On the relevance of long-tailed durations for the statistical multiplexing of large aggregations. *

N. G. Duffield
AT&T Laboratories
Room 2C-323, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
duffield@research.att.com

Abstract

In this paper we obtain large deviation bounds for the queue length in a buffer fed by an $M/G/\infty$ process with G long-tailed, and in the limit that the arrival intensity grow to infinity along with the service rate so as to keep the offered load constant. This provides an example in which long-tailed arrival distributions are no bar to obtaining statistical multiplexing gain when such arrivals are sufficiently spread out over time.

1 Introduction.

In this paper we investigate the extent to which long tailed behavior in the activity of sources make them difficult to statistically multiplex together. The word *together* is crucial here, because it is easier to achieve multiplexing gain across several sources than within a single source.

To achieve gain within a source at a buffer requires being able to allocate to it a service rate (between its peak and mean rate) in such a way that statistical fluctuations in the arrivals overflow the buffer sufficiently rarely. Much recent work has been devoted to finding large deviation approximations to the probability that, in an infinite buffer, the queue length Q exceeds a given level b. Such approximations can be written informally as

$$\mathbf{P}[Q > b] \approx e^{-h(b)\delta}, \tag{1}$$

where the scaling function h is determined by the arrival process. For a large class of short range dependent processes (see [11, 10] and references in [4]) h is linear, and so the tail is exponential. For class of long-range dependent arrival processes, including fractional Brownian motion, h is a power law and so the tail is Weibullian (see [8] and a previous lower bound in [15]).

The $M/G/\infty$ process has recently attracted some attention as a model of arrivals (see e.g. [1, 13, 18]). It is a random marked Poisson process of some intensity r the marks of which are i.i.d. random variables with common distribution G with mean σ. The arrival times are those of customers, each of which commences service immediately for a time whose duration is the corresponding mark. Let N_t be number of customers in the system at time t and define $A_t =$

*To appear in the Proceedings of the 34th Annual Allerton Conference on Communication, Control, and Computing, October 2–4, 1996.

$\int_0^t dt' N_{t'}$. The interpretation of A as an arrivals process is that each of the N_t customers present at time t deposits fluid into the buffer at unit rate (or in discrete time, packets of unit size at each timeslot). The fluid is drained out at rate s. Set $\phi = s - r\sigma > 0$ for stability.

Parulekar and Makowski [16] have determined an asymptotic of the form (1) for this model when G is short tailed. Building on their work, we find a large deviation *upper bound* when G is Pareto with power $\alpha \in (1, 2)$:

$$\limsup_{b\to\infty} \frac{\log \mathbf{P}[Q > b]}{\log b} \leq 1 - (\alpha - 1)\phi, \tag{2}$$

when the RHS is negative. We have not determined whether this bound is tight: as we discuss in section 4 below, the large deviation lower bound does not obtain in this case, at least not by use of the methods in [8]. Observe though that the exponent in (2) is different to that which would be found in the $M/G/1$ model or On-Off Model (see [5]) with the same service distribution, namely an exponent of $1 - \alpha$. That there is difference is not surprising, since large deviations in the two models occur by different mechanisms. For the $M/G/1$ model, since G is long tailed, the most likely way for an (asymptotically large) backlog of work to arise is for one of the service periods to be large. However, in the $M/G/\infty$ model, to develop a backlog of size b requires that a number $n > s$ of sources to be active during a period of duration at least $b/(n - s)$: some "conspiracy" between arrivals is required. (This observation can be used to formulate lower bounds for $\mathbf{P}[Q > b]$).

This leads as to consider a second asymptotic: that of scaling the service rate as sL and the arrival rate as rL (or equivalently, superposing L sources of arrival rate r) with $L \to \infty$. In this limit the load remains independent of L, but the proportional contribution of each active source to that load goes to zero as $L \to \infty$. Scaling a buffer level likewise as Lb so as to keep the buffer size per unit arrival rate constant, we find that the requirement for conspiracy as the means of generating large deviations in the corresponding queue length Q^L becomes manifest: applying a result from [7] we show that

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbf{P}[Q^L > Lb] = -I(b), \quad \text{with} \quad \lim_{b\to\infty} \frac{I(b)}{\log b} = (\alpha - 1)\phi. \tag{3}$$

Such asymptotics arise because for each fixed time t, the distribution of the normalized arrival process $L^{-1}A_t^L$ converges exponentially fast to its mean $r\sigma t$ as $L \to \infty$. From (3) we have the approximation

$$\mathbf{P}[Q^L > Lb] \approx e^{-LI(b)} \approx b^{-L(\alpha-1)\phi} \tag{4}$$

for first L and then b sufficiently large. For fixed b the tail probabilities are exponentially small in L, reflecting the multiplexing gain available across sources, while it decreases only algebraically in b: the effect of the long tailed service distribution. It is worth remarking that superpositions of Markovian sources or fractional Brownian Motions have the same exponential decay in L for their tail distributions, though $I(b)$ is asymptotically proportional to δb or δb^{2-2H} respectively (here H is the Hurst parameter of the fractional Brownian Motion).

That the long tailed on duration of the sources should not prevent statistical multiplexing gain even for long tailed active periods can be seen by observation similar to those made for the large b asymptotic. A small number of sources may effectively remove some of the service capacity during their long active periods, but it requires a large number of them to be active in order for a large deviation in the buffer occupancy proportional to the aggregate arrival rate. Such a heuristic does

not evidently apply to the corresponding $M/G/1$ model: in this case work arrives instantaneously in long tailed chunks, rather than being spread out over the duration of the active period: just one such arrival can use up the available service rate. Of course, from the point of view of modeling, one will chose $M/G/1$ or $M/G/\infty$ as appropriate: there is no necessary relation between them. But although both can be seen as members of the class of "arrival processes with long-tailed behavior", they have radically different properties under statistical multiplexing. In fact, rather than the decay of (4), $\mathbf{P}[Q^L > Lb]$ decays only algebraically in L as $L \to \infty$ for $M/G/1$.

The contents of the paper are as follows. In section 2 we recapitulate the large deviation lower bounds for the large buffer asymptotics for a single arrival process. In section 3 we extend a result in [8] to derive an upper bound for the large buffer asymptotics which we can apply to the long tailed $M/G/\infty$ process in section 4. We conclude in section 5 by deriving the tail asymptotics for the $M/G/\infty$ model as the arrival rate is increased to infinity while keeping the service duration and offered load constant.

Finally, we mention that although we have dealt here with loss probabilities, the duration of loss periods is also important for understanding the impact of long tailed durations. This has been studied in a related model in [1]. One expects that long tailed active period will give rises to long duration of periods of overload. However, for small overshoots of a given threshold, in bufferless models at least, the duration is overload can be largely independent of the duration of the active period beyond its mean: see [9] for more details.

2 Large deviation lower bounds

Let us formulate in more detail a large deviation result the tail asymptotics of the queue length in an infinite buffer which is fed by a stationary arrivals process and served at some constant rate s. Index time by $T = \mathbf{R}_+$ or $\mathbf{Z}_+$ and let A_t be the work arriving in the queue in the interval $[-t, 0)$ and define the (backward) excess workload process W by $W_0 = 0$ and $W_t = A_t - st$. Then we have the pathwise relation for the queue length Q at time 0 (see [3])

$$Q = \sup_{t \geq 0} W_t. \tag{5}$$

Here we shall be concerned with obtaining bounds of the form

$$-\delta_+ \leq \liminf_{b \to \infty} \frac{\log \mathbf{P}[Q > b]}{h(b)} \leq \limsup_{b \to \infty} \frac{\log \mathbf{P}[Q > b]}{h(b)} \leq -\delta_- \tag{6}$$

for some constants $\delta_\pm$ depending on the processes, and where h is a *scaling function*, namely a positive function on T which increases to $+\infty$. If $\delta_- = \delta_+ = \delta$ then we can informally write the approximation (1). Common to the works cited in the introduction is the idea of characterizing the large deviation properties of the queue length in term of those of the workload process.

Recall [6] that a *rate function* on $\mathbf{R}_+$ is a lower semicontinuous function taking values in $[0, \infty]$, and a *good* rate function is one with compact level sets. Let $(X_t : t \geq 0)$ be a real stochastic process and v a scaling function. Then the pair $(X_t, v(t))$ is said to satisfy a *large deviation principle* with rate function I if for all Borel subsets B of $\mathbf{R}_+$

$$-\inf_{x \in B^0} I(x) \leq \liminf_{t \to \infty} \frac{1}{v(t)} \log \mathbf{P}[X_t \in B] \leq \limsup_{t \to \infty} \frac{1}{v(t)} \log \mathbf{P}[X_t \in B] \leq -\inf_{x \in \bar{B}} I(x), \tag{7}$$

where B^o is the interior and $\bar{B}$ the closure of B.

According to the Gärtner-Ellis Theorem (see e.g. [6]), to conclude that $(X_t, v(t))$ satisfies and LDP it is sufficient to demonstrate the existence of the cumulant generating function (CGF, also called the log moment generating function) $\lambda(\theta) = \lim_{t\to\infty}(1/v(t)) \log \mathbf{E}[\exp(\theta v(t) X_t)]$, which in addition satisfies a (sufficient) technical condition, that it be essentially smooth. Then $(X_t, v(t))$ satisfies (7) with rate function equal to the Legendre transform λ^* of λ i.e.

$$\lambda^*(x) = \sup_{x\in\mathbf{R}}\{\theta x - \lambda(\theta)\}. \tag{8}$$

In the present case, define the transient CGF

$$\lambda_t(\theta) = \frac{1}{v(t)} \log \mathbf{E}[\exp(\theta v(t) W_t / a(t))]. \tag{9}$$

Theorem 1 *Assume*

(i) The limit $\lambda(\theta) = \lambda_t(\theta)$ exists as an extended real number for all $\theta \in \mathbf{R}$ and is essentially smooth.

(ii) $\lambda(\theta) < 0$ for some $\theta > 0$.

(iii) There exists a scaling function h such that the limit

$$f(c) = \lim_{t\to\infty} \frac{h(t)}{v(a^{-1}(t/c))} \tag{10}$$

exists and is finite for all $c > 0$ where $a^{-1}(x) = \sup\{t \in T : a(s) \le x\}$.

Then the lower bound in (6) holds with

$$\delta_+ \equiv \inf_{c>0} \frac{\lambda^*(c^+)}{f(c)} \quad \textit{which is equal to} \quad \inf_{c>0} \frac{\lambda^*(c)}{f(c)}, \tag{11}$$

if f is continuous, or the effective domain of λ^* is unbounded to the right.

Remark: One case in which (10) obtains easily is when $v \circ a^{-1}$ is *regularly varying* (see [2] for more details). In this case taking $h = v \circ a^{-1}$ we get (10) with $f(c) = c^{\tilde{f}}$ where the power $\tilde{f}$ is the *index* of the regularly varying function $v \circ a^{-1}$. Indeed, if $v \circ a^{-1}$ *is* regularly varying, this is the only choice of h which gives a non-trivial f.

Proof of Theorem 1. [8] gives us the lower bound in (6) with δ_+ as in (11). To remove the necessity of taking right limits, observe that being a convex function, λ^* is continuous on the interior of its effective domain (see Theorem 10.1 of [19]). Hence, as noted in [16], if the effective domain of λ^* contains $\mathbf{R}_+$, the second formulation of δ_+ in (11) applies. If not, the effective domain contains $[0, \bar{c})$ for some $\bar{c} \ge 0$. Due to (ii), λ^* is non-decreasing on $\mathbf{R}_+$. Hence

$$\inf_{c>0} \lambda^*(c^+)/f(c) = \inf_{0<c<\bar{c}} \lambda^*(c)/f(c) \tag{12}$$

where $\bar{c}$ is the boundary of the effective domain of λ^*. The second form of δ_+ now follows since f is continuous, and λ^* is continuous from the left at $\bar{c}$, being non-decreasing on $\mathbf{R}_+$ and lower semicontinuous. ■

3 A Further Large Deviation Upper Bound.

Turning now to large deviation upper bounds for $\mathbf{P}[Q > b]$ we first formulate a bound independently of statistical assumptions on the workload process W. We work here in discrete time: extensions to continuous time can be made using argument similar to those in [8].

Theorem 2 *Let Δ be some function $\mathbf{R}_+ \to \mathbf{R}_+$, a, v, h scaling functions and λ_t as in (9).*

(i) Assume for all t that $\lambda_t(\theta) > 0$ for $\theta < 0$.

$$\limsup_{b\to\infty} \frac{\log \mathbf{P}[Q > b]}{h(b)} \leq \max\left[-\delta_- + \limsup_{b\to\infty}(\log \Delta(b))/h(b), \limsup_{b\to\infty} \frac{1}{h(b)} \log \sum_{t>\Delta(b)} \exp(v(t)\lambda_t^*(b/a(t))\right], \tag{13}$$

where

$$\delta_- = \liminf_{b\to 0} \frac{I(b)}{h(b)}, \quad \text{with} \quad I(b) = \inf_{t>0} v(t)\lambda_t^*(b/a(t)). \tag{14}$$

(ii) If, furthermore, Δ can be chosen such that

$$\lim_{t\to\infty} \frac{\log \Delta(b)}{h(b)} = 0, \tag{15}$$

and for all $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ sufficiently small

$$\lim_{\varepsilon\to 0} \lim_{t\to\infty} \frac{1}{h(b)} \log \sum_{t>\Delta(b)} e^{-\varepsilon_1 v(t)(\lambda^*(0)-\varepsilon_2)} \leq -\delta_-, \tag{16}$$

then

$$\limsup_{b\to\infty} \frac{1}{h(b)} \log \mathbf{P}[Q > b] \leq -\delta_-. \tag{17}$$

Proof of Theorem 2. (i) For any $t_0 > 0$, by using (5) to decompose $\{Q > b\} = \cup_{t>1}\{W_t > b\}$ and applying Chebychev's inequality on each component to get $P[W_t > b] \leq e^{-v(t)b\theta/a(t)}$ $\mathbf{E}[\exp(\theta v(t) W_t/a(t))]$ for any $\theta > 0$, and so taking the infimum over $\theta > 0$ taking the infimum over $\theta > 0$ (which by the assumption on λ_t is equivalent to taking the infimum over all $\theta \in \mathbf{R}$) we get

$$\mathbf{P}[W_t > b] \leq \exp(-v(t)\lambda_t^*(b/a(t))). \tag{18}$$

Hence

$$\mathbf{P}[Q > b] \leq \Delta(b) \max_{t\leq\Delta(b)} \mathbf{P}[W_t > b] + \sum_{t>\Delta(b)} \mathbf{P}[W_t > b] \tag{19}$$

$$\leq \Delta(b)e^{-I(b)} + R(b)\ , \quad \text{where} \quad R(b) = \sum_{t>\Delta(b)} e^{-v(t)\lambda_t^*(b/a(t)}. \tag{20}$$

The result the follows by applying the principle of the largest term (see [12]). The proof of (ii) follows trivially. ■.

It is worth restating that Theorem 2 (ii) follows from (i) only when the tailsum term in (13) can be neglected w.r.t. δ_-. Two questions naturally arise from the preceding results. The first question is what are examples of processes that satisfy the hypothesis. The second is when are δ_- and δ_+ equal so that the bounds (6) are tight. All the examples considered in [8] satisfy the hypotheses of Theorem 1. What is new here are the growth conditions (15) and (16): they replace stronger conditions in [8], and as we shall see in the next section, admit a wider class of examples. Here we remark that $\delta_- = \delta_+$ in the examples of [8]. This includes, for instance, asymptotically self-similar processes with Hurst parameter $H \in [1/2, 1)$. In this case we should take $\tilde{a} = 1$, $\tilde{v} = 2 - 2H$, and Theorem 2(ii) follows by choosing $\Delta(b) = b \log b$.

Before leaving the general case, we prove the following inequality between δ_+ and δ_- which holds even in the absence of the lower bound in (6).

Theorem 3 *Assume for scaling functions v, a, h that (10) is satisfied for some f and that λ_t is pointwise convergent to λ. Then*

$$\delta_- = \liminf_{b\to\infty} \frac{I(b)}{h(b)} \leq \limsup_{b\to\infty} \frac{I(b)}{h(b)} \leq \delta_+ = \inf_{x>0} \frac{\lambda^*(x)}{f(x)}. \tag{21}$$

The proof uses the following Lemma, which is proved in [7].

Lemma 1 *Let χ_n be a sequence of convex functions on $\mathbf{R}$. If for some function χ, $\lim_{n\to\infty} \chi_n = \chi$ pointwise on the interior of the effective domain of χ, then $\lim_{n\to\infty} \chi_n^* = \chi^*$ pointwise on the interior of the effective domain of χ^*.*

Proof of Theorem 3. Only the second inequality has to be proved: the rest is definitions. Setting the dummy variable in the variational expression (14) for I to be $c = b/a(t)$, then any $x > 0$ we have for any $b, c > 0$ $I(b)/h(b) \leq f(b/c))\lambda^*_{a^{-1}(b/c)}(c)/h(b)$. Taking $b \to \infty$ for fixed c, $\limsup_{b\to\infty} I(b)/h(b) \leq \lambda^*(c)/f(c)$, for any c in the interior of the effective domain of λ^*, by Lemma 1 and (10). The result follows by taking the infimum over all $c > 0$ by virtue of (12) ■.

4 Applications to an $M/G/\infty$ arrival process.

Parulekar and Makowski [16] have determined the CGF for A_t for the discrete version of the $M/G/\infty$ process for classes of G and have used it the result of [8] to determine the tail asymptotics as in (6) and (17) with $\delta_+ = \delta_-$ for a number of models in which G is short tailed. In particular, in determining scaling functions a, v and appropriate for the CGF to exist, they take $a(t) = t$, and then derive the scaling function v directly from the model in question:

$$v(t) = -\log G_e^c(t)/dt \tag{22}$$

where G_e is the stationary excess distribution for the service time.

However, the long tailed case has features which take it out of the domain of Theorem 1 and (sometimes) Theorem 2. One difficulty is with the lower bound and stems from the form of the CGF λ. For $v(t) = o(t)$ it is shown in [17] that:

Theorem 4 *Assume $v(t) = o(t)$ with $v(t)/t, t = 1, 2, \ldots\}$ eventually monotone decreasing. Assume that there exists a mapping $\Gamma : \mathbf{N} \to \mathbf{N}$ such that (i) $\Gamma(t) < t$ for all t, (ii) $\lim_{t\to\infty} v(t)\Gamma(t)/t =$*

∞ *and (iii)* $\lim_{t\to\infty} v(t)\Gamma(t)/(t\,v(\Gamma(t))) = 0$. *Then the limiting* CGF $\lambda(\theta) = \lim_{t\to\infty}(1/v(t))\log \mathbf{E}[\exp(v(t)W_t/t)]$ *exists and is given by*

$$\lambda(\theta) = \begin{cases} -\theta\phi & \text{if} \quad \theta < 1 \\ +\infty & \text{if} \quad \theta > 1 \end{cases}, \tag{23}$$

with $\lambda(1)$ *determined by further detail of the model in question.*

Now, this λ is not essentially smooth since it fails to be steep: $|\lambda'(\theta)|$ does not approach ∞ as θ approaches the boundary of its effective domain, namely 1. So the LD lower bound cannot be established by Theorem 1, However, things are even worse. Clearly, whatever the value of $\lambda(1) \geq \lim_{x\nearrow 1}\lambda(x)$ we have

$$\lambda^*(x) = \begin{cases} +\infty & \text{if} \quad x < -\phi \\ x+\phi & \text{if} \quad x > -\phi \end{cases}, \tag{24}$$

Thus, the only exposed point of λ^* (see Section 2.3 of [6] for terminology) is $-\phi$ and so the large deviation lower bound for the pair $(W_t/t, v(t))$ obtained through the Gärtner-Ellis theorem is trivial: for open sets G not containing $-\phi$: $\liminf_{t\to\infty}\log \mathbf{P}[W_t/t \in G]/v(t) \geq -\infty$. This is not to say that there is no non-trivial LDP, just that the Gärtner-Ellis theorem cannot be used to prove it. (See Remark (d) on page 46 of [6] for an example of a process with an LDP which also cannot be derived from this Theorem).

We now turn to the upper bound. As has been pointed out in [16], the hypotheses used in the proof in [8] of the LD upper bound in (6) are violated when $v(t) \sim \log t$. Unfortunately this precludes applying the theorems to the case of long tailed G. For example, when G is Pareto i.e.

$$\lim_{t\to\infty} \frac{G^c(t)}{t^{-\alpha}} = 1, \tag{25}$$

for some $\alpha > 1$, then

$$\lim_{t\to\infty} \frac{v(t)}{(\alpha-1)\log t} = 1. \tag{26}$$

But we are able to do better on the basis of Theorem 2.

Theorem 5 *Consider the* $M/G/\infty$ *process when* G *has Pareto tail of power* $\alpha \in (0,1)$ *and suppose that* $(\alpha-1)\phi > 1$. *Then* $\delta_- = \delta_+$ *and*

$$\limsup_{b\to\infty} \frac{\log \mathbf{P}[Q > b]}{\log b} \leq 1 - (\alpha-1)\phi. \tag{27}$$

Proof of Theorem 5. To recap, we take $a(t) = t$, $v(t) = h(t) = (\alpha-1)\log t$ and hence $f(x) = 1$. To obtain convergence of λ_t we apply Theorem 4 taking, for instance,

$$\Gamma(t) = \lfloor t/(1+\log(1+\log t))\rfloor, \tag{28}$$

Since $a(t) = t$, $\lambda_t'(0) = t^{-1}\mathbf{E}A_t - s = -\phi < 0$ for all t so that assumption (i) of Theorem 2 is satisfied. It is easy to see that λ_t^* is increasing on $\mathbf{R}_+$ as a consequence. Next observe that since $f(x) = 1$ and λ^* is increasing we have $\delta_+ = \phi = \lambda^*(0)$.

Since 0 lies in the interior of the domain of λ^*, we have by Lemmas 1 and the increasing property of λ_t^* on $\mathbf{R}_+$, that, for any $\varepsilon > 0$ $\lambda_t^*(b/a(t)) \geq \lambda^*(0) - \varepsilon = \delta_+ - \varepsilon$ for t sufficiently large. So for b sufficiently large, then in (20) we have $R(b) \leq ((\alpha-1)\phi - 1)^{-1}\Delta(b)^{1-(\alpha-1)\phi}$. Thus by Theorem 2(i)

$$\limsup_{b\to\infty} \frac{\log \mathbf{P}[Q > b]}{h(b)} \leq \max[\beta - \delta_-, \beta(1-(\alpha-1)\delta_+)], \text{ where } \beta = \limsup_{b\to\infty} \frac{\log \Delta(b)}{h(b)}. \tag{29}$$

We now show that δ_- and δ_+ are equal: then the upper bound is minimized by choosing $\Delta(b)$ such that $\beta = 1/(\alpha-1)$ and (27) obtains. By Theorem 3, $\delta_+ \geq \delta_-$, so we need only establish the reverse inequality.

We are free to replace $v(t)$ and $h(t)$ by $(\alpha-1)\log t$ for t greater than some t_0, forming a sequence to which they are asymptotically equivalent. We make the change of variable $c = b/a(t)$ and set $f_b(c) = (\log b)/\log(b/c)$. Then $I(b)/h(b) = \inf_{c>0} \lambda^*_{b/c}(c)/f_b(c) = \lambda^*_{b/c_b}(c_b)/f_b(c_b)$, where the infimum is achieved at c_b. (If the infimum is not achieved, use instead for any $\varepsilon > 0$ a point c_b where the infimum is achieved to within ε, then take $\varepsilon \to 0$ at the end). Then $\liminf_{b\to\infty} I(b)/h(b) = \lim_{b\to\infty} \lambda^*_{b/c_b}(c_b)/f_b(c_b)$ along some subsequence, which we also denote by (c_b). Now (c_b) must have a subsubsequence (which we also denote by (c_b)) with one of the following properties. Either c_b is eventually bounded; or $\lim_{b\to\infty} c_b = \infty$.

In the first case, $c_b < K$ for some $K > 0$ so that $\lim_{b\to\infty} 1/f_b(c_b) \geq 1$ while since λ^* is non-decreasing $\delta_- = \lim_{b\to\infty} \lambda^*_{a^{-1}(b/c_b)}(c_b) \geq \lim_{b\to\infty} \lambda^*_{a^{-1}(b/c_b)}(0) = \lambda^*(0) = \delta_+$ and we are done.

In the case $c_b \to \infty$, either b/c_b is bounded, or it diverges. If it is bounded then so is $t_b = a^{-1}(b/c_b)$, and so for sufficiently large b on the subsubsequence $I(b)/h(b) \geq k_1 \max_{1\leq t\leq t_0} \lambda_t^*(k_2 b)/\log b$, for some constants t_0, k_1, k_2. Clearly λ_t is finite and differentiable on $\mathbf{R}$, so that λ_t^* is strictly convex, and also $x \mapsto \lambda_t^*(x)$ is increasing when $x > s$. Hence for some $k_3 > 0$ $\lambda_t^*(x) > k_3 x$ for large enough x, and so $I(b)/h(b)$ diverges as $b \to \infty$. This contradicts the finiteness of δ_+: there is no such subsequence (c_b).

Finally, we treat the case that c_b and b/c_b are divergent. Then for any $K > 0$, $b > Kc_b$ eventually. Since $b \to 1/f_b(c)$ is increasing we get $1/f_b(c_b) \geq 1 - 1/\log K$ and since K is arbitrary $\liminf_{b\to\infty} 1/f_b(c_b) \geq 1$. Finally $\lambda^*_{a^{-1}(b/c_b)}(c_b) \geq \lambda^*_{a^{-1}(b/c_b)}(0) \to \lambda^*(0)$ as $b \to \infty$ and hence $b/c_b \to \infty$. Thus along the subsubsequence $\lim_{b\to\infty} I(b)/h(b) \geq \lambda^*(0) = \delta_+$ as required. ■

5 Asymptotics for large arrival rates at fixed load.

We now turn to our second large deviation principle, in which the arrival rate is scaled as rL and the service rate as sL where L grows to infinity. Consider first a slightly more general setup in which for each $L > 0$ we have a arrival process $A^L = (A_t^L : t \geq 0)$ served at rate sL. Define the corresponding excess workload process $W_t^L = A_t^L - sLt$ and queue length $Q_t^L = \sup_{t\geq 0} W_t^L$. Finally define the family of CGF's

$$\lambda_t^L(\theta) = (v(t)L)^{-1} \log \mathbf{E}[e^{\theta v(t) W_t^L/t}], \tag{30}$$

for some scaling functions $v(t)$. (Here $a(t) = t$).

Theorem 6 *Assume*

(i) For each $\theta \in \mathbf{R}$, the limit

$$\lambda_t(\theta) = \lim_{L\to\infty} \lambda_t^L(\theta) \tag{31}$$

exists as an extended real number, uniformly in t sufficiently large, and each λ_t is essentially smooth; or

(i′) *W_t^L is an L-fold superposition of i.i.d. processes W_t, so that $\lambda_t^L = \lambda_t$.*

(ii) *For some $\theta, \varepsilon > 0$, $\lambda_t(\theta) < -\varepsilon$ for all t.*

(iii) *For all $\varepsilon > 0$ $\lim_{t_0 \to \infty} \limsup_{L \to \infty} L^{-1} \log \sum_{t > t_0} e^{-\varepsilon L v(t)} = -\infty$.*

Then with $I(b)$ as in (14)

$$\lim_{L \to \infty} L^{-1} \log \mathbf{P}[Q^L > Lb] = -I(b). \tag{32}$$

This theorem was proved under stronger conditions in [7]. Given the uniform convergence in (i), (ii) follows easily if λ_t converges to some λ satisfying Hypothesis (ii) of Theorem 1, at least on the interior of the effective domain of the latter. With an additional technical condition the result extends to continuous time. The canonical example of is when A_t^L is a homogeneous L-fold superposition of an arrival process A_t. This also applies to models with Poissonian arrivals, such as $M/G/\infty$, since scaling the arrival rate by $L \in \mathbf{N}$ in our $M/G/\infty$ models gives the same distribution for A_t^L as taking an L-fold superposition of A_t. We call I the *shape function*: it determines the shape of the log loss curve, asymptotically for large L. Comparing (32) with (14) we see that δ_- is a lower bound for the asymptotic decay rate (in b) of the loss ratio in both the single source and many source cases (when the appropriate large deviation theorems hold).

Most importantly for our present purpose, the hypotheses of Theorem 6 in some sense easier to satisfy than those of Theorem 2. One reason for this is that the transient CGF λ_t is typically essentially smooth in applications, whereas, as we have seen the previous section, the limit λ does not always have this property. The smoothness property plays different roles in each context. In Theorem 1 is is used to establish a temporal LDP for $(W^t/t, v(t))$ as $t \to \infty$. In Theorem 6 it is used to establish the spatial LDP for $(A_t^L/L - st, L)$ at each fixed t: this determines how quickly the Law of Large Numbers is obeyed as $L \to \infty$. This is proved using Gärtner-Ellis, or simply Cramer's theorem if the arrivals are i.i.d. superpositions. Also one can show that (15) and (16) in Theorem 2 imply hypothesis (iii) in Theorem 6.

Theorem 7 *Consider the $M/G/\infty$ arrival process of intensity rL, served at rate sL, where G is Pareto with power $\alpha > 1$. Then the associated workload processes W_t^L satisfy the hypotheses of Theorem 6 and (3) holds as $\mathbf{N} \ni L \to \infty$.*

Proof of Theorem 7. We establish the hypotheses of Theorem 6. For (i′), use the remark above that A_t^L has the distribution of an i.i.d. superposition. For (ii) we have seen in Theorem 5 the convergence of λ_t to λ when $v(t) = (\alpha - 1) \log t$. It is not difficult to see that for each t, λ_t is everywhere defined and hence differentiable. Since A_t is stationary and $a(t) = t$ we see that $\lambda_t'(0) = t^{-1}\mathbf{E}A_t - s = \lambda'(0) < 0$, the last inequality since λ is convex and (say) $\lambda(1/2) < 0$. Thus since λ_t is differentiable $\lambda_t(\theta_t) < 0$ for some $\theta_t > 0, t \in \mathbf{N}$. For each t let θ_t' be the supremum of such θ_t. But $\lim_{t \to \infty} \theta_t' = 0$ would contradict $\lambda(1/2) < 0$ and so (ii) holds. For (iii), by a similar argument to that used in Theorem 5, then for $\varepsilon' > \varepsilon > 1/L$ and sufficiently large t_0:

$$L^{-1} \log \sum_{t > t_0} \exp(-\varepsilon L v(t)) \leq -\varepsilon' t_0 (\alpha - 1) \tag{33}$$

from which (iii) follows. Thus the LDP (32) holds, and the asymptotic from of the shape function follows from Theorem 3 and the fact, established during the proof of Theorem 5 that $\delta_+ = \delta_-$. ■

We conclude by reviewing the behavior of the corresponding $M/G/1$ model with the same service time distribution G. Since the arrivals are Poissonian A_t^L is equal in distribution to A_{Lt} and hence

$$Q^L = \sup_{t>0}(A_t^L - cLt) \stackrel{d}{=} \sup_{t>0}(A_{Lt} - cLt) = \sup_{t\geq 0}(A_t - ct) = Q; \tag{34}$$

Thus $\mathbf{P}[Q^L > Lb] \approx k(bL)^{-(\alpha-1)}$ for some $k > 0$ and large L. This decays only algebraically in L, rather than exponentially as for $M/G/\infty$.

References

[1] V. Anantharam (1995). On the Sojourn Time of Sessions at an ATM Buffer with Long-Range Dependent Input Traffic, *Proceedings of the 34th IEEE Conference on Decision and Control, December 1995.*

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

[3] A.A. Borovkov (1984). *Asymptotic Methods in Queueing Theory.* Wiley, Chichester.

[4] C.-S. Chang (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control*, **39**:913–931.

[5] G.L. Choudhury and W. Whitt (1996). Long-tail buffe-content distributions in broadband networks. Preprint, AT&T Laboratories.

[6] A. Dembo and O. Zeitouni (1993). *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London.

[7] N.G. Duffield (1996). Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Prob.*, **33:** 840–857.

[8] N.G. Duffield and N. O'Connell (1995). Large deviations and overflow probabilities for the general single-server queue, with applications, *Math. Proc. Cam. Phil. Soc.*, **118**:363–374.

[9] N.G. Duffield and W. Whitt (1996). Recovery from rare congestion events in multi-server systems, Preprint, AT&T Laboratories.

[10] P.W. Glynn and W. Whitt (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. In: *Studies in Applied Probability* Eds. J. Galambos and J. Gani, *Journal of Applied Probability, Special Volume* **31A** 131–159

[11] G. Kesidis, J. Walrand and C.S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. *IEEE/ACM Trans. Networking*, **1**:424-428.

[12] J.T. Lewis and C.E. Pfister (1995). Thermodynamic probability theory: some aspects fo large deviations. *Russian Mathematical Surveys*, **50**:2 47-83.

[13] N. Likhanov, B. Tsybakov and N.D. Georganas (1995). Analysis of an ATM buffer with self-similar ("fractal") input traffic, *Proceedings of Infocom95,* Boston, April 1995, pp. 985–992.

[14] B.B. Mandelbrot and J.W. Van Ness (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, **10**:422–437.

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

[16] M. Parulekar and A. Makowski (1996). Tail probabilities for a multiplexer with self-similar traffic. *Proc. IEEE INFOCOM'96, San Francisco, March 26-28, 1996*, pp1452–1459.

[17] M. Parulekar and A. Makowski (1996). Tail probabilities for $M/G/\infty$ input processes(I): preliminary asymptotics. Preprint, University of Maryland.

[18] V. Paxson and S. Floyd (1995). Wide-area traffic: the failure of Poisson modelling. *IEEE/ACM Transactions on Networking*, **3**:226-244.

[19] R.T. Rockafellar (1970) *Convex Analysis.* Princeton University Press, Princeton.