Published in *INFORMS Journal on Computing,* 2000, ***12,*** pages 261-271.

Internet-Type Queues with Power-Tailed Interarrival Times and Computational Methods for their Analysis

Carl M. Harris / *Department of Systems Engineering and Operations Research*, *George Mason University, Fairfax, Virginia 22030, USA*

Percy H. Brill / *Departments of Management Science and Mathematics & Statistics, University of Windsor, Windsor, Ontario N9B 3P4, Canada; Email: brill@uwindor.ca*

Martin J. Fischer / *Mitretek Systems, 7525 Colshire Drive, McLean, Virginia 22102, USA: Email: mfischer@mitretek.org*

Subject classifications: Communications, Queues: Algorithms, Stochastic
Other key words: Probability, Queues

(Received: April 2000; revised: July 2000, accepted: July 2000)

Internet traffic flows have often been characterized as having power-tailed (long-tailed, fat-tailed, heavy-tailed) packet interarrival times or service requirements. In this work, we focus on the development of complete and computationally efficient steady-state solutions of queues with power-tailed interarrival times when the service times are assumed exponential. The classical method for obtaining the steady-state probabilities and delay-time distributions for the G/M/1 (G/M/*c*) queue requires solving a rootfinding problem involving the Laplace-Stieltjes transform of the interarrival-time distribution function. Then the exponential service assumption is combined with the derived geometric arrival-point probabilities to get both the limiting general-time state and delay distributions. However, in situations where there is a power tail, the interarrival transform is typically quite complicated and never analytically tractable. In addition, it is possible that there is only a degenerate steady-state system-size probability distribution. Thus, an alternative approach to obtaining a steady-state solution is typically needed when power-tailed interarrivals are present. We exploit the exponentiality of the steady-state delay distributions for the G/M/1 and G/M/*c* queues, using level crossings and a transform-approximation method, to develop an alternative rootfinding problem when there are power-tailed interarrival times. Extensive computational results are given.

Willinger and Paxson (1998) summarized the well-known fact that Internet traffic characteristics do not allow it to be modeled as POTS - plain old telephone systems. So noteworthy was this fact that it also appeared in the public media, see Dye (1999). Perhaps not since Erlang (1917-1918) published his original work on telephone systems, is the time more appropriate for researchers in queueing theory to develop methods for studying congestion problems occurring in the Internet.

Current research has been in two main areas. The first and by far the most active is in the characterization of the statistical nature of Internet traffic (see Adler et al (1998), Fowler, (1999); Crovella, Taqqu and Bestavros, (1998); Crovella and Bestavros, (1997); and Paxson and Floyd, (1995)). This research resulted in the traffic being classified into one or more of the following areas:

- Self-similar (or fractal) traffic traces
- Long-range dependence
- Burstiness on multiple scales
- Long- or heavy-tailed packet interarrival times and/or service requirements.

To a lesser extent researchers in queueing theory have been developing methods to analyze these very complicated Internet congestion problems (for instance, see Andersen at al (1995), Lucantoni (1993), Lucantoni et al (1994), Nuets (1991), Nuets (1989), Feldmann and Whitt (1998), Fischer and Harris (1999), Grenier et al (1999), Heyman (1998) and Heyman and Lakshman (to appear). As pointed out in Feldmann and Whitt some of these methods like the Markovian Arrival Process (MAP) come with significant implementation challenges and they state (page 246) that new models should be sought and examined.

In this paper we present two such methods and examine them in the study of G/M/c queues, where the arrival distribution may have a long tail. Because of the multi-faceted aspect of Internet traffic (long-tailed, long-range dependence, self-similarity and burstiness) it is difficult to present a unified tool that addresses all these issues. Researchers have been developing tools that address usually just one of these aspects. Our research follows this line and deals with the development of easily computable congestion tools for queues with long or heavy – tailed arrivals. Future research will deal with the other aspects.

Fowler (1999) presents an excellent summary table of invariants (various traffic statistics) appearing in the Internet. Invariants like interarrival time of network packets, connection sizes, and connection duration is present along with their protocol level and associated probability distribution. The information in the Fowler table is drawn from Willinger and Paxson (1998), Leland et al (1994), Park et al (1996), Park et al (1997), and Pitkow (1999). From the table it is immediate that long and heavy tailed distributions play a major role in characterizing these traffic invariants.

Here, we focus on modeling queues resulting from complex, non-standard arrival streams, most frequently thought to be generated in the Internet by power-tailed distributions. Even in the basic case where such arrival streams face exponential service (thus representing possible G/M/c systems), classical queueing methods begin to struggle because they normally require a rather precise transform calculus, with full solution dependent on a definitive traffic intensity $\rho < 1$ and on the interarrival-time density having a concise Laplace transform. Since one or both of these requirements may be violated under power-tailed input, we have developed two alternative

approaches to the numerical solution of these queueing problems. The application to Internet congestion problems are packet arrivals generated from individual sources – destination networks pairs (Adler et al (1998)), packets generated by users at a key board (Fowler (1999)), URL requests (Crovella and Bestavros, (1997). and Ethernet frame interarrival times (Fowler (1999)). Greiner, Jobmann, and Lipsky (1999) have provided an especially good argument for the use of power-tailed distributions in modeling telecommunication queues, both for arrivals and service times. They have also used the G/M/1 queue as the model of choice in the presentation of their results.

The usual approach to obtaining the stationary delay-time distributions and system-size probabilities for the G/M/1 and G/M/c queueing problems requires solving a rootfinding problem involving the Laplace-Stieltjes transform (LST), $A^*(z)$, of the interarrival-time distribution function A (X). In the case of the G/M/1, the appropriate form of the problem is (often called the fundamental equation of *branching processes*)

$$z = A^*[\mu(1-z)],\ z \in (0,1) \tag{1}$$

where $1/\mu$ is the expected service time (see Gross and Harris (1998), for example). The single root to this equation in (0, 1) then becomes the parameter of a geometric distribution for steady-state system sizes at the embedded arrival points. These geometric probabilities are then combined with convolutions of the exponential service distribution to derive the stationary line-delay distribution. However, it is quite possible that the interarrival-time distribution possesses a complicated non-closed-form and/or non-analytic LST and thus that the standard rootfinding problem gets quite nasty. For example, the use of a Newton-type numerical procedure would require the derivative to exist at a possible starting point, something that might not be true when working with a non-analytic transform.

There are many familiar examples of distribution functions defined over the positive reals that lack analytic LSTs (and fail to have all moments). These include the Pareto (of the Pearson Type 6 family), folded or half-Cauchy (full Cauchy is a Pearson Type 7), and any similarly slow-varying distribution, such as the CDF (from Feller (1971))

$$A(x) = 1 - \frac{1}{\ln(x\ +\ e)} \quad (x \geq 0). \tag{2}$$

The above inverse-log distribution has a density function given by

$$a(x) = \frac{1}{[\ln(x+e)]^2}\ \frac{1}{x+e} \quad (x > 0),$$

but does not have a finite mean. Note that $a(x)$ is a completely monotone density just as the Pareto and indeed has a similar shape.

The primary class of interarrival distribution functions relevant to this paper has the property that their complementary distribution functions have tails that decay algebraically in the limit as

$$\overline{A}(x) = 1 - A(x) \sim cx^{-b}$$

for some positive constants b and c as $x \to \infty$. More formally, we would write that

$$\lim_{x\to\infty} [x^i \overline{A}(x)] = \infty \quad \forall\, i \;>\; b.$$

Such distributions are thus naturally called *power-tailed* distributions. Sometimes these distributions are also said to be *fat-tailed*, *heavy-tailed*, or *long-tailed*. But we shall use the latter terms to describe the larger class of distributions in which the tail probabilities satisfy

$$\lim_{x\to\infty} [e^{ax} \overline{A}(x)] = \infty \quad \forall\, a > 0,$$

that is, their survival functions go to 0 more slowly than any exponential.

The log-normal is a frequently quoted example of a fat-tailed distribution not formally possessing a power tail but with moments that are always finite and increasing very rapidly. The kth moment of a log-normal random variable is $\exp[k\mu + k^2\ \sigma^2/2)]$ when the originating normal has mean μ and variance σ^2. We are especially interested in the log-normal as an example in the current context because it is often a feasible modeling alternative to a power-tailed distribution. Another well-known example of a heavy-tailed distribution is a Weibull with shape parameter (say β) less than 1. If α is the scale parameter its k-th moments $\alpha^{-k/\beta}\,\Gamma(1 + k/\beta)$ will grow quickly if $\alpha \le 1$ (Mood, Greybill, Boes, p.543, (1974)).

An immediate consequence of the power-tailed property is that the moment integrals eventually lack finite moments and therefore no analytic Laplace Transform exists. In the queueing context, the most difficult problem occurs when the interarrival times do not possess one single moment. The possible inability to form a nice rootfinding problem thus forces us to find alternative ways to analyze the G/M/1 and G/M/c systems under such power-tailed circumstances. We know at the beginning, however, that at least there is a legitimate steady-state line-delay distribution function for any interarrival-time CDF even when the expected-value integral diverges to 4, that is, when $\lambda = 0$ (see Section 1).

In Section 1, we offer the major theoretical foundations of our work. Then we use level-crossing methods in Section 2 to derive the key integral equation for the G/M/1 problem. Section 3 outlines the alternative numerical method for the G/M/1, wherein we use a very large discrete approximation of the interarrival CDF in solving the standard branching-process rootfinding problem. We have called this approach *TAM*, for Transform Approximation Method. We get back to level crossing in Section 4, where we set up the appropriate numerical problem for the multi-server G/M/c queue. Our numerical results follow in Section 5, and some thoughts on possible generalizations using level crossings are presented in Section 6. We close with concluding remarks in Section 7.

1. Theoretical Foundations

In the standard analysis of the G/M/1 queue (for example, see Gross and Harris, (1998)), the steady-state probability for the number of customers Q in system just before an arrival is given for all nonnegative n by

$$\Pr\{Q = n\} = q_n = (1 - r_0) r_0^{\,n},$$

where, as noted earlier, r_0 is the distinct root in (0,1) of the transform equation given by (1), while the steady-state general time probabilities are $p_n = \lambda\, q_n/\mu$ ($n > 0$) and $p_0 = 1 - \lambda/\mu$. In addition, the steady-state queue and system waiting-time distribution functions are respectively given for $x \geq 0$ by

$$W_q(x) = 1 - (1 - q_0)\, e^{-\mu(1-r_0)x} = 1 - r_0 e^{-\mu(1-r_0)x} \tag{3}$$

$$\text{with mean } W_q = \frac{r_0}{\mu(1 - r_0)},$$

and

$$W(x) = 1 - e^{-\mu(1-r_0)x} \quad \text{with mean } W = \frac{1}{\mu(1 - r_0)}.$$

These results are valid for all forms of the interarrival distribution $A(X)$; but the complexity of the numerical problem of obtaining r_0 very much depends on the algebraic nature of $A(X)$. In the event that the Stieltjes transform of $A(X)$ is available in relatively concise closed-form, we know, for example, that a standard successive-substitution routine is an effective way to find the needed root r_0 (see Gross and Harris (1998) pp 250-254).

Normally, we require that the traffic intensity ρ, defined as the ratio of the mean service time to the mean interarrival time, be less than 1 for steady state to exist. However, since we are focusing on a power-tailed input here, there is the possibility that the mean interarrival time may not exist (or call it 4). Fortunately, we do not have a problem in this case with regard to the existence of a limiting delay-time distribution function since steady state is reached by the underlying delay-time random walk whenever process-state 0 is positive recurrent (see Cohen, (1982)). We can easily see that this is the case given the regular sparse nature of a power-tailed interarrival process with an infinite expectation.

In fact, there is sort of a reverse problem here, in the sense that we could assign a value of 0 to the utilization rate ρ when the mean interarrival time $(1/\lambda)$ does not exist. The utilization rate $\rho = (1/\mu)/(1/\lambda) = (1/\mu)/4 = 0$ in a G/G/1 suggests that the server is never busy and prevents us from giving meaning to many standard queueing results; more formally, the limiting general-time state process is degenerate and puts all probability mass on state 0. This follows from the fact that $p_n = \lambda\, q_n/\mu$ ($n > 0$) and $p_0 = 1 - \lambda/\mu$. We should also add that, despite the guaranteed existence of the limiting mean line delay and system waiting time in a G/M/1 queue lacking a first interarrival time

moment, we cannot use Little's formula to find the limiting queue and system sizes because λ is essentially 0. But, again, there will always be a legitimate limiting probability distribution $\{q_n\}$ for the system sizes seen by the arriving customers. The expected value of these sizes can be found directly from the limiting mean line delay as μW_q and is $r_0/(1 - r_0)$ for the G/M/1, power-tailed interarrival times or not.

For the G/M/1 queue the steady state distribution of the waiting time exists even if the interarrival CDF A(≅) has an infinite mean. To see this fact, define $A_n(x) = A(x)/A(t_n)$ for 0#x#t_n, and $A_n(x) = 1$ for $t_n < x$, where t_n is finite (n∃1), $\{t_n\}$84 and $\int_{(0,\infty)} (1- A_n(x))\, dx > 1/\mu$ for every n∃1. Thus A_n(≅) is a finite truncation of A(≅) at t_n and the mean of A_n(≅) is finite and exceeds $1/\mu$ for every n∃1. Moreover $A_n(x)$ 9 A(x) for every x∃0 as n6 4. By Corollary 7.5, page 80 of Asmussen (1987), derived starting with the Lindley recursion, the steady state distribution of the waiting time for the queue A_n(≅)/M/1 with mean service time $1/\mu$ exists for every n∃1. Letting n 6 4 implies that the steady state distribution of the waiting time exists for the queue A(≅)/M/1. A similar argument applies to the G/M/c queue, with $1/\mu$ replaced by $1/(c\mu)$.

For the G/M/c system, the transform equation and rootfinding problem of (1) is slightly modified to incorporate the number of servers *c*, as

$$z = A^*[c\mu(1 - z)]. \tag{4}$$

The resultant queue and system waiting-time distribution functions are given respectively in terms of the root r_0 of (4) by

$$W_q(x) = 1 - [1 - W_q(0)]\, \exp(-c\,\mu\,(1 - r_0)x) \tag{5}$$

and

$$W_q = [1 - W_q(0)]\,/(c\,\mu\,(1 - r_0)). \tag{6}$$

It easily follows from (5) that the conditional distribution function of the line delay, given that there is a delay $x > 0$, has an exponential distribution with rate parameter $c\,\mu\,(1 - r_0)$. But one of the more difficult parts of completing the G/M/c solution is obtaining $W_q(0)$ for (5) and then the entire set of steady-state system probabilities. See Gross and Harris (1998) for the relevant details of the classical approach.

2. Integral Equation for the Absolutely Continuous Portion of the G/M/1 Delay-Time CDF

Since we know that any stationary G/M/1 queue, power-tailed interarrival times or not, has the limiting delay distribution function given in (3), it follows that the derivative of the absolutely continuous portion of $W_q(x)$ defined over the positive reals has form

$$f(x) = \mu r_0 (1 - r_0)\, e^{-\mu(1-r_0)x}.$$

For ease of computation, we henceforth write $\gamma = \mu(1 - r_0)$ and $K = r_0\gamma = \mu r_0 (1 - r_0)$, and thus have

$$f(x) = Ke^{-\gamma x} \quad (x > 0). \tag{7}$$

Brill (1979, 1988) has shown that any such function $f(x) = W_q'(x)\,(x > 0)$ satisfies the equation

$$f(x) = \mu \int_{y=x}^{\infty} \overline{A}(y - x) f(y) dy \quad (x > 0), \tag{8}$$

where $A(x)$ = Pr{interarrival time # x}, $\overline{A}(x) = 1 - A(x)$, and μ = exponential service rate. The derivation of (8) given in the above references is by means of an embedded level-crossing technique and by a Acontinuous@ level-crossing method, respectively. Most importantly, no assumptions whatsoever are made about the existence of the moments of the interarrival-time distribution in either of these derivations. Thus it is not necessary for A(X) to possess even a mean or variance, and hence (8) holds for any G/M/1 queue in which A(X) is Pareto, folded Cauchy, or any other power-tail distribution.

We now substitute (7) into (8), and obtain a relatively simple integral equation for the exponential rate γ, given by

$$\int_{u=0}^{\infty} \overline{A}(u) e^{-\gamma u} du = \frac{1}{\mu}. \tag{9}$$

In (9), $\overline{A}(u)$ and μ are assumed to be known, while $\gamma = \mu(1 - r_0)$ is the problem=s unknown. Then, once we have solved (9), the full form of the limiting delay-time CDF is given by

$$W_q(x) = 1 - \frac{K}{\gamma} e^{-\gamma x} = 1 - (1 - \frac{\gamma}{\mu}) e^{-\gamma x} = 1 - r_0 e^{-\mu(1-r_0)x} \quad (x \geq 0). \tag{10}$$

Note that the left-hand side of (9) is in fact the Laplace transform of the complementary CDF, evaluated at γ. Thus we have replaced the role of the Laplace-Stieltjes transform in the rootfinding; much like one can change the computation of moments by integrating with respect to the complementary CDF. Equations (9) and (1) are equivalent, as seen by a quick integration by parts or equivalently by expressing the Laplace transform of $\overline{A}$ (≅) in terms of the Laplace transform of the derivative of A(≅). We introduce and use (9) because: (i) it is derived in one step by balancing up and downcrossing rates, whereas the classical derivation of (1) is lengthy (e.g., Gross and Harris(1998), pp 248-250); (ii) its use is consistent with the derivation of equations (15) - (18) for the G/M/c queue given in section 4, and the level crossing approach leads to intuitive and conceptual interpretations about the model such as that given in the last paragraph of section 5; (iii) it is readily amenable to state dependent generalizations, such as those discussed in section 6. Either equation

can be solved by search techniques, Newton=s method, MAPLE computer software, etc. Although the degree of difficulty or ease in solving (9) and (1) is generally the same, we have found examples with heavy-tailed interarrival distributions where it is faster to solve (9) than (1) when using the same technique.

In our subsequent computations below, we use the unit (one-parameter) Pareto, folded Cauchy, inverse-log distributions, and log-normal all as illustrations. Our version of the Pareto distribution will have CDF $A(x)$, $x \exists 0$, and density $a(x)$, $x > 0$, given by

$$A(x) = 1 - \frac{1}{(1+x)^{\beta}} \quad \text{and} \quad a(x) = \frac{\beta}{(1+x)^{\beta+1}},$$

while the CDF and density for the folded unit Cauchy are

$$A(x) = \frac{2}{\pi} \arctan(x) \quad \text{and} \quad a(x) = \frac{2}{\pi(1+x^2)}.$$

Using this form for the Pareto, Equation (9) becomes

$$\int_{u=0}^{\infty} \frac{e^{-\gamma u}}{(1+u)^{\beta}}\, du = \frac{1}{\mu}. \tag{11}$$

Note that the Pareto has all moments up to $\lceil \beta - 1 \rceil$, where $\lceil x \rceil$ = least integer $\geq x$. Thus if the shape parameter $\beta < 1$, $A(X)$ has no mean or variance; but (11) yields a solution for γ nonetheless. The final form of the derivative of the absolutely-continuous portion of the delay-time CDF is given by $f(x) = Ke^{-\gamma x}$, where γ is obtained from (11) and $K = \gamma(1 - \gamma/\mu)$. As in (3), the mean line delay W_q is given by $r_0/[\mu(1 - r_0)] = 1/\gamma - 1/\mu = K/\gamma^2$.

If instead, the interarrival times were assumed to have the folded Cauchy distribution, then we would write Equation (9) as

$$\int_{u=0}^{\infty} [1 - \frac{2}{\pi} \arctan(u)] e^{-\gamma u}\, du = \frac{1}{\mu} \tag{12}$$

and once again calculate $K = \gamma(1 - \gamma/\mu)$ and $W_q = K/\gamma^2$ using the solution γ to (12). For inverse-log interarrival times as in (2), Equation (9) becomes

$$\int_{u=0}^{\infty} \frac{e^{-\gamma u}}{\ln(u+e)}\, du = \frac{1}{\mu}. \tag{13}$$

This turns out to be a highly efficient way to work with inverse-log interarrivals, preferable to using its density in (1).

3. Transform Approximation Method

A most feasible alternative approach to the numerical portion of power-tailed G/M/1 and G/M/c problems is the use of the transform approximation portion of the Harris and Marchal (1998) transform-matching method, as briefly outlined in Fischer and Harris (1999). Henceforth we refer to this as the transform approximation method (TAM). To use TAM, we take the Laplace-Stieltjes transform B call it $A_N^*(z)$ B of an equiprobable N-point approximation (N Asuitably large@) of the interarrival distribution A(X), such that

$$A_N^*(z) = \frac{1}{N} \sum_{k=1}^{N} e^{-zx(k)} \tag{14}$$

where the $x(k)$, $k = 1, 2, ..., N$, are chosen to cover the outcome space of the original interarrival random variable with $A(x(k)) = (k - 0.5)/N$. [That is, $\lim_{N\to\infty} A_N^*(z) = A^*(z)$ in the sense of the fundamental definition of Stieltjes integration.] Then we go back through all of the standard G/M/1 or G/M/c numerics using the approximate Stieltjes integral $A_N^*(z)$ in either Equation (1) or (4), thus not requiring any numerical Riemann integration on (0, 4). Note, however, that the value of N might be quite large (maybe as high as 10^6), so that we cannot say that the Stieltjes transform of the A(X) is in the "concise form" we require to make the rootfinding problem of Equation (1) attractive.

These approximate Stieltjes integration calculations can be done very quickly, for example, with a short Visual Basic function macro in Excel. However, the problem with this method, in contrast to the level-crossing approach, is that there are two stages of approximation, the first coming in finding an appropriate value of N and the second resulting from the numerical rootfinding of (1) or (4). The rest of our observations about the overall queueing problems themselves remain the same. But it should be emphasized that the use of TAM is a very competitive computational method. For a fixed value of N, even as large as 10,000 or more, the computational speed is impressive.

For the three main distributions with which we are working, the Pareto (P), folded Cauchy (C), and inverse log (L), the transform approximation of Equation (14) may be written out by appropriately inverting each distribution function. The results are

$$\text{for } P\text{: } A_N^*(z) = \frac{1}{N} \sum_{k=1}^{N} e^{-z\left([(k-0.5)/N]^{-1/\beta}-1\right)},$$

$$\text{for } C\text{: } A_N^*(z) = \frac{1}{N} \sum_{k=1}^{N} e^{-z\tan[\pi(k-0.5)/2N]},$$

$$\text{for } L\text{: } A_N^*(z) = \frac{1}{N} \sum_{k=1}^{N} e^{-z\left[e^{N/(k-0.5)}-e\right]}.$$

All the numerical work later presented in the paper has been carried out for both level crossing and TAM. We are able to get precisely the same answers by manipulating the parameters of the respective numerical routines. By the same token, it is difficult to compare the two methods because they are both rootfinding problems essentially built around approximate Laplace transforms comprised of sums of a comparably large number of exponential terms. We say essentially because the level-crossing approach requires the numerical solution for the functional inverse of the Laplace transform of $\overline{A}(.)$ evaluated at $1/\mu$, and its performance is a function of the particular numerical method used.

It is important to note that TAM is especially well suited for heavy-tailed situations where large interarrival-time data sets are available and there has yet been no distribution fitting. Then Equation (14) can be used precisely as outlined above and the entire numerical procedure is unchanged. Thus, overall, the use of TAM would contrast sharply in approach and how it does its calculations with attempts to find an alternative member of another class of distributions as a replacement fit for the heavy-tailed interarrival distribution (e.g., see Feldman and Whitt (1998); Harris and Marchal (1998); Greiner, Jobmann, and Lipsky (1999)).

4. Integral Equation for the Absolutely Continuous Portion of the G/M/c Delay-Time CDF

For the G/M/c queue, a level-crossing approach analogous to the methods in Brill (1979, 1988) leads to the system of integral equations

$$f_c(x) = c\mu \int_{y=x}^{\infty} \overline{A}(y-x)\, f_c(y)dy \quad (x>0), \tag{15}$$

and

$$g_{c\text{-}1}(x) + (c-1)\mu \int_{-\infty}^{x} g_{c\text{-}1}(y)\, dy = c\mu \int_{y=0}^{\infty} \overline{A}(y-x) f_c(y)dy + g_{c-2}(0)\, \overline{A}(-x) \quad (x < 0), \tag{16}$$

$$g_i(x) + i\mu \int_{-\infty}^{x} g_i(y)\, dy = (i+1)\mu \int_{-\infty}^{x} g_{i+1}(y)\, dy + g_{i\text{-}1}(0)\, \overline{A}(-x) \quad (i = 1,\ldots,c-2; x<0), \tag{17}$$

$$g_0(x) = \mu \int_{-\infty}^{x} g_1(y)dy \quad (x<0). \tag{18}$$

We briefly outline the derivation of Equations (15) - (18). Let the system state be {V(t),i}, t>0, i = 0,...,c. When i =c, V(t) denotes the age in the system at instant t, of the last customer that entered service. When i 0 {0,...,c-1}, -V(t) denotes the remaining interarrival time at instant t, i.e. until the next arrival to the system and there are i customers in service. Note that when i = c, V(t) > 0 and when i < c, V(t) < 0. In both cases the corresponding sample path has slope +1. The steady state distribution of {V(t)} is the same as that of the waiting time (Brill (1988)). Let $f_c(\cong)$ denote the derivative of the absolutely continuous portion of the steady state mixed joint CDF of {V(t),c}. Let $g_i(\cong)$ (i = 1,...,c-1) denote the derivative of the absolutely continuous portion of the steady state mixed joint CDF of {V(t),i}. In equation (15) the left-hand side represents the continuous sample

path (SP) upcrossing rate of level x when i = c. The right-hand side of (15) represents the jump SP downcrossing rate of level x, due to service completions when there are c customers in service. Equating SP up and downcrossing rates of level x > 0 when i = c yields equation (15). In equation (16) the left-hand side represents the SP exit rate from state set {(-4,x),c-1}. The first term on the left is the continuous SP upcrossing rate of level x. The second term is the service completion rate when there are c-1 customers in service and the remaining time until the next arrival exceeds -x. The right-hand side of (16) represents the SP entrance rate into state set {(-4,x),c-1}. The first term is the downward jump SP rate from positive waiting states, i.e. when there are c customers in service, into {(-4,x),c-1} in which there are c-1 customers in service and the remaining time until the next arrival exceeds -x. The second term is the rate at which there are arrivals when there are c-2 customers in service and the next total interarrival time exceeds -x, thereby generating an entrance into {(-4,x),c-1}. Equation (16) then follows from the law of conservation of exit and entrance rates for {(-4,x),c-1}. The equations in (17) are derived in a similar manner. In practice these equations can be written down quickly by inspection of the sample path.

Once more, we know that for all G/M/c queues $f_c(x) = Ke^{-\gamma x}$ $(x > 0)$ for some positive K and γ, so that γ is the solution to

$$\int_0^\infty \bar{A}(u)e^{-\gamma u}\, du = \frac{1}{c\mu}. \tag{19}$$

Now let w_0 = Pr{the wait in line of an arrival = 0}. From the level-crossing approach, the rate at which zero-waiting customers arrive is given by $\sum_{i=0}^{c-1} g_i(0)$, and the rate at which non-zero-waiting customers arrive is given by

$$\begin{aligned} c\mu \int_0^\infty A(y)\, Ke^{-\gamma y} dy &= -c\mu K \int_0^\infty \bar{A}(y)e^{-\gamma y} dy + \frac{c\mu K}{\gamma} \\ &= -K + \frac{c\mu K}{\gamma} \end{aligned}$$

by applying (19). Hence

$$w_0 = \frac{\sum_{i=0}^{c-1} g_i(0)}{\sum_{i=0}^{c-1} g_i(0) + c\mu K/\gamma - K}\ . \tag{20}$$

Now, from level-crossing theory, $g_{c-1}(0) = f_c(0) = K$.

For the G/M/2 queue, we obtain upon solving (15) - (18) for $g_0(0)$, the formula

$$w_0 = \frac{K\ +\ KB}{K\ +\ KB\ +2\mu K/\gamma\ -K} = \frac{1\ +\ B}{B\ +\ 2\mu/\gamma} \tag{21}$$

where

$$B = \frac{1 - 2\mu \int_0^\infty e^{-\mu t} \int_0^\infty a(y+t)\, e^{-\gamma y} dy\, dt}{\int_0^\infty e^{-\mu t} a(t)\, dt} \tag{22}$$

and $a(x) = dA(x)/dx, x > 0$. We outline the derivation of the expression for B given in (22). Note that $g_0(0)$ is a constant on the right-hand side of (16), and is also a term in (20). First we differentiate with respect to x<0 on both sides of (16) and form a first order differential equation for $g_1(x)$, x<0. Applying the boundary conditions $e^{\mu x}$ $g_1(x)$ 6 $g_1(0)$ = K as x80 and $e^{\mu x}$ $g_1(x)$ 6 0 as x9-4 , we then obtain $g_0(0)$ = KB with B given by (22). The value of K is then obtained using the normalizer $w_0 + \int_0^\infty f_c(x)dx = 1,$ which yields

$$K = \gamma(1 \,!\, w_0). \tag{23}$$

All moments about 0 of the wait of an arrival can be found as

$$E[W_q^{\,n}] = \int_0^\infty y^n K e^{-\gamma y} dy = \frac{n!K}{\gamma^{n+1}} \quad (n \geq 1). \tag{24}$$

In particular, the mean and variance are

$$W_q = K/\gamma^2,\ Var[W_q] = K(2\gamma - K)/\gamma^4. \tag{25}$$

Note that since $w_0 > 0$, it follows from (23) that $\gamma > K$.

The computational procedure for G/M/c is thus:

$ Use (19) to solve for γ.
$ Use (15) - (18) to solve for $g_i(0)$, $i = 0, \ldots, c-1$, noting that $g_{c-1}(0) = K$.
$ Use (20) to solve for w_0.
$ Use (23) and (25) to obtain the value of K and the moments of the steady-state line waiting time.

In the particular case when $c = 2$, Step (ii) yields $g_0(0) = KB$ and Step (iii) yields w_0, using (21) and (22).

To provide a numerical illustration of the problem in a multi-server setting, we shall do a set of calculations for a Cauchy/M/2 system, where the values for the mean service rate, except for the final four rows, will be the same as those found in Table I for the Cauchy/M/1 example (thus here E[*S*] = previous 2E[*S*]). As noted above, the equation for γ that thus comes out in (19) for the Cauchy/M/2 is precisely the same as that found in (12), with the latter=s right-hand side replaced by

$1/(c\mu)$. The results are displayed in Table V.

5. Numerical Results

In all of Table I and in Table II, numerical values of the level-crossing parameter γ were found using MAPLE and its numerical solver, or by using our own coding of Newton=s method. The latter occasionally was faster for problems in which the result for r was coming out very close to 1. Both of these numerical approaches work especially well in our problem because the key integral of Equation (9) is monotone decreasing and convex in γ for all of our examples. The results presented in these tables were found using both TAM (Table I) and our level crossing method (Table II). The roots found via TAM in Table 1 were also verified using our level crossing method and the roots of Table II were verified with TAM. We also did some further selective checks of our answers using long simulation runs of the relevant Markovian G/M/1 delay-time process.

The most important observations to be made immediately from our numerical experiments are that all our approaches work beautifully and that the resultant experimental queues are thus completely analyzed. We never came across a problem that we could not solve, Pareto or Cauchy, single or multiple servers. In addition, all the results are consistent with our earlier assertion that the delay process in all of these queues exists, independent of the number of interarrival-time moments possessed. It is also worth noting that infinite mean interarrival times create the anomalous situation in which the traffic intensity is (essentially) always 0 while the limiting mean line delay increases without bound as the mean service time grows.

All of the single-server Cauchy and inverse-log examples, as well as those for the Pareto with $\beta < 1$, are especially interesting since these are precisely the queues having infinite mean interarrival times. In these situations, randomly generated sequences of interarrival times would not satisfy the strong law of large numbers. Yet, we get the surprising result that not only does the line delay process $\{W_n\}$ have a steady state (which we have already discussed), but that the rate of its convergence into the Acorrect neighborhood@ of values is quite rapid and follows with the value of the root r_0, much as would the delay sequence behave in terms of ρ when it exists. To give some numerical meaning to this assertion, in Figure 2 we have taken the Cauchy example from Table II where $E[S] = 1$ and made simulation single runs of ever-increasing length, beginning with 100,000 and going up to 10 million. As we know from the table, the actual (numerically derived) answer is $W_q = 1.4342$. In these runs, the furthest away from the actual answer was 1.4077 (at 500,000) or 1.15%, and at end of the full complement of 10^7, we were within 0.15%. Real sharp precision is taking a while, but fairly good answers are available almost immediately.

To get a sense of how the mean line delay moves with the variability of the Pareto, we can make comparisons across the β groupings, keeping ρ constant and examining resultant differences in W_q in Table I. We should indeed expect mean line delay to go down as the variability of the interarrivals decreases, with their mean kept constant. This is precisely what happens. For example, a value of $\rho = 0.3$ appears in both the $\beta = 1.5$ and 2.5 data sets, and we see respective mean line delays of 0.8965 and 0.1467. For $\beta = 2.5$ and 3.5, we have the repeat appearance of $\rho = 0.75$, with respective mean delays of 2.9437 and 1.3491. And, of course, once we assume that the mean

interarrival time exists, we have a legitimate steady-state general-time probability distribution with all $p_n > 0$ and Little=s laws are at once applicable.

There are a number of further observations we can make from the data of Table I. First, as in the classical G/M/1 model, we see that the root $r_0 = 1 ! \gamma/\mu$ is always greater than $e^{-1/\rho}$ in those cases where the first interarrival moment exists. Also as expected, the difference between ρ, $e^{-1/\rho}$ and r_0 goes to 0 as ρ (if it exists) and r_0 increase to 1.

Figure 1 offers a plot of the mean service time versus the estimated mean line delay for the data of Table II, and we see the sort of monotone increasing and convex behavior we should be expecting. If we had instead used r_0 as the abscissa for this figure, we would have gotten a Asteeper@ curve, still monotone increasing and convex. Remember, however, that there is nothing comparable to the steady-state property that $\rho < 1$, so that the mean delay will exist for all mean service times and there is thus no asymptotic behavior.

Table III contains results for six G/M/1 problems using the inverse-log distribution of (2) for the interarrival times (using both level crossing and TAM). As we noted earlier, the use of the level-crossing equation (9) works especially well for this case. The pattern of the results is very similar to that of the Cauchy/M/1, though the new values of W_q stay uniformly lower than those for the Cauchy example. Table IV, in turn, presents comparisons between a Pareto with shape parameter .95 and a similarly looking log-normal whose originating normal has mean 0 and standard deviation 1.5, thus a mean of $\exp(1.5^2/2) = 3.0802$ and second moment of $\exp(1.5^2) = 9.4877$. The critical point here is that the relative mean delays track quite well for low values of log-normal model=s traffic intensity, but begin to diverge quite sharply as the mean service time grows (see Figure 3). That is, the impact of the Pareto=s fat tail becomes more and more pronounced with increasing mean service time.

Figure 4 illustrates these ideas even more extensively. In that figure the coefficient of variation (= standard deviation of interarrival time / mean interarrival time) is equal to 3 for all cases. We used TAM to evaluate Wq for the Pareto, Log Normal, Weibull and Gamma distributions, the Pareto being a power tailed distribution, the Log Normal and Weibull heavy tailed, and the Gamma exponential tailed. We see the effect of the power tail in that it clears out the queue and results in smaller delays than the Log Normal and Weibull. The gamma, with an exponential tail, has the largest delay.

Next, we have run an illustrative set of cases for $c = 2$, with interarrival times assumed to be distributed as a folded-Cauchy, and the results are displayed in Table V using level crossing. Some of the values for the mean service rate here are exactly the same as those found in Table II for the Cauchy/M/1 example (thus E[S] now = 2 H previous E[S]). As noted earlier, the equation for γ that thus comes out in (18) for the Cauchy/M/2 is precisely the same as that found in (12), with the latter=s right-hand side replaced by $1/(c\mu)$. Once again, results are confirmed by the use of the transform approximation method.

It is interesting to note in Table V that the values in the columns for γ, r_0, and W_q are monotonic. However, the values for K increase to a maximum at E[S] = 1.2 (row 3), and then decrease. From level-crossing theory, the value K represents the arrival rate to the system when there

are exactly c !1 servers occupied (in this case, exactly one server occupied). If E[S] (and therefore r_0) are small, then fewer arrivals will find exactly one server occupied, since the system will be empty much of the time. However, as E[S] (and r_0) increases gradually, more and more arrivals will find exactly one server occupied. Then, as E[S] (and r_0) increase further, more arrivals will have to wait in line, so fewer arrivals will find exactly one server occupied. This analysis intuitively explains the variation in the K values as E[S] and r_0 increase, that is, from small values to a maximum, and then gradually decreasing. The same would be true for $c > 2$ servers and arrivals then finding exactly c !1 servers occupied. There may be a possible application of this observation. That is, are there situations where it is important to maximize the proportion of arrivals that find exactly c !1 servers occupied, or equivalently, where arrivals just fill the c available servers and there are no customers waiting in line?

It is well know that phase type distributions can be used to approximate any probability density over the reals. Feldmann and Whitt (1998) consider this fact in developing a hyperexponential distribution approximation to distributions with heavy tails. They show that a hyperexponential distribution can be fitted to any power tailed distribution, and present a fitting algorithm that is recursive over time. The computational problem rests in the development of a good fitting algorithm that requires a large number of exponential distributions. They relate their development to the use a phased-based distribution as a heavy tail approximation and state that one would also be faced with a non-trivial fitting problem.

A natural question is how does that approach compare with the ones discussed in this paper. We have compared the Weilbull example given on pages 249-251 of Feldmann and Whitt (1998) and TAM. We solved the root finding problem for the Weibull/M/1 queue using the Laplace-Stieltjes transform of a hyperexponential distribution and the parameters given in Table I of their paper. In this comparison the mean interarrival time is 1.0 and the mean service time is 0.8. The G/M/1 root via Feldmann and Whitt was 0.9804 and resulted in $W_q = 40.09$. Using TAM (with N= 10000) we got the root to be 0.9808 and $W_q = 40.98$. Thus, both methods fundamentally got the same result in less than one second on the same personal computer.

There is an important difference that one has to consider when comparing these methods. When using a Phased-based (or hyperexponential) approximation one has to use or develop a fitting algorithm. From a practical point of view that is not a major problem but it constrains the accuracy of the solution and potentially increases the run time. Using TAM no fitting algorithm is required and the time required evaluating (14) is minimal. Thus, TAM is easier to implement and overall is faster.

Feldmann and Whitt also point out that when their hyperexponential approximation was used to approximate a long-tailed service time distribution it inherited many of the same difficulties as the original distribution. We are currently in the process of applying TAM to this class of problems and our results will be reported in future papers.

Comparisons with the level crossing methods are not as straight forward. Since the level crossing method and TAM get the same results on the G/M/c problem, accuracy should not be an issue. We have used MAPLE software numerical procedures to generate numbers with level

crossing. The fitting algorithm when using a hyperexponential approximation or phased-based distribution has to be developed. Given that Feldmann and Whitt recently published their algorithm, it seems reasonable to assume that an off the shelf fitting algorithm for this class of problems does not exist. Thus, it is plausible that our level crossing method would be easier to implement.

6. Extensions and Generalizations

In this work, we have advocated two alternative numerical approaches to the solution of G/M/c-type queues when the interarrival times have a power-tailed distribution, one using level-crossing balance arguments and the second employing approximate Stieltjes integration in the TAM approach. Each method has its strengths and weaknesses. It turns out that one of the major benefits of using level crossing is that its use allows the standard G/M/c paradigm to be generalized to much larger classes of possible models without altering the basic numerical ideas. A specific and potentially useful idea is to allow the introduction of state dependence into the models presented in Sections 2 and 3.

First, assume that the interarrival times depend on the system workload. Let W_n, S_n, and U_{n+1} denote the wait in line and service time of customer n, and the interarrival time between customers n and $n + 1$, respectively. Then U_{n+1} will depend on $W_n + S_n$ ($n \exists 1$). Denote this model by G(X)/M/1.

The derivative f(X), of the absolutely continuous part of the CDF of the waiting time now satisfies the integral equation

$$f(x) = \mu \int_{y=x}^{\infty} \overline{A}(y - x; y)\, f(y) dy \quad (x > 0), \tag{26}$$

where $\overline{A}(\bullet, y)$ represents the complimentary distribution function and its general dependence on the system workload through the variable y. As a fairly general type of dependence, consider

$$\overline{A}(\bullet, y) = \overline{A_i}(\bullet) \quad (x_i < y \leq x_{i+1}), \tag{27}$$

where $0 \equiv x_0 < x_1 < \ldots < x_M < x_{M+1} \equiv \infty$ and M is a positive integer. Then (26) expands into a system of $M + 1$ integral equations for f(X), given by

$$\begin{aligned} f(x) = \mu \int_{y=x}^{x_{i+1}} \overline{A_i}(y-x) f(y) dy \\ + \mu \sum_{j=i+1}^{M} \int_{y=x_j}^{x_{j+1}} \overline{A_j}(y-x) f(y) dy \quad (i = 0, \ldots, M; x_i < x \leq x_{i+1}). \end{aligned} \tag{28}$$

In (28), A_i(X) will have the form of the power-tailed distributions considered earlier. The probability that an arrival has a wait of 0 is given by

$$w_0 = \frac{f(0)}{f(0) + \mu \sum_{i=0}^{M} \int_{x_i}^{x_{i+1}} A_i(y) f(y) dy}. \tag{29}$$

The normalizing condition is

$$w_0 + \int_0^{\infty} f(x)\, dx = 1. \tag{30}$$

We can further generalize this model by assuming additionally that the service rate μ depends on the workload, that is, S_{n+1} depends on $W_n + S_n$. Thus let

$$\mu(y) = \mu_i \ , \quad x_i < y \leq x_{i+1}, \ \text{i= 0,…,M} \tag{31}$$

where the $\{x_i\}$ are defined as in (27). The system state can be defined as ($V(t)$, i), where $V(t)$ is the system time of the customer in service at instant $t \geq 0$, and μ_i is the corresponding service rate. In this case the system of integral equations for f(X) expands further, and the level-crossing approach can be used to write down the appropriate system.

7. Concluding Remarks

The rapid emergence of the Internet has created some interesting and challenging telecommunications modeling problems. As we have noted, one of the unusual factors is the possible presence of interarrival streams generated by power-tailed distributions. In such circumstances, the standard solution methods for the G/M/1 and G/M/c queues do not work well because of the failure of the interarrival times to possess all moments. To get around this problem, we have used both

level-crossing methods and transform approximation for numerically analyzing G/M/1 and G/M/c queues with power-tailed interarrival times. We have demonstrated the method's use by presenting several detailed examples using Pareto, folded Cauchy, and inverse-log distributions, and have compared our results with other published methods.

Our numerical work has shown that the power-tailed interarrival distribution can force such queues to behave very awkwardly. For example, in cases with an effective infinite interarrival-time mean, congestion occurred even when the expected load could not be defined. We have thus derived the full form of the exponential steady-state line-delay distribution, even though there will not be a nondegenerate distribution of system sizes. The long-tailed nature of the power-tailed input process helps to clear out congestion when a large interarrival time has occurred. Overall, our methods can be applied for all fat-tailed G/M/c queues, independent of their shape and the number of actual moments possessed.

Our procedures always generate a complete analysis of the steady-state delays of the queues we study. The approach is both analytic and numerical, and also utilizes some known closed-form mathematical results from the classical queueing literature. Each method is computationally quick and their accuracy can be well controlled. Throughout, it is important to remember that the delay process is always well behaved, even though the sequence of interarrival times may not satisfy a law of large numbers.

As a result of our examination we feel that any numerical procedure that captures the long-tailed nature of the interarrival times could be used to study these G/M/c queues. These queues are not as sensitive to solution methods as the dual problem – long tailed service. Our future work will turn to similar numerical work for M/G/1 and M/G/c queues. In those queues it is well known that long tailed service times can significantly impact congestion.

Remembrance

Dr. Brill and Dr. Fischer would like to point out the very special place that this paper has in our professional lives. We have been very fortunate to be associated with Dr. Carl Harris for a very long time. This is the last paper that Carl wrote. During a meeting at George Mason University in December 1999, Carl suggested that we collaborate on work that resulted in this paper. Carl organized and led the work. He wrote up the initial version of the paper and submitted it to the INFORMS Journal on Computing. Tragically, within a week after he submitted the paper Carl passed away. It was truly a very sad event, and it has been difficult trying to carry this work forward. We feel very honored to have worked with such a great contributor to our field over many years, and to have had the opportunity to work so closely with him on his final paper.

Acknowledgments

We thank Dr. David Kelton; editor-in-chief of JoC for his encouragement, in carrying out the revision based on the reviewers' reports. We thank the associate editor and referees for their quick, thorough, reviews of the initial submission, under trying circumstances. Their comments and suggestions have helped us to improve the paper. In addition to their home institutions, the authors wish to express their sincere appreciation to the Natural Sciences and Engineering Research Council of Canada and the National Science Foundation of the United States (under Grant No. DMI-0071185) for providing the critical support necessary for this research. Dr. Martin Fischer would like to thank the support he received from Mitretek Systems. The authors would also like to thank Professor Donald Gross and the rest of the members of the George Mason Internet Traffic Engineering Project for their help in the preparation of this paper.

β	E[S]	ρ	r_0	W_q
0.5	1	DNE	0.2912	0.4109
0.5	3	DNE	0.5265	3.3362
0.5	5	DNE	0.6400	8.8895
0.5	10	DNE	0.7729	34.0243
0.5	20	DNE	0.8686	132.1718
0.5	30	DNE	0.9074	294.0271
1.5	0.2	0.1	0.2546	0.0683
1.5	0.6	0.3	0.5991	0.8965
1.5	1	0.5	0.8127	4.3398
1.5	1.4	0.7	0.9373	20.9256
1.5	1.6	0.8	0.9730	57.683
1.5	1.8	0.9	0.9935	273.0939
2.5	0.1	0.15	0.2286	0.0296
2.5	0.2	0.30	0.4231	0.1467
2.5	0.3	0.45	0.5905	0.4325
2.5	0.4	0.60	0.7339	1.1034
2.5	0.5	0.75	0.8548	2.9437
2.5	0.6	0.90	0.9521	11.9188
3.5	0.05	0.125	0.1670	0.0100
3.5	0.10	0.250	0.3199	0.0470
3.5	0.20	0.500	0.5903	0.2882
3.5	0.30	0.750	0.8181	1.3491
3.5	0.35	0.875	0.9157	3.8006
3.5	0.38	0.950	0.9682	11.5749

Table I. TAM Results for an Assortment of Pareto/M/1 Queues (DNE =Does Not Exist).

E[*S*]	**γ**	r_0	W_q
0.2	4.3175	0.1365	0.0316
0.4	1.8210	0.2716	0.1491
0.6	1.0114	0.3931	0.3887
0.8	0.6264	0.4989	0.7963
1.0	0.4108	0.5892	1.4342
1.2	0.2788	0.6655	2.3869
1.4	0.1934	0.7292	3.7703
1.6	0.1362	0.7821	5.7429
1.8	0.0969	0.8256	8.5215
2.0	0.0694	0.8611	12.4032
4.0	0.0029	0.9886	346.0810
6.0	0.0001	0.9993	8112.469
8.0	$0.532H10^{-5}$	0.99996	187876.4
10.0	$0.230H10^{-6}$	0.999998	4347782.9

Table II: Results for an Assortment of Cauchy/M/1 Queues (Level Crossing).

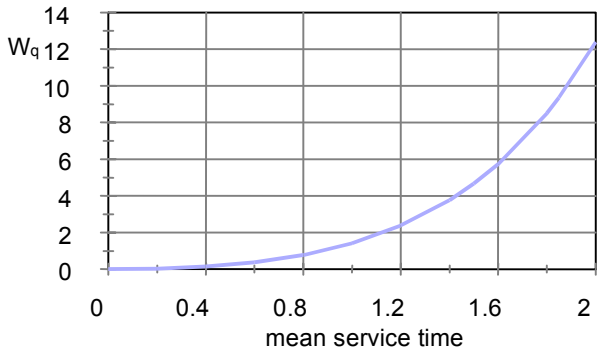

Figure 1: Mean Delays for Cauchy/M/1

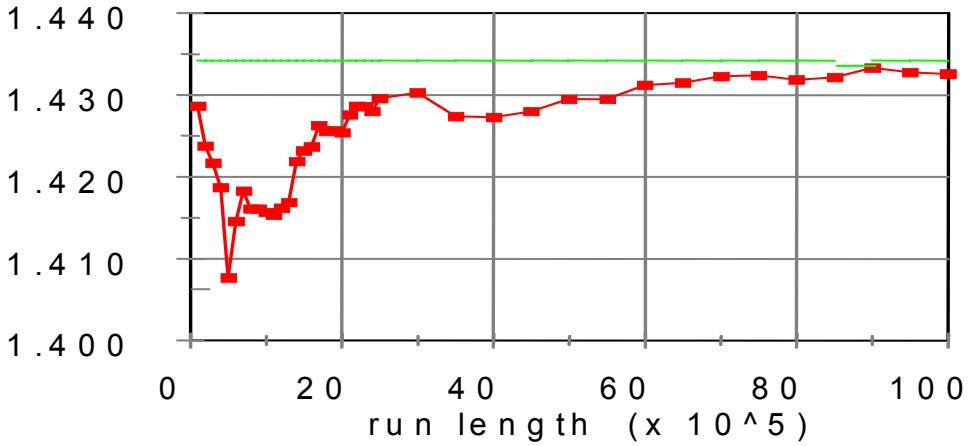

Figure 2. Simulation of W_q as Function of Run Length

E[*S*]	*(*	r_0	W_q
0.25	3.6854	0.0787	0.0213
0.50	1.7207	0.1397	0.0812
1.00	0.7690	0.2310	0.3004
2.00	0.3254	0.3492	1.0730
4.00	0.1307	0.4772	3.6517
8.00	0.0509	0.5926	11.6379

Table III: Results for an Assortment of Inverse-Log/M/1 Queues.

[*S*]	*(P*	*(L-N*	Δ_{L-N}	W_q(P)	W_q(L-N)

0.50	1.3240	1.3153	0.1623	0.2553	0.2603
0.75	0.7341	0.7069	0.2435	0.6120	0.6647
1.00	0.4612	0.4251	0.3247	1.1683	1.3524
1.25	0.3105	0.2695	0.4058	1.9706	2.4600
1.50	0.2185	0.1749	0.4870	3.0767	4.2162
1.75	0.1585	0.1137	0.5681	4.5595	7.0444
2.00	0.1176	0.0725	0.6493	6.5034	11.7862
2.25	0.0888	0.0443	0.7305	9.0138	20.3363
2.50	0.0679	0.0245	0.8116	12.2200	38.3894

Table IV. Comparing Some Log-Normal(μ=0, Φ=1.5)/M/1 vs. Pareto(.95)/M/1 Queues.

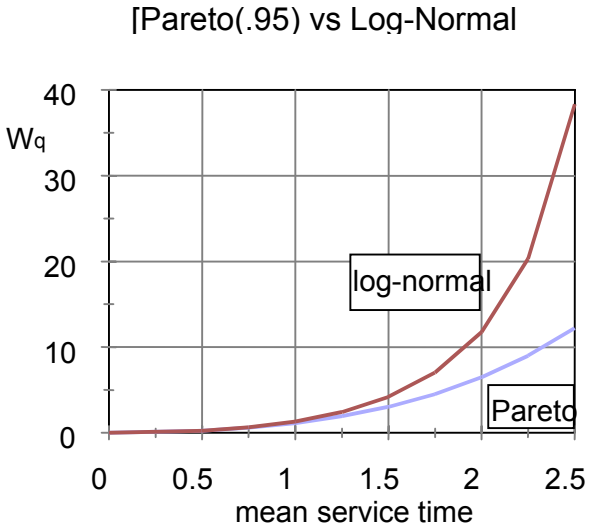

Figure 3. Comparative Mean Delays

Published in 71.

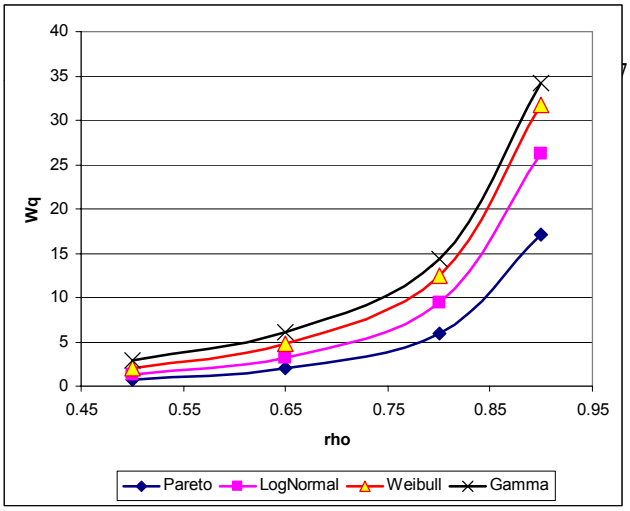

Figure 4. Effect of Power, Heavy and Exponential Tail Arrivals

E[*S*]	**(**	r_0	***K***	W_q
0.4	4.3175	0.1365	0.1419	0.0316
0.8	1.8210	0.2716	0.2062	0.0622
1.2	1.0114	0.3931	0.2164	0.2116
1.6	0.6264	0.4989	0.1999	0.5095
2.0	0.4108	0.5892	0.1726	1.0229
2.4	0.2788	0.6655	0.1429	1.8391
2.8	0.1934	0.7292	0.1150	3.0754
3.2	0.1362	0.7821	0.0907	4.8896
3.6	0.0969	0.8256	0.0704	7.4998
4.0	0.0694	0.8611	0.0540	11.2024
4.4	0.0500	0.8900	0.0410	16.4046
4.8	0.0362	0.9132	0.0310	23.6651
5.2	0.0262	0.9318	0.0232	33.7455
5.6	0.0190	0.9467	0.0173	47.6960

Table V: Results for Illustrative Cauchy/M/2 Queues.

References

Adler, R.J., Feldman, R.E., Taqqu, M.S. (1998), *A Practical Guide to Heavy Tails – Statistical Techniques and Applications*. Birkhauser, Boston.

Andersen, A. T., Jensen, A., Nielsen, B.F. (1995), "Modeling and Performance Study of Packet – Traffic with Self-Similar Characteristics over Several Time Scales with Markovian Arrival Processes (MAP)," *Twelfth Nordic Teletraffic Seminar, NTS* 12, 269-283.

Asmussen,S.(1987), *Applied Probability and Queues*, John Wiley, New York.

Brill, P.H. (1979), "An Embedded Level Crossing Technique for Dams and Queues," *Journal of Applied Probability* 16, 174-186.

Brill, P.H. (1988), "Single Server Queues with Delay-dependent Arrival Streams," *Probability in the Engineering and Informational Sciences* 2, 231-247.

Cohen, J.W. (1982), *The Single Server Queue*, rev. ed., North-Holland, Amsterdam.

Crovella, M.E. and A. Bestavros (1997), "Self-Similarity in the Word Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking* 5, 835-847.

Crovella, M.E., M.S. Taqqu and A. Bestavros (1998), "Heavy-Tailed Probability Distributions in the World Wide Web," in *A Practical Guide to Heavy Tails*: *Statistical Techniques and Applications*, R. Adler, R. Feldman, and M.S. Taqqu, editors. Birkhaüser, Boston.

Dye, L. (1999), " Old Rules Does Not Apply in New Age: Experts Find Erlang Phone Formula Fails to Explain Digital Traffic Jams," *LA Times Syndicate.*

Erlang, A.K. (1917-1918), " Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges," *Post Office Electrical Engineer's Journal,* 10, 189 – 197.

Feldman, A. and W. Whitt (1998), "Fitting Mixtures of Exponentials to Long-Tail Distributions to Analyze Network Performance Models,"*Performance Evaluation* 31, 245-279.

Feller, W. (1971), *An Introduction to Probability Theory*, 2nd ed., John Wiley, New York.

Fischer, M.J. and C.M. Harris (1999), "A Method for Analyzing Congestion in Pareto and Related Queues," *The Telecommunication Review* 10, Mitretek Systems, McLean, Virginia, 15-28.

Fowler, T.B. (1999), "A Short Tutorial on Fractals and Internet Traffic," *The Telecommunication Review* 10, Mitretek Systems, McLean, Virginia, 1-14.

Greiner, M., M. Jobmann, and L. Lipsky (1999), "The Importance of Power-Tail Distributions for Modeling Queueing Systems," *Operations Research* 47, 313-326.

Gross, D. and C.M. Harris (1998), *Fundamentals of Queueing Theory*, 3rd ed., John Wiley, New York.

Harris, C.M. and W.G. Marchal (1998), "Distribution Estimation Using Laplace Transforms," *INFORMS Journal on Computing* 10, 448-458.

Heyman, D. (1998), "Some Issues in Performance Modeling of Data Teletraffic," *Performance Evaluation,* 34, 227-247.

Heyman, D. and Lakshman, "Long Range Dependence and Queueing Effects for VBR Video," (to appear in Performance Evaluation).

Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994), On the Self-similar nature of Ethernet traffic (extended version)", *IEEE/ACM Trans. Networking* 2(1), 1-13.

Lucantoni, D.M. (1993), " The BMAP/G/1 Queue: a Tutorial," *Models and Techniques for Performance Evaluation of of Computer and Communication Systems* 2, 1-15.

Lucantoni, D.M., Choudhury, G.L. and Whitt,W. (1994), " The Transient BMAP/G/1 Queue," *Stochastic Models* 10 145-182.

Mood, A.M., Greybill, F.A., Boes, D.C. (19740, *Introduction to the Theory of Statistics, third ed*., McGraw-Hill, New York.

Nuets, M. (1991), "Modeling Data Traffic Streams," *TETETRAFFIC and DATATRAFFIC in a period of change,* ITC-13, A. Jensen and V.B.Iversen (Eds.) Elsevier Science Publishers B.V. North Holland.

Nuets, M. (1989) , *Structured Stochastic Matrices of M/G/1 Type and Their Applications,* Marcel Dekker, New York.

Paxson, V. and S. Floyd (1995), "Wide-Area Traffic: the Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking* 3, 226-244.

Park, K., Kim, G., and Crovella, M. (1996), On the Relationship Between File Sizes, Transport Protocols, and Self-similar Network Traffic. *Proceedings 1996 International Conference on Network Protocols*, 171-180.

Park, K., Kim, G., and Crovella, M.E. (1997), On the Effect of Traffic Self-similarity on Network Performance, *Proceedings of the SPIE Conference on Performance and Control of Network Systems* 3231, 296-310.

Pitkow, J. (1999), "Summary of WWW Characterizations", Zerox Palo Alto Research Center, Palo Alto, CA.

Whitt, W. (1974), "The Continuity of Queues," *Advances in Applied Probability* 6, 175-183.

Willinger, W. and V. Paxson (1998), "Where Mathematics Meets the Internet," *Notices of the AMS* 45, 961-970.

Willinger, W., V. Paxson and M.S. Taqqu (1998), Self-Similarity and Heavy Tails: Structural Modeling of Network Traffic," in *A Practical Guide to Heavy Tails*: *Statistical Techniques and Applications*, R. Adler, R. Feldman, and M.S. Taqqu, editors. Birkhauser, Boston.