T U M

I N S T I T U T F Ü R I N F O R M A T I K

The Importance of Power-tail Distributions for Telecommunication Traffic Models

Michael Greiner / Manfred Jobmann / Lester Lipsky

TUM-I9521
Juni 1995

T E C H N I S C H E U N I V E R S I T Ä T M Ü N C H E N

TUM-INFO-06-1995-I9521-350/1.-FI

Typescript: LaTeX/Postscript

Druck: Mathematisches Institut und
Institut für Informatik der
Technischen Universität München

The Importance of Power-tail Distributions for Telecommunication Traffic Models

Michael Greiner, Manfred Jobmann and Lester Lipsky[‡]

Institut für Informatik
Technische Universität München, Germany

Keywords

Power-tail distributions, heavy/fat-tail distributions, stable distributions, regularly varying functions, asymptotically self-similar functions, truncated tails, G/M/1 queues, renewal processes, buffer overflow, utilization.

1 Introduction

Recent measurements of traffic on communications networks have shown some extraordinary behavior which may prove critical for understanding the performance of broadband, and indeed, other information networks. For instance, Leland, et.al., [LELA94] have measured and analyzed the arrival of many millions of packets on ETHERNET networks at Bellcore, while Beran, et.al., [BERA95] have measured and analyzed several millions of frames from Variable-Bit-Rate (VBR) video services. The data collected have been measured accurately enough to give reliable numbers for the number of packet arrivals in intervals of as little as 0.01 seconds. The data displayed in both papers show enormous instability of arrival rates. No matter how large one takes for measurement intervals, the number of arrivals per unit time varies widely. This has been described as "self-similar", and "fractal" behavior. These, and other papers, have argued that $r(k)$, the auto-correlation function lag-k of the number of arrivals per time interval, must go to zero so slowly that $\sum_{k=1}^{\infty} r(k) = \infty$. They imply that any realistic model of such traffic must include very long-range correlation effects. However, this may not necessarily be true, since a renewal process (no correlation of interarrival times) where the interarrival times have a power tail distribution (i.e., distribution functions which behave as $1 - F(x) \Rightarrow 1/x^\alpha$ for large x) could generate such data [LIPS95]. Georganas [GEOR94] has argued that if service time distributions which have power tails with infinite means ($\alpha \leq 1$) are considered, then buffer sizes may become a serious problem. His analysis requires that the measured arrival rate goes to zero as $x \to \infty$, so it is not clear what a realistic system would see. In any case, even if $\alpha > 1$ buffer sizes may have to be very large indeed to avoid data loss.

[‡]Permanent Address: Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269-3155

While the warnings given to designers of future systems are no doubt correct, the statistical analyses have not revealed how either simulation or analytic techniques can be applied to study the performance of such systems. On the other hand, it has been shown that "self-similar" data can be generated by a renewal process where the interarrival times come from a single power-tail distribution with a finite mean (but *infinite variance*) [LIEF94]. The simplest model for this would be a GI/M/1 queue. Alternatively, the results in [LIKH95] indicate that a Poisson process with a "disbursed" batch of packets whose number is distributed by a power tail, can also generate self-similar data. In its simplest version, this can be transformed into an M/G/1 queue, where 'G' is a power-tail distribution.

In this paper, we describe in detail the properties of power-tail distributions, and then present an analytic class of well-behaved distributions (a sub-class of which are *Phase Distributions* which can be used in Markov Chain models) that have truncated power tails, and in the limit become power-tail distributions. This class was first used in [LIPS86] to explain the long-tail behavior of measured CPU times at Bellcore in 1985 [LELA85]. It was also used to show what might happen in data-retrieval systems which have power-tail file sizes [GARG92], and even to explain the distribution of medical insurance claims [LOWR93]. We then use these distributions to study the behavior of steady-state GI/M/1 queues, as a model for telecommunications networks. We then present the results of a parametric study of the affects of different α's on the *geometric parameter, s* [LIPS92] as a function of the *utilization parameter*, ρ, where

$$\rho := [arrival\ rate] \cdot [mean\ service\ time]. \tag{1}$$

The variance of a power-tail distribution is infinite if $\alpha \leq 2$, but our calculations show that the steady-state performance of these queues becomes worse only gradually as α drops below 2 with ρ fixed. The performance only becomes disastrous as α approaches 1 from above (i.e., the mean still exists). We also present calculations for distributions with truncated power tails, and show that they too can yield extraordinarily large mean queue lengths.

Of course, all this is done assuming steady-state behavior. But this may require inordinately many arrivals before such large queue lengths could be seen in reality. Discrete event simulation models must necessarilly suffer from the same problem. We present an argument showing that the closer α is to 1, the more arrivals must occur before any system's steady state can be approached. It remains for the future to yield appropriate descriptions of transient behavior.

2 Power-tail Distributions

First we display some standard definitions which will be useful in this paper. In all cases the time variables are in the range $0 \leq x < \infty$.

2.1 Definitions

Let X be a random variable representing the time for a process to complete. Then the *Probability Distribution Function* (PDF) is one-sided, and defined by:

$$F(x) := Pr(X \leq x)$$

with $F(0_{-}) = 0$, and $\lim_{x\to\infty} F(x) = 1$. The *Reliability Function* is

$$R(x) := Pr(X > x) = 1 - F(x),$$

and the *probability density function* (pdf) for the process, if it exists, is:

$$f(x) := \frac{d\,F(x)}{dx} = -\frac{d\,R(x)}{dx} \cdot$$

The ℓ^{th} *moment* of the distribution (or the *expectation of* X^ℓ), if it exists, is defined by:

$$\overline{x^\ell} := \mathrm{E}(X^\ell) := \int_0^\infty x^\ell \cdot f(x)\,dx. \tag{2}$$

The *variance* of the distribution is defined by

$$\sigma^2 := \mathrm{E}([X - \mathrm{E}(X)]^2) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \overline{x^2} - \bar{x}^2,$$

and the dimensionless quantity, the *coefficient of variation*, is defined by

$$C^2 := \frac{\sigma^2}{\bar{x}^2} \cdot$$

The conditional probability for the process to complete, given that it has not finished by time x is given by

$$F(t;x) := Pr(X \leq x + t \,|\, X > x) = \frac{F(x+t) - F(x)}{R(x)},$$

with appropriate definitions for $R(t;x)$ and $f(t;x)$. t is often referred to as the *residual time*. Let T_x be the random variable for the time remaining, conditioned on x. Then the *mean time remaining, given that the task is still running at time* x, is

$$\mathrm{E}(T_x) = \int_0^\infty t \cdot f(t;x)\,dt \;=\; \frac{1}{R(x)} \int_0^\infty t \cdot f(x+t)\,dt \tag{3}$$

and the *mean residual time* (the mean service time remaining, given that it is not known when service began) is given by the well-known formula

$$\overline{x_r} := \int_0^\infty \mathrm{E}(T_x) \cdot \frac{R(x)}{\mathrm{E}(X)}\,dx = \frac{\mathrm{E}(X^2)}{2 \cdot \mathrm{E}(X)} = \bar{x} \cdot \left[\frac{C^2+1}{2}\right].$$

The right-most term in the above expression shows clearly that if $C^2 > 1$ ($C^2 < 1$), then the mean residual time is greater (less) than the mean time for service.

There is a broad class of distributions which are variously called *heavy-, fat-, long-tailed* or *sub-exponential distributions*. They are categorized by their behavior for large x, namely their reliability functions all go to 0 slower than any exponential. In other words, if

$$\lim_{x\to\infty} \frac{x^n \cdot e^{-ax}}{R(x)} = 0 \quad \forall\, a > 0 \text{ and } \forall\, n \geq 0, \tag{4}$$

then $R(x)$ has a heavy (fat, long) tail. Sub-exponential distributions are defined as those satisfying

$$\lim_{x\to\infty} \frac{R(x+t)}{R(x)} = 1 \quad \forall\, t\,. \tag{5}$$

It is not hard to show that every distribution satisfying (4) is sub-exponential. For the purposes of this paper, we define functions that do not satisfy (5) as *well behaved*. Tighter definitions can be found, but are not necessary for this exposition.

Many heavy-tailed distributions are not uniquely determined by their moments. According to [FELL71] (p.514), if the moments of a distribution satisfy

$$\limsup_{\ell\to\infty} \frac{1}{\ell} \left[\mathrm{E}(X^\ell)\right]^{1/\ell} < \infty \tag{6}$$

then $F(x)$ is uniquely determined. All well behaved distributions [those which do not satisfy (4)] satisfy (6), but heavy tails do not generally satisfy this limit test. A commonly used class of distributions are the *Weibull Distributions*, defined by: $R(x) = \exp(-\lambda \cdot x^c)$. Weibull distributions with $c < 1$ are heavy tailed [they satisfy (4)], and do not satisfy (6). It has been shown that if $c \leq 1/2$, then there exist other distributions with the same moments [LIEF??].

2.2 Some Properties

The sub-class of heavy-tailed distributions of interest here are the *Power-tail Distributions*. They have the following property for large x:

$$R(x) \Longrightarrow \frac{c}{x^\alpha} \quad \text{where } \alpha > 0. \tag{7}$$

Alternatively,

$$\lim_{x\to\infty} x^\ell \cdot R(x) = \begin{cases} 0 & \text{for} \quad \ell < \alpha \\ \infty & \text{for} \quad \ell > \alpha\,. \end{cases} \tag{8}$$

Distributions which have this asymptotic behavior are sometimes referred to as *Pareto Distributions*, but this is inappropriate since the specific form of Pareto distributions $[f(x) = c\, x^{\mu-1}/(1+x)^{\alpha+\mu}]$ are of significance only in so far as they have the same asympotic behavior as all other power-tail distributions. A more appropriate name, if one must name functions after people, would be *Lèvy*, or *Lèvy-Pareto* distributions, since Lèvy defined and found the class of *stable* functions which have these power tails (where only $0 < \alpha < 2$ is meaningful). See [FELL71] and below.

Straight-forward differentiation yields the asymptotic form for the pdf,

$$f(x) \Longrightarrow \frac{\alpha\, c}{x^{\alpha+1}}\,\cdot \tag{9}$$

Then from (2) and elementary calculus, it follows that all moments for $\ell \geq \alpha$ are infinite! Thus, if $\alpha \leq 2$ then $F(\cdot)$ has infinite variance, and if $\alpha \leq 1$ then $F(\cdot)$ has infinite mean. One can ask what an infinite mean signifies, since an infinite time-span cannot be experienced in any service, and a finite sum of finite numbers must necessarily be finite. An understanding of this relies on ideas statisticians have been trying to explain to the rest of us for many years. For well behaved distributions, it is expected that the *average* of a sequence of numbers drawn from the same distribution should approximate the mean for that distribution. In general, let $x_1, x_2, \cdots, x_n, \cdots$ be such a sequence. Suppose $\mathrm{E}(X^\ell)$ exists for all $\ell \geq 0$, and let

$$s_n^{(\ell)} := \frac{1}{n}\sum_{k=1}^{n} x_k^\ell. \tag{10}$$

Then

$$\left| s_n^{(\ell)} - \mathrm{E}(X^\ell) \right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{11}$$

From a practical point of view, this means that an increasing number of terms in the sum will lead to an ever decreasing difference (with statistical fluctuations, of course). If one wishes to improve the estimate of $\mathrm{E}(X^\ell)$ by a factor of 2, then one must include 4 times as many numbers. But if $\mathrm{E}(X^\ell)$ is infinite, then (11) loses its meaning, and s_n^ℓ increases unboundedly with n.

Mathematically, (11) comes from the *Central Limit Theorem*, which states the following (see [FELL71], p.259):

> Let $X_1, X_2, ...$ be mutually independent continuous random variables with a common distribution, $F(\cdot)$, with finite mean, $\bar{x}$, and variance, σ^2. Then the distribution of their averaged sum,
>
> $$S_n := \frac{1}{n}\sum_{k=1}^{n} X_k \tag{12}$$
>
> tends to the Normal distribution, with the same mean, but with variance σ^2/n.

The theorem applies only to the first moment, but if $F(\cdot)$ has all finite moments, then the theorem can be applied in turn to each moment, making (11) valid for all ℓ. If, however, $\mathrm{E}(X^\ell) = \infty$ for all $\ell \geq \alpha > 2$, the statement, *tends to the Normal distribution* has to be softened somewhat. For n large enough, the distribution for S_n looks normal (pun) for $0 < x < \bar{x} + a \cdot \sigma$, where a is large and grows with n. But (8) still applies. (Also remember that the Normal distribution actually extends from $-\infty$ to $+\infty$, whereas we are dealing with distributions which are only non-zero for $x \geq 0$. What happens here is that the Normal distribution which S_n tends toward, has negligibly small probabilities for $x < 0$. I.e., n must be large enough so that $\exp(-n\bar{x}/2\sigma^2) \ll 1$.)

Note that if $F(\cdot)$ is already a Normal distribution, then S_n is also Normal, but with smaller variance. This is a special case of distributions which are said to be *stable*. The definition of a stable distribution might be stated in the following way:

> Let $X, X_1, X_2, \ldots$ be mutually independent continuous random variables with a common distribution, $F(\cdot)$. $F(\cdot)$ ist called *stable* iff for each $n \in \mathbb{N}$ S_n has the same distribution as $c_n X + d_n$ for some constants $c_n \in \mathbb{R}^+$ and $d_n \in \mathbb{R}$.

From the statement of the central limit theorem, the Normal Distribution can be the only stable distribution with finite variance. But what about distributions with infinite variance? P. Lèvy showed that all stable distributions, other than the Normal, have power tails with exponent $0 < \alpha \leq 2$, one for each power. This leads to modification of (11) with $\ell = 1$, to read:

$$|s_n - \mathrm{E}(X)| = O\left(\frac{1}{n^\beta}\right) \quad \text{where} \quad \beta = 1 - \frac{1}{\alpha} \quad \text{for} \quad 1 < \alpha \leq 2. \tag{13}$$

If $\alpha > 2$ then $\beta = 1/2$ and we get (11) for $\ell = 1$, but if $\alpha \leq 1$ then (13) is meaningless.

Observe that (13) implies that if $\alpha < 2$, many more samples have to be taken to gain the same estimate of $\bar{x}$ as was the case for $\alpha = 2$. For instance, suppose that $\alpha = 1.5$. Then $\beta = 1/3$, and it now takes 8 times as many points to improve the estimate by a factor of 2. This is not a trivial difference. (The following discussion is not to be taken as a serious attempt at a quantitative estimate of any sum's accuracy. Innumerable counter-examples abound. However it does serve to give a reasonable qualitative idea of how the number of samples needed grows with desired accuracy.) Suppose that one would like to get a k-digit estimate of $\bar{x}$. Then one might write $O(n^{-\beta}) \sim 10^{-k}$. Upon manipulating this formula one can come up with a *guesstimate* of the number of points needed, namely

$$n(k, \beta) := c \cdot 10^{k/\beta}, \tag{14}$$

where c is some scaling constant. (For finite variance, $c = \sqrt{\sigma^2 / \bar{x}^2}$.) One can say that with this many sample points, it is "highly probable" that the measured average will be within k digits of the mean.

Suppose $c \approx 1$, then for 2-digit accuracy (fluctuations of a graph of less than this order would be perceived by the unaided eye as a "fairly smooth" curve), if $\alpha \geq 2$ then $n(2, 0.5) = 10^{2/0.5} = 10,000$ points. On the other hand, if $\alpha = 1.5$, then $n(2, 1/3) = 1,000,000$ points! The number of points needed for a given accuracy is extremely sensitive to α. For an α of 1.4, $\beta = 1/3.5$ and $n = 10,000,000$. For $\alpha = 1.1$, $\beta = 1/11$, yielding $n = 10^{22}$. With this kind of instability, an Ampere of current ($\approx 1.6 \times 10^{19}$ electrons per second) could not be measured with any accuracy. For $\alpha = 1.05$, the temperature of a gram-molecular-weight of a substance would fluctuate too much to be measured. (Avogadro's number is 6.023×10^{23} molecules, while $n(2, 1.05) = 10^{42}$.) This argument will be used below in the discussion of "self-similarity".

2.3 A Simple Example

One might ask why power tail distributions have been seen so infrequently in the past. Part of this has to do with the way we have been examining data, and part has to do with the size of the samples normally examined. If the number of samples is small, then extraordinary samples (e.g., a job requiring an extremely large service time) are often *blamed on the weather*. Also, if the customer population is restricted (e.g., a closed system) and small compared to the queue sizes one would expect over a long period of time for the equivalent unbounded (open) system, then the effects of the tail will not be recognized.

In 1985, Leland and Ott [LELA85] examined the CPU times of over 6 million jobs that were executed at BELLCORE over a 6 months period. The usual analysis of such data reorders the times into size place, and plots the fraction of jobs which have time less than or equal to x. In the limit, this should approach the distribution function for CPU times. A better function is the fraction of jobs greater than x, which approaches the reliability function. Both functions are monotonic, and look very well behaved. In the 1960's many computer installations did just this, and concluded rightly that the distribution of CPU times could not be purely exponential. They then invariably fit their data to hyperexponential distributions (a weighted sum of two or more exponentials). (See [TRIV82] who reproduces data from the University of Michigan [ROSI65]. A report from the University of Minnesota [MINN67] shows similar data.) What Leland and Ott plotted instead, was the mean time remaining for those jobs greater than x. That is, they evaluated the equivalent of (3). They found that this function increased linearly with x for 5 to 6 orders of magnitude. In other words, the expected time remaining for a job increased linearly with the amount of time it had already run. This is one of the important properties of power-tail distributions.

Another way to expose the tail is to plot the reliability function on *log-log* paper. From (7) we can see that

$$\log(R(x)) \Longrightarrow \log(c) - \alpha \cdot \log(x)\,.$$

This is again a straight line, now with slope $-\alpha$. We demonstrate this with a simple example. Consider the reliability function,

$$R(x) = a \cdot e^{-x} + \frac{1-a}{(1+x)^2} \quad \text{for} \quad 0 \le a \le 1. \tag{15}$$

It is easy to show that this has a mean value of 1 for all a, but for $a < 1$ it has a power-tail with $\alpha = 2$, and thus has infinite variance. Figure 1 shows this function for $a = 0.0, 0.5, 0.8,$ and 1.0. For $a = 1$, we have the pure exponential function, but the other three curves look very similar to the first, so one would expect no surprises, even though they actually have infinite variance. Figure 2 shows the log of the same functions plotted against $\log(x)$. Here it is clear that the tails are different. All three power-tail functions show the straight line described previously, with negative slope $\alpha = 2$. It would be interesting to reexamine the CPU data published in the 1960's in this light.

Returning to residual times, for x large enough, the asymptotic form for $R(x)$ given in (7) can be substituted directly into (3) giving

$$\mathrm{E}(T_x) \Longrightarrow \frac{x}{\alpha - 1}\,.$$

This clearly shows the linear behavior found by [LELA85]. Using (15) we get the relation

$$\mathrm{E}(T_x) = \frac{a \cdot e^{-x} + \frac{1-a}{1+x}}{a \cdot e^{-x} + \frac{1-a}{(1+x)^2}} = (1+x)\left[\frac{(1+x)\cdot a \cdot e^{-x} + 1 - a}{(1+x)^2 \cdot a \cdot e^{-x} + 1 - a}\right].$$

The formulas for large x depend on whether $a = 1$ or not, namely

$$\mathrm{E}(T_x) \left\{ \begin{array}{ccl} = & 1 & \text{for} \quad a = 1 \\ \Rightarrow & x+1 & \text{for} \quad a < 1\,. \end{array} \right. \tag{16}$$

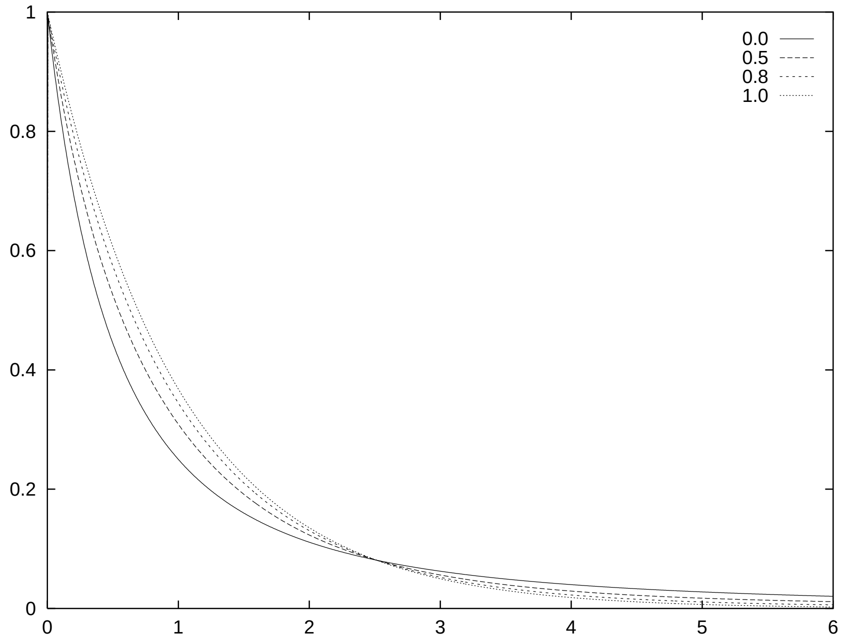

Figure 1: **Four reliability functions, three with power tails and infinite variance.** These are all taken from Equation (15), for $a = 0$, 0.5, 0.8 and 1.0 . The mutual crossing at $x \approx 2.51$ is a peculiarity of this particular equation. Note that the heavier the tail (smaller a), the more likely it is that $X < \bar{x} = 1$ (smaller $R(1)$). But if X exceeds $x \approx 3$, then the process is likely to last much longer.

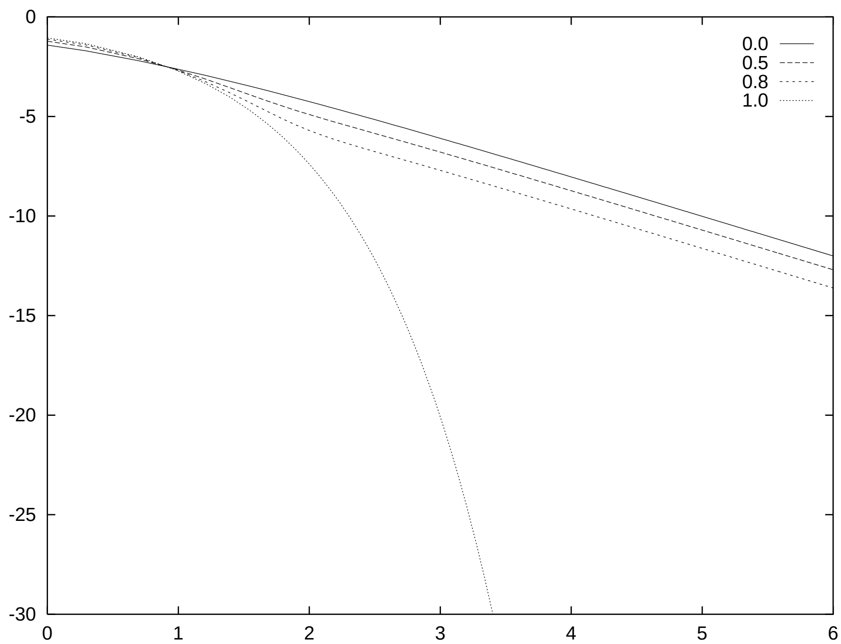

Figure 2: **The same four functions, now plotted on a log-log scale.** No matter how small $(1 - a)$ is, as long as it is greater than 0, the power tail term will dominate for large enough x.

For the exponential case ($a = 1$), the mean time remaining is always 1, a consequence of its memoryless property. But for $a < 1$, the mean time remaining is *more* than the time already spent.

One might ask whether very long residual times are *likely* to happen. Consider the possibility that 20 units of time would elapse without a completion occurring. The probability for this to occur is $R(20)$. For $a = 1$, this is $R(20) = 1/e^{20} = 1/485,165,195$. Only one event in 500 million would last this long. So, for a system which lasts for a million events, one could fairly say that this event is so unlikely that it can be ignored. However, for $a = 0.8$, $R(20) = 1/2205$. In some sense, this is also an unlikely event. But in a system which lasts for 1,000,000 events, this type of event is almost surely going to occur, not once, but *many times*, so it cannot be ignored. And when it does occur, expect it to last another 21 units of time. For $a = 0.5$, an event that lasts more than 20 units of time is even more likely ($R(20)$ is now $1/882$).

2.4 A Robust Function

It is puzzling why power-tail distributions should show up in various aspects of computer system performance. In this section we present a model which mimics in a simple way what could be causing this phenomenon. It thus gives some insight as to why power tails occur, and gives a rationale for use of the term *self-similar*. At the same time it will provide us with a functional form (first introduced in [LIPS86]) which has a power tail, can be truncated, and can be used for simulation and Markov modelling. (Depending on the base function used, this class of distributions can be *matrix-exponential* [LIPS92], or even *phase distributions* [NEUT81].)

First consider the following scenario. Suppose a "typical" computer user chooses to run a program whose CPU time is best described by a distribution function, $F_0(x)$, with a mean of 1 second. After receiving the result, he decides, with probability 1/2, to run the program again, but with modifications which increase its CPU time by a factor of 2. After receiving the second result, he decides (again with probability 1/2) whether to run the program yet again, with more modifications which increase its CPU time by another factor of 2. Even if this looping continued indefinitely, only 1/2 the users would run their programs more than once, only 1 in 4 users would run their programs more than twice, and less than 1 in a thousand would run their programs more than 10 times. On average, each user will only run his program twice. So, the frequent user is not common, yet the mean CPU time per job grows unboundedly. The mean time is given by:

$$\bar{x} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 4 + \frac{1}{16} \cdot 8 + \cdots + \cdots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots + \cdots = \infty.$$

Of course it would take an infinite amount of time, and an infinite number of users for this sum to be complete. But what would be seen over time, is a user behavior which seems to stabilize (an average of two runs per user), but with the infrequent arrival of very big jobs, which get bigger, and cause the mean CPU time to grow ever bigger as well. This is a reasonable qualitative description of power-tail behavior generally, where ℓ^{th} *moment* ($\ell \geq \alpha$) replaces *mean CPU time*.

A formal mathematical description of the above process is as follows. Let $X_0, X_1, \cdots X_n, \cdots$ be random variables representing the time for the n^{th} rerun of a program which will run at least that many times (X_0 is the initial run, X_1 is the first rerun, etc.). Let $F_n(x)$ be the distribution function for X_n, with reliability function, $R_n(x)$, and density function, $f_n(x)$. Next, let $0 < \theta_n \leq 1$ be the conditional probability that a program will be run at least one more time, given that it ran n times ($\theta_0 := 1$). Last, define

$$\gamma_n := \mathrm{E}(X_n)/\mathrm{E}(X_{n-1}).$$

For the example just given, we have for $n > 0$, $\theta_n = 1/2$, $\gamma_n = 2$, and $\mathrm{E}(X_0) = 1$.

For notational convenience, we define $\theta(n) := \theta_1\theta_2 \cdots \theta_n$ $[\theta(0) = 1]$, and $\gamma(n) := \gamma_1\gamma_2 \cdots \gamma_n = \mathrm{E}(X_n)/\mathrm{E}(X_0)$. Then

$$\Theta_N := \sum_{n=0}^{N-1} \theta(n)$$

where Θ_N is the expected number of times a user will run a program (original or modified), with up to $N-1$ modifications. The random variable, Y_N, given by

$$Y_N := \frac{1}{\Theta_N} \sum_{n=0}^{N-1} \theta(n) X_n$$

represents the CPU time of a program, among those that have not run more than N times. The distribution function for Y_N is given by

$$F_{Y_N}(x) = \frac{1}{\Theta_N} \sum_{n=0}^{N-1} \theta(n) F_n(x),$$

with mean

$$\mathrm{E}(Y_N) = \frac{\mathrm{E}(X_0)}{\Theta_N} \sum_{n=0}^{N-1} \theta(n)\gamma(n).$$

These formulas are far too rich in parameters for our expository purposes, so we make some simplifying assumptions. We point out, however, that the power-tail behavior we will demonstrate is valid for this general expression as long as $\lim_{n\to\infty} \theta(n) = 0$ and $\gamma(n) \to \infty$, while $\theta_n \geq a > 0$ and $\gamma_n \geq b > 1$, for the same infinite set of n's, for some a and b.

Assume that $\theta_n = \theta$ and $\gamma_n = \gamma$ for all $n > 0$. Then $\theta(n) = \theta^n$, and $\gamma(n) = \gamma^n$. Consequently,

$$\Theta_N = \sum_{n=0}^{N-1} \theta^n = \frac{1-\theta^N}{1-\theta}.$$

Next assume that all the $F_n(x)$'s are the same shape as $F_0(x)$, and that $F_0(x)$ is well behaved. That is

$$F_n(x) = F_0(x/\gamma^n), \quad \text{and thus} \quad \mathrm{E}(X_n) = \gamma^n \mathrm{E}(X_0) \;\; \forall\, n.$$

The corresponding formula for $R_n(x)$ is obvious, but

$$f_n(x) = \gamma^{-n} f_0(x/\gamma^n).$$

The density function for Y_N becomes

$$f_{Y_N}(x) = \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \left(\frac{\theta}{\gamma}\right)^n f_0(x/\gamma^n), \tag{17}$$

with reliability function

$$R_{Y_N}(x) = \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \theta^n R_0(x/\gamma^n). \tag{18}$$

We will call these functions *truncated power-tail distributions* for reasons which will become clear presently. They are well behaved (or Phase, or matrix-exponential) if $R_0(x)$ is, and converge to

$$f(x) := \lim_{N\to\infty} f_{Y_N}(x) = (1-\theta) \sum_{n=0}^{\infty} \left(\frac{\theta}{\gamma}\right)^n f_0(x/\gamma^n), \tag{19}$$

and

$$R(x) := \lim_{N\to\infty} R_{Y_N}(x) = (1-\theta) \sum_{n=0}^{\infty} \theta^n R_0(x/\gamma^n). \tag{20}$$

Although $R_{Y_N}(\cdot)$ is well behaved for all N as defined in Subsection 2, the limit function, $R(x)$, is not.

The moments of $F_{Y_N}(\cdot)$ are easy to find. From the definition,

$$\mathrm{E}(Y_N^\ell) = \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \left(\frac{\theta}{\gamma}\right)^n \int_0^\infty x^\ell f_0(x/\gamma^n)\, dx.$$

We make the substitution, $x = u\gamma^n$ and get

$$\mathrm{E}(Y_N^\ell) = \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \left(\frac{\theta}{\gamma}\right)^n \gamma^{n(\ell+1)} \int_0^\infty u^\ell f_0(u)\, du = \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \left(\theta\gamma^\ell\right)^n \mathrm{E}(X_0^\ell), \tag{21}$$

and finally,

$$\mathrm{E}(Y_N^\ell) = \frac{1-\theta}{1-\theta^N} \cdot \frac{1-(\theta\gamma^\ell)^N}{1-\theta\gamma^\ell} \cdot \mathrm{E}(X_0^\ell). \tag{22}$$

As long as $\theta\gamma^\ell < 1$, the limit can be taken to get

$$\mathrm{E}(Y^\ell) := \lim_{N\to\infty} \mathrm{E}(Y_N^\ell) = \frac{1-\theta}{1-\theta\gamma^\ell} \cdot \mathrm{E}(X_0^\ell). \tag{23}$$

But if $\theta\gamma^\ell \geq 1$ the limit diverges (infinite moments). We identify α by the relation,

$$\theta\gamma^\alpha = 1\,, \quad \text{or} \quad \alpha := -\frac{\log(\theta)}{\log(\gamma)}\,. \tag{24}$$

This is the same α as in (7). Then we have the typical power-tail relation for moments:

$$\mathrm{E}(Y^\ell) < \infty \quad \Longleftrightarrow \quad \ell < \alpha. \tag{25}$$

We next show that $R(x)$ asymptotically satisfies a property which matches Feller's definition of a *regularly varying function*. This could also be used as the definition for an *asymptotically self-similar function*. We then show (as does Feller) that such functions must have power tails. Consider (20), evaluated at $x = \gamma t$,

$$R(\gamma t) = (1-\theta)\sum_{n=0}^{\infty} \theta^n R_0(t/\gamma^{n-1}) = (1-\theta)\sum_{n=-1}^{\infty} \theta^{n+1} R_0(t/\gamma^n)$$

$$R(\gamma t) = \theta(1-\theta)\sum_{n=0}^{\infty} \theta^n R_0(t/\gamma^n) + (1-\theta)R_0(\gamma t) = \theta R(t) + (1-\theta)R_0(\gamma t).$$

But $R_0(t)$ is well behaved, and drops off at least as fast as some negative exponential, so for t large enough, $R_0(\gamma t)$ must be small compared to the sum, therefore,

$$R(\gamma t) \longrightarrow \theta R(t) \quad \text{as} \quad t \longrightarrow \infty.$$

This can be done any number of times, so we have, for large t

$$R(\gamma^k t) = \theta^k R(t).$$

Let $u = \gamma^k$, solve for k [$k = \log(u)/\log(\gamma)$], and substitute for it to get

$$\theta^k = e^{k\log(\theta)} = e^{\log(\theta)\log(u)/\log(\gamma)}.$$

But from the discussion following (23), $\alpha = -\log(\theta)/\log(\gamma)$, so

$$\theta^k = e^{-\alpha\log(u)} = e^{\log(u^{-\alpha})} = u^{-\alpha}.$$

Therefore

$$R(\gamma^k t) = R(ut) = R(t)/u^\alpha.$$

Let t be large enough, but fixed, and let $x = ut$, then finally

$$R(x) = R(t)/(x/t)^\alpha = \frac{R(t)t^\alpha}{x^\alpha} = \frac{c}{x^\alpha}\,. \tag{26}$$

Feller defines a *regularly varying function* as one which satisfies

$$R(tx) \longrightarrow \psi(x)R(t) \quad \text{as} \quad t \longrightarrow \infty\,.$$

He then goes on to show that for a monotone decreasing function [which $R(x)$ is], $\psi(x) = x^{-\alpha}$.

In the following section, we will use this family of functions to study the behavior of steady-state G/M/1 queues. For that purpose, we let

$$R_0(x) = e^{-\mu x}, \quad \text{where} \quad \mu = \frac{1-\theta}{1-\gamma\theta}\,.$$

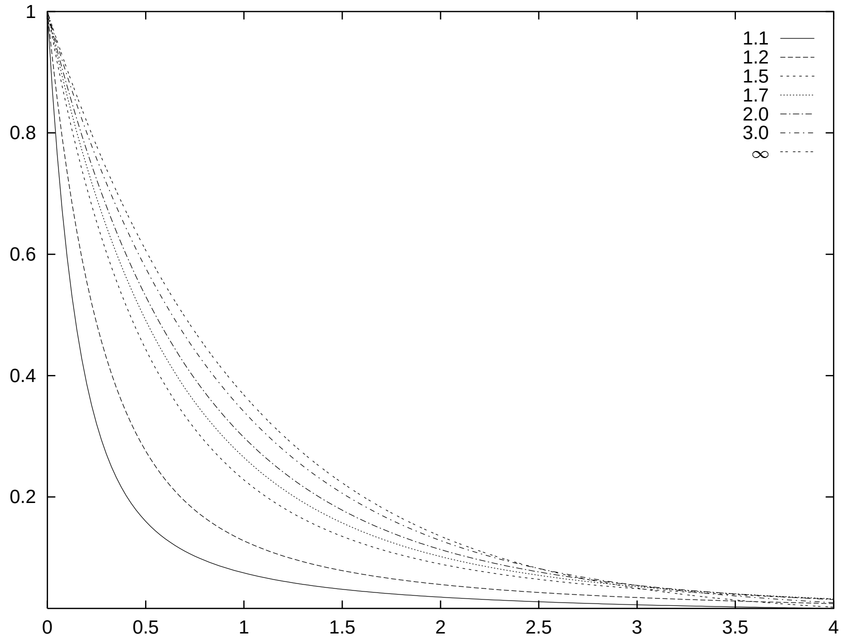

Figure 3: $\boldsymbol{R(x)}$ **as a function of** $\boldsymbol{x}$ **for** $\boldsymbol{\alpha \in \{1.1, 1.2, 1.5, 1.7, 2.0, 3.0, \infty\}}$ **and** $\boldsymbol{\theta = 0.5}$**.** For $x \approx 3$ the curves start to cross until finally for $x \approx 216$ the curves for $\alpha = 1.1$ and $\alpha = 1.2$ cross and complete the reordering of the curves. (Compare with Figure 1.)

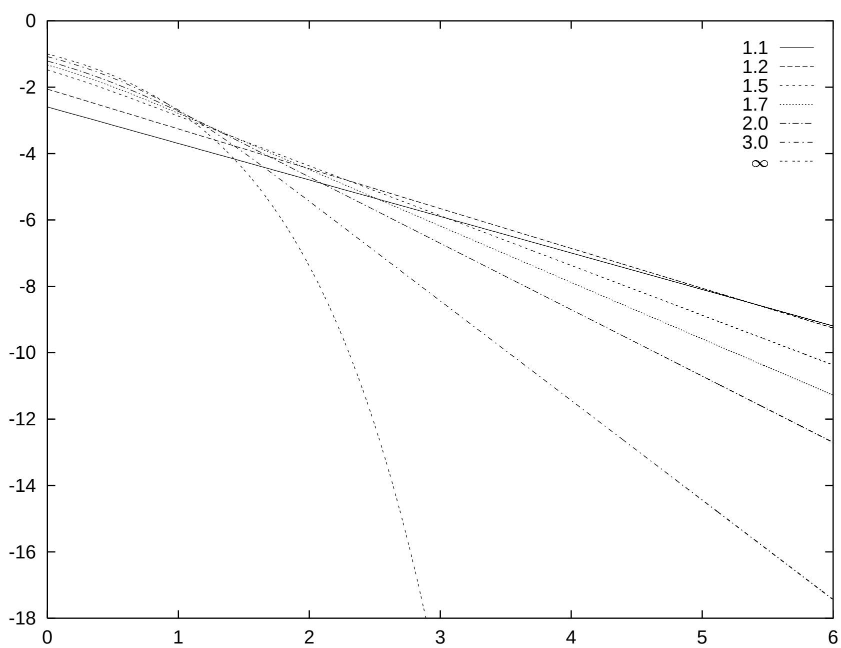

Figure 4: $\boldsymbol{R(x)}$ **as a function of** $\boldsymbol{x}$ **for** $\boldsymbol{\alpha \in \{1.1, 1.2, 1.5, 1.7, 2.0, 3.0, \infty\}}$ **and** $\boldsymbol{\theta = 0.5}$**, now plotted on a log-log scale.** As α approaches ∞, $R(x)$ approaches e^{-x} for finite x.

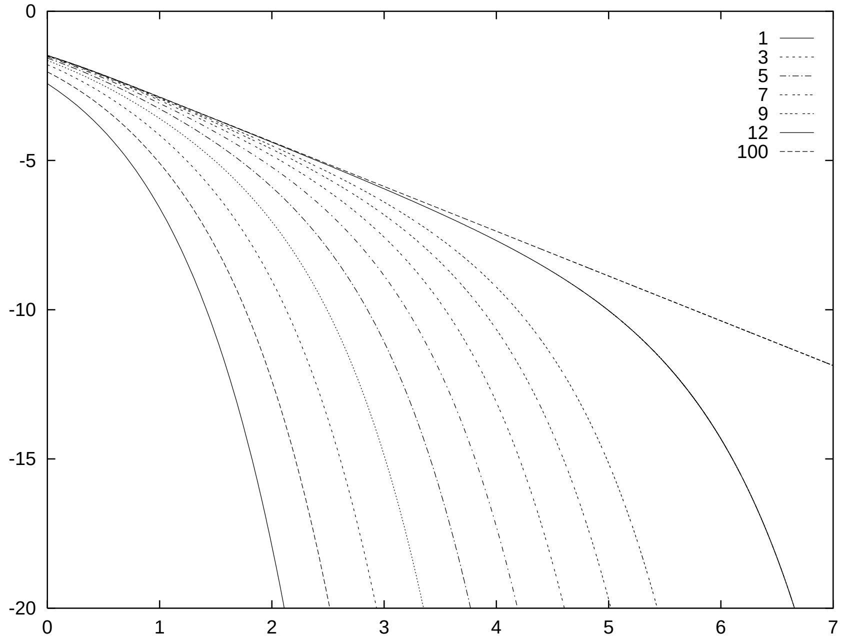

Figure 5: **Truncated power-tail reliabilities R_{Y_N} for $N \in \{1, 2, \ldots, 9, 12, 100\}$, $\alpha = 1.5$ and $\theta = 0.5$, plotted on a log-log scale.** For $N = 100$, $\log(R(\cdot))$ is a straight line for many orders of magnitude of change in x. Even with only 12 terms, $\log(R(\cdot))$ looks like a straight line for a factor of 100 change in x.

Since $\mathrm{E}(X_0) = 1/\mu$, from (23), $\mathrm{E}(Y) = 1$ as long as $\gamma\theta < 1$. Figure 3 shows $R(x)$ versus x for various values of α and $\theta = 0.5$[1]. Figure 4 displays the same functions on a log-log scale, showing their linear behavior for large x. We see clearly that a larger α yields a larger negative slope for $R(x)$. Also, for x large enough, $R(x)$ bounds e^{-x} from above, and the larger α is, the closer $R(x)$ is to e^{-x}.

Finally, in Figure 5 we display some functions with truncated tails. Here we see how the tail fills in for increasing N. In some sense, this mimics the way data points accumulate for real systems. For a given set of data, there is a largest member, and very few other elements of comparable size. Therefore, the fraction of samples that are nearly that large drops to 0 very rapidly with increasing x. As more samples are added, a few will be much larger than all previous ones, and the *tail fills in*. Thus we can map, at least qualitatively, the increase in number of samples with increase in N. Note that for finite N, $\mathrm{E}(Y_N) < 1$.

3 The G/M/1 Queue As Model

Our purpose in this paper is to examine the impact which interarrival processes that exhibit power-tail behavior can have on a telecommunications network. To show that the impact can be great, it is only necessary to examine the simplest of such systems. The data in [LELA94] clearly indicate that we can expect packets to arrive in unusual patterns. Their analysis of irregularity of the number of packets arriving per unit time is consistent

[1] The curves change only slightly when using another value for θ, e.g. $\theta \in \{0.25, 0.75\}$.

with a *Hurst Parameter*, H, with value $H = 0.8$. The relationship between H and α is given by [LIKH95] as

$$H = (3 - \alpha)/2, \quad \text{or} \quad \alpha = 3 - 2 \cdot H\,. \tag{27}$$

This, then corresponds to $\alpha = 1.4$. Although there is much time spent in various papers discussing the connection between H, α, and the autocorrelation coefficients, $r(k)$, for the *number of arrivals* in successive equal time intervals, it can be shown [LIPS95] that a renewal process with power-tailed interarrival times has the appropriate properties even though the *interarrival times* themselves are uncorrelated. Therefore the $r(k)$'s will play no part in what follows.

In their paper, Leland, et.al., have shown that their data (some 29 million packets over a 24 hour period) remains unstable over 5 orders of magnitude of time intervals, namely from number of arrivals in 0.01 seconds, to number in 100 seconds. They then argued that this showed a manifest self-similarity for the process. We claim that if they had been able to increase the time intervals by two orders of magnitude the instability would disappear. In principle, if $\alpha > 1$ then there exist intervals large enough to get smooth data. On the other hand, we concede that the number of sample points needed to achieve this stability could exceed the lifetime of the system being investigated. In Subsection 2 we argued that a power-tail distribution with $\alpha = 1.4$ would require sample sizes on the order of 10^7 to yield visual smoothness. Unfortunately, the largest intervals displayed in [LELA94] contained "only" 33,000 or so packets per 100 second interval. Therefore, according to (14), if they had increased their intervals by two orders of magnitude, their graphs would have achieved visual smoothness. Unfortunately, even a set of 29 million data points is not sufficient to yield more than two or three intervals, so an honest test cannot be made. But from our guesstimate arguments, it is clear that for all $\alpha > 1$ the data will stabilize in principle. For $\alpha < 1$ the data will grow *more* erratic with increased interval width. Only for $\alpha = 1$ will this self-similarity property repeat itself for all orders of magnitude.

3.1 The Arrival Process

Whatever terminology is used for wildly varying arrivals, there are (at least) two inequivalent simple process classes which could generate them. First, one could imagine files whose sizes are distributed according to some power-tail distribution, and instead of being sent whole, are broken up into much smaller packets. Suppose that such files are sent out according to a Poisson distribution, the individual packets disburse somewhat (separate randomly in time from each other), and are recognized at the destination as *bursts* of packets. This is a *compound Poisson process.* Such models have been used to generate arrivals, but as far as we know, no-one has used a generator with a power tail. Likhanov, et.al., [LIKH95] have come close, however. Instead of trying to deal with the disbursion of packets, they argue that the Poisson arrival of bursts is equivalent to an M/G/1 queue, where the service times are power tails. (Actually, they assume a large number of independent sources, and then reprove the well-known theorem that the merging of many independent arrival processes approaches a Poisson process whose arrival rate is the sum of the individual arrival rates.) Garg, et.al., [GARG92] presented an M/G/1 model in studying data retrieval systems where file sizes came from truncated power-tail distributions. They also studied what happened as the tail behavior extended to larger and larger x.

The other simple process which could generate wild data is simply a renewal process where the interarrival times have a power tail. [LIEF94] have demonstrated that indeed one can have very wild data with arbitrarilly large bursts. After all, if there are long intervals with no arrivals, then there must be other intervals which have far more than their share. The data in [LELA94] does not seem to show extremely long periods with no arrivals. Therefore, a simple renewal model may have to be merged with a small background Poisson process. Unfortunately, the merging of two (as opposed to very many) renewal processes is *not* a renewal process (unless both are Poisson). This would then make the well-known solution of a steady-state GI/M/1 queue inappropriate. But a system with a non-renewal arrival process is a very difficult problem indeed. We will therefore solve the GI/M/1 queue as an approximation to the more appropriate system.

3.2 Finding The Geometric Parameter, s

It is well known that the steady-state probability for finding k customers in an open GI/M/1 queue, $\pi(k)$, is given by [LIPS92]

$$\begin{aligned}\pi(0) &= 1-\rho \\ \pi(k) &= (1-s)\cdot\rho\cdot s^{k-1}, \quad k>0\end{aligned}$$

where s is the *geometric parameter* satisfying the equation

$$s = B^*[\lambda(1-s)]. \tag{28}$$

$B^*(z)$ is the Laplace transform of the interarrival distribution, and λ is the service rate of the exponential server. Let $\bar{x}$ be the mean interarrival time. Then

$$\rho = \frac{1}{\lambda\bar{x}}$$

and the mean queue length (including the one being served) for the process is

$$\bar{q} := \sum_{k=0}^{\infty} k\cdot\pi(k) = \frac{\rho}{1-s}\cdot$$

For telecommunications systems, the arrival probabilities are of importance. The probability that an arriving packet will find exactly $k>0$ other packets already there is given by

$$a(k) = (1-s)\cdot s^k = \pi(k)\cdot\frac{s}{\rho}\cdot$$

Suppose that a no-loss system is required. One might have a primary buffer of K slots, and a *backup buffer* of unbounded size (e.g., a *disc-array sub-system*). Then the probability that an arriving packet will have to be stored in the backup buffer is

$$Pr(K) = \sum_{k=K}^{\infty} a(k) = (1-s)\sum_{k=K}^{\infty} s^k = s^K.$$

These equations all show the important role s plays in GI/M/1 queues. We see that the smaller s is, the better system performance we can expect. Equivalently, the closer s is to 1, the bigger $\bar{q}$ and $Pr(K)$ will be, giving less desirable performance. When s is close to 1, it is better to look at $Pr(K)$ as a function of $t := 1 - s$, for then

$$Pr(K) = s^K = e^{K\log(s)} = e^{K\log(1-t)} \approx e^{-tK} \quad \text{for} \quad t \ll 1.$$

There are some general statements one can make. For instance, when $\rho = 1$, so does s. If $R(0) = 1$ (a *non-defective* distribution) then $s = 0$ when $\rho = 0$. Also, only for the M/M/1 queue does $s = \rho$. We say that if $s > \rho$ then system performance is worse, and if $s < \rho$ then system performance is better than one could ask for. It has been shown [LIPS92] that the slope of the curve, s versus ρ at $\rho = 0$ is $\bar{x}f(0)$. So if this is less than (greater than) 1, then for small ρ, performance is better (worse) than the equivalent M/M/1 queue. At the other end, at $\rho = 1$, the slope is $2/(C^2+1)$. If $C^2 > 1$ ($C^2 < 1$) then performance is worse (better). It is also known [LIPS92] that near $\rho = 1$ performance depends only on the moments of the interarrival time distribution, and thus on α and θ. In particular, if $\alpha \leq 2$ then $C^2 = \infty$ and the slope is 0. This means that s will remain close to 1 even as ρ decreases.

In general, for small ρ performance depends only on the behavior of $f(x)$ for small x. Given (17), the values of $f(x)$ and its derivatives at $x = 0$ depend on $f_0(x)$. For the function chosen here [$\mu e^{-\mu x}$ with $\mu = (1-\theta)/(1-\theta\gamma)$], $f(0) > 1$ for all θ and all α. A different function could have been chosen which would have yielded a smaller s for small ρ (e.g., $\mu^2 x e^{-\mu x}$). But the performance for $\rho \to 1$ would be the same. This shows the difficulty in selecting test functions in exploring the general performance of systems. Without more knowledge of a particular system, no model can be relied upon to give an accurate picture of the performance for small or intermediate ρ. This will be discussed further in the next sections, when calculation results are presented.

3.3 Behavior of Queues based on f_{Y_N} (Truncated Tails)

We return now to (28) and our explicit test function

$$f_{Y_N}(x) = \mu \cdot \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \left(\frac{\theta}{\gamma}\right)^n \exp(-\mu x/\gamma^n), \tag{29}$$

where $\mu = (1-\theta)/(1-\theta\gamma)$, $\theta\gamma < 1$ and $\theta\gamma^\alpha = 1$, i.e. $\gamma = (1/\theta)^{1/\alpha}$. Its mean is given by

$$\bar{x}_N := \mathrm{E}(Y_N) = \frac{1-(\theta\gamma)^N}{1-\theta^N} \quad [\text{cf. } (21)].$$

As long as N is finite our test function is well behaved in that all its moments are finite, and it drops off exponentially for large x (see Subsection 2).

From (28) and the definition of the Laplace Transform, the following non-linear equation must be solved for its smallest positive root.

$$s = \int_0^\infty e^{-\lambda(1-s)x} f_{Y_N}(x)\, dx = \mu \frac{1-\theta}{1-\theta^N} \sum_{n=0}^{N-1} \frac{\theta^n}{\lambda(1-s)\gamma^n + \mu}.$$

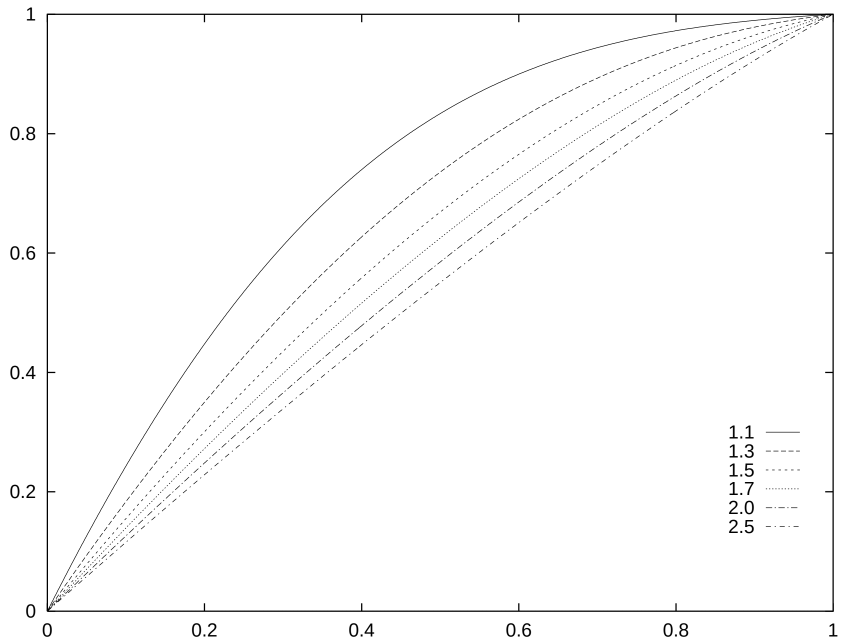

Figure 6: **ρ-s-Diagram for various values of α where $\theta = 0.5$ and $N = 10$.**

Note that $s = 1$ is always an extraneous root of this equation. It has no relevance to our discussion; its physical significance comes from the fact that $\int_0^\infty f_{Y_N}(x)\,dx = 1$. It can be eliminated by subtracting 1 from both sides and using $1 = (1-\theta)/(1-\theta^N)\sum_{n=0}^{N-1}\theta^n$. Then we can solve instead, $g(s) = 0$, where

$$g(s) = \frac{1-\theta}{1-\theta^N}\sum_{n=0}^{N-1}\frac{(\gamma\theta)^n}{(1-s)\gamma^n + \mu/\lambda} - 1. \tag{30}$$

Figures 6 to 15 show the characteristic behavior of the geometric parameter s, i.e. the solution of $g(s) = 0$, as a function of ρ, for various values of the parameters α, θ and N. The figures are grouped into three classes:

- **Fixed θ and N; α variable** (Figures 6 and 7): The smaller α is, the larger is s. For α small enough, and N large enough, performance can be very bad (s close to 1).
- **Fixed α and θ; N variable** (Figures 8-10): The larger N is, the larger is s. For larger values of α the values of s stabilize more rapidly. If we think of N as somehow related to the sample size in a real system, then these figures show that for smaller α more samples must be included (larger N) in order for the system to experience the full effect of the tail [$s(\rho;\,N) \Rightarrow s(\rho;\,\infty)$]. We also see that for α close to, or bigger than 2.0, performance is not disasterous, even for large N.
- **Fixed α and N; θ variable** (Figures 11-15): For small N we have a somewhat unexpected result: For fixed ρ and N, s is not necessarilly a monotonic function of θ. (fig. 11, 14). For larger N (e.g. $N \geq 12$ for $\alpha = 1.1$ or $N \geq 5$ for $\alpha = 1.5$) the

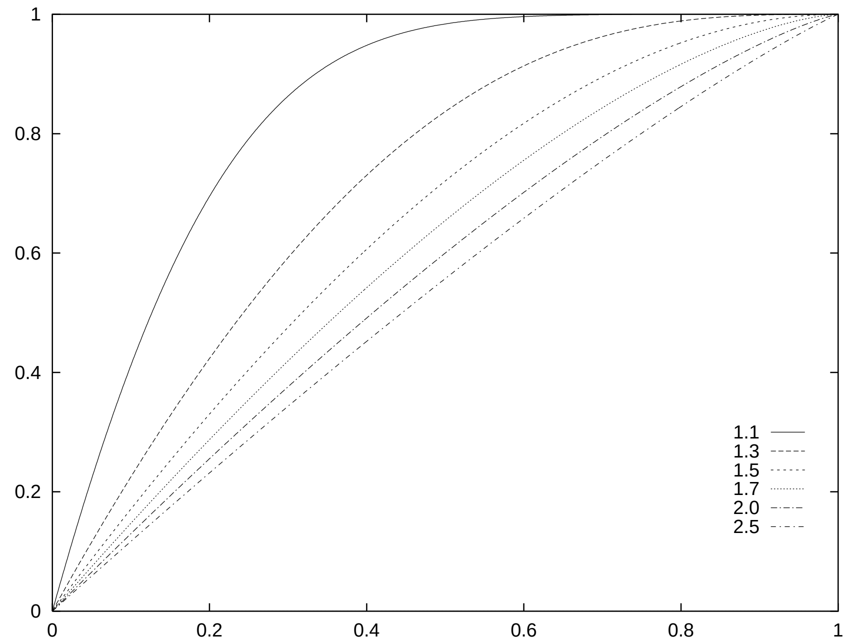

Figure 7: **ρ-s-Diagram for various values of α where $\theta = 0.5$ and $N = 30$.**

largest value s^* of s (for fixed ρ) will be obtained at different values of θ (fig. 12, 13, 15). However, we can see that s^* is a monotonic non-decreasing function of N. Also, θ loses importance as N increases. That is, the range of possible values for s as θ varies from 0 to 1, decreases with increasing N. So in the limit $N \to \infty$ it is no longer necessary to consider different values of θ, except perhaps if θ is close to 1 (see Figure 13 and next subsection).

We see then, that even though f_{Y_N} is well behaved, s can be very close to 1 even for moderate ρ. Part of this is due to the largeness of C^2, and thus the almost horizontal slope at $\rho = 1$. But for small α, and large N s stays close to 1 even when ρ is as small as 0.5. For fixed N, the smaller α is the smaller ρ will be before s deviates from 1. Also, for smaller α, N must be larger before approaching its limiting value, corresponding to the expected experimental result that smaller α requires more events to experience the full impact of a power tail, as implied by (14).

3.4 Behavior of Queues based on f

For $N \to \infty$ $\bar{x} \equiv \mathrm{E}(Y) = 1$ and (30) reduces to

$$g(s) = (1-\theta) \sum_{n=0}^{\infty} \frac{(\gamma\theta)^n}{(1-s)\gamma^n + \mu/\lambda} - 1. \tag{31}$$

From a numerical point of view it can be advantageous to solve for $t := 1 - s$ instead of s in the vicinity of $\rho = 1$. So we get the alternative equation

$$\tilde{g}(t) := g(1-s) = (1-\theta) \sum_{n=0}^{\infty} \frac{(\gamma\theta)^n}{t\gamma^n + \mu/\lambda} - 1 = (1-\theta) \sum_{n=0}^{\infty} \frac{(1-t)(\gamma\theta)^n - \theta^n \mu/\lambda}{t\gamma^n + \mu/\lambda} \stackrel{!}{=} 0\,. \tag{32}$$

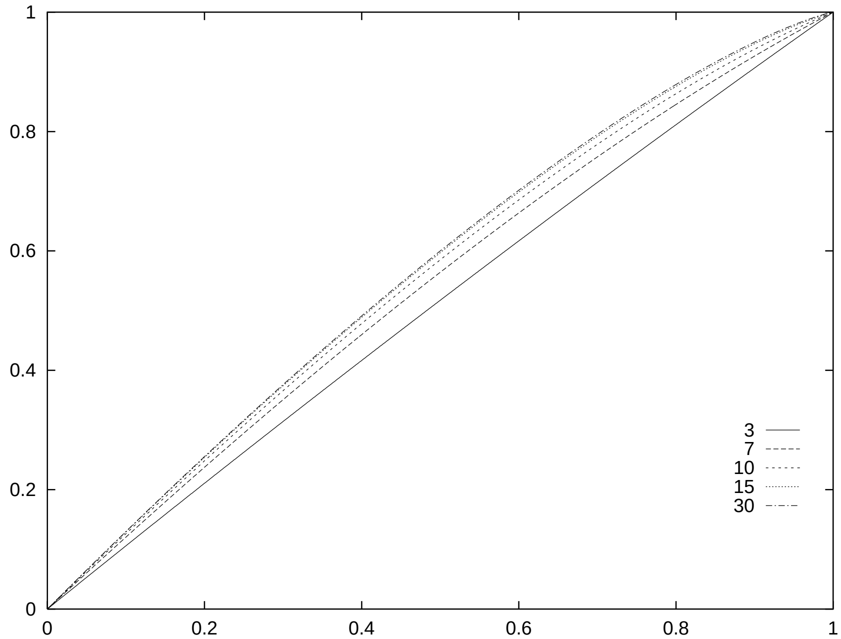

Figure 8: **ρ-s-Diagram for various values of N where $\alpha = 2.0$ and $\theta = 0.5$.**

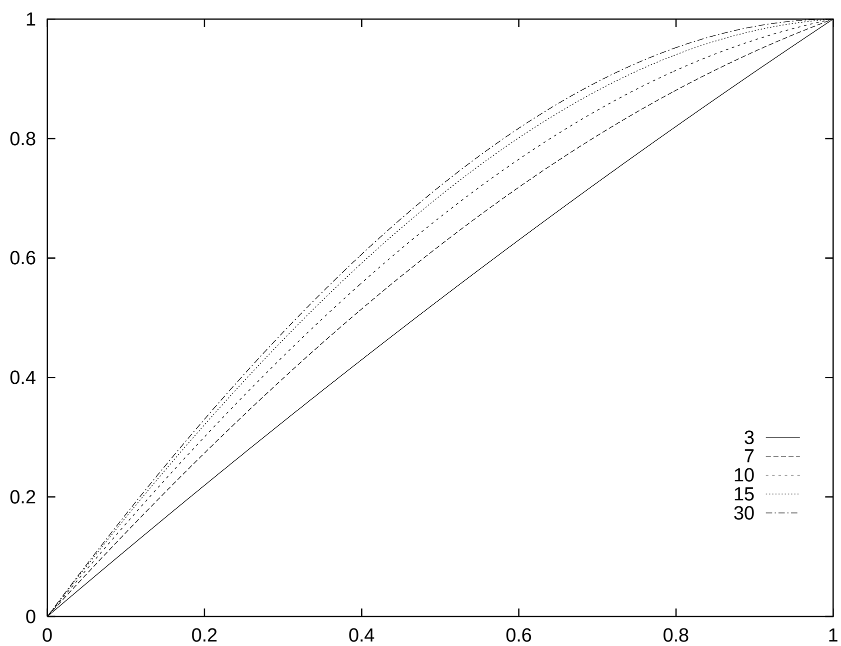

Figure 9: **ρ-s-Diagram for various values of N where $\alpha = 1.5$ and $\theta = 0.5$.**

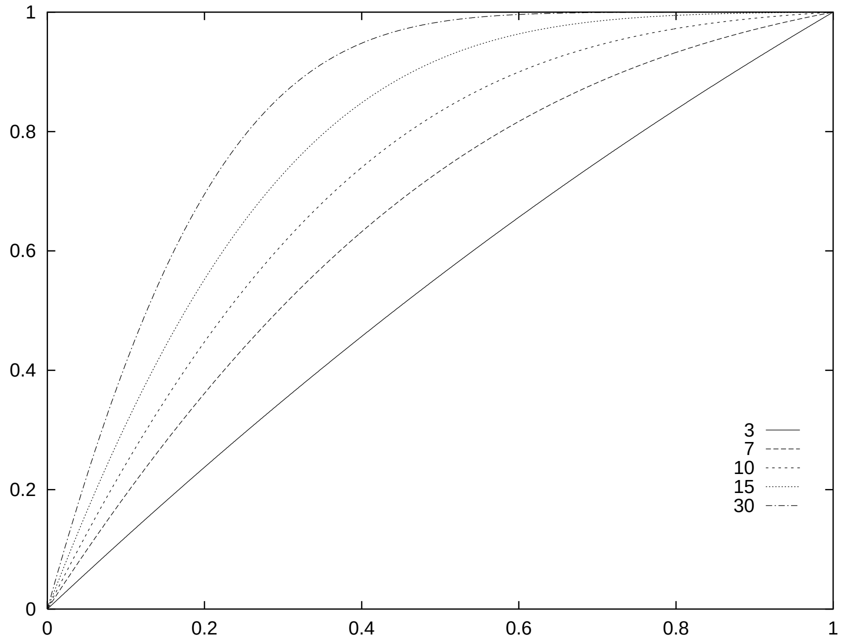

Figure 10: **ρ-s-Diagram for various values of N where $\alpha = 1.1$ and $\theta = 0.5$.**

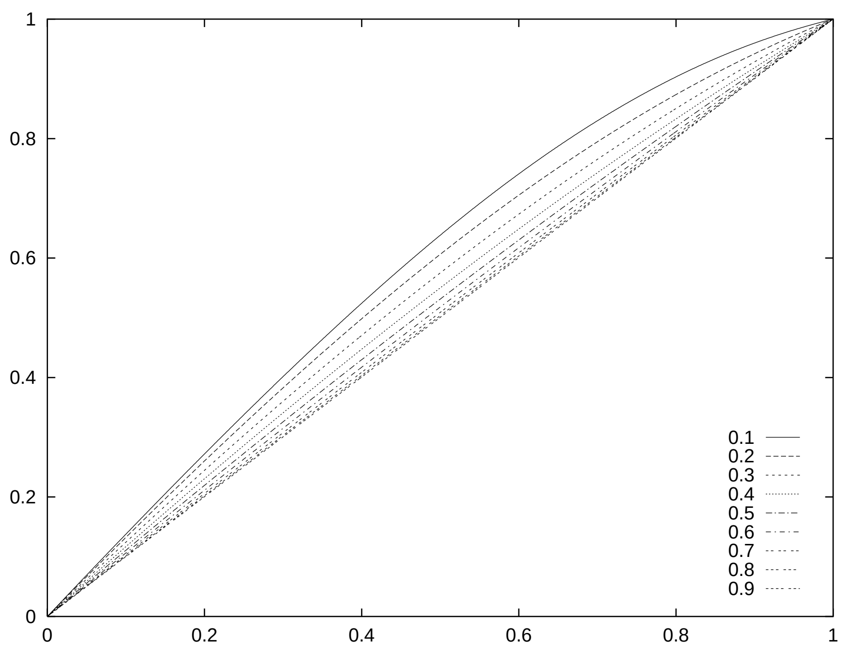

Figure 11: **ρ-s-Diagram for various values of θ where $\alpha = 1.5$ and $N = 3$.**

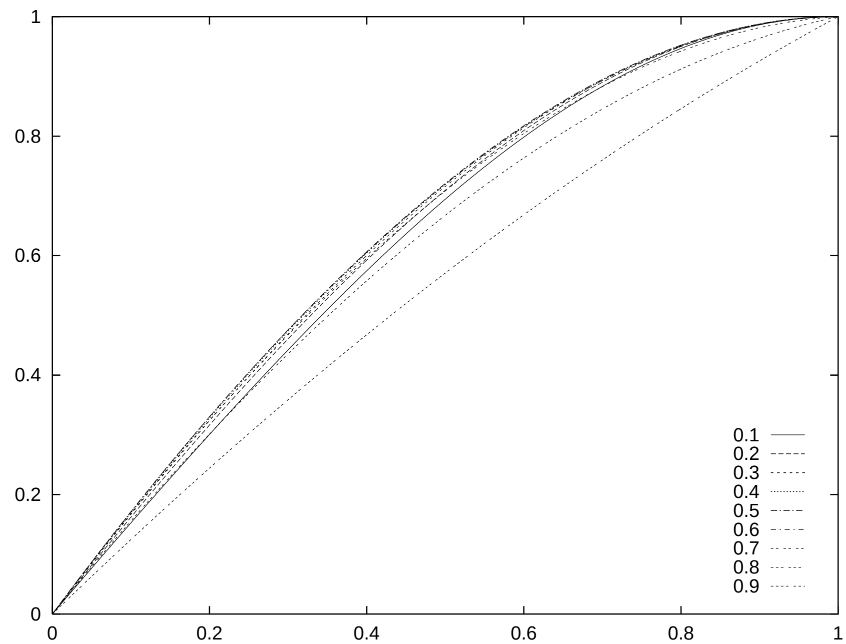

Figure 12: **ρ-s-Diagram for various values of θ where $\alpha = 1.5$ and $N = 30$.**

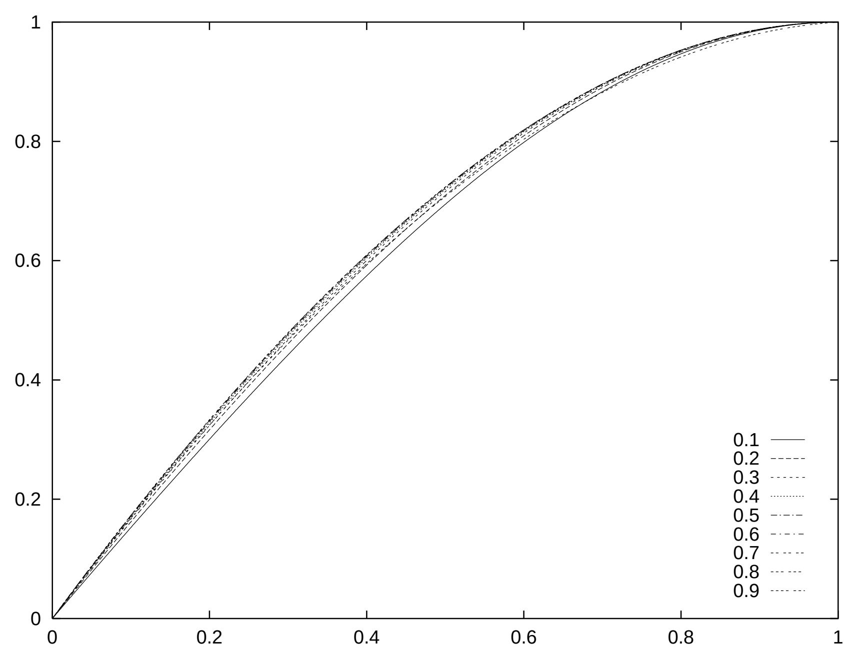

Figure 13: **ρ-s-Diagram for various values of θ where $\alpha = 1.5$ and $N = 100$.**

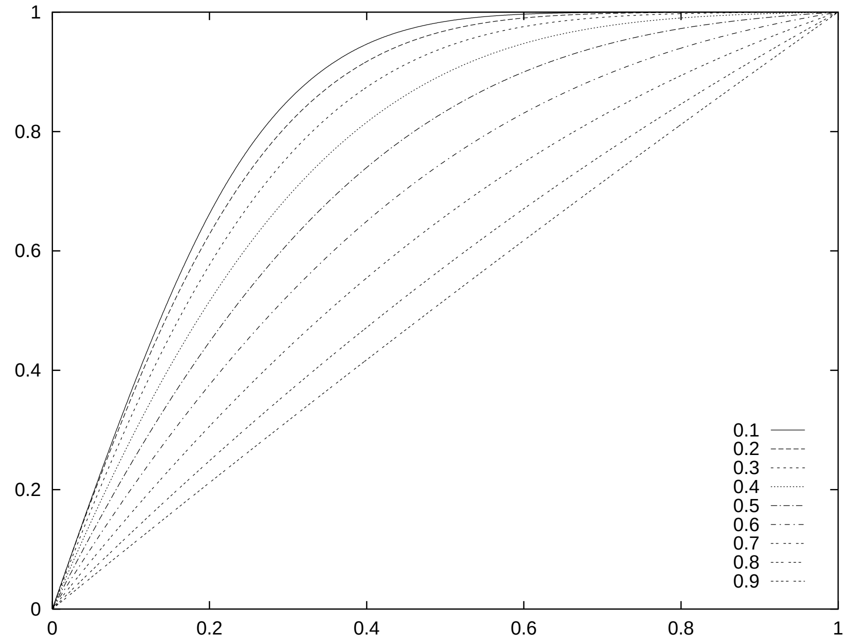

Figure 14: **ρ-s-Diagram for various values of θ where $\alpha = 1.1$ and $N = 10$.**

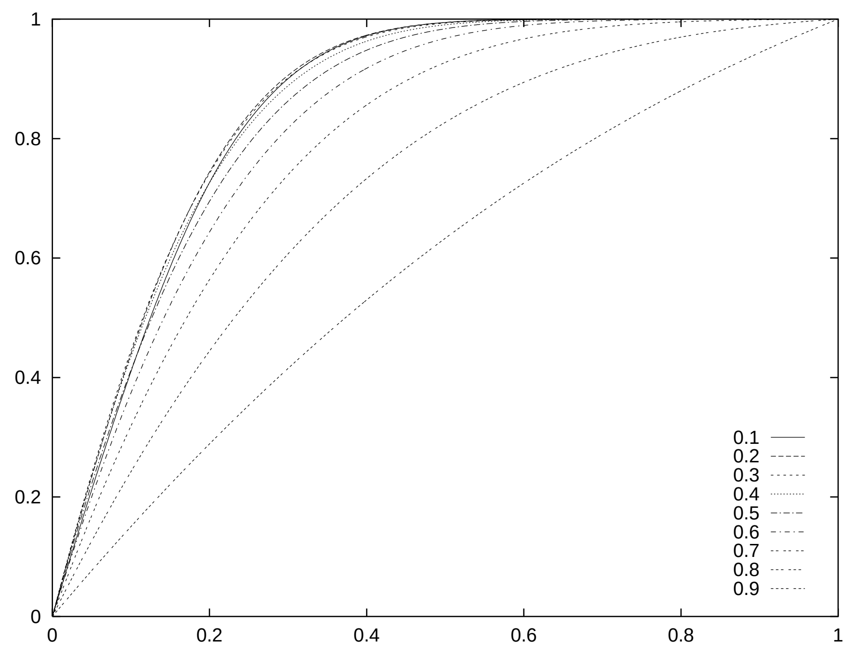

Figure 15: **ρ-s-Diagram for various values of θ where $\alpha = 1.1$ and $N = 30$.**

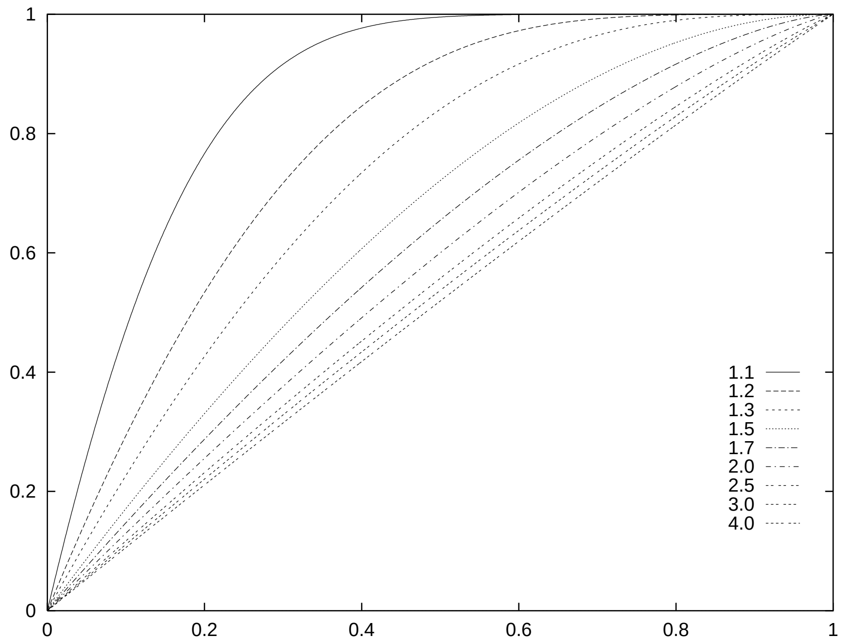

Figure 16: **ρ-s-Diagram for different values of α where $\theta = 0.5$ and $N \to \infty$.**

Figure 16 shows s for various values of α. As mentioned in the previous subsection (see p.19) the graphs differ only slightly for different values of θ. Note that this figure does not differ significantly from Figure 7, showing that a function with a truncated tail can cause almost as bad performance as its limit function. Also, the transition in going from small N to large N to ∞ is smooth and convergent. We also see that as α increases, performance improves (smaller s for the same ρ), and for large α approaches that of the M/M/1 queue ($s = \rho$). Again, the transition is smooth. For instance, even though $f(x)$ has an infinite variance for $\alpha = 2$, there is no abrupt change in behavior as one goes from $\alpha < 2$ to $\alpha > 2$.

Next, in Figure 17 we display for some values of α the value of $\rho = 1/(\lambda \bar{x})$ where t is 0.01 (0.001, 0.0001), and ρ is computed from equation (32). This figure attempts to show how small ρ can be and still have a value for $s = 1 - t$ close to 1, depending on α. Its significance can best be shown by the following.

Example: For $\alpha = 1.1$ we get $\rho \approx 0.46$ when $t = 0.01$, so

$$Pr(K = 10) = (1 - 0.01)^{10} = 0.99^{10} \approx 90.4\%$$

and $Pr(50) \approx 60.5\%$. In other words: Even if the utilization is less than a half the probability that the primary buffer of size 10 (50) is full when a new packet arrives is greater than 90% (60%). To keep the probability of overflow to below 10% would require a primary buffer of size $K = 229$. For $\alpha = 1.5$ these probabilities occur for $\rho \approx 0.91$. On the other hand, for $\alpha = 2.5$, we get the more reasonable result that ρ is greater than 0.98, i.e., a system must be nearing saturation (ρ close to 1) to get high probabilities of overflow. Remember that for $\alpha \leq 2$ Y has infinite variance [see (25)].

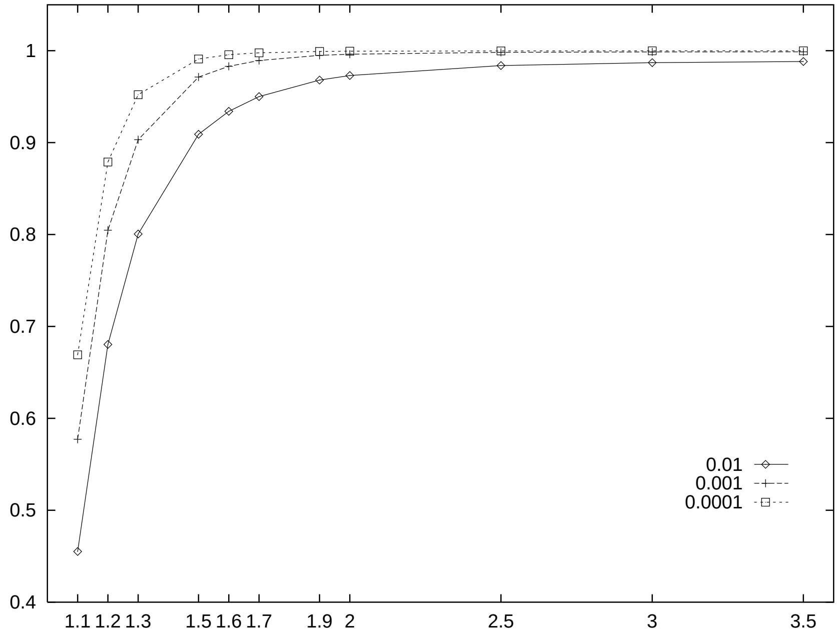

Figure 17: $\boldsymbol{\alpha}$-$\boldsymbol{\rho}$-**Diagram for** $\boldsymbol{t \in \{0.01, 0.001, 0.0001\}}$ **and** $\boldsymbol{\theta = 0.5}$ **.**

4 Conclusion/Summary

In the first part of this paper we gave a detailed description of the properties of power-tailed distributions, also known as Pareto, or Lèvy-Pareto distributions. They form a proper subclass of the so-called heavy-tailed, or sub-exponential distributions. The hierarchy of distributions is summarized in Figure 18. We then showed how application of the Central-limit Theorem must be modified when dealing with power tails. In particular we showed that the number of events which must transpire before steady-state solutions can be relevant grows unboundedly large as α approaches 1 from above (for $\alpha \leq 1$ there never can be a steady state). We then introduced a class of well behaved distributions which can be used to analytically model processes which have power tails, or truncated power tails. They are also useful for discrete event simulations. One of these distributions was used to model a steady-state G/M/1 queue. We showed that steady-state behavior (as represented by the geometric parameter, s) varies smoothly with $\alpha > 1$. It also varies smoothly as the truncated tail is filled in ($N \to \infty$). We reiterate the final remark of the introductory section: *To fully understand the real impact which power-tail distributions will have on telecommunications (and other) systems, appropriate descriptions of transient behavior must be developed.*

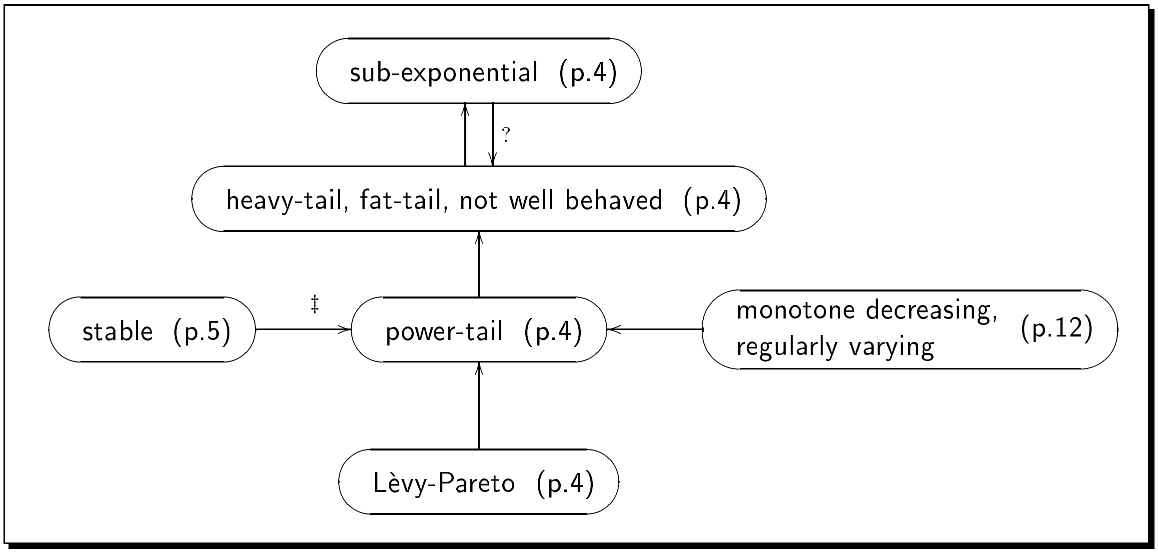

Figure 18: **Overview of the relations between various sub-classes of distributions.** "→" denotes the relation "is a sub-class of".

Appendix - Asymptotic behavior of g

In Section 3.4 we examined the properties of a G/M/1 queue by finding the unique root of (28) for $s = 1 - t$ in the open interval, (0, 1). We eliminated the extraneous root at $s = 1$ ($t = 0$) by using the expression

$$\tilde{g}(t) := \frac{1 - t - B^*(\lambda t)}{t}$$

which for our test function yields (32). It can be shown for any function [LIPS92], that $\tilde{g}(0) = 1/\rho - 1 = (1-\rho)/\rho$. A power-series expansion can be used to explore its behavior near $\rho = 1$. If the appropriate moments are finite, then for ρ close to 1 (i.e. t close to 0), one can write

$$\tilde{g}(t) \approx \tilde{g}(0) + \tilde{g}'(0) \cdot t + \tilde{g}''(0) \cdot t^2/2 + \cdots,$$

where for $k > 0$ (see [LIPS92])

$$\tilde{g}^{(k)}(0) = -k!\Psi[(-\lambda \mathbf{V})^{k+1}] = \frac{(-1)^k}{k+1} \cdot \frac{\mathrm{E}(X^{k+1})}{(\rho\bar{x})^{k+1}}.$$

For t small enough, $\tilde{g}(t) = 0$ can be solved approximately as a polynomial root-finding problem. However, if $\alpha \leq 2$, then the second moment is infinite, and the equation doesn't hold. We used instead,

$$\tilde{g}(t) = \frac{1-\rho}{\rho} + c \cdot t^{\alpha-1} + \cdots, \tag{33}$$

where c is some constant which could depend upon ρ. We now prove this equation, and give some explicit properties to c. Higher terms in the series can also be found in this way.

‡except for the normal distribution

Define

$$h(t\,;\beta) := \frac{\tilde g(t) - \tilde g(0)}{t^\beta}.$$

Clearly, if $\tilde g'(0)$ exists, then $h(0_+\,;1) = \tilde g'(0)$, $h(0_+\,;\beta < 1) = 0$, and $h(0_+\,;\beta > 1) = \infty$. Even if $\tilde g'(0)$ is infinite, there may exist some $\beta \neq 1$ such that

$$0 < \lim_{t \to 0} |h(t\,;\beta)| < \infty. \tag{34}$$

This is sometimes called the *generalized derivative*. For the power-tail distribution used in this paper we can write

$$\begin{aligned} t^\beta h(t\,;\beta) &= (1-\theta)\frac{\lambda}{\mu}\sum_{n=0}^{\infty}\frac{(\theta\gamma)^n}{1+(\lambda/\mu)t\gamma^n} - (1-\theta)\frac{\lambda}{\mu}\sum_{n=0}^{\infty}(\theta\gamma)^n \\ &= (1-\theta)\frac{\lambda}{\mu}\sum_{n=0}^{\infty}(\theta\gamma)^n\left[\frac{1}{1+(\lambda/\mu)t\gamma^n} - 1\right] = -(1-\theta)\left(\frac{\lambda}{\mu}\right)^2\sum_{n=0}^{\infty}\frac{t(\theta\gamma^2)^n}{1+(\lambda/\mu)t\gamma^n}. \end{aligned}$$

Recall from (24) that $\theta\gamma^\alpha = 1$, and let $\beta = \alpha - 1 + \varepsilon$, then

$$h(t\,;\alpha - 1 + \varepsilon) = -\frac{(1-\theta)}{t^\varepsilon}\left(\frac{\lambda}{\mu}\right)^2\sum_{n=0}^{\infty}\frac{(t\gamma^n)^{2-\alpha}}{1+(\lambda/\mu)t\gamma^n}.$$

Before taking the limit for $t \to 0$, we anticipate our results to simplify the above expression somewhat. As long as $1 < \alpha < 2$, the infinite sum converges, but $t^\varepsilon \to \infty$ if $\varepsilon < 0$, and $t^\varepsilon \to 0$ if $\varepsilon > 0$. Therefore, in order to satisfy constraint (34), we must have $\varepsilon = 0$. We assume this in what follows.

Let $\tilde h(t\,;\alpha - 1) := h(t\,\mu/\lambda\,;\alpha - 1)$. Then we have

$$\tilde h(t\,;\alpha - 1) = -(1-\theta)\left(\frac{\lambda}{\mu}\right)^\alpha\sum_{n=0}^{\infty}\frac{(t\gamma^n)^{2-\alpha}}{1+(t\gamma^n)}.$$

If $\tilde h(t\,;\alpha - 1)$ has a finite limit as $t \to 0$, then any sequence $\tilde h(t_\ell\,;\alpha - 1)$ will converge to the same limit if $\{t_\ell\}$ is a monotonic decreasing sequence with $t_\ell \to 0$. Suppose for now that the limit exists, and let $t_\ell = t_0/\gamma^\ell$, where $t_0 > 0$, but otherwise unspecified. Then

$$\tilde h(t_\ell\,;\alpha - 1) = -(1-\theta)\left(\frac{\lambda}{\mu}\right)^\alpha\sum_{n=0}^{\infty}\frac{(t_0\gamma^{n-\ell})^{2-\alpha}}{1+(t_0\gamma^{n-\ell})} = -(1-\theta)\left(\frac{\lambda}{\mu}\right)^\alpha\sum_{n=-\ell}^{\infty}\frac{(t_0\gamma^n)^{2-\alpha}}{1+(t_0\gamma^n)}.$$

We take the limit and get:

$$\lim_{\ell \to \infty}\tilde h(t_\ell\,;\alpha - 1) = -(1-\theta)\left(\frac{\lambda}{\mu}\right)^\alpha\sum_{n=-\infty}^{\infty}\frac{(t_0\gamma^n)^{2-\alpha}}{1+(t_0\gamma^n)}.$$

From the ratio test for convergence, it is clear that the doubly infinite sum converges as long as $1 < \alpha < 2$. That means the limit exists for each $t_0 > 0$. In order for the original equation to have a (unique) finite limit, the doubly infinite sum must be independent of t_0. We have computed the sum for many values of t_0, γ and α, and have found that for fixed α, γ, the sums for different values of t_0 agree to 14 significant decimal digits, the full

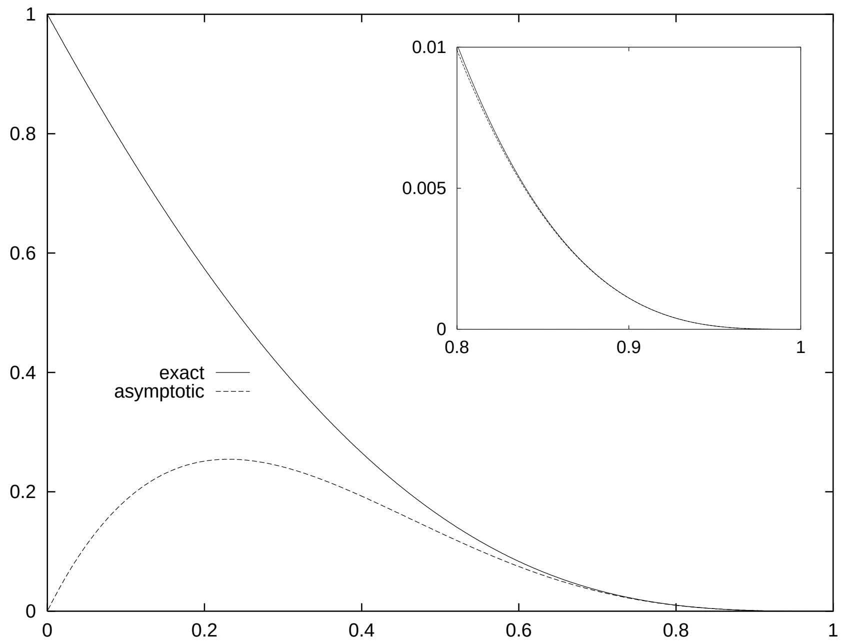

Figure 19: **$\boldsymbol{\rho}$-$\boldsymbol{t}$-Diagram both for the exact and the asymptotic equation in $\boldsymbol{t}$ where $\boldsymbol{\alpha = 1.3}$ and $\boldsymbol{\theta = 0.5}$.** From the insert, it is clear that the asymptotic equation is an excellent approximation for ρ as small as 0.9 .

precision of our calculations. Thus we assume that the limit indeed exists, and is equal to (we arbitrarilly set t_0 to 1, and recall that $\lambda/\mu = (1-\theta\gamma)/[(1-\theta)\rho]$, with $\theta = 1/\gamma^\alpha$)

$$c \equiv c(\alpha\,,\gamma) = -\frac{(\gamma^\alpha - \gamma)^\alpha}{\gamma^\alpha \rho^\alpha (\gamma^\alpha - 1)^{\alpha-1}} \sum_{n=-\infty}^{\infty} \frac{(\gamma^n)^{2-\alpha}}{1+\gamma^n}$$

since $\lim\limits_{t\to 0} \tilde{h}(t\,;\alpha-1) = \lim\limits_{t\to 0} h(t\,;\alpha-1) \overset{(33)}{=} (c \cdot t^{\alpha-1})/t^{\alpha-1} = c\,.$

When $\tilde{g}(t)$ is set to 0 from (33), we get for t,

$$t^{\alpha-1} = \left[\frac{\rho\gamma(\gamma^\alpha - 1)}{\gamma^\alpha - \gamma}\right]^{\alpha-1} \frac{\gamma(1-\rho)}{\gamma^\alpha - \gamma}\,\Gamma(\alpha\,,\gamma), \tag{35}$$

where for convenience, we have defined:

$$\frac{1}{\Gamma(\alpha\,,\gamma)} := \sum_{n=-\infty}^{\infty} \frac{(\gamma^n)^{2-\alpha}}{1+\gamma^n}.$$

This equation shows how $s = 1-t$ approaches 1 as ρ approaches 1, for $1 < \alpha < 2$. It is also extremely useful as an initial guess for finding the exact root of $g(s) = 0$ [or $\tilde{g}(t) = 0$] by numerical means. In fact, the closer ρ is to 1, the harder it is to find the exact root *unless* this asymptotic expression is used. Figures 19 and 20 visualize the quality of approximation by (35) in the vicinity of $\rho = 1$. Clearly, the smaller α is, the better is the approximation even for smaller ρ .

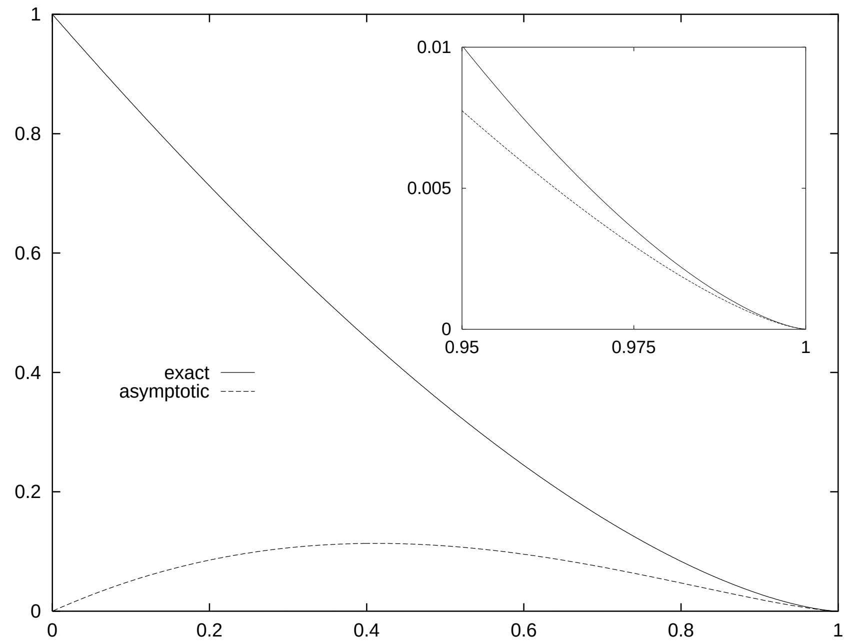

Figure 20: **$\boldsymbol{\rho}$-$\boldsymbol{t}$-Diagram both for the exact and the asymptotic equation in $\boldsymbol{t}$ where $\boldsymbol{\alpha = 1.7}$ and $\boldsymbol{\theta = 0.5}$.** Here the asymptotic equation is only good for $\rho > 0.98$.

As one last comment, we note that the doubly infinite sum diverges for $\alpha = 1$ and $\alpha = 2$, thus $\Gamma(1\,,\gamma) = \Gamma(2\,,\gamma) = 0$. Interestingly, $\Gamma(\alpha\,,\gamma)$ is actually symmetric about $\alpha = 1.5$. By replacing the dummy variable n with $-n$ in the sum, and then manipulating a little, it can be shown that for $1 \leq \alpha \leq 2$

$$\Gamma(\alpha\,,\gamma) = \Gamma(3 - \alpha\,,\gamma),$$

or for $0 \leq \delta \leq 1$

$$\Gamma(1 + \delta\,,\gamma) = \Gamma(2 - \delta\,,\gamma).$$

That is $\Gamma(1.1\,,\gamma) = \Gamma(1.9\,,\gamma)$, $\Gamma(1.2\,,\gamma) = \Gamma(1.8\,,\gamma)$, etc.

References

[BERA95] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, "Variable-Bit-Rate Video Traffic and Long-Range Dependence", *IEEE/ACM Trans. on Networking*, 1995.

[FELL71] William Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley and Sons, New York, 1971.

[GARG92] Sharad Garg, Lester Lipsky and Maryann Robbert, "The Effect of Power-Tail Distributions on the Behavior of Time Sharing Computer Systems", *1992 ACM SIGAPP Symposium on Applied Computing*, Kansas City, MO, March, 1992.

[GEOR94] Nicolas D. Georganas, "Self-Similar ("Fractal") Traffic in ATM Networks". In: Ralf Steinmetz (Ed.), Lecture Notes in Computer Science, Vol. 868, *Multimedia: Advanced Teleservices and High-Speed Communication Architectures*, Proc. 2nd Intern. Workshop, IWACA'94, Heidelberg, Germany, Sept. 1994.

[LELA85] Will E. Leland and Teunis Ott, "Analysis of CPU times on 6 VAX 11/780's at BELLCORE", *Proceedings of the International Conference on Measurement and Modeling of Computer Systems*, April 1986.

[LELA94] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel V. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", *Proc. of IEEE/ACM Trans. on Networking*, **2**,1, Feb. 1994.

[LIEF??] Appie van de Liefvoort, *???Reference on the moment problem???*.

[LIEF94] Appie van de Liefvoort and Hai Feng Weng, "On the Instability ('Fractal' Behavior) of Arrival Counts Generated By Power-Tail Renewal Processes", Department Report, Computer Science and Telecommunications Program, University of Missouri-Kansas City, 1994.

[LIKH95] Nikolai Likhanov, Boris Tsybakov and Nicolas D. Georganas, "Analysis of an ATM Buffer with Self-Similar ("Fractal") Input Traffic", Proc. IEEE INFOCOM'95, Boston, April 1995.

[LIPS86] Lester Lipsky, "A Heuristic Fit of an Unusual Set of Data", Bell Communications Research Report, January 1986.

[LIPS92] Lester Lipsky, *QUEUEING THEORY: A Linear Algebraic Approach*, MacMillan and Company, New York, 1992.

[LIPS95] Lester Lipsky and Pierre Fiorini, "Auto-Correlation of Counting Processes Associated with Renewal Processes", Technical Report, Booth Research Center, University of Connecticut, August 1995.

[LOWR93] Walter Lowrie and Lester Lipsky, "A Model For The Probability Distribution of Medical Expenses", Proceedings of *CONFERENCE OF ACTUARIES IN PUBLIC PRACTICE*, 1993.

[MINN67] Report of the University of Minnesota Computing Center, 1967?

[NEUT81] Marcel F. Neuts, *MATRIX-GEOMETRIC SOLUTIONS IN STOCHASTIC MODELS, an Algorithmic Approach*, John Hopkins University Press, Baltimore and London, 1981.

[ROSI65] R. F. Rosin, "Determining a Computer Center Environment", *CACM*, 1965.

[TRIV82] Kishor S. Trivedi, *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, Prentice-Hall,Englewood Cliffs, NJ, 1982.