A_1^ > x] \in \mathcal{S}$:*

$$\mathbb{P}[Z > x] \stackrel{x \rightarrow \infty}{\sim} p \frac{\rho}{1-\rho} \mathbb{P}[(r-1)A_1^* > x]. \quad (3.14)$$

Proof:

Formulas (3.5) and (3.12) yield (3.13), if we prove that: $\mathbb{P}[W + B_1^* > x] \sim \mathbb{P}[W > x] + \mathbb{P}[B_1^* > x]$, and: $\mathbb{P}[W + B_1(A_1^*) > x] \sim \mathbb{P}[W > x] + \mathbb{P}[B_1(A_1^*) > x]$. The first relation is a direct consequence of Theorems 3.13 and 7.5; the second one follows similarly, but we must check that: $\sup_x \mathbb{P}[B_1^* > x] / \mathbb{P}[B_1(A_1^*) > x] < \infty$. By definition:

$$\mathbb{P}[B_1(A_1^*) > x] = \frac{1}{\mathbb{E}[A_1]} \mathbb{E} \left[\int_0^{A_1} \mathbb{1}_{\{B_1(t) > x\}} dt \right] = \frac{1}{\mathbb{E}[A_1]} \mathbb{E} \left[\int_0^{A_1} \mathbb{1}_{\{B_1(A_1 - u) > x\}} du \right].$$

In view of Condition (3.1), we have: $B_1(A_1 - u) \geq B_1 - (R-1)u$, and $A_1 \geq B_1/(R-1)$. So we get:

$$\begin{aligned} \mathbb{P}[B_1(A_1^*) > x] &\geq \frac{1}{\mathbb{E}[A_1]} \mathbb{E} \left[\int_0^{B_1/(R-1)} \mathbb{1}_{\{B_1 - (R-1)u > x\}} du \right] \\ &= \frac{1}{(R-1)\mathbb{E}[A_1]} \mathbb{E} \left[\int_0^{B_1} \mathbb{1}_{\{v > x\}} dv \right] \\ &= \frac{\mathbb{E}[B_1]}{(R-1)\mathbb{E}[A_1]} \mathbb{P}[B_1^* > x], \end{aligned}$$

which completes the proof of Formula (3.13). That $\mathbb{P}[Z > x] \in \mathcal{S}$ now follows from Theorem 7.5 and Corollary 7.6. \spadesuit

Remark 3.16 If $\mathbb{P}[A_1 > t] \sim l(t)/t^{1+\epsilon}$ (where $l(t)$ is a slowly varying function, and $\epsilon > 0$), then $\mathbb{P}[A_1^* > t] \sim l(t)/(\alpha t^\epsilon)$ (see Lemma 7.8). So in case $r_1(t) \equiv r > 1$, we get:

$$\mathbb{P}[Z > x] \stackrel{x \rightarrow \infty}{\sim} p \frac{\rho}{1-\rho} \frac{l(x/(r-1))}{\alpha \epsilon [x/(r-1)]^\epsilon} \stackrel{x \rightarrow \infty}{\sim} p \frac{\rho}{1-\rho} \frac{(r-1)^\epsilon l(x)}{\alpha \epsilon x^\epsilon}.$$

This result has already been proved independently by Choudhury and Whitt ([12], Theorem 2) and Heath et al. ([27], last formula). Actually, the general, subexponential case with a constant input rate has later been treated by Jelenkovic and Lazar ([29], Theorem 9).

Remark 3.17 In the above proof, we checked that $\mathbb{P}[B_1^* > x] = O(\mathbb{P}[B_1(A_1^*) > x])$ in full generality. The converse is not necessarily true, which may lead to very different tail behaviours for W and Z . For example, take:

$$r_1(t) = 1 + \frac{1}{1+t} \quad \text{and} \quad \mathbb{P}[A_1 > t] = \frac{1}{(1+t)[1 + \ln(1+t)]^{1+\delta}}, \quad t \geq 0,$$

with $\delta > 0$. Then standard calculations yield $\alpha = 1/\delta$ and:

$$\begin{cases} \mathbb{P}[B_1 > x] = \mathbb{P}[A_1 > e^x - 1] = \frac{e^{-x}}{(1+x)^{1+\delta}}, & x \geq 0 \\ \mathbb{P}[B_1(A_1^*) > x] = \mathbb{P}[A_1^* > e^x - 1] = \frac{1}{(1+x)^\delta}, & x \geq 0. \end{cases}$$

The distribution of B_1 thus belongs to the class $\mathcal{S}(1)$ (see Klüppelberg, [33] Theorem 2.1 and [32] Theorem 3.2). If moreover we assume that S_1 is an exponential variable and $\lambda\alpha < 1$, then Theorem 2 of Pakes [44] shows that:

$$\mathbb{P}[W > x] \stackrel{x \rightarrow \infty}{\sim} M \mathbb{P}[B_1 > x] \quad (\text{with } M \text{ a positive constant}).$$

On the other hand: $\mathbb{E}[B_1]\mathbb{P}[B_1^* > x] \sim \mathbb{P}[B_1 > x]$ ([44], Lemma 5), and we deduce from Formula (3.5) and Theorem 7.5 that:

$$\mathbb{P}[Z > x] \stackrel{x \rightarrow \infty}{\sim} (1-p)\mathbb{P}[B_1(A_1^*) > x].$$

Thus W has an exponential tail, whereas the tail of Z is polynomial.

Let us now assume that S_1 has an exponential distribution, and consider the busy period P_1 . If in addition: $r_1(t) \equiv r > 1$, then, using the notations of Theorem 3.6, we have: $\gamma[\omega, \theta] = \alpha[\omega + (r-1)\theta]$ (where $\alpha[\cdot]$ is the Laplace-Stieltjes transform of A_1). In consequence, $\pi[\omega]$ is the solution of:

$$\pi[\omega] = \alpha[\omega + (r-1)(\omega + \lambda(1 - \pi[\omega]))] = \alpha[r\omega + (r-1)\lambda(1 - \pi[\omega])].$$

This equation characterizes the busy period of an M/G/1 queue with Poisson rate $\lambda(r-1)/r$ and typical service time rA_1 (see for example [16] Section II.4.4). Hence the following result is a straightforward consequence of a theorem proved by De Meyer and Teugels [40].

Theorem 3.18 *Assume that $\lambda\beta (= \lambda(r-1)\alpha) < 1$. Then $\mathbb{P}[P_1 > t]$ is regularly varying if and only if $\mathbb{P}[A_1 > t]$ is regularly varying, and if either is the case:*

$$(1 - \lambda\beta)\mathbb{P}[P_1 > t] \stackrel{t \rightarrow \infty}{\sim} \mathbb{P}[rA_1 > (1 - \lambda\beta)t].$$

Proof:

The proof is based on the characterization of regularly varying tails in terms of the series expansion of the associated Laplace-Stieltjes transforms in the neighbourhood of 0 (see Lemma 7.8). The case of $\mathcal{R}_{-\nu}$ with integral ν is especially difficult to address.



Notice that the generalization of Theorem 3.18 to the subexponential case is still an open problem.

The following two remarks express views that are to a certain extent opposing, but nevertheless both true.

Remark 3.19 *The results for buffer content and busy period distribution in Theorems 3.13, 3.15 and 3.18 (and their extensions to multiple sources in the next two sections) provide vital insight into the effect of long-tailed activity period distributions on the performance of modern communication networks. It appears that a long-tailed activity period distribution gives rise to an even worse tail behaviour of the buffer content (e.g., if $\mathbb{E}[A_1] < \infty$, $\mathbb{E}[A_1^2] = \infty$ then the mean buffer content is not even finite). And the tail behaviour of the busy period is of the same nature as that of the activity period. Since*

the output process of the buffer (that may feed into another buffer) is an on/off process with on-periods the busy periods, this suggests that heavy tails, and hence long-range dependence, propagate through a network. This matches conclusions drawn in Section IV.C of [23].

Remark 3.20 *The direct applicability of results like Theorems 3.13, 3.15 and 3.18 is limited. In practice buffers are not infinite. Moreover, traffic control will be exercised: sources may not be admitted, and inflow rates may be reduced. A warning is also in order when one would use, e.g., (3.12) as an approximation. It has been observed by Abate et al. [1] that such approximations typically yield much less accurate results than approximations based on the dominating exponential term in a model with exponential tails. Therefore [1], and also [18], propose particular classes of regularly varying service time distributions for which the LST is so manageable that one can obtain a series representation of the corresponding M/G/1 waiting time distribution.*

4. SUPERPOSITION OF ON/OFF SOURCES

In this section, we study the behaviour of a fluid queue with an infinite buffer and outflow rate equal to one, fed by $N > 1$ independent on/off sources. For $1 \leq i \leq N$, we assume that source i , when active, has an input rate $r_i \geq 1$. For all the other values (parameters, variables or processes) related to source i , $1 \leq i \leq N$, we also add an index i to the notations introduced in Section 3.

First we derive the conditions for the long-range dependence of the cumulate input traffic. They may be easily deduced from Theorem 3.9. The stationary regime is obtained by superposing independent versions of the stationary on/off sources (see Subsection 3.2 for the construction of the stationary sources and the specific notations introduced there). Hence the global traffic is characterized by the input rate process: $r^*(t) := \sum_{i=1}^N r_i^*(t)$.

Theorem 4.1 *Assume that for all i , the distribution of A_{i1} or S_{i1} is non-lattice. Then the stationary process $(r^*(t))_{t \geq 0}$ is long-range dependent if and only if $\mathbb{E}[A_{i1}^2] = \infty$ or $\mathbb{E}[S_{i1}^2] = \infty$ for some i , and then:*

$$\int_0^t \text{Cov}(r^*(0), r^*(u)) du \stackrel{t \rightarrow \infty}{\sim} \sum_{i=1}^N r_i^2 p_i (1 - p_i) \left[p_i \int_0^t \mathbb{P}[A_{i1}^* > u] du + (1 - p_i) \int_0^t \mathbb{P}[S_{i1}^* > u] du \right].$$

Proof:

Since all the sources are independent: $\text{Cov}(r^*(0), r^*(t)) = \sum_{i=1}^N \text{Cov}(r_i^*(0), r_i^*(t))$ for all $t \geq 0$. Hence the result is a direct consequence of Theorem 3.9. \spadesuit

From now on, we assume that for $1 \leq i \leq N$, the silence periods S_{in} of source i are exponential variables, of parameter λ_i . Then obviously the periods S_n of total silence

are themselves i.i.d. exponential variables, of parameter $\lambda := \sum_{i=1}^N \lambda_i$. Proceeding, the periods A_n of “global activity” (when at least one on/off source is active), and the associated input processes $r_n(t)$, $t \geq 0$, are i.i.d. too. Finally, we have:

$$\forall n \geq 1, \forall t \geq 0: \quad 1 \leq r_n(t) \leq R := \sum_{i=1}^N r_i.$$

Hence the superposition of these on/off sources gives rise to a *global, fluid source* in the sense of the definition given in Section 3. The notations introduced there will here refer to this global source. In particular, B_n will denote the total, net input produced during the global activity period A_n .

The parameters $\alpha = \mathbb{E}[A_1]$ and $\beta = \mathbb{E}[B_1]$ of the global source can be easily calculated via the auxiliary parameters p (stationary probability of silence) and ρ (traffic intensity). Since p is the stationary probability that all the sources are silent, then:

$$p = \prod_{i=1}^N p_i = \prod_{i=1}^N \frac{1}{1 + \lambda_i \alpha_i}.$$

But p is also equal to $1/(1 + \lambda\alpha)$ (with λ as defined above), which yields the value of α . Moreover, the global traffic intensity ρ is obviously the sum of the traffic intensities generated by the various sources, that is:

$$\rho = \sum_{i=1}^N r_i(1 - p_i) = \sum_{i=1}^N r_i \frac{\lambda_i \alpha_i}{1 + \lambda_i \alpha_i}.$$

But we also have: $\rho = \lambda(\alpha + \beta)/(1 + \lambda\alpha)$, which now yields the value of β .

In view of Theorems 3.2 and 3.5, the stability condition of the fluid queue is $\rho < 1$. Our goal will now be to relate the stationary behaviour of the buffer to the characteristics of the on/off sources, and especially to study the effect of one or several sources with long-tailed active periods. A natural approach consists in trying to apply the results exposed in Section 3, but first we must derive the characteristics of the global source (in particular the laws of A_1 and B_1). They are studied in Subsection 4.1. As we shall see, this approach may work for the analysis of W , but concerning the laws of Z and P_1 , more direct approaches may be preferable. Subsection 4.2 contains a specific characterization of the busy period distribution. Then a detailed discussion of the tail behaviour of the buffer content is presented in Subsection 4.3, in case only *one source* does not have exponentially tailed activity periods. It is currently impossible to make a complete analysis of the model in the general case; partial results and bounds are mentioned in Subsection 4.4 for the case of several sources with long-tailed on-period distributions.

4.1 Characteristics of the global, fluid source

The description of the global source is originally due to Cohen [15], and has been recently completed by the same author [17]. Here we adopt the convention that *all the sources are initially silent*.

If $I_i(t)$ denotes the indicator function of {source i silent at time t }, then $I(t) := \prod_{i=1}^N I_i(t)$ is the analogous indicator for the global source. Now set:

$$h_i(t) := r_i \int_0^t [1 - I_i(u)] du.$$

It is the total amount of fluid generated by source i between 0 and t ; hence the global source generates: $h(t) := \sum_{i=1}^N h_i(t)$.

Cohen's formulas rely on the following lemma. It is valid for any fluid source with exponential off-periods.

Lemma 4.2 For $Re(\omega) > 0$, $Re(\theta) \geq 0$:

$$\int_0^\infty e^{-\omega t} \mathbb{E}[e^{-\theta h(t)}; I(t) = 1] dt = \frac{1}{\omega + \lambda(1 - \mathbb{E}[e^{-\omega A_1 - \theta(A_1 + B_1)}])}. \quad (4.1)$$

A similar formula applies to each source i , $1 \leq i \leq N$. Since $B_{i1} = (r_i - 1)A_{i1}$, we get:

$$\int_0^\infty e^{-\omega t} \mathbb{E}[e^{-\theta h_i(t)}; I_i(t) = 1] dt = \frac{1}{\omega + \lambda_i(1 - \alpha_i[\omega + r_i\theta])}, \quad (4.2)$$

where $\alpha_i[\omega] := \mathbb{E}[e^{-\omega A_{i1}}]$.

Proof:

If t belongs to the n th silence period, then: $e^{-\theta h(t)} \mathbb{1}_{\{I(t) = 1\}} = e^{-\theta(A_1 + B_1 + \dots + A_{n-1} + B_{n-1})}$. The result follows easily from the decomposition over all the silence periods. ♠

Now if we remark that for all $t \geq 0$:

$$\mathbb{E}[e^{-\theta h(t)}; I(t) = 1] = \prod_{i=1}^N \mathbb{E}[e^{-\theta h_i(t)}; I_i(t) = 1],$$

then we can relate the law of (A_1, B_1) to those of A_{i1} , $1 \leq i \leq N$, thanks to an inversion formula for Laplace-Stieltjes transforms.

Proposition 4.3 For $Re(\omega) > 0$, $Re(\theta) \geq 0$, and any choice of $x > 0$:

$$\begin{aligned} & \frac{1}{\omega + \lambda(1 - \mathbb{E}[e^{-\omega A_1 - \theta(A_1 + B_1)}])} \\ &= \int_0^\infty e^{-\omega t} \left\{ \prod_{j=1}^N \frac{1}{2\pi i} \int_{x-i\infty}^{x+i\infty} \frac{e^{yt} dy}{y + \lambda_j(1 - \alpha_j[y + r_j\theta])} \right\} dt. \end{aligned} \quad (4.3)$$

A key issue consists in extending the validity of (4.3) to a domain $\operatorname{Re}(\theta) \geq 0$, $\operatorname{Re}(\omega) > -\delta - \operatorname{Re}(\theta)$, for some $\delta > 0$, so that we can take $\omega = -\theta$ and thus obtain the Laplace-Stieltjes transform of B_1 . Actually, this would directly yield the Laplace-Stieltjes transform of W , since the Pollaczek-Khintchine Formula (3.4) is equivalent to:

$$\mathbb{E}[e^{-\theta W}] = \frac{1 - \lambda\beta}{1 - \lambda\beta\mathbb{E}[e^{-\theta B_1^*}]} = \frac{-\theta(1 - \lambda\beta)}{-\theta + \lambda(1 - \mathbb{E}[e^{-\theta B_1}])}.$$

Such an extension is justified by Cohen in the case when all the Laplace-Stieltjes transforms $\alpha_i[\omega]$, $1 \leq i \leq N$, have negative abscissae of convergence [17]; but this means that all the active periods A_{i1} , $1 \leq i \leq N$, have exponential moments, which we will precisely not assume in the sequel. (We will later see however that the extension can still be justified when all the sources *but one* have exponentially tailed active periods.) Also notice the absence of a formula for $\mathbb{E}[e^{-\theta B_1(A_1^*)}]$, which prevents us from investigating the distribution of Z via Theorem 3.5.

4.2 Characterization of the busy period

The characterization obtained for the busy period in Theorem 3.6 is still valid, but is not adapted to the superposition of on/off sources. A specific system of equations, explicitly involving the Laplace-Stieltjes transforms $\alpha_i[\omega]$, $1 \leq i \leq N$, has been obtained by Kaspi and Rubinovitch [30]. This subsection is devoted to the presentation of their results.

Again we shall denote by $(P_n)_{n \geq 1}$ the i.i.d. sequence of the successive busy periods of the fluid queue; for $1 \leq i \leq N$, we will denote by $(P_{in})_{n \geq 1}$ an i.i.d. sequence of busy periods starting with an activity period of source i . We introduce the associated Laplace-Stieltjes transforms:

$$\pi[\omega] := \mathbb{E}[e^{-\omega P_1}], \quad \pi_i[\omega] := \mathbb{E}[e^{-\omega P_{i1}}], \quad 1 \leq i \leq N, \quad \omega \geq 0.$$

Obviously, since silence periods are exponentially distributed, we have:

$$\pi[\omega] = \sum_{i=1}^N \frac{\lambda_i}{\lambda} \pi_i[\omega].$$

In the following theorem, a restrictive assumption is made on the input rates of the on/off sources. At the end of the subsection, we discuss how this assumption might be removed.

Theorem 4.4 *If $r_i = 1$, $1 \leq i \leq N$, then for all i :*

$$P_{i1} \simeq A_{i1} + \sum_{j \neq i} (P_{j1} + \dots + P_{jK_{ij}}), \quad (4.4)$$

where sequences $(P_{jn})_{n \geq 1}$, $j \neq i$, are independent of each other and of $(A_{i1}, K_{ij}, j \neq i)$, and given A_{i1} , the conditional distribution of $(K_{ij}, j \neq i)$ is that of independent Poisson variables of parameters $\lambda_j A_{i1}$, $j \neq i$.

In terms of Laplace-Stieltjes transforms, for $\omega > 0$, the vector $(\pi_i[\omega])_{1 \leq i \leq N}$ is the unique solution in $(-\infty, 1]^N$ of the system of equations:

$$\pi_i[\omega] = \alpha_i[\omega + \sum_{j \neq i} \lambda_j(1 - \pi_j[\omega])], \quad 1 \leq i \leq N. \quad (4.5)$$

Furthermore, P_1 is a.s. finite if and only if $\rho \leq 1$, and then:

$$\mathbb{E}[P_{i1}] = \frac{\rho_i}{\lambda_i(1 - \rho)} \quad (= \infty \text{ if } \rho = 1), \quad 1 \leq i \leq N, \quad \mathbb{E}[P_1] = \frac{\rho}{\lambda(1 - \rho)}. \quad (4.6)$$

Remark 4.5 The traffic intensities ρ_i of the various sources, and the global traffic intensity $\rho = \sum_i \rho_i$, have been calculated at the beginning of this section.

Proof:

The argument bears some analogy with that used in the proof of Theorem 3.6, but it is more subtle. Assume that at time 0 the queue is empty, and an active period $A_{i1} = x$ of source i is initiated, whereas all the other sources are silent. If these sources do not start an activity period before time x , then $P_{i1} = x$ (remember that $r_i = 1$). Otherwise, a source $j \neq i$ starts sending input at some time $x_1 < x$. Then *imagine* that we stop the activity of source i at this time, and set the remaining input $x - x_1$ temporarily aside: source i thus enters a silence period, and later behaves according to its usual alternating structure. Hence everything works as if at time x_1 a busy period of type P_{j1} were initiated, and at the end of this busy period all the sources are silent. At this time, imagine that source i immediately starts a (residual) activity period of length $x - x_1$, and that possible interruptions are handled as the first one. Kaspi and Rubinovitch [30] argue that this procedure will not affect the distribution of the total busy period P_{i1} . Since the successive interruptions by source j , $j \neq i$, form an i.i.d. sequence $(P_{jn})_{n \geq 1}$, and the number of interruptions is clearly a Poisson variable K_{ij} of parameter $\lambda_j x$, then: $P_{i1} = x + \sum_{j \neq i} (P_{j1} + \dots + P_{jK_{ij}})$; this is Equation (4.4). The right-hand side of (4.5) is obtained by integrating this formula with respect to x according to the distribution of A_{i1} .

For the question of uniqueness, as well as the finiteness of P_1 and the values of $\mathbb{E}[P_{i1}]$, $1 \leq i \leq N$, see the (quite technical) proof in [30]. In [30], a complicated expression is given for $\mathbb{E}[P_{i1}]$, where the numerator actually amounts to $\prod_{j \neq i} (1 + \rho_j)$ (following the authors' notations), which yields (4.6).



It might be asked whether the general case $r_i \geq 1$, $1 \leq i \leq N$, could be treated with a similar argument. Then it is necessary to take into account two kinds of interruptions: those occurring during the first activity period $A_{i1} = x$ (only due to sources $j \neq i$), and those occurring during the processing of the net input $B_{i1} = (r_i - 1)x$ (for which all the sources may be responsible). Actually it seems that the problem can be solved by combining the arguments of Theorems 3.6 and 4.4. A paper is in preparation on the subject; the results will be used to study the tail behaviour of P_1 in case at least one source has a regularly varying activity period distribution.

4.3 Heavy-tailed buffer content with $N - 1$ exponential sources

In this subsection, we show the dominating influence of one source, with regularly varying activity periods, on the stationary buffer content, when all the other sources have exponentially distributed activity periods. This is the only case in which a quite complete analysis of tail behaviours has been achieved. Three different works were independently devoted to this issue, each of them based on a specific approach (Boxma [9], Jelenkovic and Lazar [29], Rolski, Schlegel and Schmidt [50]). For the sake of clarity, the following theorem has not been formulated under the weakest assumptions; refinements in [9, 29, 50] are exposed below, along with a sketch of their methods of proof. We emphasize the singularity of the result on $\mathbb{P}[Z > x]$, which does not require that $r_1 \geq 1$, nor even that sources have exponential silence periods.

Theorem 4.6 *Assume that for $2 \leq i \leq N$, the activity periods A_{in} of source i are exponentially distributed, whereas $\mathbb{P}[A_{11} > x]$ is regularly varying. Suppose that $\rho < 1$, and set: $c := 1 - \sum_{i=2}^N \rho_i$. Then $\mathbb{P}[B_1 > x]$, $\mathbb{P}[W > x]$ and $\mathbb{P}[Z > x]$ are regularly varying too, and we have:*

$$\mathbb{P}[B_1 > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda_1 p_1}{\lambda p} \mathbb{P}[(r_1 - c)A_{11} > x], \quad (4.7)$$

$$\mathbb{P}[W > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda_1 (r_1 - c) \alpha_1}{c - \lambda_1 (r_1 - c) \alpha_1} \mathbb{P}[(r_1 - c)A_{11}^* > x], \quad (4.8)$$

$$\mathbb{P}[Z > x] \stackrel{x \rightarrow \infty}{\sim} p_1 \frac{\rho_1}{c - \rho_1} \mathbb{P}[(r_1 - c)A_{11}^* > x]. \quad (4.9)$$

The latter formula is valid for $r_1 > c$ and general distributions of the variables S_{i1} , $1 \leq i \leq N$.

Corollary 4.7 *Under the conditions of Theorem 4.6, if $\mathbb{P}[A_{11} > x] \sim l(x)/x^{1+\delta}$ as $x \rightarrow \infty$ (where $l(x)$ is a slowly varying function, and $\delta > 0$), then:*

$$\mathbb{P}[B_1 > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda_1 p_1}{\lambda p} \frac{(r_1 - c)^{1+\delta} l(x)}{x^{1+\delta}}, \quad (4.10)$$

$$\mathbb{P}[W > x] \stackrel{x \rightarrow \infty}{\sim} \frac{\lambda_1}{c - \lambda_1 (r_1 - c) \alpha_1} \frac{(r_1 - c)^{1+\delta} l(x)}{\delta x^\delta}, \quad (4.11)$$

$$\mathbb{P}[Z > x] \stackrel{x \rightarrow \infty}{\sim} p_1 \frac{\rho_1}{c - \rho_1} \frac{(r_1 - c)^\delta l(x)}{\alpha_1 \delta x^\delta}. \quad (4.12)$$

Proof of Corollary 4.7:

The corollary is a direct consequence of Theorem 4.6, in view of Lemma 7.7 (Appendix B). ♠

For a superposition of sources, Theorem 4.6 is the only complete result available about the tail behaviour of the buffer content. That is why we are going to spend

some time commenting it. Before sketching the approaches of the different authors, it is worth giving an interpretation of the formulas. First notice that $\sum_{i=2}^N \rho_i$ is the traffic intensity generated by the exponential sources, that is the average portion of capacity that the fluid queue devotes to these sources. Hence, c is the average capacity available for source 1.

Now imagine that the capacity of the queue is c instead of 1, and that it is only fed by source 1. The net input produced during the activity period A_{11} is now:

$$B_{11} = (r_1 - c)A_{11},$$

and consequently we get: $\beta_1 = (r_1 - c)\alpha_1$, and: $B_{11}^* \simeq B_{11}(A_{11}^*) \simeq (r_1 - c)A_{11}^*$. Moreover, by a time-scaling argument, it is not difficult to see that Formulas (3.12) (from Theorem 3.13) and (3.14) (from Theorem 3.15) must be adapted by putting c instead of 1 in the denominators. In particular, when $\mathbb{P}[A_{11} > x]$ is regularly varying, then $\mathbb{P}[A_{11}^* > x]$ is regularly varying too, hence subexponential, and we find:

$$\begin{aligned} \mathbb{P}[W > x] &\stackrel{x \rightarrow \infty}{\sim} \frac{\lambda_1(r_1 - c)\alpha_1}{c - \lambda_1(r_1 - c)\alpha_1} \mathbb{P}[(r_1 - c)A_{11}^* > x], \\ \mathbb{P}[Z > x] &\stackrel{x \rightarrow \infty}{\sim} p_1 \frac{\rho_1}{c - \rho_1} \mathbb{P}[(r_1 - c)A_{11}^* > x]. \end{aligned}$$

These are precisely Formulas (4.8) and (4.9). The conclusion is clear: as far as the tails of the buffer content (at the beginning of activity periods, or in continuous time) are considered, the contribution of the exponential sources is equivalent to a simple *reduction of the queue capacity* according to their traffic intensities. This remark is due to Jelenkovic and Lazar [29]; dealing with the asymptotics of $\mathbb{P}[Z > x]$, they also remark that the exponential sources might equivalently be replaced by sources with *constant input rates* ρ_i , $2 \leq i \leq N$.

Proof of Boxma [9]:

In this paper, the results are formulated as in Corollary 4.7, and only the asymptotics of $\mathbb{P}[W > x]$ are described, though the approach would yield the tail of B_1 as well. A restrictive assumption is that the exponent δ is assumed *non-integer*. On the other hand, the results are improved in the following way: first, it is only assumed that the Laplace-Stieltjes transforms $\alpha_i[\omega] = \mathbb{E}[e^{-\omega A_{i1}}]$, $2 \leq i \leq N$, are *rational*, thus allowing for more general distributions than the exponential one (like hyperexponential or Erlang distributions); second, a partial *converse* of Theorem 4.7 is established: if W satisfies property (4.11), then $\mathbb{P}[A_{11} > x] \sim l(x)/x^{1+\delta}$.

The approach is based on Cohen's Formula (4.3). Thanks to applications of the residue theorem to the integrals:

$$\int_{x-i\infty}^{x+i\infty} \frac{e^{yt} dy}{y + \lambda_j(1 - \alpha_j[y + r_j\theta])}, \quad 2 \leq j \leq N, \quad (4.13)$$

the following expression is eventually obtained: for $\text{Re}(\omega) > 0$, $\text{Re}(\theta) \geq 0$:

$$\begin{aligned} & \frac{1}{\omega + \lambda(1 - \mathbb{E}[e^{-\omega A_1 - \theta(A_1 + B_1)}])} \\ &= \sum_{k_2, \dots, k_N} \frac{\prod_{j=2}^N 1/(1 - \lambda_j \alpha'_j [\theta r_j + u_{jk_j}])}{(\omega - u_{2k_2} - \dots - u_{Nk_N}) + \lambda_1(1 - \alpha_1[\theta r_1 + (\omega - u_{2k_2} - \dots - u_{Nk_N})])}, \end{aligned}$$

where $u_{jk} = u_{jk}(\theta)$ is the k th pole of the integrand in (4.13), and the sum is over all the poles.

After multiplication of both sides by $-\theta(1 - \lambda\beta)$, it is then shown that the relation may be continued analytically so that we can take $\omega = -\theta$. As already mentioned, we thus obtain $\mathbb{E}[e^{-\theta W}]$, and the rest of the proof consists in expanding it in a Taylor series in terms of the corresponding expansion of $\alpha_1[\theta]$. The forms of these expansions are directly related to tail behaviours in case of regular variation (see Lemma 7.8), so the conclusion is easy.

Proof of Jelenkovic and Lazar [29]:

Here the authors concentrate on the tail of Z , and the result is formulated as in Theorem 4.6. The assumptions are weakened in the following way. First, sources i , $2 \leq i \leq N$, need not have activity periods of exponential distribution: it is only required that they give rise to an *exponential-tailed buffer content* Z' , when they are fed into a fluid queue of capacity $c' > \sum_{i=2}^N \rho_i$, in the absence of source 1 (actually, the relevant hypothesis is that $\mathbb{P}[Z' > x] = o(\mathbb{P}[A_{11}^* > x])$). Second, $\mathbb{P}[A_{11} > x]$ need not be regularly varying, but only of *intermediate regular variation*, that is:

$$\lim_{\mu \downarrow 1} \liminf_{x \rightarrow \infty} \frac{\mathbb{P}[A_{11} > \mu x]}{\mathbb{P}[A_{11} > x]} = \lim_{\mu \downarrow 1} \limsup_{x \rightarrow \infty} \frac{\mathbb{P}[A_{11} > \mu x]}{\mathbb{P}[A_{11} > x]} = 1.$$

This property propagates to A_{11}^* , and implies that $\mathbb{P}[A_{11}^* > x] \in \mathcal{S}$.

The proof of Jelenkovic and Lazar is in two parts, one for the upper bound, one for the lower bound. It is based on the following characterization of Z :

$$Z \simeq \sup_{t \geq 0} \int_0^t [r^*(u) - 1] du.$$

Their treatment of the upper bound may be interpreted as follows.

First choose $\epsilon > 0$ such that: $\rho_1 < c_\epsilon := 1 - (1 + \epsilon) \sum_{i=2}^N \rho_i$, and modify the fluid queue such that a portion c_ϵ of its capacity is *reserved* for source 1, and a portion $c'_\epsilon := 1 - c_\epsilon$ is *reserved* for the other sources. This actually gives rise to two independent and stable queues, of stationary buffer contents Z_ϵ (for the queue fed by source 1 only) and Z'_ϵ (for the queue fed by the other sources), and a sample-path argument easily shows that: $Z \leq_{st} Z_\epsilon + Z'_\epsilon$, due to the loss of efficiency induced by this system. Moreover, we know from Theorem 3.15 (Theorem 9 of [29]) that: $\mathbb{P}[Z_\epsilon > x]/\mathbb{P}[(r_1 - c_\epsilon)A_{11}^* > x] \rightarrow$

$p_1\rho_1/(c_\epsilon - \rho_1)$, hence $\mathbb{P}[Z'_\epsilon > x] = o(\mathbb{P}[Z_\epsilon > x])$ by initial hypothesis. In consequence (see Theorem 7.5 and Corollary 7.6): $\mathbb{P}[Z_\epsilon + Z'_\epsilon > x] \sim \mathbb{P}[Z_\epsilon > x]$. Finally, we obtain: $\limsup_{x \rightarrow \infty} \mathbb{P}[Z > x]/\mathbb{P}[(r_1 - c_\epsilon)A_{11}^* > x] \leq p_1\rho_1/(c_\epsilon - \rho_1)$, and thanks to intermediate regular variation, the upper bound comes by letting ϵ tend to 0.

The proof for the lower bound in [29] does not bear a similar interpretation. In the below remark, we rather present new heuristics in the same spirit.

Remark 4.8 *Imagine that the fluid from source 1 (say, fluid 1) and the fluids from the other sources (say, fluid 2) go to two different buffers (say resp. buffers 1 and 2), and fluid 1 is processed only when buffer 2 is empty (thus we give “priority” to fluid 2). The total amount of fluid in the buffers is not affected by this procedure; in particular, if Z_1 denotes the stationary buffer 1 content, then: $Z \geq_{st} Z_1$. Now add a second queue of capacity c_ϵ ($\epsilon \in (0, 1)$), fed by an on/off source that emits at rate 1 during the periods when buffer 2 is empty. Denote by Z''_ϵ its stationary buffer content (stability can be easily checked). It is not difficult to see that at any time t , $Z_1(t) + Z''_\epsilon(t)$ is larger than $Z_{-\epsilon}(t)$, the buffer content of a single queue of capacity c_ϵ , fed by source 1. Finally we obtain: $Z_1 \geq_{st} Z_{-\epsilon} - Z''_\epsilon$, hence $\mathbb{P}[Z > x] \geq \mathbb{P}[Z_{-\epsilon} - Z''_\epsilon > x]$, where $Z_{-\epsilon}$ and Z''_ϵ are independent. To conclude, notice that $\mathbb{P}[Z_{-\epsilon} - Z''_\epsilon > x] \sim \mathbb{P}[Z_{-\epsilon} > x]$ (see Lemma 7.3), and let ϵ tend to 0.*

Proof of Rolski, Schlegel and Schmidt [50]:

In this work, the authors focus their attention on the tail of B_1 , from which they later deduce the form of $\mathbb{P}[W > x]$. They assume that all the activity periods of sources $2, \dots, N$ follow an *exponential law*. Like in [29], they assume that $\mathbb{P}[A_{11} > x]$ is of *intermediate regular variation*.

The proof is based on an accurate account of the activity periods A_{1n} of source 1 that are contained in the global activity period A_1 , and of their contribution to B_1 . For this, the Markov structure of the “background” process $I'(t) = (I_i(t))_{2 \leq i \leq N}$ is strongly exploited (remember that $I_i(t) = \mathbb{1}_{\{\text{source } i \text{ silent at time } t\}}$).

An important step of the proof consists in establishing the following result. Set $n(t) = r_1 - 1 + \sum_{i=2}^N r_i \mathbb{1}_{\{I_i(t) = 0\}}$: it is the net input rate at time t if t belongs to an activity period of source 1. Consider a typical period A_{11} starting at time 0, $I'(0)$ being arbitrarily distributed. Then thanks to the intermediate regular variation of $\mathbb{P}[A_{11} > x]$, it is shown that:

$$\mathbb{P} \left[\int_0^{A_{11}} n(t) dt > x, I'(A_{11}) = e_I \right] \stackrel{x \rightarrow \infty}{\sim} \left(\prod_{i \in I} p_i \right) \mathbb{P}[(r_1 - c)A_{11} > x],$$

where I is a given subset of $\{2, \dots, N\}$, and $e_I = (\mathbb{1}_{\{i \in I\}})_{2 \leq i \leq N}$. Hence only the stationary characteristics of $I'(t)$ (relating to sources $2, \dots, N$) appear in the limit.

For $N = 2$, an interesting interpretation of Formula (4.7) is obtained:

$$\mathbb{P}[B_1 > x] \stackrel{x \rightarrow \infty}{\sim} \theta_1 \sum_{n=1}^{\infty} (1 - \theta_2) \theta_2^{n-1} \mathbb{P}[(r_1 - c)(A_{11} + \dots + A_{1n}) > x], \quad (4.14)$$

where θ_1 is the probability that A_1 contains at least one activity period A_{11} of source 1, and θ_2 is the stationary probability that the gap between two consecutive periods A_{1n} and A_{1n+1} be bridged by an activity period of source 2. Assuming that $\mathbb{P}[A_{11} > x]$ is subexponential, it is not difficult to show that the right-hand term in (4.14) is equivalent to $(\theta_1/(1 - \theta_2))\mathbb{P}[(r_1 - c)A_{11} > x]$, with $\theta_1/(1 - \theta_2) = (\lambda_1 p_1)/(\lambda p)$.

4.4 Several long-tailed on-period distributions - Lower bounds

Very little is known about the superposition of several sources with long-tailed on-period distributions. So far, the only precise asymptotics that we found in the literature were obtained by Jelenkovic and Lazar [29]. As a preliminary stage of their work, they obtain the speed of convergence of $\mathbb{P}[I(t) = 1]$ to p in the subexponential case (Theorem 1 of [29]): if $p_i > 1/2$ and $\mathbb{P}[A_{i1}^* > t] \in \mathcal{S}$ for $1 \leq i \leq N$, then:

$$\mathbb{P}[I(t) = 1] - p \stackrel{t \rightarrow \infty}{\sim} p \sum_{i=1}^N (1 - p_i) \mathbb{P}[A_{i1}^* > t].$$

Then, under more constraining assumptions which we will not present here, they show (Theorem 3 of [29]) for N identical sources:

$$\mathbb{P}[A_1^* > t] \stackrel{t \rightarrow \infty}{\sim} N \frac{1 - p_1}{1 - p_1^N} \mathbb{P}[A_{11}^* > t].$$

This formula might be a starting point for further research on the superposition of several sources with subexponential on-period distributions. Nothing similar is yet available for the tails of the buffer content distributions, which of course are of primary interest. This seems to be currently one of the most challenging issues in the modelling of fluid queues.

Nevertheless, an interesting insight into the tail behaviour of the buffer content may be obtained via *lower bounds* for $\mathbb{P}[Z > x]$, which do not even require that the silence periods be exponentially distributed. The most obvious method is based on the remark that for all x and i : $\mathbb{P}[Z > x] \geq \mathbb{P}[Z^i > x]$, where Z^i is the buffer content of the fluid queue fed by source i alone; then in case $\mathbb{P}[A_{i1}^* > t] \in \mathcal{S}$ and $r_i > 1$ for some i , Theorem 3.15 provides a subexponential lower bound for $\mathbb{P}[Z > x]$. Jelenkovic and Lazar's approach in [29] (or the equivalent arguments of Remark 4.8) yields the more accurate (asymptotic) lower bound: $\mathbb{P}[Z_{-\epsilon}^i > x]$, where $Z_{-\epsilon}^i$ differs from Z^i in that the capacity of the queue is reduced to $c_{-\epsilon}^i := 1 - (1 - \epsilon) \sum_{j \neq i} \rho_j$ ($\epsilon > 0$), and it is only required that $r_i > 1 - \sum_{j \neq i} \rho_j$; if not only $\mathbb{P}[A_{i1}^* > t] \in \mathcal{S}$ but it is of intermediate regular variation, then ϵ can be taken equal to 0.

However, even this finer method fails to provide long-tailed lower bounds if, whenever $\mathbb{P}[A_{i1}^* > t] \in \mathcal{L}$, we have: $r_i \leq 1 - \sum_{j \neq i} \rho_j$. For this case, Choudhury and Whitt [12] propose an interesting argument, which does not provide sharp, lower bounds, but at

least can be used to detect conditions leading to subexponential tails. Their result may be formulated as follows.

Proposition 4.9 *Let $I \subset \{1, \dots, N\}$ such that $r_I := \sum_{i \in I} r_i > 1$. Then:*

$$\mathbb{P}[Z > x] \geq \prod_{i \in I} (1 - p_i) \mathbb{P}[(r_I - 1)A_{i1}^* > x].$$

Proof:

Consider the on/off sources in stationary regime, and denote by T_i^* the remaining activity time of source i at time 0 (with $T_i^* = 0$ if $I_i^*(0) = 1$). Set $T_I^* = \min_{i \in I} T_i^*$. Then by means of elementary theory of point processes, the authors show that for all x : $\mathbb{P}[Z > x] \geq \mathbb{P}[(r_I - 1)T_I^* > x]$, which yields the announced lower bound. ♠

Corollary 4.10 *If $r_I > 1$ and $\mathbb{P}[A_{i1}^* > t] \in \mathcal{L}$ for all $i \in I$, then for all $\epsilon > 0$: $e^{\epsilon x} \mathbb{P}[Z > x] \rightarrow \infty$ as $x \rightarrow \infty$. Moreover, if $\mathbb{P}[A_{i1} > t] \in \mathcal{R}$ for all $i \in I$, then the lower bound of the former Proposition is regularly varying.*

Proof:

The first statement is a direct consequence of Lemma 7.3. The second one is a consequence of Lemma 7.7, since obviously a product of regularly varying functions is regularly varying too. ♠

5. AN INFINITE NUMBER OF SOURCES

Consider the case of N independent *identical* on/off sources. The off-periods are negative exponentially distributed with mean $1/\lambda_i = N/\lambda$, $i = 1, \dots, N$; the on-periods A_{ij} of each source i have the same yet unspecified distribution with mean α_1 and LST $\alpha_1[\cdot]$, and the inflow rate is r . Now let $N \rightarrow \infty$ and $\lambda_i \rightarrow 0$, such that λ is kept fixed. Then activations of sources (i.e., arrivals of messages) occur according to a Poisson process with rate λ , and the number of simultaneously active sources is distributed as the number of customers in an $M/G/\infty$ system with service time distribution being the activity period distribution of the sources. The superposition of on/off sources again gives rise to one global fluid source, with $\exp(\lambda)$ distributed silence periods and with periods of “global activity” which we denote by A_n . The buffer content at the beginning of global activity periods is again distributed as the waiting time in an $M/G/1$ queue, with arrival rate λ and with service times the net increments B_n during A_n , $n = 1, 2, \dots$. It is easily seen that $\mathbb{E}[A_1] = (e^{\lambda\alpha_1} - 1)/\lambda$ and that $\mathbb{E}[B_1] = r\alpha_1 e^{\lambda\alpha_1} - (e^{\lambda\alpha_1} - 1)/\lambda$ (use the reasoning that in Section 4 led to $\mathbb{E}[A_1]$ and $\mathbb{E}[B_1]$, or let $N \rightarrow \infty$ in those expressions). The stability condition of the fluid queue is $\rho = r\lambda\alpha_1 < 1$.

In the present section we survey the literature for such an $M/G/\infty$ -related fluid model of on/off sources. Like in Section 4, a key reference for this section is a paper

of J.W. Cohen [15], presenting a detailed analysis of the superposition of a finite, as well as infinite, number of on/off sources. He assumes that the inflow rate r of each source equals the outflow rate 1 of the buffer, but most of his results are easily extended to the case $r \geq 1$. Cohen [15] does not pay special attention to long-tailed distributions, but his results can be used to obtain insight into the effect of long-tailed (in particular, regularly varying) activity period distributions on the steady-state distributions of the buffer content W at embedded points in time and Z at arbitrary time, and on the steady-state distribution of the busy period P . This is the topic of Subsection 5.1. In that subsection we assume that $r > 1$. In Subsection 5.2 this assumption is dropped, and lower and upper bounds are presented for the steady-state buffer content distribution.

5.1 Explicit results for the buffer content and busy period distributions

The central question in this subsection is: if the activity period distributions of the (identical) sources are long-tailed, what can then be said about the tail behaviour of the distributions of W , Z and P ? In the case of regularly varying tails of activity period distributions, a useful starting point for studying $\mathbb{P}[W > x]$ is the following result of Cohen ([15], Formula (2.2.7)) for the joint LST of A_1 and B_1 (he takes inflow rate $r = 1$):

Proposition 5.1 For $\operatorname{Re}(\theta) \geq 0$, $\operatorname{Re}(\omega) > 0$,

$$\frac{1}{\omega + \lambda(1 - \mathbb{E}[e^{-\omega A_1 - \theta(A_1 + B_1)}])} = \int_{t=0}^{\infty} e^{-\omega t} e^{-\lambda I(\theta, t)} dt, \quad (5.1)$$

with (cf. [15], Formula (2.2.9)): for $\operatorname{Re}(\theta) \geq 0$, $t \geq 0$, and any choice of $y > 0$,

$$\begin{aligned} I(\theta, t) &:= \frac{1}{2\pi i} \int_{y-i\infty}^{y+i\infty} e^{ut} \frac{1 - \alpha_1[\theta + u]}{u^2} du \\ &= t(1 - \alpha_1[\theta]) + \mathbb{E}[A_{11} e^{-\theta A_{11}}] - \mathbb{E}[(A_{11} - t)e^{-\theta A_{11}} \mathbb{I}_{\{A_{11} \geq t\}}]. \end{aligned} \quad (5.2)$$

Formula (5.1) is obtained by considering a superposition of independent alternating renewal processes. The proposition can be seen as a limiting form of Proposition 4.3 with $\lambda_j = \lambda/N$. We successively show how (5.1) yields the tail behaviour of the distribution of A_1 , of B_1 and (via the M/G/1 relation) of W . In [8] the following result is proved:

Theorem 5.2 Assume that $\rho < 1$, and that $\mathbb{P}[A_{11} > x] \sim l(x)/x^{1+\delta}$ when $x \rightarrow \infty$ (where $l(x)$ is a slowly varying function, and $\delta \in (0, 1)$), then, with $p := e^{-\lambda\alpha_1}$ the probability of total silence:

$$\mathbb{P}[A_1 > x] \stackrel{x \rightarrow \infty}{\sim} \frac{1}{p} \mathbb{P}[A_{11} > x]. \quad (5.3)$$

Proof:

Take $\theta = 0$ in (5.1) and multiply by ω . Using (5.2) yields (apply partial integration in the last step):

$$\begin{aligned} \frac{\omega}{\omega + \lambda(1 - \mathbb{E}[e^{-\omega A_1}])} &= \int_{t=0}^{\infty} \omega e^{-\omega t} \exp\{-\lambda\alpha_1 + \lambda \int_{x=t}^{\infty} \mathbb{P}[A_{11} > x] dx\} dt \\ &= 1 + e^{-\lambda\alpha_1} \int_{t=0}^{\infty} e^{-\omega t} d\{e^{\lambda \int_{x=t}^{\infty} \mathbb{P}[A_{11} > x] dx}\}. \end{aligned} \quad (5.4)$$

After some scaling, an LST of a proper probability distribution appears in the RHS of (5.4). Application of Lemma 7.8 now yields the behaviour of the lefthand side of (5.4) for $\omega \rightarrow 0$, and subsequently the behaviour of $\mathbb{E}[e^{-\omega A_1}]$ for $\omega \rightarrow 0$. Another application of Lemma 7.8 gives the statement of the theorem. \spadesuit

Remark 5.3 *One can easily check that the reverse statement of the theorem also holds; a regularly varying behaviour of $\mathbb{P}[A_1 > x]$ implies a regularly varying behaviour of $\mathbb{P}[A_{11} > x]$. Furthermore, at the expense of additional calculations one can also handle the case of a (non-integer) $\delta > 1$.*

Remark 5.4 *The distribution of A_1 equals the busy period distribution of an $M/G/\infty$ queue, as has been observed in Remark 2.4 of [15]. 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 . A by-product of Theorem 5.2 (plus Remark 5.3) is hence that the busy period of an $M/G/\infty$ queue has a regularly varying tail of index $-\nu$ iff the tail of the service time distribution is regularly varying of index $-\nu$. Jelenkovic and Lazar [29] prove that this statement remains true when “regularly varying” is replaced by “subexponential”.*

The study of the tail behaviour of B_1 is more important, and more difficult, than that of A_1 . It is undertaken in [8, 29, 52]; in all three papers the restrictive assumption is made that $\mathbb{P}[A_{11} > x]$ is regularly varying, and the last paper considers the Pareto distribution within the class of regularly varying functions. In each of these papers, starting-point is Formula (2.2.19) of [15], that follows from (5.1) and (5.2) after some manipulations (take $r = 1$ for simplicity):

$$\begin{aligned} \frac{\theta - \lambda(1 - \alpha_1[\theta])}{\theta - \lambda(1 - \mathbb{E}[e^{-\theta B_1}])} &= 1 - \lambda \int_{t=0}^{\infty} \left\{ \int_{x=t}^{\infty} e^{-\theta(x-t)} d\mathbb{P}[A_{11} < x] \right\} e^{-\lambda I(\theta, t)} dt, \quad \text{Re } \theta \geq 0. \end{aligned} \quad (5.5)$$

A careful analysis of the behaviour of the right-hand side of (5.5) for $\theta \rightarrow 0$ allows yet another application of Lemma 7.8, leading to the conclusion that $\mathbb{P}[B_1 > x]$ is regularly

varying of index $-\nu$ and hence that $\mathbb{P}[W > x]$ is regularly varying of index $1 - \nu$. Below we cite the resulting precise statement for $\mathbb{P}[B_1 > x]$ and $\mathbb{P}[W > x]$ from Theorem 6 respectively Theorem 12 of Jelenkovic and Lazar [29].

Theorem 5.5 *Assume that $\rho < 1$, and that $\mathbb{P}[A_{11} > x] \sim l(x)/x^{1+\delta}$ when $x \rightarrow \infty$ (where $l(x)$ is a slowly varying function) then (with again $p = e^{-\lambda\alpha_1}$ the probability of total silence):*

$$\mathbb{P}[B_1 > x] \stackrel{x \rightarrow \infty}{\sim} \frac{1}{p} \mathbb{P}[(r - 1 + \rho)A_{11} > x], \quad (5.6)$$

$$\mathbb{P}[W > x] \stackrel{x \rightarrow \infty}{\sim} \rho \left(\frac{1}{1 - \rho} - \frac{1}{r} \right) \mathbb{P}[(r - 1 + \rho)A_{11}^* > x]. \quad (5.7)$$

Note that (5.7) immediately follows from (5.6) using (3.12).

Remark 5.6 *Jelenkovic and Lazar [29] formulate a result for $\mathbb{P}[Z > x]$ that is very similar to (5.7). They can prove that result under the assumption that a certain conjecture is true.*

Remark 5.7 *Via a renewal-theoretic approach, Cohen [15] derives the LST of the busy-period distribution, for the case $r = 1$. He shows that this distribution coincides with the busy-period distribution of an $M/G/1$ queue with arrival rate λ and service times A_{1j} . This result could also have been obtained by starting from the busy-period results of Rubinovitch [51] for $N < \infty$, and then letting $N \rightarrow \infty$ (cf. our Section 4, Formula (4.5)).*

We have seen that Proposition 4.3, resp. Proposition 5.1, are key results for the analysis of a fluid queue fed by a finite, resp. an infinite number of sources. Following the approach that led Cohen in [15, 17] to those propositions, one arrives at the following generalization.

Proposition 5.8 *Consider a fluid queue fed by a superposition of N independent sets of sources, with:*

(i) N_1 sets of infinite numbers of sources as described in this section, the j -th set ($1 \leq j \leq N_1$) having silence rate λ_j , activity period distribution with LST $\alpha_j[\cdot]$ and inflow rate r_j ; and:

(ii) $N - N_1$ sets of identical sources, the j -th set ($N_1 < j \leq N$) containing $n_j < \infty$ on/off sources with $\exp(\lambda_j)$ distributed silence periods, activity periods of LST $\alpha_j[\cdot]$ and inflow rate r_j .

Let λ be the silence rate for this superposition of sources. For $\text{Re}(\theta) \geq 0$, $\text{Re}(\omega) > 0$, and any choice of $y > 0$:

$$\frac{1}{\omega + \lambda(1 - \mathbb{E}[e^{-\omega A_1 - \theta(A_1 + B_1)}])} = \int_{t=0}^{\infty} e^{-\omega t} \left[\prod_{j=1}^N \Psi_j(\theta, t) \right] dt, \quad (5.8)$$

with

$$\Psi_j(\theta, t) := \begin{cases} \exp\left(-\lambda_j \frac{1}{2\pi i} \int_{y-i\infty}^{y+i\infty} e^{ut} \frac{1 - \alpha_j[\theta r_j + u]}{u^2} du\right), & 1 \leq j \leq N_1, \\ \left[\frac{1}{2\pi i} \int_{y-i\infty}^{y+i\infty} \frac{e^{ut}}{u + \lambda_j(1 - \alpha_j[\theta r_j + u])} du \right]^{n_j}, & N_1 < j \leq N. \end{cases}$$

Note that, when all the sets of infinite numbers of sources have the same inflow rate, one can immediately aggregate those N_1 sets into one set of an infinite number of sources.

By allowing just one set of sources to have activity period distributions with a non-exponential tail, one may use Proposition 5.8 combined with the method exposed in [8, 9], to study the tail behaviour of B_1 and W .

5.2 Bounds

So far we have almost always assumed that the inflow rate of each source exceeds the outflow rate of the buffer. If this assumption does not hold, then the buffer content may decrease, and even become zero, during a global activity period. For the case of a single source this leads to a trivial model, but for multiple sources it may lead to serious analytical complications regarding the net increment B_1 (see also Remark 3.3). In this subsection we discuss results for which the above flow assumption does not necessarily hold.

Brichet et al. [10] derive lower and upper bounds for $\mathbb{P}[Z > x]$ in the case of a fluid queue fed by an infinite number of identical on/off sources with inflow rate equal to one. Their starting point is a set of known lower and upper bounds for $\mathbb{P}[Z > x]$ for $N < \infty$. They specify the limits of these bounds for $N \rightarrow \infty$, when the outflow rate $C = N\nu + \gamma\sqrt{N}$ (this represents a heavy traffic condition, as ν denotes the fraction of time a source is on). Note that the inflow rate of each source is here much smaller than the outflow rate of the buffer. Brichet et al. prove that, if either one of the silence and activity period distributions is regularly varying, then the determining factor in both limiting bounds for $\mathbb{P}[Z > \xi\sqrt{N}]$, for fixed $\xi > 0$, is $\exp(-R\xi^{2(1-H)})$; i.e., both bounds are Weibullian. Here R is a constant and H is determined by the power laws of the silence and activity period distributions, in exactly the same way as below (2.4) and (2.5) in this paper. The Weibullian nature of the bounds for this heavy-traffic case with inflow rates much smaller than the outflow rate contrasts with the regularly varying behaviour that was proven for $\mathbb{P}[W > x]$ in Theorem 5.5 and that was conjectured for $\mathbb{P}[Z > x]$ in Theorem 13 of [29] (see Remark 5.6). The Weibullian nature of the bounds *does* agree with the tail behaviour obtained by Norros [42] for the fluid queue fed by Fractional Brownian Motion: see (2.2) and (2.3). Brichet et al. show via some scalings and a limiting operation, how the FBM model of Norros is related to their fluid queue, thus providing a further physical motivation for the FBM model.

Duffield [20] studies the workload of a buffer fed by N sources which may be of a very general nature (not necessarily on/off fluid) and which may display long-range dependence. More specifically, he assumes that the input process of each source has a power-law large-deviation scaling (as is the case in FBM). Letting N approach infinity while keeping the offered load constant, he proves under very general conditions that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \mathbb{P}[Z > Nx] = -M(x), \quad (5.9)$$

where the shape function $M(x)$ is expressed in terms of the cumulant generating functions of the input traffic. He studies the asymptotics of $M(x)$ for $x \rightarrow \infty$, and he discusses the economies of scale which are available through the statistical multiplexing of large numbers of sources with large deviations scalings. For FBM with Hurst parameter H , he shows that $M(x) = \delta_H x^{2(1-H)}$, with δ_H some constant, confirming the logarithmic accuracy of (2.3).

Remark 5.9 *In case the inflow rate r is less than 1 in our fluid queue fed by an infinite number of on/off sources, it is important to determine the distribution of the time periods during which the number of active sources exceeds $c = 1/r$. Guillemin et al. [26] do exactly that, for $\exp(1)$ distributed activity periods (hence, for an $M/M/\infty$ system). They use the Markovian nature of the system and a uniformization technique. They also present several asymptotic results.*

Remark 5.10 *Parulekar and Makowski [45] consider the distribution of the number of busy servers b_n at time n in a discrete-time $M/G/\infty$ system. They remark that the stationary version $\{b_n^*, n = 0, 1, \dots\}$ is short-range (long-range) dependent iff the second moment of the service time distribution is finite (infinite). Subsequently they consider a discrete-time single server queue with constant release rate c cells per slot, and feed it with the traffic stream $\{b_n, n = 0, 1, \dots\}$; i.e., b_n cells of unit length arrive at time n . They consider the tail probability of the buffer content of the latter single server queue. They show that the appropriate large deviations scaling is related to the forward recurrence time of the service time distribution.*

For the case of subexponential service time distribution of the $M/G/\infty$ system, Liu et al. [39] derive asymptotic upper and lower bounds for the queue length distribution in the single server queue fed by the $M/G/\infty$ system; they show that these bounds are tight in some instances.

6. CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

In this survey we have discussed fluid queues fed by a number of on/off sources. We have mainly concentrated on the relation between the tail behaviour of the activity periods of the sources and the tail behaviour of the buffer content and busy period, exposing among other things how heavy-tailed (and long-range dependent) behaviour of the former quantities gives rise to heavy-tailed behaviour of the latter quantities.

The subject matter of this paper is in the midst of its development. Presently several directions are open for further research. One direction is to weaken the distributional restrictions: e.g., extension of regular variation results to results for intermediate regularly varying or subexponential distributions; admittance of non-exponential silence periods. A second direction concerns the study of the restriction that the inflow rate of each source is at least equal to the outflow rate of the buffer; see the beginning of Subsection 5.2. In [9] a conjecture is formulated concerning conditions under which regular variation of one or more activity period distributions does (not) propagate to regular variation of the buffer content process, when inflow rates do not necessarily exceed the outflow rate; see also the results in [29]. A third direction is a study of the interplay of multiple sources with long-tailed activity period distributions. Section 5 presents results for an infinite number of identical sources; in the case of heterogeneous sources, one may expect the source with the heaviest tail to dominate. A fourth direction is to obtain a much better quantitative insight into the tail behaviour of W and Z . In this respect, numerical results and series developments for $\mathbb{P}[W > x]$ and $\mathbb{P}[Z > x]$ would be most useful (cf. [1, 18]).

Other major issues are: extension of the results for more general sources (see [28] for a study of a source that alternates between various states according to a Markov process); extension of the results to networks (here busy period results will be useful, as indicated in Remark 3.19); admission control and transmission rate control.

In our survey we have put an emphasis on fluid queues, fed by on/off sources. An advantage of these models is that, to a considerable extent, one can make use of classical queueing results. However, as we have seen above, this approach also has its limitations. Hence the possibilities of quite different models, like FBM and non-linear deterministic maps (see Subsection 2.2), deserve a careful further study.

7. APPENDIX

Appendix A: Subexponentiality

The literature contains various notions of non-exponential tail behaviour of probability distributions. An important class is the class of *subexponential* distributions, which contains, a.o., the Pareto, Weibull and lognormal distributions. The introduction of subexponentiality is due to Chistyakov [11]. A related notion is the “moderate growth” of a distribution tail, introduced by Smith [53]; we use the term “long-tailed” that is presently more common. We keep a formulation based on random variables rather than distribution functions.

Definition 7.1 X has a subexponential tail, denoted by $\mathbb{P}[X > t] \in \mathcal{S}$, if

$$\frac{\mathbb{P}[X + X' > t]}{\mathbb{P}[X > t]} \rightarrow 2 \quad \text{as } t \rightarrow \infty,$$

where X and X' are i.i.d.

X is long-tailed, to be denoted by $\mathbb{P}[X > t] \in \mathcal{L}$, if for all real u :

$$\frac{\mathbb{P}[X > t + u]}{\mathbb{P}[X > t]} \rightarrow 1 \quad \text{as } t \rightarrow \infty.$$

Lemma 7.2 *If $\mathbb{P}[X > t]$ is subexponential, then it is long-tailed.*

Proof:

See Chistyakov [11], or Athreya and Ney [5], or Embrechts *et al.* [22] (Lemma 1). ♠

The following properties of long-tailed distributions thus hold for subexponential tails.

Lemma 7.3 *If $\mathbb{P}[X > t] \in \mathcal{L}$, then:*

- (i) for all $\epsilon > 0$, $e^{\epsilon t} \mathbb{P}[X > t] \rightarrow \infty$ as $t \rightarrow \infty$;
- (ii) for any non-negative random variable Y independent of X :
 $\mathbb{P}[X - Y > t] / \mathbb{P}[X > t] \rightarrow 1$ as $t \rightarrow \infty$.

Proof:

See Chistyakov [11] or Embrechts *et al.* [22] (Lemma 1) for property (i) ((ii) is easy to prove). ♠

The introduction of subexponential distributions was motivated by the tail-analysis of convolutions. Their nice properties are presented below.

Lemma 7.4 *Let $(X_n)_{n \geq 1}$ be an i.i.d. sequence. If $\mathbb{P}[X_1 > t] \in \mathcal{S}$, then:*

- (i) for $n \geq 1$: $\mathbb{P}[X_1 + \dots + X_n > t] / \mathbb{P}[X_1 > t] \rightarrow n$ as $t \rightarrow \infty$;
- (ii) $\forall \epsilon > 0, \exists K > 0 / \forall t \geq 0, \forall n \geq 1$: $\mathbb{P}[X_1 + \dots + X_n > t] / \mathbb{P}[X_1 > t] \leq K(1 + \epsilon)^n$.

Proof:

See Athreya and Ney ([5], IV.4). ♠

The above property (i) may be partially extended to the convolution of different distributions.

Theorem 7.5 *Assume that $\mathbb{P}[X > t] \in \mathcal{S}$, and $\mathbb{P}[Y > t] = O(\mathbb{P}[X > t])$.*

(i) If $\mathbb{P}[Y > t] \in \mathcal{L}$ or $\mathbb{P}[Y > t] = o(\mathbb{P}[X > t])$, then:

$$\mathbb{P}[X + Y > t] \stackrel{t \rightarrow \infty}{\sim} \mathbb{P}[X > t] + \mathbb{P}[Y > t].$$

Moreover, $\mathbb{P}[X + Y > t] \in \mathcal{S}$, and $(\theta\mathbb{P}[X > t] + (1 - \theta)\mathbb{P}[Y > t]) \in \mathcal{S}$ for all $\theta \in (0, 1)$.

(ii) If $\mathbb{P}[Y > t] \in \mathcal{L}$ and $\mathbb{P}[Y > t] \neq o(\mathbb{P}[X > t])$, then $\mathbb{P}[Y > t] \in \mathcal{S}$.

Proof:

See Embrechts *et al.* [22] (Proposition 1) and Cline [13] (Corollary 3.2 and Theorem 3.4). ♠

Corollary 7.6 *If $\mathbb{P}[X > t] \in \mathcal{S}$ and $\mathbb{P}[Y > t] \sim K\mathbb{P}[X > t]$ as $t \rightarrow \infty$ (where K is a positive constant), then $\mathbb{P}[Y > t] \in \mathcal{S}$.*

Proof:

Since $\mathbb{P}[X > t] \in \mathcal{L}$ and $\mathbb{P}[Y > t] \sim K\mathbb{P}[X > t]$, then obviously $\mathbb{P}[Y > t] \in \mathcal{L}$ and $\mathbb{P}[Y > t] \neq o(\mathbb{P}[X > t])$. The conclusion is easy. (See also the direct proof of Pakes in [44] (Lemma 2).) ♠

Unfortunately, it is not known how to conclude from its LST that a distribution is subexponential. A very useful relation between the tail behaviour of a distribution and the behaviour of its LST near the origin exists for the class of regularly varying distributions, which forms a subclass of the class of subexponential distributions. The next appendix is devoted to regular variation.

Appendix B: Regular variation

Regular variation is an important concept in probability theory and various other fields. The main reference text is the book [7]. A measurable positive function f is called regularly varying of index ζ if, for all $x > 0$,

$$f(xt)/f(t) \rightarrow x^\zeta, \quad t \rightarrow \infty,$$

(cf. [7], p. 18). One writes $f \in \mathcal{R}_\zeta$. When $\zeta = 0$, one speaks of a *slowly varying function*; this could for instance be a constant, or a logarithmic function. In this paper, a slowly varying function is denoted by $l(\cdot)$. We shall say that a stochastic variable $X \geq 0$ has a regularly varying tail when $\mathbb{P}[X > t]$ is a regularly varying function; an example is provided by 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. As observed in Subsection 2.2, X is now called heavy-tailed. This case is known to give rise to long-range dependence, see below (2.5).

The first lemma shows the equivalence between the regular variation of $\mathbb{P}[X > t]$ and that of $\mathbb{P}[X^* > t]$ (when $\mathbb{E}[X] < \infty$); it also shows that the set of distributions with regularly varying tails is closed under convolutions.

Lemma 7.7 (i) *For all $\nu > 0$, $\mathbb{P}[X^* > t] \in \mathcal{R}_{-\nu}$ if and only if $\mathbb{P}[X > t] \in \mathcal{R}_{-\nu-1}$, and if either is the case then:*

$$\mathbb{P}[X^* > t] \stackrel{t \rightarrow \infty}{\sim} \frac{t}{\nu \mathbb{E}[X]} \mathbb{P}[X > t].$$

(ii) *If X_1 and X_2 are two independent, nonnegative random variables such that $\mathbb{P}[X_1 > t] \in \mathcal{R}_{-\nu_1}$ and $\mathbb{P}[X_2 > t] \in \mathcal{R}_{-\nu_2}$, then $\mathbb{P}[X_1 + X_2 > t] \in \mathcal{R}_{-\min(\nu_1, \nu_2)}$, and moreover:*

$$\mathbb{P}[X_1 + X_2 > t] \stackrel{t \rightarrow \infty}{\sim} \mathbb{P}[X_1 > t] + \mathbb{P}[X_2 > t].$$

Proof:

(i) See [7], Sections 1.5.6 and 1.7.3, as part of the Karamata theorem and the Monotone Density Theorem.

(ii) See [25], VIII.8, Proposition. ♠

The next lemma links the behaviour of $\mathbb{P}[X > t]$ for $t \rightarrow \infty$ to the behaviour of its LST $\phi(s)$ for $s \rightarrow 0$. This is part of Theorem 8.1.6 on p. 333/334 of [7], originally due to Bingham and Doney. Assume that the first n moments μ_1, \dots, μ_n of X are finite. Define

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

Lemma 7.8 *Let $\nu = n + \psi$ with $0 < \psi < 1$. The following are equivalent:*

$$\phi_n(s) \sim s^\nu l(1/s), \quad s \rightarrow 0, \tag{7.1}$$

$$\mathbb{P}[X > t] \sim \frac{(-1)^n}{\Gamma(1-\nu)} t^{-\nu} l(t), \quad t \rightarrow \infty. \tag{7.2}$$

In Theorem 8.1.6 of [7] the somewhat more complicated cases $\psi = 0, 1$ are also discussed.

Appendix C: Proofs

Proof of Lemma 3.10:

By definition:

$$\begin{aligned} \frac{1}{r^2} \text{Cov}(r^*(0), r^*(u)) &= \mathbb{P}[I^*(0) = I^*(u) = 0] - \mathbb{P}[I^*(0) = 0]\mathbb{P}[I^*(u) = 0] \\ &= \mathbb{P}[I^*(0) = 0]\mathbb{P}[I^*(0) = 1](\mathbb{P}[I^*(u) = 0|I^*(0) = 0] - \mathbb{P}[I^*(u) = 0|I^*(0) = 1]), \end{aligned}$$

which yields Formula (3.10). By a renewal argument, we get:

$$\begin{aligned} &\mathbb{P}[I^*(u) = 0|I^*(0) = 0] \\ &= \mathbb{P}[I^*(u) = 0, T^* > u|I^*(0) = 0] + \mathbb{P}[I^*(u) = 0, T^* \leq u|I^*(0) = 0] \\ &= \mathbb{P}[A_1^* > u] + \int_0^u \mathbb{P}[A_1^* + S_0 \in dv]\mathbb{P}[I(u-v) = 0]. \end{aligned}$$

Hence:

$$\begin{aligned} &\int_0^t \mathbb{P}[I^*(u) = 0|I^*(0) = 0]du \\ &= \int_0^t \mathbb{P}[A_1^* > u]du + \int_0^t \int_0^u \mathbb{P}[A_1^* + S_0 \in dv]\mathbb{P}[I(u-v) = 0] du \\ &= \int_0^t \mathbb{P}[A_1^* > u]du + \int_0^t \mathbb{P}[A_1^* + S_0 \leq w]\mathbb{P}[I(t-w) = 0]dw. \end{aligned}$$

Similarly:

$$\begin{aligned} &\mathbb{P}[I^*(u) = 0|I^*(0) = 1] \\ &= \mathbb{P}[I^*(u) = 0, T^* > u|I^*(0) = 1] + \mathbb{P}[I^*(u) = 0, T^* \leq u|I^*(0) = 1] \\ &= 0 + \int_0^u \mathbb{P}[S_1^* \in dv]\mathbb{P}[I(u-v) = 0]. \end{aligned}$$

Hence:

$$\int_0^t \mathbb{P}[I^*(u) = 0|I^*(0) = 1]du = \int_0^t \mathbb{P}[S_1^* \leq w]\mathbb{P}[I(t-w) = 0]dw.$$

Finally, since:

$$\mathbb{P}[A_1^* + S_0 \leq w] - \mathbb{P}[S_1^* \leq w] = \mathbb{P}[S_1^* > w] - \mathbb{P}[A_1^* + S_0 > w],$$

we thus obtain (3.11). ♠

Proof of Lemma 3.12:

By the basic theory of regenerative processes (see e.g. Theorem 1.2, Chapter V of Asmussen [3]), we have: $\mathbb{P}[I(t) = 0] \rightarrow 1 - p$ as $t \rightarrow \infty$. Given some $\epsilon > 0$, we can find some t_0 such that: $|\mathbb{P}[I(t) = 0] - (1 - p)| \leq \epsilon(1 - p)$ for all $t \geq t_0$. Consequently, for

$t \geq t_0$:

$$\begin{aligned}
 & \left| \int_0^t f(u) \mathbb{P}[I(t-u) = 0] du - (1-p) \int_0^t f(u) du \right| \\
 & \leq \int_0^{t-t_0} f(u) |\mathbb{P}[I(t-u) = 0] - (1-p)| du + (2-p) \int_{t-t_0}^t f(u) du \\
 & \leq \epsilon(1-p) \int_0^t f(u) du + t_0(2-p) \sup_{u \geq t-t_0} f(u) \\
 & \leq 2\epsilon(1-p) \int_0^t f(u) du \quad \text{if } t \geq t_1, \text{ for some } t_1 > t_0.
 \end{aligned}$$

The proof is complete. ♠

REFERENCES

1. J. Abate, G.L. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems*, 16:311–338, 1994.
2. D. Anick, D. Mitra, and M.M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Techn. J.*, 61:1871–1894, 1982.
3. S. Asmussen. *Applied Probability and Queues*. John Wiley & Sons Ltd, 1987.
4. S. Asmussen. Aspects of matrix Wiener-Hopf factorization in applied probability. *Math. Scientist*, 14:101–116, 1989.
5. K.B. Athreya and P.E. Ney. *Branching Processes*. Springer-Verlag, 1972.
6. J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43:1566–1579, 1995.
7. N.H. Bingham, C.M. Goldie, and J.L. Teugels. *Regular Variation*. Cambridge University Press, Cambridge, 1987.
8. O.J. Boxma. Fluid queues and regular variation. *Performance Evaluation*, 27-28:699–712, 1996.
9. O.J. Boxma. Regular variation in a multi-source fluid queue; accepted for publication in ITC 15. Report BS-R9614, CWI, 1996.
10. F. Brichet, J. Roberts, A. Simonian, and D. Veitch. Heavy traffic analysis of a storage model with long range dependent on/off sources. *Queueing Systems*, 23:197–215, 1996.
11. V.P. Chistyakov. A theorem on sums of independent, positive random variables and its applications to branching processes. *Theory Probability Appl.*, 9:640–648, 1964.
12. G.L. Choudhury and W. Whitt. Long-tail buffer-content distributions in broadband networks. Submitted to *Performance Evaluation*, 1995.

13. D.B.H. Cline. Convolutions of distributions with exponential and subexponential tails. *J. Austr. Math. Soc. (Series A)*, 43:347–365, 1987.
14. J.W. Cohen. Some results on regular variation for distributions in queueing and fluctuation theory. *Journal of Applied Probability*, 10:343–353, 1973.
15. J.W. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Processes and their Applications*, 2:31–58, 1974.
16. J.W. Cohen. *The Single Server Queue*. North-Holland Publ. Cy., Amsterdam, 2nd edition, 1982.
17. J.W. Cohen. On the effective bandwidth in buffer design for the multi-server channels. Report BS-R9406, CWI, 1994.
18. J.W. Cohen. On the $M/G/1$ queue with heavy-tailed service time distributions. Report BS-R9702, CWI, 1997.
19. M.E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *Performance Evaluation Review*, 24:160–169, 1996.
20. N.G. Duffield. Economies of scale in queues with sources having power-law large deviation scalings. *Journal of Applied Probability*, 33:840–857, 1996.
21. N.G. Duffield and N. O’Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cam. Phil. Soc.*, 118:363–374, 1995.
22. P. Embrechts, C.M. Goldie, and N. Veraverbeke. Subexponentiality and infinite divisibility. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 49:335–347, 1979.
23. A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4:209–223, 1996.
24. A. Erramilli, R.P. Singh, and P. Pruthi. Chaotic maps as models of packet traffic. In J. Labetoulle and J.W. Roberts, editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proceedings of the 14th International Teletraffic Congress*, pages 329–338. North-Holland Publ. Cy., Amsterdam, 1994.
25. W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. John Wiley & Sons Ltd, 2nd edition, 1971.
26. F. Guillemin, G. Rubino, B. Sericola, and A. Simonian. Transient characteristics of an $M/M/\infty$ system applied to statistical multiplexing on an ATM link. Rapport de recherche 2386, INRIA, Novembre 1994.
27. D. Heath, S. Resnick, and G. Samorodnitsky. Heavy tails and long range dependence in on/off processes and associated fluid models. Technical Report 1144, School of Operations Research and Industrial Engineering, Cornell University, January 1996. To appear in *Mathematics of Operations Research*.

28. P.R. Jelenković and A.A. Lazar. Subexponential asymptotics of a Markov-modulated G/G/1 queue. Submitted to *Journal of Applied Probability*, 1995.
29. P.R. Jelenković and A.A. Lazar. Multiplexing on/off sources with subexponential on-periods. CTR Technical Report 457-96-23, Columbia University, New-York, NY 10027-6699, USA, July 1996.
30. H. Kaspi and M. Rubinovitch. The stochastic behavior of a buffer with non-identical input lines. *Stochastic Processes and their Applications*, 3:73–88, 1975.
31. O. Kella and W. Whitt. A storage model with a two-state random environment. *Operations Research*, 40(S2):S257–S262, 1992.
32. C. Klüppelberg. Subexponential distributions and integrated tails. *Journal of Applied Probability*, 25:132–141, 1988.
33. C. Klüppelberg. Subexponential distributions and characterizations of related classes. *Probability Theory and Related Fields*, 82:259–269, 1989.
34. L. Kosten. Stochastic theory of a multi-entry buffer (I). *Delft Progress Report*, 1:10–18, 1974.
35. L. Kosten. Stochastic theory of a multi-entry buffer (II). *Delft Progress Report*, 1:44–50, 1974.
36. L. Kosten. Stochastic theory of data handling systems with groups of multiple sources. In W. Bux and H. Rudin, editors, *Performance of Computer-Communication Systems*, pages 321–331. North-Holland Publ. Cy., Amsterdam, 1984.
37. L. Kosten and O.J. Vrieze. Stochastic theory of a multi-entry buffer (III). *Delft Progress Report*, 1:103–115, 1975.
38. 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*, 2:1–15, 1994.
39. Z. Liu, Ph. Nain, D. Towsley, and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long-range dependent process. Technical Report CMPSCI 97-16, Univ. of Massachusetts at Amherst, 1997.
40. 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. *Journal of Applied Probability*, 17:802–813, 1980.
41. D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Advances in Applied Probability*, 20:646–676, 1988.
42. I. Norros. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.
43. I. Norros. On the use of fractional Brownian motion in the theory of connectionless

- networks. *IEEE Journal on Selected Areas in Communications*, 13:953–962, 1995.
44. A.G. Pakes. On the tails of waiting-time distributions. *Journal of Applied Probability*, 12:555–564, 1975.
 45. M. Parulekar and A.M. Makowski. Tail probabilities for $M/G/\infty$ input processes (I): Preliminary asymptotics. Report, Univ. of Maryland, Inst. for Systems Research, 1996.
 46. V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.
 47. N.U. Prabhu. *Stochastic Storage Processes*. Springer-Verlag, 1980.
 48. P. Pruthi. *An Application of Chaotic Maps to Packet Traffic Modeling*. Ph.D Thesis, Royal Institute of Technology, Dept. of Teleinformatics, Kista, Sweden, 1995.
 49. J.W. Roberts, U. Mocci, and J. Virtamo. *Broadband Network Traffic - Final Report of Action COST 242*. Springer-Verlag, 1996.
 50. T. Rolski, S. Schlegel, and V. Schmidt. Asymptotics of Palm-stationary buffer content distributions in fluid flow queues. Preprint, November 1996.
 51. M. Rubinovitch. The output of a buffered data communication system. *Stochastic Processes and their Applications*, 1:375–382, 1973.
 52. A. Simonian and D. Veitch. A storage model with high rate and long range dependent on/off sources. Technical report, FT-CNET, 1996.
 53. W.L. Smith. On the tails of queue time distributions. Mimeo Series 830, Dept. of Stat., Univ. of N.-Carolina, Chapel Hill, 1972.
 54. T.E. Stern and A.I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability*, 23:105–139, 1991.
 55. 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.
 56. W. Willinger, M.S. Taqqu, R. Sherman, and D.V. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5:71–86, 1997.