ELSEVIER

Operations Research Letters 28 (2001) 233–242

operations research letters

www.elsevier.com/locate/dsw

A note on queues with $M/G/\infty$ input

Michel Mandjes [1]

Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974, USA

Received 1 May 2000

Abstract

We consider a fluid queue fed by sessions, arriving according to a Poisson process; a session has a heavy-tailed duration, during which traffic is sent at a constant rate. We scale Poisson input rate Λ, buffer space B, and link rate C by n, such that we get $n\lambda$, nb, and nc, respectively. Then we let n grow large. In this regime, the overflow probability decays exponentially in the number of sources n; we examine the specific situation in which b is also large.

In Duffield (Queueing Syst. 28 (1998) 245–266) this setting is considered. A crucial role was played by the function $v(t) := -\log \mathbb{P}(D^\star > t)$ for large t, $D^\star$ being the *residual* session duration. Duffield covers the case that $v(\cdot)$ is regularly varying of index strictly between 0 and 1 (e.g., Weibull); this note treats slowly varying $v(\cdot)$ (e.g., Pareto, Lognormal).

The proof adds insight into the way overflow occurs. If $v(\cdot)$ is slowly varying then, during the trajectory to overflow, the input rate will exceed the link rate only slightly. Consequently, the buffer will fill 'slowly', and the typical time to overflow will grow 'faster than linearly' in the buffer size. This is essentially different from the 'Weibull case', where the input rate will significantly exceed the link rate, and the time to overflow is essentially proportional to the buffer size. In both cases there is a substantial number of sessions that remain in the system during the entire path to overflow.

Keywords: Queues; Approximations; Communications

1. Introduction

This note focuses on an infinite buffer drained at constant rate C per unit time. Sessions arrive according to a discrete-time Poisson process, i.e., the number of sessions arriving in the ith time slot are i.i.d. Poisson random variables with mean Λ. A session remains at the resource during a random time, distributed as D; the holding times of the individual sessions are i.i.d. During its holding time, a session generates traffic at a constant rate r. This model is called a *queue with $M/G/\infty$ input*. We are interested in the steady-state probability that the buffer content exceeds level B.

[1] The author is also with CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands, and Faculty of Mathematical Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

E-mail address: michel@research.bell-labs.com (M. Mandjes).

0167-6377/01/$ - see front matter
PII: S0167-6377(01)00074-8

It is noted that the $M/G/\infty$ input model naturally models the long-range dependent effects detected in network traffic [7,13,14]. By assuming the session durations D long-tailed, the influence of long-range dependence can be assessed. Parulekar and Makowski [12], Duffield [4], and Liu et al. [9] investigate *large-buffer asymptotics*, or, more precisely, logarithmic tail asymptotics of the queue length distribution. They find upper and lower bounds, that match in some cases. Also recent work by Likhanov and Mazumdar [8] should be mentioned; there the authors focus on a special type of long-tailed distributions.

An interesting new direction was chosen in Duffield [5]. He derives so-called *large-aggregation asymptotics* rather than large-buffer asymptotics. The procedure followed in [5] is as follows. (i) First input and resources are scaled by a parameter n, i.e., capacity $C \equiv nc$, buffer threshold $B \equiv nb$, and Poisson rate $\Lambda \equiv n\lambda$. (ii) Then under this regime, a general theorem on the overflow probability [3] can be invoked: the decay rate of the loss probability can be expressed as a variational problem, for general λ, b, and c. (iii) For the special case of large b, the variational problem requires asymptotics of the cumulant generating function of the traffic generated by an $M/G/\infty$ input process, as derived in [11].

In this note we will follow the procedure described above. In [5] a crucial role was played by the function $v(t) := -\log \mathbb{P}(D^\star > t)$, $D^\star$ being the *residual* session duration. If $v(t)$ is regularly varying with index h in $[0,1)$ the durations could be thought of as long tailed. Ref. [5] covers the case of $h \in (0,1)$, but the solution of the important case of $h = 0$ (with for instance Pareto and Lognormal durations) is not complete. We will exactly point out what part is lacking, and how to solve that.

The solution found for $h = 0$ has important consequences for the intuition on the way the queue builds up. For large b, the duration of the busy period preceding overflow will typically grow *faster* than proportional in b. The input rate during such a trajectory will be only slightly larger than the link rate. Essentially, overflow is caused by a number of sessions that remain in the system during the entire busy period. This is a crucial difference with exponential-like duration distributions; there sessions present at the beginning of the busy period will typically have left at overflow. These effects are due to the fact that for durations with $h = 0$ it is relatively 'low-cost' to be extremely long. Also for $h \in (0,1)$ overflow is due to a number of sessions staying in the system during the entire path to overflow, but the time to overflow is essentially linear in b and the input rate is substantially larger than the link rate.

Section 2 gives preliminaries, relying on [5,11]. In Section 3, the proof for $h = 0$ is given. Section 4 gives a refinement for the case $D^\star$ is distributed Pareto. Section 5 gives the intuition behind the results.

2. Preliminaries

As inidicated in Introduction, we consider a buffered resource in discrete time. Sessions arrive according to a discrete-time Poisson process, i.e., the number of sessions arriving in the ith time slot are i.i.d. Poisson random variables with mean Λ. A session remains at the resource during a random time, distributed as D; the holding times of the individual sessions are i.i.d. We assume that D had a finite mean—this makes sure that the *residual* session duration time has a proper distribution

$$\mathbb{P}(D^\star = t) = \frac{\mathbb{P}(D \geqslant t)}{\mathbb{E}\, D}.$$

During their holding time, a session generates traffic at a constant rate r; like in [5] we can adapt to the case at which the session transmits is not constant, but is averaging on a faster time scale than the long-tailed session durations.

We are interested in the probability of the buffer content exceeding level B, denoted by $p(B,C)$. We rescale the resources by the number of sources: $C \equiv nc$ and $B \equiv nb$. Also, we scale the Poisson rate of arrivals: $\Lambda \equiv n\lambda$. We assume that the system is stable:

$$\frac{\rho}{c} < 1, \quad \text{with } \rho := \lambda r \mathbb{E}\, D.$$

In the scaled model we define

$$p_n(b,c) := \text{steady-state probability that the buffer content exceeds level } nb.$$

Define also

$$A(t) := \begin{cases} \text{Traffic generated by an } M/G/\infty \text{ input, with Poisson arrival rate } \lambda \\ \text{and sessions i.i.d. distributed as } D, \text{ transmitting at rate } r, \\ \text{in steady state, during a time interval of length } t. \end{cases}$$

The next proposition stems from Duffield [5] and characterizes, for general buffer size b, the decay rate of the overflow probability as the solution of a variational problem. The exact conditions are given in [5, Theorem 3].

Proposition 2.1 (Decay rate for general b). *A variational problem to compute the decay rate is given by*

$$I(b) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(b,c) = \inf_{t\in\mathbb{N}} w(t) \bar{J}_t\left(\frac{b}{t} + c\right), \tag{1}$$

for an increasing positive function $w(\cdot)$ and

$$\bar{J}_t(x) := \sup_\theta \left(\theta x - \frac{\log \mathbb{E}e^{\theta A(t) w(t)/t}}{w(t)}\right).$$

The function $w(\cdot)$ in Proposition 2.1 is usually called a *scaling function*. It was introduced in [3,6] to enable large-deviations analysis in situations where there was no exponential decay in the buffer size. The optimizing t, say $t_b^\star$, could be considered as the typical duration of a busy period preceding overflow [15].

For the model with $M/G/\infty$ input, it was proposed in [11] to use $w(t) = v(t)$, where $v(t)$ is the extension of $-\log \mathbb{P}(D^\star > t)$ to $\mathbb{R}_+$. In the sequel we use this scaling. The following lemma gives the asymptotics of the log moment generating function under this scaling. It is due to Parulekar and Makowski [11].

Lemma 2.2 (Cumulant function). *Suppose $v(t)/t$ is monotone decreasing in the limit, and suppose there is a mapping $\Gamma : \mathbb{N} \to \mathbb{N}$ satisfying* (i) $\Gamma(t) < t$ *for all* $t \in \mathbb{N}$, (ii) $v(t)\Gamma(t)/t \to \infty$, *and* (iii) $v(t)\Gamma(t)/(tv(\Gamma(t))) \to 0$. *Then*

$$\lim_{t\to\infty} \frac{\log \mathbb{E}e^{\theta A(t) v(t)/t}}{v(t)} = (\lambda r \mathbb{E}D)\theta = \rho\theta$$

if $\theta < r^{-1}$; the limit equals ∞ if $\theta > r^{-1}$.

The next definition defines a class of long-tailed distributions. We will assume that the session duration is of this type.

Definition 2.3 (Subexponentially varying distribution). Suppose the function $v(\cdot)$ is regularly varying of index h (at infinity), that is,

$$\frac{v(yt)}{v(t)} \to y^h, \quad t \to \infty$$

for all $y > 0$. If $v(\cdot)$ is regularly varying of index $h \in [0,1)$, we say that $D^\star$ has a subexponentially varying distribution, or $D^\star \in \mathscr{V}$. The class of subexponentially varying distributions with $h = 0$ is called $\mathscr{V}_0$.

In the last definition we used the concept of regular variation, see for instance Bingham et al. [2, Section 1.4].

For standard long-tailed distributions D (like Pareto, Lognormal, and Weibull) there exists a Γ with the properties mentioned in Lemma 2.2. This was verified in Section 4.1 of Duffield [5]. It is also not hard to verify that these distributions are in $\mathscr{V}$; in fact Pareto and Lognormal are in $\mathscr{V}_0$, whereas Weibull is in $\mathscr{V} \setminus \mathscr{V}_0$. The exact definition of the distributions mentioned are given in [12]—at the moment it suffices to know that $v(t)$ looks roughly like $\log t$ for Pareto sessions, like $(\log t)^2$ for Lognormal sessions, and like t^β for Weibull sessions ($\beta \in (0,1)$).

3. Analysis

We first state the main theorem, which is Theorem 4 in Duffield [5].

Theorem 3.1 (Decay rate for large b). *Under the assumptions of Proposition* 2.1, *Lemma* 2.2, *for all* $D^\star \in \mathscr{V}$,

$$\lim_{b\to\infty} \frac{I(b)}{v(b)} = \phi \quad if\ h = 0, \quad and \quad \lim_{b\to\infty} \frac{I(b)}{v(b)} = (hr)^{-h} \left(\frac{\phi}{1-h}\right)^{1-h} \quad if\ h \in (0,1).$$

Here

$$\phi := \frac{c - \lambda r \mathbb{E}D}{r} = \frac{c-\rho}{r}.$$

Let $t_b^\star$ be the optimizing argument for t in (1). In his proof of this theorem, particularly in the lower bound, Duffield [5] needed that $t_b^\star$ essentially grows linearly in b. Lemma 3.2 states that the time to overflow is indeed *at least* linear in b, for large b. It was proven in case (iii) of [5, p. 258].

Lemma 3.2 (Time to overflow is not sublinear in b). *For all* $D^\star \in \mathscr{V}$ *there exists a positive* d_- *such that*

$$\liminf_{b\to\infty} \frac{t_b^\star}{b} > d_-.$$

Duffield's [5] proof of Theorem 4 (which is equivalent to our Theorem 3.1 above) used that the time to overflow is, in addition to Lemma 3.2, also *at most* linear. However, for $h = 0$ it is not clear whether this statement holds. This can be explained as follows.

Equivalently to $t_b^\star/b < d_+$ eventually (for some finite d_+), we can say that there cannot be a subsequence $s_b := b/t_b^\star$ such that s_b goes to zero. To draw this conclusion, [5] used that $s_b \to 0$ would imply $v(b)/v(b/s_b) \to 0$. However, it holds for $h > 0$, but *not* for $h = 0$. Consider the following counterexample: $v(b) = \log b$ and $s_b = (\log b)^{-1}$. It is easily verified that $v(\cdot)$ is slowly varying, so $h = 0$. However,

$$\lim_{b\to\infty} \frac{v(b)}{v(b/s_b)} = \lim_{b\to\infty} \frac{\log b}{\log(b \log b)} = 1 \neq 0.$$

In other words, for $h = 0$ it is not clear whether there is a d_+ such that

$$\limsup_{b\to\infty} \frac{t_b^\star}{b} < d_+,$$

in fact we will explain in Section 5 that such a d_+ does not exist. Therefore, for the case $h = 0$ the proof in Duffield [5] has to be adapted. Below we present an alternative proof.

Proof of Theorem 3.1. As said, we focus on $h=0$. The proof of the upper bound is given in [5]. We give the lower bound. Choose an arbitrary $\delta\in(0,r^{-1})$. Due to Proposition 2.1

$$I(b)=\inf_{t\in\mathbb{N}} v(t)\sup_{\theta}\left(\theta\left(\frac{b}{t}+c\right)-\frac{\log \mathbb{E}e^{\theta A(t)v(t)/t}}{v(t)}\right)$$

$$\geqslant \inf_{t\in\mathbb{N}} v(t)\left((r^{-1}-\delta)\left(\frac{b}{t}+c\right)-\frac{\log \mathbb{E}e^{(r^{-1}-\delta)A(t)v(t)/t}}{v(t)}\right).$$

If b is large, then $t_b^\star$ will be large due to Lemma 3.2; in fact $t_b^\star > bd_-$ eventually. So, applying Lemma 2.2 for any $\varepsilon>0$ for b large enough

$$I(b)\geqslant \inf_{t\in\mathbb{N}} v(t)\left((r^{-1}-\delta)\left(\frac{b}{t}+c\right)-\rho(r^{-1}-\delta)(1+\varepsilon)\right). \tag{2}$$

This yields

$$\liminf_{b\to\infty}\frac{I(b)}{v(b)}\geqslant \liminf_{b\to\infty}\inf_{t>bd_-}\left(\frac{v(t)}{v(b)}\right)(r^{-1}-\delta)\left(\frac{b}{t}+c-\rho(1+\varepsilon)\right)$$

$$\geqslant (r^{-1}-\delta)\liminf_{b\to\infty}\left(\inf_{t>bd_-}\left(\frac{v(t)}{v(b)}\right)\inf_{t>bd_-}\left(\frac{b}{t}+c-\rho(1+\varepsilon)\right)\right).$$

Due to the fact that $v(\cdot)$ is slowly varying (and monotone increasing),

$$\frac{v(t)}{v(b)}\geqslant 1-\eta$$

for arbitrary positive η, b large enough, and $t>bd_-$. Noticing that $b/t\geqslant 0$, and letting $\delta\downarrow 0$ and $\varepsilon\downarrow 0$, we get

$$\liminf_{b\to\infty}\frac{I(b)}{v(b)}\geqslant\frac{c-\rho}{r}=\phi,$$

which proves the stated. □

4. Refinement

In this section, we give a refinement of Theorem 3.1, for the case of Pareto distributed durations.

Theorem 4.1. *For Pareto(α) on-times, i.e.,* $\mathbb{P}(D^\star>t)=Kt^{-\alpha+1}$,

$$I(b)-v\left(\frac{b\log b}{c-\rho}\right)\left(\frac{c-\rho}{r}\right)\left(1+\frac{1}{\log b}\right)\to 0$$

for $b\to\infty$. *Equivalently*,

$$I(b)-(\alpha-1)\left(\frac{c-\rho}{r}\right)(\log b+\log\log b)$$

$$\to -(\alpha-1)\left(\frac{c-\rho}{r}\right)(\log(c-\rho)-1)-\left(\frac{c-\rho}{r}\right)\log K.$$

Proof. Our proof consists of a lower bound and an upper bound. Starting point of both is the variational problem (1). The lower bound is derived by 'guessing' the θ and performing the optimization over t—in

the upper bound the value of t is 'guessed', and given this t the supremum over θ is computed. Notice that for the Pareto distribution Lemma 2.2 applies (with $\Gamma(n) := \lfloor n/(1+\log(1+\log n))\rfloor$, see [5]).

Lower bound. Start with bound (2), and plug in $v(t) = -\log K + (\alpha-1)\log t$. The infimum over $t \in \mathbb{N}$ is of course larger than the infimum over $t > 0$. Differentiation to t gives the first order condition:

$$f(t) := (\alpha-1)\left(\frac{b}{t} + c - \rho(1+\varepsilon)\right) + (\log K - (\alpha-1)\log t)\frac{b}{t} = 0. \tag{3}$$

We now show that the optimizing t, i.e., $t_b^\star$, looks like $b\log b/(c-\rho)$. To conclude this, first plug

$$t = t_b^+ := (1+\varepsilon)\frac{b\log b}{c-\rho}$$

into (3), $\varepsilon > 0$. It is not hard to show that

$$f(t_b^+) = (\alpha-1)c\varepsilon + g_\varepsilon(b),$$

with a function $g_\varepsilon(b) \to 0$ as $b \to \infty$. Hence $f(t_b^+)$ is positive for large b. Also, with $t = t_b^- := (1-\varepsilon)(b\log b)/(c-\rho)$,

$$f\left(t_b^-\right) = -(\alpha-1)c\varepsilon + h_\varepsilon(b),$$

with $h_\varepsilon(b) \to 0$ for $b \to \infty$. This implies that $f(t_b^-)$ is negative for large b. In other words: between t_b^- and t_b^+ the first derivative shifts from negative to positive. Hence, the minimum is achieved somewhere between these epochs:

$$t_b^\star = (1+\varepsilon^\star)\frac{b\log b}{c-\rho}, \quad \text{with } \varepsilon^\star \in [-\varepsilon, \varepsilon].$$

We arrive at

$$I(b) \geqslant \left(r^{-1} - \delta\right) v\left(\frac{b\log b}{c-\rho}(1+\varepsilon^\star)\right)\left(\left(\frac{c-\rho}{(1+\varepsilon^\star)\log b} + c\right) - \rho(1+\varepsilon)\right).$$

Now let $\delta \downarrow 0$ and $\varepsilon \downarrow 0$.

Upper bound. Define

$$f_t(\theta) := \frac{\log \mathbb{E}\mathrm{e}^{\theta A(t) v(t)/t}}{v(t)}.$$

Make the following observations:

- It is straightforward that $f_t'(0) = \rho$ for all t. Also it is well known that $f_t(\cdot)$ is a convex function. Therefore $f_t(\theta) \geqslant \rho\theta$ for positive θ. Let on the interval $\theta \in [0, r^{-1} + \delta]$ the function $h_t(\theta)$ be defined as $\rho\theta$.
- The probability that k sessions are present at time 0 has a Poisson distribution with mean λ/μ. Therefore

$$\mathbb{E}\mathrm{e}^{\theta A(t) v(t)/t} \geqslant \sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k \frac{\mathrm{e}^{-\lambda/\mu}}{k!}\mathbb{P}(D^\star > t)^k \mathrm{e}^{\theta r v(t) k},$$

yielding

$$f_t(\theta) \geqslant \frac{\lambda}{\mu}\frac{\mathrm{e}^{(\theta r - 1)v(t)} - 1}{v(t)}.$$

This expression goes to ∞ for all $\theta > r^{-1}$ as $t \to \infty$, and is exponentially increasing in θ. This means that for arbitrary $\zeta > 0$ and t large enough $f_t(\theta)$ will be larger than $\theta(\zeta + c)$, with $\theta > r^{-1} + \delta$. Define for these θ the function $h_t(\theta)$ as $\theta(\zeta + c)$.

Define

$$t_b := \left\lceil \frac{b \log b}{c-\rho} \right\rceil .$$

Clearly,

$$I(b) \leqslant v(t_b) \sup_\theta \left(\theta \left(\frac{b}{t_b} + c \right) - f_{t_b}(\theta) \right) \leqslant v(t_b) \sup_\theta \left(\theta \left(\frac{b}{t_b} + c \right) - h_{t_b}(\theta) \right).$$

Now take b large, such that $b/t_b < \zeta$. Then the supremum will be attained for $\theta = r^{-1} + \delta$, leading to

$$I(b) \leqslant v(t_b)(r^{-1} + \delta) \left(\frac{b}{t_b} + c \right) - \rho(r^{-1} + \delta).$$

Notice that

$$v(t_b) - v\left(\frac{b \log b}{c-\rho} \right) \to 0 \quad \text{and} \quad v(t_b) \frac{b}{t_b} - v\left(\frac{b \log b}{c-\rho} \right) \frac{c-\rho}{\log b} \to 0$$

as $b \to \infty$. The upper bound follows by letting $\delta \downarrow 0$. □

Remark (On − off input)**.** Consider n sources feeding into a buffer that is emptied at rate nc. The sources have peak rate r and mean rate ρ. Let $I(b)$ be the decay rate of the probability of the buffer exceeding level nb. Then for Pareto(α) on-times, i.e., $\mathbb{P}(A^\star > t) = Kt^{-\alpha+1}$, it was proven by Mandjes and Borst [10] that

$$\frac{I(b)}{v(b)} \to \frac{c-\rho}{r-\rho} \quad \text{as } b \to \infty.$$

This can be refined, similarly to the above as

$$I(b) - v\left(\frac{b \log b}{c-\rho} \right) \left(\frac{c-\rho}{r-\rho} \right) \left(1 + \frac{1}{\log b} \right) \to 0$$

for $b \to \infty$. Equivalently,

$$I(b) - (\alpha - 1) \left(\frac{c-\rho}{r-\rho} \right) (\log b + \log\log b)$$

$$\to -(\alpha - 1) \left(\frac{c-\rho}{r-\rho} \right) (\log(c-\rho) - 1) - \left(\frac{c-\rho}{r-\rho} \right) \log K.$$

This result was used in [16], where we studied the queueing behavior of systems with *heterogeneous* on–off input with regularly varying on-times.

5. Interpretation

This section treats the intuition behind the analysis for $h = 0$.

Time to overflow is superlinear. As shown above, for $h = 0$, it was not clear whether there is a d_+ such that $t_b/b < d_+$ eventually. Duffield [5] showed in case (ii) in the proof of the lower bound of Theorem 4 of [5] that there exists such a d_+ for $h \in (0,1)$. This touches on a crucial distinction between the cases $h = 0$ and $h \in (0,1)$.

- Let us first consider $b/t_b^\star$ for b large and $h = 0$. In the proof of Theorem 3.1 we saw that $t_b^\star$ is such that if we choose $b/t_b^\star = 0$ the lower bound is attained. This suggests that $t_b^\star$ is *superlinear* in b. On the other hand, $t_b^\star$ is such that $v(t_b^\star)/v(b) \to 1$ as $b \to \infty$. This means that $t_b^\star = bf(b)$ for some 'subpolynomial'

function $f(\cdot)$ (i.e., $f(b)<b^{\varepsilon}$ for all $\varepsilon>0$ and b large enough). One could think of $t_b^{\star} \sim b\log b$ (cf. the proof of Theorem 4.1).

- Now in the case $h\in(0,1)$, $t_b^{\star}$ will be essentially linear in b, as derived in [5]. In other words: the time to overflow will be proportional to the buffer size.

Heuristic. Due to the long tail, for $D^{\star}\in\mathscr{V}$ it is 'low-cost' to have overflow due to sessions that remain in the system during the entire path to overflow. As the buffer is large, in addition to these sessions, traffic is generated at a rate $\Lambda r\mathbb{E}D$. This gives rise to the following heuristic:

$$p(B,C)\approx\max_{K\in\mathbb{N}}\mathbb{P}\left(D^{\star}\geqslant\frac{B}{Kr+\Lambda r\mathbb{E}D-C}\right)^{K},$$

where the optimizing K can be interpreted as the most likely number of sessions that stay in the system during the complete trajectory to overflow. The above heuristic can be considered as a variant of the *reduced load* equivalence proven in [1].

The justification of this heuristic is the following. Put $K\equiv nk$, and use the scaling $\Lambda\equiv n\lambda$, $B\equiv nb$, and $C\equiv nc$. Then, following the heuristic, with $\rho:=\lambda r\mathbb{E}D$,

$$\begin{aligned}\frac{1}{n}\log p_n(b,c)&\approx-\min_{k:kr+\rho>c}kv\left(\frac{b}{kr+\rho-c}\right)\\&\approx-\min_{k:kr+\rho>c}k(kr+\rho-c)^{-h}v(b).\end{aligned}\tag{4}$$

The minimum is attained for

$$k^{\star}=\frac{c-\rho}{r(1-h)}.\tag{5}$$

Inserting this $k^{\star}$ into (4) indeed leads to the decay rate of Theorem 3.1.

- Interestingly, for $h=0$ we get from (5) that the input rate is roughly equal to c:

  $$k^{\star}r+\rho=\frac{c-\rho}{r}r+\rho=c.$$

  This is in agreement with the superlinear time to overflow. During the path to overflow, the input rate only slightly exceeds link rate c. One could think for instance of $t_b^{\star}\sim b\log b$, such that the input rate is in the order of $c+(\log b)^{-1}$.

- If $h>0$ the input rate is strictly larger than c, leading to a time to overflow that is essentially linear in b:

  $$t_b^{\star}\sim\frac{1}{c-\rho}\frac{1-h}{h}b.$$

  Notice that for $h\uparrow 1$ the time to overflow will be extremely short.

Summarizing, if $h\in(0,1)$ long sessions are so 'low-cost' that the most likely path to overflow is such that some sessions remain in the system during the entire busy period preceding overflow. If $h=0$ long sessions have even 'lower cost': some sessions remain in the system during the entire path to overflow, but at the same time the net input rate is only slightly positive (which requires *extremely* long sessions to cause overflow).

Path to overflow. In Fig. 1 we show the decay rate as a function of the buffer size, i.e., $I(b)$. It reflects the crucial differences caused by the shape of the session duration distribution. For Pareto sessions and b large, $I(b)$ looks like $\log b$, for Weibull sessions like b^{β}, and for exponential sessions it is proportional to b. Lognormal sessions would lead to an $I(b)$ curve that is similar to $(\log b)^2$. Fig. 2 depicts the most likely time to overflow. We see the superlinear behavior for Pareto sessions.

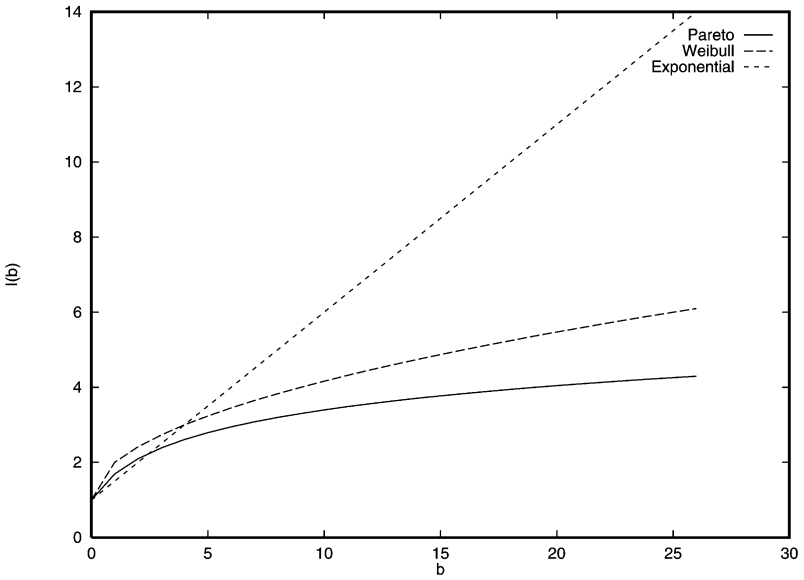

Fig. 1. Loss curve (decay rate $I(b)$ as function of b).

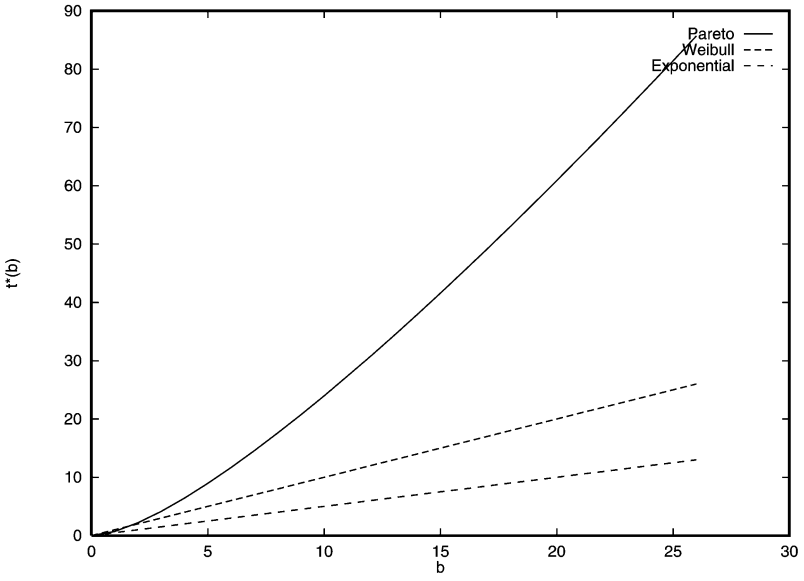

Fig. 2. Time to overflow ($t_b^\star$ as function of b).

References

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

[2] N. Bingham, C. Goldie, J. Teugels, Regular Variation, Encyclopedia of Mathematics and its Applications, Vol. 27, Cambridge University Press, Cambridge, 1987.

[3] N. Duffield, Economies of scale in queues with sources having power-law large deviation scalings, J. Appl. Probab. 33 (1996) 840–857.
[4] N. Duffield, On the relevance of long-tailed durations for the statistical multiplexing of large aggregations, Proceedings of the Annual Allerton Conference on Communications, Control and Computation, 1996.
[5] N. Duffield, Queueing at large resources driven by long-tailed $M/G/\infty$-modulated processes, Queueing Syst. 28 (1998) 245–266.
[6] N. Duffield, N. O'Connell, Large deviations and overflow probabilities for the general single server queue, with applications, Proc. Cambridge Philos. Soc. 118 (1995) 363–374.
[7] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Trans. Networking 2 (1994) 1–15.
[8] N. Likhanov, R. Mazumdar, Cell loss asymptotics in buffers fed by heterogeneous long-tailed sources, Proceedings IEEE Infocom, 2000.
[9] Z. Liu, P. Nain, D. Towsley, Z.-L. Zhang, Asymptotic behavior of a multiplexer fed by a long-range dependent process, J. Appl. Probab. 36 (1999) 105–118.
[10] M. Mandjes, S. Borst, Overflow behavior in queues with many long-tailed inputs, Adv. Appl. Probab. 32 (2000) 1150–1167.
[11] M. Parulekar, A. Makowski, Tail probabilities for a multiplexer driven by $M/G/\infty$ input processes (I): preliminary asymptotics, Queueing Syst. 27 (1997) 271–296.
[12] M. Parulekar, A. Makowski, $M/G/\infty$ input process: a versatile class of models for network traffic, Proceedings IEEE Infocom, 1997.
[13] V. Paxson, S. Floyd, Wide area traffic: the failure of Poisson modeling, IEEE/ACM Trans. Networking 3 (1995) 226–244.
[14] W. Willinger, M. Taqqu, R. Sherman, D. Wilson, Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level, IEEE/ACM Trans. Networking 5 (1997) 71–86.
[15] D. Wischik, Sample path large deviations for queues with many inputs, Ann. Appl. Probab. (2001), to appear.
[16] B. Zwart, S. Borst, M. Mandjes. Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off sources, submitted.