Performance Evaluation 40 (2000) 47–70

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

Performance implications of very large service-time variances [☆]

Daniel P. Heyman [*]
AT&T Labs, Holmdel, NJ 07733, USA

Abstract

Measurements of file sizes transported on the World-Wide-Web have led some researchers to propose describing them by probability distributions with infinite variance. The M/G/1 queue often arises as a performance model for components of the WWW, and the service times correspond to file sizes; the infinite variance of the file sizes becomes the variance of the service times. In this paper the effects of very large service-time variances on some performance measures for the M/G/1 queue are explored via numerical examples and analytic arguments.

The first main conclusion is that it is the form of the service-time distribution over a wide finite range that controls the steady-state queueing performance, so distributions with very large finite variances can yield the same behavior as distributions with infinite variances. The second main conclusion is that very large service-time variances cause the rate of approach to steady-state performance to be so slow that steady-state performance measures are not likely to be of engineering interest. A third conclusion is that a common device of using the probability that the work in an infinite queue exceeds the level b to approximate the probability that a finite buffer of size b overflows may be very inaccurate. The approximation works better for the fat-tailed distributions studied than for the others.

The most important engineering implication of these results is that when service times have a very large variance (such as file transfers on the WWW), performance criteria other than steady-state measures have to be used. ©2000 Elsevier Science B.V. All rights reserved.

Keywords: M/G/1 queue; Fat-tailed distributions; Pareto distribution; Numerical Laplace transform inversion

1. Introduction

The purpose of this paper is to explore the effects of very large service-time variances on some performance measures for the M/G/1 queue. The relevance of these results is that measurements on the WWW indicate that file sizes and file transfer times follow a distribution with power-law tails and a very large (some think infinite) variance [1,2]. When many subscribers share a network resource, a limiting operation (that is described in Section 2) gives rise to an M/G/1 queueing model, where the service time

[☆] An earlier version of this paper was presented at the Performance and Control of Network Systems II Conference, in: Wai Sum Lai, Robert B. Cooper (Eds.), Vol. 3530, Boston, Mass., 2–4 November 1998.

[*] Present address: AT&T Laboratories, Room D5-3C35, 200 Laurel Avenue, Middletown, NJ 07748, USA.
E-mail address: dph@buckaroo.att.com (D. P. Heyman)

0166-5316/00/$ – see front matter ©2000 Elsevier Science B.V. All rights reserved.
PII: S0166-5316(99)00069-3

inherits the distribution of the file sizes. Consequently, the M/G/1 queue with a very large variance for the service times may be appropriate for sizing network elements handling WWW traffic. The goal of this paper is to discover which properties of distributions with power-law tails and large variances are important for performance analysis.

The M/G/1 queue is often the first queueing model one encounters with some distribution in it other than the exponential. Among the first results one learns is the Pollaczek–Khintchine formula for the steady-state mean queue size, from which it follows that the variance of the service times has a first-order effect on the mean delay. Moreover, the *mean* delay is infinite when the *variance* of the service times is infinite, so service-time distributions with very large variances may have a profound effect on the delay distribution. The empirical evidence that appropriate models for WWW resources should use service times with very large variances suggested the following three questions to me.

- When the service times have power-law tails and a very large variance, does the steady-state queue-size distribution (in the range of engineering interest) differ significantly from the same model with an infinite variance for the service times?
- Does the queue-size distribution approach its steady-state value quickly enough for the latter to be relevant for engineering decisions?
- Is the steady-state $P\{\text{buffer contents} > b\}$ for an M/G/1 queue with an infinite buffer a good approximation to the overflow probability for the same queue with a finite buffer of size b?

The rest of the paper is devoted to formulating these questions precisely, and using theoretical analyses and numerical examples to answer them. For the first two questions, we examine how fat tails affect the complementary distribution of the steady-state work-in-system and the convergence rate of the expected queue length at time t, respectively. For those who cannot wait to learn the answers, they are no, no, and sometimes.

The rest of this paper is organized as follows. Section 2 contains some background information, including a heuristic derivation of the M/G/1 queueing model and some basic facts about fat-tailed distributions. The next three sections explore the three questions above (in turn), and my conclusions are given in the final section.

2. Background material

In this section, I explain how the M/G/1 queue with service-times having power-law tails arises from WWW traffic with file sizes with power-law tails. Then some examples of fat-tailed distributions are given and compared.

2.1. Derivation of the M/G/1 model

Consider a source that alternates between busy and idle phases. The sojourn times in the busy phases are i.i.d. random variables, and so are the sojourn times in the idle state, and all of the sojourn times are independent of one another. When the source is busy, it generates traffic at a constant rate, r say. When the source is idle, it does not generate any traffic. Let F be the distribution function of an arbitrary sojourn time in the idle state, and let its mean be m/λ, where m is a scaling factor that will come into play in a moment. Let Y be the (random) length of an arbitrary sojourn time in a busy phase, $S = rY$, and G be the distribution function of S. Then S is the amount of work (in units of say, bits, or bytes, or cells, or packets) that arrive in an arbitrary busy phase. Let the mean and variance of S be

$1/\mu$ and σ^2, respectively. The mean time between starts of busy phases is $E(Y) + m/\lambda$. In our context where the arrivals are files, processing times typically are a linear scaling of the file sizes. Since r can be made unity with a suitable choice of dimensions for file sizes and processing times, we will take $r = 1$.

When m is very large, this arrival process is approximately a process where packages of work having distribution function G arrive with spacings having distribution function F; as we will now show. With the assumptions above, the packages of work have distribution function G, the times between work arrival epochs have distribution function $F * G$, where $*$ denotes convolution, and the spacings between these epochs are i.i.d. We want to show that when time is scaled by the factor $1/m$, the distribution of the times between work arrival epochs converges to F as $m \to \infty$. First, let X be a generic sojourn time in the idle state and $X_m = mX$, with $E(X) = 1/\lambda$. Let $\tilde{F}$ and $\tilde{F}_m$ be the Laplace–Stieltjes transforms of X and X_m, respectively. Then

$$\tilde{F}_m(s) = E(\mathrm{e}^{-sX_m}) = E(\mathrm{e}^{-smX}) = \tilde{F}(ms). \tag{1}$$

Now we want to stretch out the time between arrivals while keeping the amount of work brought by each arrival unchanged. A way to envision this is to scale the interarrival times by a new (scaled) time variable (denoted by τ) that is related to the old (unscaled) time variable (denoted by t) via $\tau = t/m$. In scaled time, the distribution of the time between initiations of busy phases is $F(\tau) * G(m\tau)$, so the mean number of initiations by time τ (the renewal function, $M(\tau)$ say) is

$$M(\tau) = \sum_{k=1}^{\infty} [F(\tau) * G(m\tau)]^{*k},$$

where the superscript denotes k-fold convolution. Taking Laplace–Stieltjes transforms and using (1) yields

$$\tilde{M}(s) = \frac{\tilde{F}(s)\,\tilde{G}(s/m)}{1 - \tilde{F}(s)\,\tilde{G}(s/m)} \to \frac{\tilde{F}(s)}{1 - \tilde{F}(s)} \quad (m \to \infty),$$

completing the demonstration.

Suppose we superpose m statistically independent sources of the type described above, and steady-state conditions prevail. The distribution of the busy phases are the same in all sources, but the distributions of the idle phases need not be. The assumption above that the mean idle phase is m/λ implies that the sum of the arrival rates is λ for every m. Let F_{jm} the distribution of the idle phases of source j when there are m sources. We require that for any $\varepsilon > 0$,

$$F_{jm}(t) \le \varepsilon, \quad j = 1, 2, \ldots, m$$

for all $t > 0$ when m is sufficiently large. These assumptions allow one to invoke the Palm–Khintchine Theorem [3] which states that when $m \to \infty$, the arrival epochs of the superposed process form a Poisson process with rate λ, no matter what the forms of F_{jm} are. Transmission pipes, multiplexers and switching devices are often modeled as a single server that works at a constant rate when work is available, and a buffer to handle arrival spurts is frequently assumed to be present. If the traffic (i.e. the packages of work) is processed by a single server working at unit rate, the queueing process is precisely an M/G/1 queue where G is the distribution function of the service times. Notice that the only requirement we need place on G is that it be the distribution of a non-negative random variable.

A detailed rigorous derivation of the M/G/1 model with integer-valued sojourn times is given in Likhanov et al. [4]. Our result generalizes Theorem 1 of Jelenkovic and Lazar [5] who start with the assumption that the idle phases are exponentially distributed. Related work is in Heath et al. [6] who examine the effects of fat-tails in both the inter-arrival and service times on the queue lengths in an on/off model. Investigations of the M/G/1 queue with fat-tailed service times go back at least as far as 1968, when a paper by Harris [7] appeared.

We obtained the M/G/1 queue arrival process by keeping the arrival rate fixed as the number of sources tends to infinity, scaling the idle phases and keeping the busy phases unscaled. In physical terms, each source initiates file transfers at a lower rate as the number of sources increases and the file sizes remain the same. Willinger et al. [8] use the same individual source model that we use. In obtaining their limit result (as $m \to \infty$) they keep the mean of the idle phases as is, and scale both the idle and busy phases. Naturally, they obtain a different model for many sources. Among the differences is that the arrival rate increases with m, and tends to infinity as m does. Under which conditions these models should be used is an important question that will not be examined here.

2.2. Some power-law-tailed distributions

Let G be the distribution function of some random variable that lives on the non-negative portion of the real line. The "tail" of G is usually not defined rigorously, but its meaning seems intuitively clear and the term is commonly used in technical discourse. I use the term *body* to describe the part of G that is not in the tail. Let $G^{c}(x) = 1 - G(x)$ for every $x \geq 0$; it is the *complementary* distribution function (c.d.f.). G is called *long-tailed* if [9]

$$\lim_{x\to\infty} \frac{G^{c}(x-y)}{G^{c}(x)} = 1$$

for all real numbers y. Three examples of long-tailed distributions are the Pareto family, the lognormal distribution and the Weibull family with shape parameter <1. The latter two have finite moments of all orders, so the long-tailed property is not synonymous with infinite moments.

The Pareto family is defined by

$$G^{c}(x) = \left(\frac{r-1}{r}\right)^{r} x^{-r}, \quad x \geq \frac{r-1}{r}.$$

The nth moment exists when $r > n$. For large x, $G^{c}(x)$ looks like const/x^{r}, which is what is usually meant by "having a fat tail". G has a *power-law tail* if

$$G^{c}(x) \approx x^{-r} L(x) \;\; (x \to \infty), \quad r > 0,$$

where L is a slowly varying function (which means that $L(xt)/L(x) \to 1$ as $x \to \infty$ for any $t > 0$) and $\approx$ means that the ratio of the left-side and the right-side goes to 1. In the general literature, distributions with this property are called *regularly varying with index r*.

The lognormal and Weibull distributions are difficult to use as the service-time distribution in an M/G/1 queue because neither has a closed-form Laplace transform. The Laplace transform of a Pareto distribution is an incomplete gamma function, which is not easy to work with, but the main reason for not wanting to use a Pareto distribution for service times is that lower bound and rate at which the tail probabilities decay cannot be chosen independently, and this affects those moments that exist. Abate et al. [10] introduced the

Pareto mixture of exponentials (PME) family to circumvent these difficulties. The two PME distributions used in this paper are called G_2 and G_3, and are defined by

$$G_2^{\mathrm{c}}(x) = \frac{1-(1+2x)\mathrm{e}^{-2x}}{2x^2}, \quad x > 0,$$

and

$$G_3^{\mathrm{c}}(x) = \frac{16[1-(1+3x/2+9x^2/8)\mathrm{e}^{-3x/2}]}{9x^3}, \quad x > 0.$$

Manifestly,

$$G_2^{\mathrm{c}}(x) \approx \frac{1}{2x^2} \quad \text{and} \quad G_3^{\mathrm{c}}(x) \approx \frac{16}{9x^3} \quad (x \to \infty),$$

so they are fat tailed. These distributions have mean 1. The variance of G_2 is infinite. The variance of G_3 is 5/3, and the third moment is infinite . The Laplace transforms contain only elementary functions.

Boxma and Cohen [11] extended the PME family. Let G_{bc} be the distribution they specify in Eq. (1.7) of their paper, with their parameters s and δ set equal to 1. The complementary distribution is

$$G_{\mathrm{bc}}^{\mathrm{c}}(x) = (2x+1)\mathrm{e}^{x}\mathrm{Erfc}(\sqrt{x}) - 2\sqrt{x/\pi}, \quad (x > 0),$$

where

$$\mathrm{Erfc}(x) = \frac{2}{\pi}\int_x^{\infty} \mathrm{e}^{-u^2}\,\mathrm{d}u$$

is the complementary error function. This distribution function has mean 1, infinite variance, and the Laplace transform is

$$\tilde{G}_{\mathrm{bc}}^{\mathrm{c}}(s) = \frac{s}{(1+\sqrt{s})^2}$$

which contains only elementary functions. This distribution will be called BC for short. From an expansion for Erfc, it can be shown that

$$G_{\mathrm{bc}}^{\mathrm{c}}(x) \approx \frac{1}{\sqrt{\pi x^3}} \quad (x \to \infty).$$

Abate and Whitt [12] show that BC is a beta mixture of exponentials.

2.3. *Exponential damping*

These power-law-tailed distributions will be compared to three distributions that do not have power-law tails. The first two are exponentially damped versions of G_{bc}. Abate et al. [13] introduce the notion of damping fat-tailed distributions by multiplying the c.d.f. by $\mathrm{e}^{-\delta x}$ for some suitably chosen $\delta > 0$; the probability that is lost by this operation is placed at the origin, and a renormalization is done to preserve the mean value. To be precise, let $\tilde{g}$ be the Laplace transform of a given density function with mean 1, and $\tilde{d}$ be the Laplace transform of the damped density function, then

$$\tilde{d}(s) = \tilde{g}\left(\frac{s}{m_\delta} + \delta\right) + 1 - \tilde{g}(\delta), \tag{2}$$

where

$$m_\delta = -\frac{\mathrm{d}}{\mathrm{d}s}\tilde{g}(\delta).$$

For BC,

$$\tilde{g}(s) = \frac{1+2\sqrt{s}}{(1+\sqrt{s})^2} \quad \text{and} \quad m_\delta = (1+\delta)^{-3}.$$

I use $\delta = 10^{-8}$ and $\delta = 10^{-4}$ to damp BC. The former will be referred to as *slight damping*, and latter as *moderate damping*. With the former value, $\mathrm{e}^{-\delta x} \geq 0.9999$ for $x \leq 10\,000$ this produces a distribution that has an exponential tail (and a variance of 15 003.0) that is practically indistinguishable from BC for $x \leq 10\,000$. If this distribution produces queueing results that are close to those produced by BC, then the infinite variance of the latter is not a key property in the performance context. The tail probabilities of G_2 decay more slowly than the tail probabilities of BC, and G_2 has an infinite variance. Thus, if slightly damped BC induces larger queues (in the region of engineering interest) than G_2 does, then the asymptotic tail behavior (which determines if the variance is finite or infinite) is not a critical feature of the service-time distribution. More discussion of exponential damping is given in Section 3.3.

With moderate damping we obtain tail probabilities that are within 1% of the tail probabilities of BC for $x \leq 100$. The variance is 161.0, which is much smaller than the variance obtained with slight damping, so it should produce different results than BC when probabilities that the service time exceeds x (for $x > 100$) are important. The third distribution is a gamma with mean 1 and squared coefficient of variation 5/3, which matches the first two moments of G_3. The power-law tail is an important feature of the service-time distribution if the gamma and G_3 distributions produce very different queueing behavior.

2.4. Some comparisons among these distributions

Here we compare the c.d.f.'s and variance–time curves of the six distributions introduced above.

Fig. 1 shows the c.d.f.'s, which were computed by numerical Laplace transform inversion using the program EULER [14]. The tail probabilities are expressed as the probability that a random variable S exceeds t, where S is to be thought of as the service time in an M/G/1 queue.

In Fig. 1 there is no discernible difference between BC and the slightly damped version for 10 000 mean service-times, so the difference is not likely to matter when used in the queueing model. The moderately damped version matches the original c.d.f. past 100 mean service-times, and is not far from the original at 1000 mean service-times. The other distributions fall off much faster.

Let $A(t)$ be the amount of work that has arrived by time t in an on–off source as described above, with $m = 1$. Assume that steady-state conditions prevail. Since the mean on times are one with all the distributions we are considering, taking $\lambda = \rho/(1-\rho)$ will make ρ the proportion of time in the on phase. Let $\tilde{V}$ be the Laplace transform of Var[$A(t)$], and $\tilde{F}$ and $\tilde{G}$ be the Laplace transforms of the off and on distributions respectively. A special case of a formula derived by Krishnan [15] (see Eq. (A.7)) yields

$$\tilde{V}(s) = \frac{2\tilde{Q}(s)}{s^2},$$

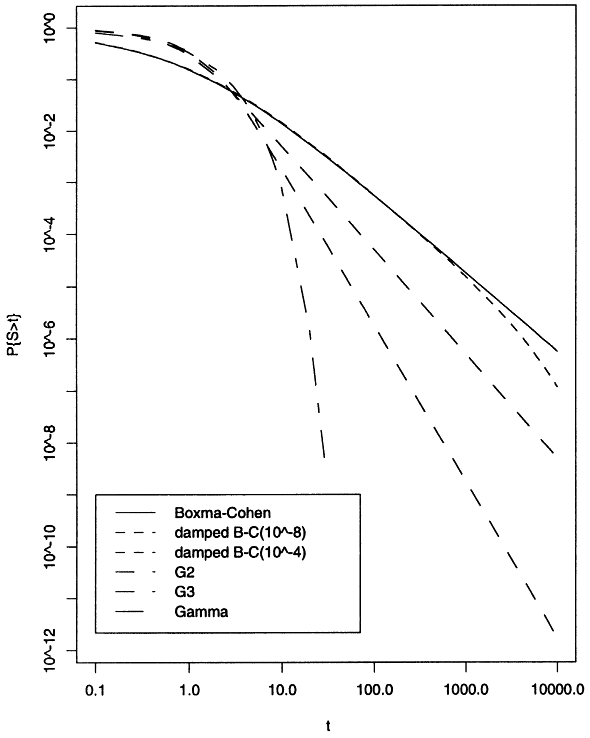

Fig. 1. Tail probabilities for six distributions.

where

$$\tilde{Q}(s) = \frac{\lambda}{s(1+\lambda)^2} - \frac{\lambda[1-\tilde{F}(s)][1-\tilde{G}(s)]}{s^2(1+\lambda)[1-\tilde{F}(s)\tilde{G}(s)]}.$$

The graph of Log(Var[$A(t)$]) vs. Log(t) is called the *variance–time plot*. When Var[$A(t)$] $\approx$ const $\times\, t^{\beta}$, the standard deviation of $A(t)$ grows as $\sqrt{(\mathrm{const})}t^{\beta/2}$, and $H = \beta/2$ is called the *Hurst parameter*. Thus, H is half of the slope of the variance–time plot for large values of time. The variance–time plot was constructed by numerically inverting $\tilde{V}$ with EULER. The results are shown in Fig. 2.

In Fig. 2, the slightly damped BC curve is indistinguishable from the undamped curve; both have slope 1.5, which corresponds to $H = 0.75$. The moderately damped version has a slightly smaller slope. The slope of the G_2 curve is 1.10, yielding $H = 0.55$. The other two curves are indistinguishable from each other, and have a slope of 1.00 ($H = 0.5$) which is anticipated by theory. (Krishnan [15] shows that finite variance implies $H = 0.5$.)

The conclusions to be drawn from Figs. 1 and 2 are that the slightly damped version of BC is essentially the same as the original distribution for arguments up to 10 000 times the mean, and the other distributions are likely to produce shorter delays when they describe service times. Since we are concerned with delay probabilities, these conclusions can be made quantitatively by calculating the probabilities of interest. This is done in the next section.

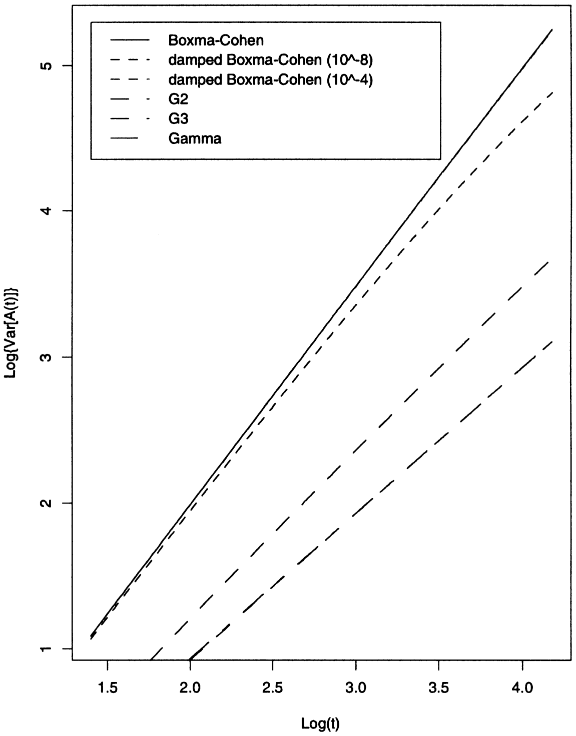

Fig. 2. Variance–time plot for six distributions.

3. Tails of the delay distribution

Now we examine the steady-state probability that the delay in queue (not including service time) exceeds t, which in symbols is $P\{W > t\}$. This distribution has intrinsic interest because delay is an important performance measure; there is also another reason to examine it. Since W also has the interpretation of the amount of work in the queue (as used in Section 2.1), several authors (e.g. [16]) use $P\{W > t\}$ as a proxy for the probability that a finite buffer of size t overflows. (This probability is defined precisely in Section 5.) This is done because it is often easier to calculate $P\{W > t\}$ exactly, or to approximate it, than it is calculate the overflow probability itself. In Section 5 we will see that this may not be a good idea, but for now we explore the effects of the service-time distribution on $P\{W > t\}$.

3.1. Numerical results

Let $\tilde{W}$ be the Laplace transform of $P\{W > t\}$. From the Pollaczek–Khintchine formula (e.g. [3, pp. 7–66] we have

$$\tilde{W}(s) = \frac{1}{s} - \frac{1-\rho}{s+\rho-\rho\tilde{G}(s)}, \tag{3}$$

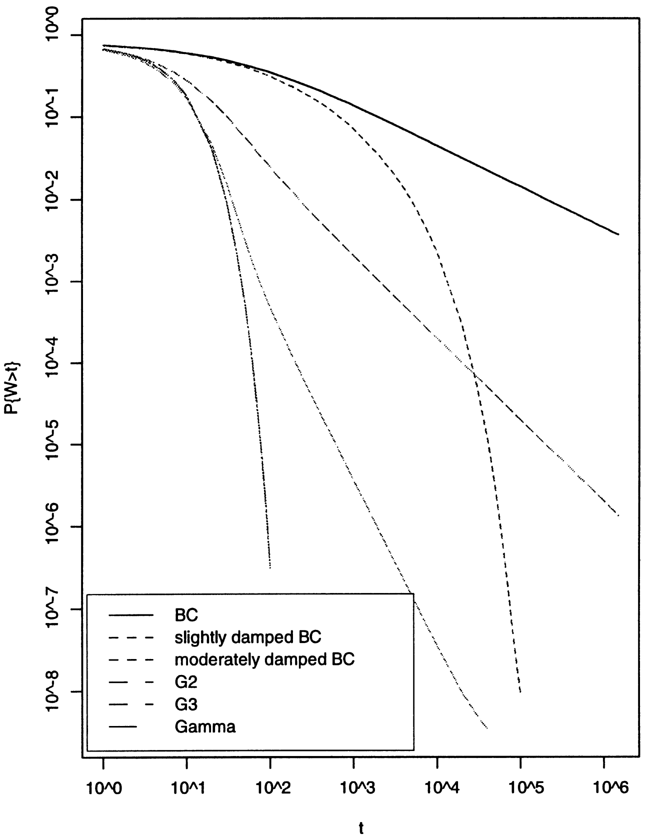

Fig. 3. $P\{\text{delay} > t\}$ for six distributions, $\rho = 0.8$ (log–log scale).

where the traffic intensity ρ equals the arrival rate λ because all of our service-time distributions have mean 1. From (3) one can deduce the Pollaczek–Khintchine formula for the mean delay,

$$E(W) = \rho \frac{c_s^2 + 1}{2(1-\rho)}, \tag{4}$$

where c_s^2 is the squared coefficient of variation of the service times. Since $c_s = \infty$ for G_{bc} and G_2, $E(W) = \infty$ for them. With $\rho = 0.8$,

$E(W) =$30 008 for slightly damped G_{bc};
$E(W) =$322.0 for moderately damped G_{bc};
$E(W) =$5.33 for G_3 and gamma.

For BC, G_2 and G_3, $\text{Var}(W) = \infty$.

Eq. (3) was numerically inverted via EULER for the six G's of interest; the results are shown in Fig. 3. In Fig. 3 BC and slightly damped BC are indistinguishable over the displayed range, which goes up to 1 million mean service times. Thus, the delay distribution is controlled (over this range) by those t where $G_{\text{bc}}^{\text{c}}(t)$ and $\mathrm{e}^{-10^{-8}t} G_{\text{bc}}^{\text{c}}(t)$ are approximately equal; from Fig. 1, this includes the interval [0, 10 000]. Thus, agreement of the bodies of the service-time distributions is important, and disagreement of the tails is not important. The G_3 and gamma curves become very different when $t > 10$, so again the form in the

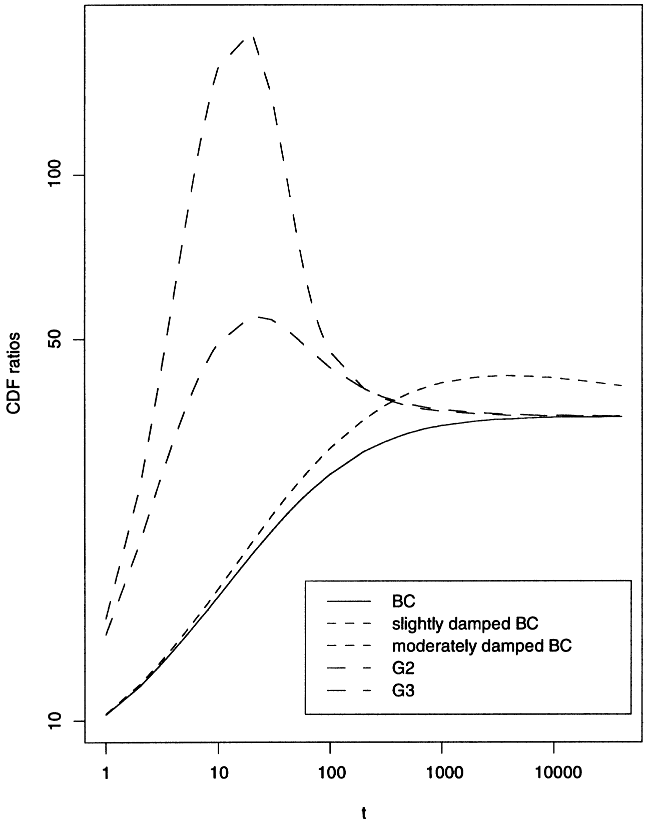

Fig. 4. $P\{\text{delay} > t \quad \rho = 0.8\}/P\{\text{delay} > t; \quad \rho = 0.1\}$ (log–log scale).

body of the service-time distribution is important. The slope (on a log–log scale) of the BC and slightly damped BC curves are $-1/2$, and the slope of the G_2 curve is -1. When the same curves are drawn when $\rho = 0.1$, the curves are similar in form to the curves in Fig. 3 and are shifted downwards. To compare the curves for $\rho = 0.1$ to those for $\rho = 0.8$, the ratios of the c.d.f.'s, excluding the gamma distribution (because that ratio is much larger than the others) are shown in Fig. 4. There is an analytic way to explain these empirical results. It is also interesting that moderately damped BC is close to BC until 100, larger than G_2 until almost 100 000, and always larger than G_3.

3.2. Analytic explanations

The following continuity argument shows that some damped BC will provide an arbitrarily close approximation of the tail probabilities induced by BC. It is clear from (2) that for every s, $\tilde{d}(s) \to \tilde{g}(s)$ as $\delta \to 0$. Let $\tilde{W}_\delta$ be the Laplace transform of $P\{W > t\}$ when δ is the damping coefficient; $\tilde{W}_0 = \tilde{W}$. In (3), $\tilde{W}$ is a continuous function of $\tilde{G}$, so $\tilde{W}_\delta(s) \to \tilde{W}(s)$ as $\delta \to 0$. From the continuity theorem [17, Chapter XIII] for Laplace transforms, the tail probabilities converge as well. This argument does not explain why $\delta = 10^{-8}$ works as well as it does, or even that it will work at all.

A more detailed analysis of the close agreement between BC and slightly damped BC can be obtained as follows. Pakes [18] proved that for a GI/G/1 queue with traffic intensity $\rho < 1$ in which the generic service-time random variable S satisfies $E(S)=1$, $E(\mathrm{e}^{sS})=\infty$ for all $s>0$, and the c.d.f. of S has the *subexponential property*, then

$$P\{W>t\}\approx\frac{\rho}{1-\rho}\int_t^\infty G^{\mathrm{c}}(x)\mathrm{d}x\quad(t\to\infty).\tag{5}$$

The subexponential property is somewhat esoteric.[2] Fortunately, we have this claim in [10].

Since power-law-tailed distributions have the property that $E\mathrm{e}^{sS}=\infty$, because exponentials go to infinity faster than polynomials, we can expect (5) to hold in in our investigations. The condition $E(S)=1$ and the identity $\int_0^\infty G^{\mathrm{c}}(x)\,\mathrm{d}x=E(S)$ imply that (5) can be written as

$$P\{W>t\}\approx\frac{\rho}{1-\rho}\left[1-\int_0^t G^{\mathrm{c}}(x)\,\mathrm{d}x\right]\quad(t\to\infty).\tag{5a}$$

Fig. 1 shows that the area under the G^{c} curves agree for BC and slightly damped BC, so (5a) shows that the tails of the waiting times will also agree. Eq. (5a) gives the essential property that is needed for two service-time c.d.f.'s to produce similar tails of the delay distributions; namely, if G^{c} and H^{c} are c.d.f.'s, and W_G and W_H are the steady-state waiting times they produce, then

$$\int_0^t G^{\mathrm{c}}(x)\mathrm{d}x\overset{\mathrm{app}}{=}\int_0^t H^{\mathrm{c}}(x)\,\mathrm{d}x\Rightarrow P\{W_G>t\}\overset{\mathrm{app}}{=}P\{W_H>t\}\quad(t\to\infty).$$

The asymptotic slopes in Fig. 3 can be explained by Eq. (15) in [10], which is the following.

When $E(S)=1$ and $P\{S>x\}\approx\alpha_r x^{-r}$ as $x\to\infty$ for $r>1$, then

$$P\{W>t\}\approx\frac{\rho}{1-\rho}\frac{\alpha_r}{r-1}t^{-(r-1)}\quad\text{as }t\to\infty.\tag{6}$$

The factor $\rho/(1-\rho)$ explains why the $\rho=0.1$ curves asymptotically are a downward shift of the $\rho=0.8$ curves for BC, G_2 and G_3, with a ratio of 36, as shown in Fig. 4. Since slightly damped BC is very close to BC, it also satisfies (6); moderately damped BC is close to BC for a while, so it does not quite satisfy (6). The gamma service-time distribution leads to an exponential tail for the waiting time, so there is no reason for it to satisfy (6). It does not, and it is excluded from Fig. 4 because the ratio reaches 5 million when $t=30$.

For BC, $r=3/2$ so (6) predicts that the BC curve in Fig. 3 should eventually have a slope of $-1/2$, which it does for $t\geq 1000$. For $G_2, r=2$ so (6) predicts that the G_2 curve in Fig. 3 should eventually have a slope of -1, which it does for $t\geq 100$. For $G_3,\ \ r=3$ so (6) predicts that the G_3 curve in Fig. 3 should eventually have a slope of -2, which it does for $t\geq 300$. It may be surprising that (6) obtains the correct slope for such small values of t. It might be more surprising that it produces a very accurate approximation for BC and G_3, and a good approximation for G_2. The exact and approximate values are shown in Fig. 5.

3.3. *Effects of truncation*

Slightly damped and moderately damped BC are noticeably different in Figs. 1–4. Here we investigate the effects of damping in more detail. Since file sizes are bounded by the size of the memory

[2] For practical purposes, the subexponential property can be regarded as equivalent to $E\mathrm{e}^{sS}=\infty$ for all $s>0$.

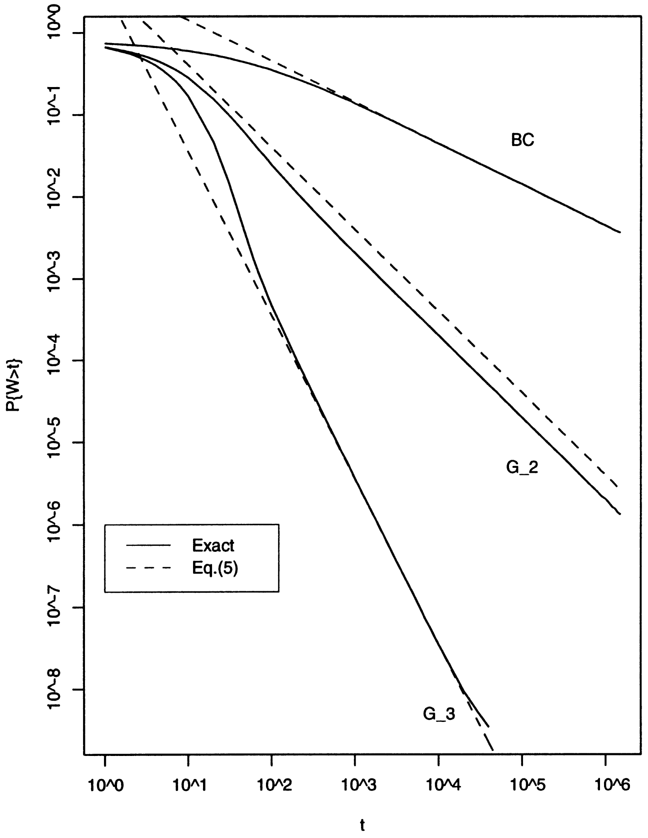

Fig. 5. Tail probabilities and approximations for three distributions.

that contains the files, it would seem appropriate to place an upper bound on file sizes. This would induce a bound on file transfer times in the model in Section 2.1. Truncating the distributions we have used leads to analytic difficulties that are best avoided, so we use damping to approximate truncating. Since

$$e^{10^{-2}} \overset{\mathrm{app}}{=} 1 - 10^{-2} = 0.99 \quad \text{and} \quad e^{-10} = 4.5 \times 10^{-5},$$

a damping factor of 10^{-k} begins to have a strong effect on $P\{S > t\}$ when t reaches 10^{k-2} and effectively truncates the tail probability when t reaches 10^{k+1}. Although damping isn't the same as truncating, it has an advantage because it models some uncertainty about the true upper bound.

Fig. 6 shows the tails of the service-time distributions for damping factors 10^{-k}, $k = 4, 5, 6, 7,$ and 8 applied to BC. Moderately damped BC begins to differ from the others when t is near 5000, and the others (especially 6, 7, and 8) remain close up to $t = 100\,000$. The tails of the delay distribution are shown in Fig. 7. The separation of the curves is much larger in Fig. 7 than in Fig. 6; i.e. the effect of damping (truncation) is larger on queueing performance than on the service-time distribution. Table 1 gives more detail about the curves in Fig. 7. From Table 1 we see that the damping factor (truncation level) can have a significant impact on the value of t needed to achieve a given $P\{W > t\}$. The last two columns of Table 1 show that slightly damped BC is indistinguishable from BC.

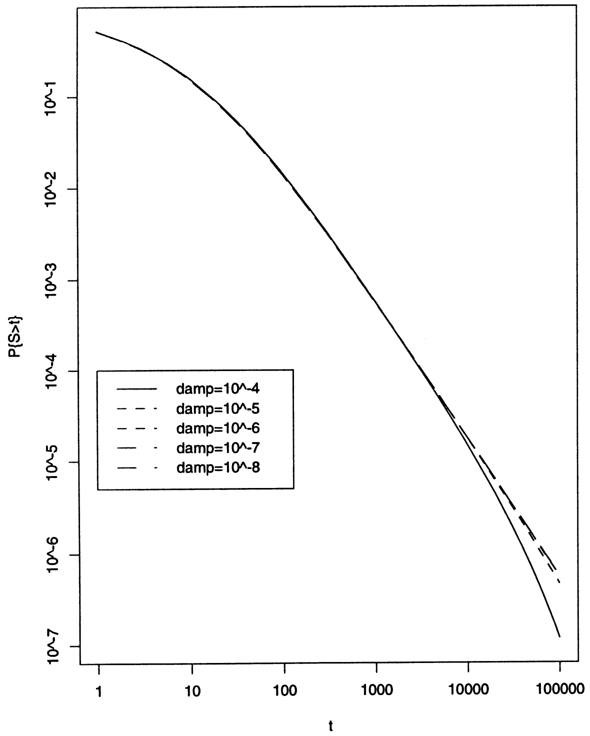

Fig. 6. $P\{S > t\}$ for five damping factors.

4. Rate of approach to the steady state

Now we examine the rate at which the steady-state is achieved in an M/G/1 queue as a function of the service-time distribution. This is done by numerically inverting the Laplace transform of the number in the system (in queue plus in service) at time t when the system is empty at time 0.

4.1. Numerical results

The Laplace transform of the expected number in the system (in queue plus in service) at time t when the system is empty at time 0 is given by Eq. (4.52) in [19], which is intricate and will not be given here. The transform was numerically inverted with EULER; the results are shown in Fig. 8. Since the

Table 1
$P\{W > t\}$ for several damping factors

t	Damping factor					
	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	0
1000	0.0071	0.1130	0.1300	0.1344	0.1370	0.1370
10 000	2.3×10^{-3}	0.0199	0.0350	0.0416	0.0449	0.0449
100 000	9.6×10^{-9}	0.0006	0.0060	0.0110	0.0143	0.0143
1 000 000	0.0000	0.0000	0.0002	0.0019	0.0045	0.0045

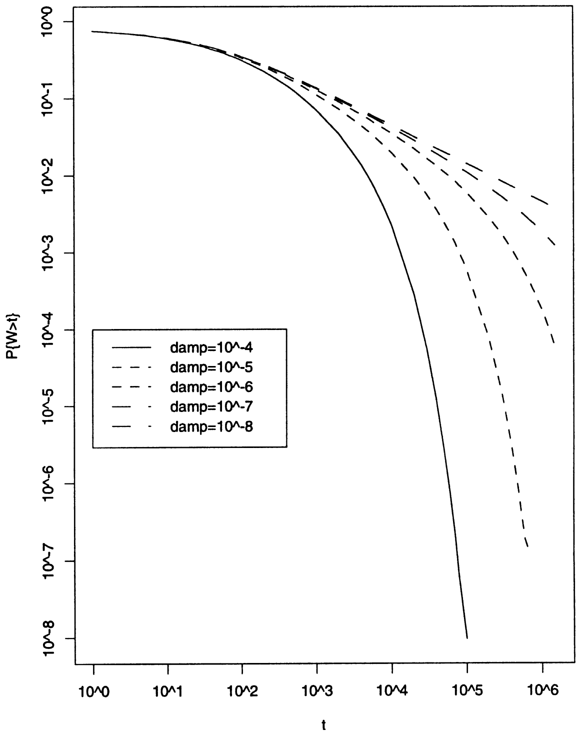

Fig. 7. $P\{W > t\}$ for five damping factors.

number-in-system and waiting-time processes have the same regeneration epochs (when they reach 0), their approach to the steady-state should be similar.

Natural time units are shown in Fig. 8; they were obtained in the following way. Consider a residential WWW service. Most subscribers currently have modems that work no faster than 28.8 Kbps. The X.2 modems work at twice that rate, and some subscribers may have ISDN access, which is 128 Kbps. To be very conservative, assume all users have ISDN access. An estimate of the mean download file size is 20 000 bytes [20], which takes 160/128 s to download at the ISDN rate. Round this down to 1 s. That is the mean service-time used in Fig. 8, and time is measured in mean service-times.

In Fig. 8 the G_3 and the gamma curves coincide and reach their steady-state value of 5.0666 ... rapidly. The value of G_3 is 4.95 at 5 min and 5.00 at 6 min, 40 s. The BC and slightly damped BC curves also coincide. The BC, damped BC, and G_2 curves approach their limiting value (finite for damped BC and infinite for BC and G_2) so slowly that the engineering value of steady-state performance measures is dubious. Even though the G_2 curve is approaching infinity and the moderately damped BC curve is approaching 648, the former curve is only 12% of the latter curve when $t = 15\,000$; the transient values can be misleading about the steady-state values. All of the curves appear to be approaching their limiting value in a concave way.

4.2. Analytic explanation

An analytic explanation for the qualitative behavior in Fig. 8 is based on an approximate model for the time-dependent behavior of the virtual-delay process of an M/G/1 queue developed by Gaver

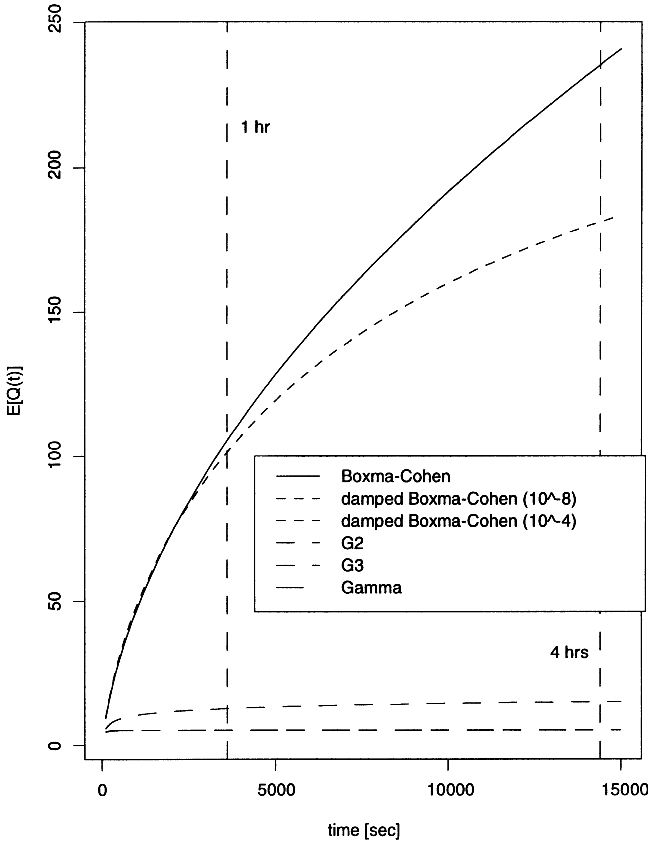

Fig. 8. Approach to steady-state system length.

[21]. The approximate model is Brownian motion (with drift) with a reflecting barrier at the x-axis. We restrict our attention to the situation where the system is empty at time 0. Let $F(x,t)$ be the probability that the Brownian motion is no larger than x at time t. Then F satisfies the partial differential equation

$$\dot{F} = -\mu F' + \frac{\sigma^2}{2} F'' \tag{7a}$$

with boundary equation

$$F(x,t) = 0 \quad \text{for } x \leq 0, \quad t \geq 0. \tag{7b}$$

In (7a), the dot denotes partial differentiation with respect to t and the prime denotes partial differentiation with respect to x. The parameters μ and σ^2 are

$$\mu = \rho - 1 \quad \text{and} \quad \sigma^2 = \rho(c_s^2 + 1) \tag{8}$$

when the mean service time is 1.

One can solve (7a) and (7b) via Laplace transforms and estimate the curves in Fig. 8, but there is a more insightful way to use (7a) and (7b). Gaver shows that by scaling time and space (i.e. the amount of virtual delay) to

Table 2
Mean of scaled diffusion

$\tau(p)$	1.0	1.4	2.15	4.2
Mean	0.425	0.450	0.475	0.495
% limit (100p)	85	90	95	99

$$\tau = \frac{\mu^2}{\sigma^2}t \quad \text{and} \quad \xi = \frac{\mu}{\sigma^2}x, \tag{9}$$

(7a) is converted to the dimensionless form

$$\dot{F} = -F' + \frac{1}{2}F'' \tag{10}$$

which can be solved once, and the probabilities in the natural units recovered by rescaling via (9). Let $W_{\text{scaled}}(\tau)$ be the scaled process governed by (10) and (7b). Gaver shows that

$$\int_0^\infty \mathrm{e}^{-s\tau} E[W_{\text{scaled}}(\tau)]\mathrm{d}\tau = \frac{1}{s(1+\sqrt{1+2s})}, \tag{11}$$

which can be numerically inverted using EULER. It is easy to obtain

$$\lim_{\tau\to\infty} E[W_{\text{scaled}}(\tau)] = \tfrac{1}{2}$$

from (11), so $E[W_{\text{scaled}}(\tau)]$ can be conveniently expressed as a percentage of its limiting value. Doing so yields Table 2.

Thus, from Table 2, (8) and (9) we see that for the mean of the unscaled diffusion to be within $(100 \times p)$% of its steady-state value,

$$t \geq \frac{c_s^2+1}{(1-\rho)^2}\tau(p)$$

is required, where $\tau(p)$ is read from Table 2. Convergence is slower for larger traffic intensities, and is very slow for very large c_s. To achieve 99% of the steady-state value,

$$t \geq 4.2\frac{c_s^2+1}{(1-\rho)^2}$$

is required.

If the diffusion model were a good approximation, it would explain why convergence is rapid with the gamma and G_3 service times ($c_s^2 = 5/3$) and why convergence is very slow with the other service times (c_s^2 is very large). Figs. 9 and 10 show that the diffusion model gives an estimate of the rate of convergence to the steady-state that is faster than the exact rate of convergence.

The results given in this section demonstrate that with fat-tailed service-times and infinite buffers, the steady-state queue lengths are a very poor indicator of the queue lengths that would be observed in finite time spans of engineering interest. The same conclusion was reached by Lipsky and Hatem [22] and by Greiner et al. [23]. Finite buffers force queue lengths to be smaller than they would have been if an infinite buffer were used, which causes the system to empty (and thus, regenerate) more often. Consequently, the M/G/1/N queue may converge fast enough; this should be investigated.

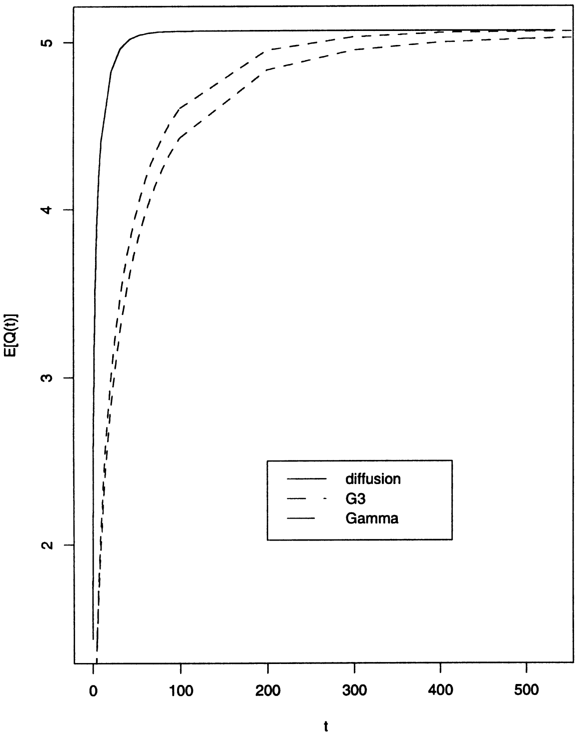

Fig. 9. Exact and approximate convergence rates for moderately damped BC service times.

5. Finite buffers

The M/G/1 models considered in the previous sections have an infinite buffer, which is taken to be an approximation of a finite buffer that is large enough to keep the overflow rate small. The probability that a buffer of size b overflows is sometimes approximated by the waiting-time distribution $P\{W > b\}$ of an infinite-buffer system (e.g. [16]). In this section we investigate the efficacy of this approximation. Previous investigations of this kind include [24–26]. In those studies, the arrival process was varied and the service times were fixed. Here, we fix the arrival process and vary the service-time distribution, and obtain somewhat different results than the previous studies did. We share with those studies some results that show that the approximation can be poor.

Let π be the steady-state distribution of the standard embedded Markov chain (with an infinite buffer) for the number present at departure epochs. Its probability generating function ($\hat{\pi}$ say) is given by the Pollaczek–Khintchine formula

$$\hat{\pi}(z) = (1-\rho)\frac{(z-1)\tilde{G}(\rho-\rho z)}{z-\tilde{G}(\rho-\rho z)}$$

when the mean service time is 1. Let

$$\pi^{\mathrm{c}}(b) = \sum_{i=b+1}^{\infty} \pi(i),$$

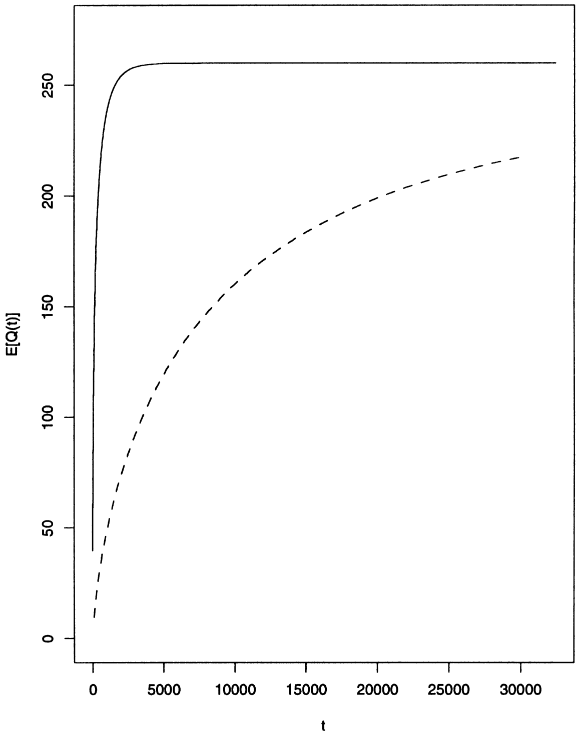

Fig. 10. Exact (solid) and approximate (dotted) convergence rates for gamma and G_3 service times.

where $\pi(i)$ is the ith element of π. Let π_b be the steady-state distribution of the number of customers present at departure epochs when the buffer is of size b. (So π is shorthand for π_∞.) The upper-Hessenberg form of the transition matrix for the embedded Markov chain for the M/G/1 queue leads to the following fact (see, e.g. exercise 752 in [3]);

$$\pi_b(i) = \frac{\pi(i)}{1 - \pi^{\mathrm{c}}(b)}, \quad i = 0, 1, \dots, b. \tag{12}$$

Let p_b be the steady-state distribution of the number of customers present just before an arrival epoch when the buffer is of size b. The exact overflow probability is $p_b(b + 1)$; it is given in Eqs. (9)–(13) in [27] which is

$$p_b(b + 1) = 1 - \frac{1}{\rho + \pi_b(0)}. \tag{13}$$

Numerical values of the overflow probability can be obtained from numerical inversion of $\hat{\pi}$, (12) and (13). The numerical inversion was done via the algorithm LATTICE-POISSON [28].

5.1. Numerical results

The efficacy of using $P\{W > b\}$ to approximate $p_b(b+1)$ is illustrated by the ratio $P\{W > b\}/p_b(b+1)$ is shown in Fig. 11 for various values of b when $\rho = 0.8$. In Fig. 11 we see that the ratio gets very large very quickly for the gamma distribution, gets very large more slowly for the G_3 and moderately damped

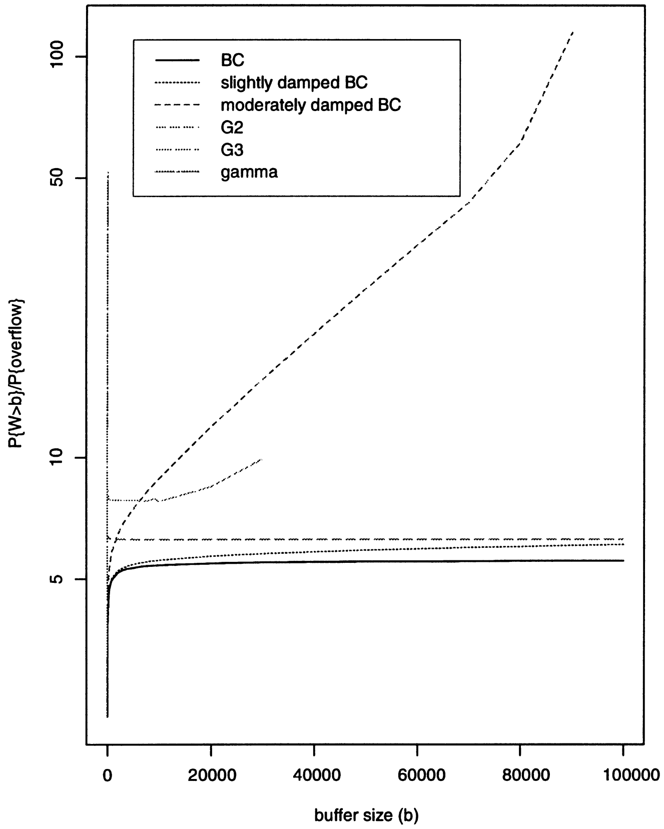

Fig. 11. Comparison of $P\{W > b\}/p_b(b + 1)$ when $\rho = 0.8$.

BC distributions, approaches 6 quickly for the G_2 distribution, and the other two ratios approach 6 slowly. The approximation works best for the distributions with the largest variances. It is very poor for gamma and moderately damped BC distributions; for the other distributions it may be a matter of context if the error by a factor of 6 is acceptable or not.

An analytic explanation of the curves in Fig. 11 can be obtained from (13) by substituting $\pi(0) = 1 - \rho$ in (12) with $i = 0$, substituting this in (13), and rearranging terms to get

$$\frac{\pi^{\mathrm{c}}(b+1)}{p_b(b+1)} = \frac{1-\rho\pi^{\mathrm{c}}(b+1)}{1-\rho} \overset{\mathrm{app}}{=} \frac{1}{1-\rho}. \tag{14}$$

Let X be the random variable representing the steady-state occupancy (number in queue plus in service) in the infinite-buffer model, so $\pi^{\mathrm{c}}(b) = P\{X > b\}$. If $\pi^{\mathrm{c}}(b + 1)$ is a good approximation of $P\{W > b\}$, then $P\{W > b\}/P\{X > b\}$ is close to 1. In this case, (14) can be rewritten as

$$\frac{\pi^{\mathrm{c}}(b+1)}{p_b(b+1)} \overset{\mathrm{app}}{=} \frac{P\{W > b\}}{p_b(b+1)} \overset{\mathrm{app}}{=} \frac{1}{1-\rho}.$$

This implies that the ratios shown in Fig. 11 should be about 5 when $\rho = 0.8$.

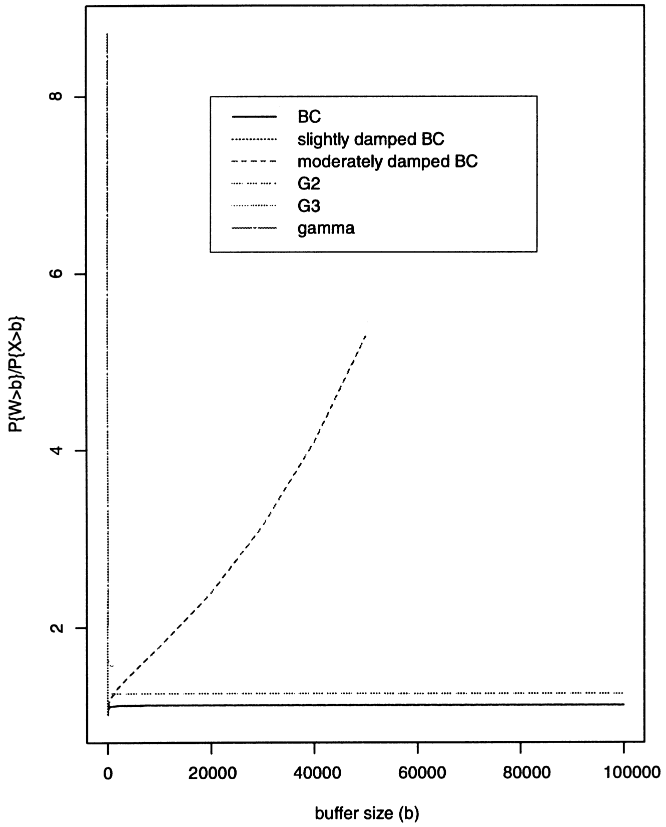

Fig. 12. Comparison of $P\{W > b\}/P\{X > b\}$ when $\rho = 0.8$.

This explanation for the behavior of the approximation provides a relation between the exact and the approximate probabilities. Rearranging the equality in (14) yields that approximation for $p_b(b+1)$ in terms of $P\{W > b\}$ proposed by Gouweleeuw and Tijms [29], namely

$$p_b(b+1) \overset{\text{app}}{=} \frac{(1-\rho)P\{W > b\}}{1-\rho P\{W > b\}}. \tag{15}$$

The ratio $P\{W > b\}/P\{X > b\}$ is plotted for the six distributions we are considering in Fig. 12. The approximation is good for the power-law-tailed distributions and not for the other two, as shown in Fig. 12. The ratios that are close to one in Fig. 12 (BC, slightly damped BC, and G_2) correspond to the distributions that were close to 6 in Fig. (11). The quality of the approximation tested in Fig. 12 appears to explain the quality of the approximation tested in Fig. 11.

6. Conclusions

There are two main conclusions that should be drawn from this study. The first is that for power-law-tailed service-time distributions with infinite variances, it is not the infinite variance per se that affects estimates of the performance measures; it is the shape of the distribution over a large part of it's support that matters. This was demonstrated by showing that the BC distribution (which has infinite variance) and the slightly damped BC distribution (which agrees with BC from 0 to 10 000 mean holding-times, but has finite variance) yield indistinguishable buffer-overflow probabilities for buffer sizes ranging from 1 to 15 000 mean service-times of work, and indistinguishable delay probabilities over the same range. Moreover, the G_2 distribution has an infinite variance and noticeably different overflow and delay probabilities from the previously mentioned pair, and the moderately damped BC distribution yields overflow and delay probabilities that are in between those induced by BC and G_2. In addition to showing that slight damping can have a very small effect, we showed that more aggressive damping can have a small effect on the service times and a large effect on the delay distribution. We argued that this is a good approximation of the effect of truncating the service-time distribution.

The second main conclusion is that when the service times have a very large variance, the rate of convergence to steady-state performance measures will be so slow that these measures are unlikely to be of engineering interest. This was demonstrated by the computations graphed in Fig. 8, and the scaling of a diffusion model approximation of the M/G/1 queue. It is possible that a finite-buffer version of the M/G/1 queue may converge fast enough for the steady-state probabilities to be relevant; this should be investigated.

A more qualified conclusion is that approximating the probability that a buffer of size b overflows by the probability that the work in the system exceeds b is suspect. The approximation appears to be very poor *unless* the service times have a distribution with a very large variance. In that case, the approximation over estimates the true loss probability by a factor of approximately $1/(1-\rho)$, where ρ is the traffic intensity.

A modeling implication of the first conclusion is that fitting a distribution with infinite support to data should be done with practical considerations in mind. It has never been suggested that voice telephone calls could last indefinitely, but using a truncated distribution instead would have increased the analytic complexity of the performance models considerably, and changed the results very little. This study was motivated by data on file sizes, which are inherently bounded by the largest memory available. The data indicate that large file sizes obey a power-law-tail distribution, but the small file sizes do not; a mixture of two distributions seems to be appropriate. Fitting an unbounded Pareto distribution to the observed hyperbolic shape of the tail of the empirical histogram can be a poor thing to do if the infinite moments of the fitted distribution will cause problems in choosing the mixing parameters, e.g. matching the mean and variance of the data. This is an especially poor thing to do because the Pareto distribution can be truncated to a finite support easily.

The engineering implication of the second conclusion is that when service times have a very large variance (such as file transfers on the WWW), performance criteria other than steady-state measures have to be used. What these performance criteria should be is an open question that deserves more attention. The third conclusion implies that using an infinite buffer model may not be a good approximation, even when the physical buffer is large. This means that performance models should explicitly model finite buffers except when the performance in the finite-buffer model can be related to the performance measures in the infinite-buffer model, as in (14) and (15).

Acknowledgements

I thank Gagan Choudhury for bringing the Boxma–Cohen distribution to my attention, for suggesting that I use the damped versions, and for sharing his transform inversion programs with me. Yoni Levy carefully read the first draft and helped clear up several garbled passages. Sid Resnick provided useful comments on the second draft. The referees suggested several valuable improvements.

Appendix A. Krishnan's formula

Since Krishnan's paper [15] is unpublished, his result used in Section 2.4 is derived here (with the author's permission). Consider the on–off model (or alternating renewal process) described in Section 2.1. I assume knowledge of the facts about alternating renewal processes as given, say, in Sections 4 and 5 of [3]. Let ν_f be the mean of the off-period distribution (F) and ν_g be the mean of the on-period distribution (G). For simplicity, assume that F and G have densities, and denote them by f and g, respectively. If F or G has mass at the origin, append this mass to the density function with a delta function. Laplace transforms will be denoted by a tilde.

Let $X(t)$ be the rate at which traffic is being generated at time t, so $X(t) = 1$ when t falls in an on-period and $X(t) = 0$ when t falls in an off-period. Assume that X is in the steady-state at time 0, so

$$p_1 \overset{\text{def}}{=} P\{X(t) = 1\} = \frac{\nu_g}{\nu_g + \nu_f}, \quad t >= 0.$$

For each sample path of the stochastic process X, ω say, define the random variable $Z(t, \omega)$ by

$$Z(t, \omega) = \int_0^t X(u, \omega)\, \mathrm{d}u, \quad t > 0.$$

It is the amount of traffic that arrives during (0, t]. In alternating renewal process terminology, it is the amount of time in the on state by time t. From here on, the argument ω will be supressed. Let

$$m_z(t) = E[Z(t)] \quad \text{and} \quad \sigma_z^2 = \text{Var}[Z(t)];$$

these are the objects we want formulas for. Clearly

$$m_z(t) = p_1 t, \quad t > 0,$$

so our task is to compute the variance. Define the product moment $R_z(\cdot)$ by

$$R_z(\tau) = E[Z(t)Z(t+\tau)]; \tag{A.1}$$

it is the same for any t by stationarity. Then

$$\sigma_z^2(t) = E\left\{\int_0^t \int_0^t X(u)X(v)\, \mathrm{d}u\, \mathrm{d}v\right\} - \{E[X(t)]\}^2 = \int_0^t \int_0^t E\{X(u)X(v)\}\, \mathrm{d}u\, \mathrm{d}v - (p_1 t)^2$$
$$= 2\int_0^t R_z(\tau)(t-\tau)\, \mathrm{d}\tau - (p_1 t)^2, \tag{A.2}$$

where the change of variable $\tau = v - u$ is used to obtain the last equation. We will obtain the Laplace transform of the variance from the Laplace transform of R_z.

We start by defining the probability

$$a(\tau) = P\{X(t+\tau) = 1 | X(t) = 1\}.$$

Conditioning on the value of $Z(t)$ in (A.1) yields

$$R_z(\tau) = p_0 a(\tau). \tag{A.3}$$

Towards obtaining $a(\tau)$ define the probabilities

$$b_{11}(\tau) = P\{X(t+\tau) = 1 | X(t) = 1, X(t-) = 0\}$$

and

$$b_{01}(\tau) = P\{X(t+\tau)=1|X(t)=0, X(t-)=1\}$$

for $\tau > 0$ and any t. Conditioning on the length of the first event yields

$$b_{11}(\tau) = 1 - G(\tau) + \int_0^\tau g(u) b_{01}(\tau - u)\,\mathrm{d}u$$

and

$$b_{01}(\tau) = \int_0^\tau f(u) b_{01}(\tau - u)\,\mathrm{d}u.$$

Taking Laplace transforms on both sides and solving simultaneously yields

$$\tilde{b}_{11}(s) = \frac{1-\tilde{g}(s)}{s[1-\tilde{g}(s)\tilde{f}(s)]} \tag{A.4}$$

and

$$\tilde{b}_{01}(s) = \tilde{f}(s)\tilde{b}_{11}(s). \tag{A.5}$$

We obtain a renewal equation for a by noticing that the conditioning event implies that the remaining life of the on period in progress at time t is the equilibrium excess of a generic on period, so its density, $\hat{g}$ say, is given by $\hat{g}(\tau) = [1-G(\tau)]/\nu_g$. Thus,

$$a(\tau) = 1 - \int_0^\tau \hat{g}(u)\,\mathrm{d}u + \int_0^\tau \hat{g}(u) b_{01}(\tau - u)\,\mathrm{d}u$$

which, upon taking Laplace transforms on both sides and using (A.4) and (A.5) yields

$$\tilde{a}(s) = \frac{s\nu_g(1-\tilde{g}\tilde{f}) - (1-\tilde{g})(1-\tilde{f})}{s^2\nu_g(1-\tilde{g}\tilde{f})} \tag{A.6}$$

where the argument s is suppressed on the right-side for notational ease. Taking Laplace transforms on both sides of (A.3), using (A.4)–(A.6), and then substituting into the Laplace transformed version of (A.2) produces the final result

$$\tilde{\sigma}_z^2(s) = \frac{2}{s^2}\left[\frac{\nu_g\nu_f}{s(\nu_g\nu_f)^2} - \frac{(1-\tilde{g})(1-\tilde{f})}{s^2(\nu_g+\nu_f)(1-\tilde{g}\tilde{f})}\right]. \tag{A.7}$$

References

[1] M.E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic, evidence and possible causes, IEEE/ACM Trans. Networking 5 (1997) 835–846.

[2] J. Aracil, R. Edell, P. Varaiya, An empirical Internet traffic study, Dept. of EECS, University of California, Berkeley, CA, 1997.

[3] D.P. Heyman, M.J. Sobel, Stochastic Models in Operations Research, Vol. I, McGraw-Hill, New York, 1982.

[4] N. Likhanov, B. Tsybakov, N.D. Georganas, Analysis of an ATM buffer with self-similar (“fractal”) input traffic, Proceedings of IEEE Infocom’95, 1995.

[5] P.R. Jelenkovic, A.A. Lazar, Multiplexing on–off sources with subexponential on periods: part II, in: V. Ramaswami, P.E. Wirth (Eds.), Teletraffic Contributions for the Information Age, Elsevier, Amsterdam, 1997.

[6] D. Heath, S. Resnick, G. Samorodnitsky, Heavy tails and long-range dependence in on/off processes and associated fluid models., Math. Oper. Res. 23 (1998) 145–165.

[7] C.M. Harris, The Pareto distribution as a queue service discipline., Oper. Res. 16 (1968) 307–313.

[8] W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson, Self-similiarity through high-variability: statistical analysis of Ethernet traffic at the source level., Comput. Commun. Rev. 25 (1995) 100–113.

[9] E. Willekens, J.L. Teugels, Asymptotic expansions for waiting time probabilities in an M/G/I queue with long-tailed service times., Queueing Systems 10 (1992) 295–311.

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

[11] O.J. Boxma, J.W. Cohen, The M/G/1 queue with heavy-tailed service time distribution., IEEE J. Select. Areas Commun. 16 (1998) 749–763.

[12] J. Abate, W. Whitt, Modelling service-time distributions with non-exponential tails: beta mixtures of exponentials, Stochastic Models 15 (1999) 517–546.

[13] J. Abate, G.L. Choudhury, W. Whitt, Calculation of the GI/G/I waiting time distribution and its cumulants from Pollaczeck's formulas, Archiv für Elektronik and Übertragungtechnik 47 (1993) 311–321.

[14] J. Abate, W. Whitt, The Fourier-series method for inverting transforms of probability distributions., Queueing Systems 10 (1992) 5–88.

[15] K.R. Krishnan, The Hurst parameter of non-Markovian on–off traffic sources, unpublished note, 1995.

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

[17] W. Feller, An Introduction to Probability Theory and its Applications, Vol. II, Wiley, New York, 1996.

[18] A.G. Pakes, On the tail of waiting-time distributions., J. Appl. Prob. 12 (1975) 555–564.

[19] J.W. Cohen, The Single Server Queue, North-Holland, Amsterdam, 1969.

[20] D.J. Ewing, R.S. Hall, M.F. Schwartz, A measurement of Internet file transfer traffic, Technical Report CU-CS-571-92, Department of Computer Science, University of Colorado, Boulder, CO, 1992.

[21] D.P. Graver Jr., Diffusion approximations and models for certain congestion problems., J. Appl. Prob. 5 (1968) 607–623.

[22] L. Lipsky, J.E. Hatem, Buffer problems in telecommunication networks, Fifth International Conference on Telecommunication Systems, Nashville, TN, 1997.

[23] M. Greiner, M. Jobman, L. Lipsky, The importance of power-tail distributions for modelling queueing systems., Oper. Res. 47 (1999) 313–326.

[24] P. Johri, Estimating cell-loss rates in high-speed networks with leaky bucket controlled sources., Int. J. Commun. Systems 8 (1995) 303–311.

[25] D.P. Heyman, Some issues in performance modelling of data teletraffic., Performance Evaluation 34 (1998) 227–247.

[26] F. Huebner, On the accuracy of approximating loss probabilities in finite queues by probabilities to exceed queue levels in infinite queues, internal AT& T report, February 1998.

[27] R.B. Cooper, Introduction to Queueing Theory, 2nd ed., North-Holland, New York, 1981.

[28] J. Abate, W. Whitt, Numerical inversion of probability generating functions., Oper. Res. Lett. 12 (1992) 245–251.

[29] F.N. Gouweleeuw, H.C. Tijms, Computing loss probabilities in discrete-time queues., Oper. Res. 46 (1998) 149–154.

Daniel P. Heyman did his undergraduate work in electrical and industrial engineering at Rensselaer Polytechnic Institute, and received his Masters and Ph.D. degrees in operations research from Syracuse University and the University of California at Berkeley. He has worked at Bell Labs and Bellcore, and currently is a principal technical staff member in the Network Design and Performance Analysis Department of AT&T Labs. He has published more than 60 papers and has held several elected and appointed positions in the Operations Research Society of America.