BACHELOR

Heavy-tailed distributions

Simonis, B.J.

Award date:
2004

Bachelor project

Technische Universiteit Eindhoven

Faculteit Wiskunde

Barbara Simonis (535317)

3rd November 2004

Contents

1 Introduction

For a long time it was assumed that we could model network traffic as a Poisson process. But by careful statistical analysis of large sets of actual data it was revealed that this is a wrong assumption. Network traffic has certain properties that can not be found in a Poisson process. In fact by looking at these properties an interesting class of distributions had to be studied, i.e. heavy-tailed distributions.

In this report we can find overviews of several articles about self-similar/heavy-tailed processes. Also some theory about these topics is given.

Section 2 explains how people who worked with Ethernet traffic changed their attention from Poisson processes to self-similar processes. In section 3 a brief overview of the theory concerning heavy-tailed processes is given. In section 4 a survey is given concerning the tail behavior of the waiting (or sojourn) time distribution in the M/G/1 queue with regularly varying service time distribution.

2 Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements

This is a brief overview of [1]. In this article it is shown how a careful statistical analysis of large sets of actual traffic measurements can reveal new features of network traffic. These features have gone unnoticed by the literature and seem to have serious implications for predicted network performance.

2.1 From Megabit to Gigabit Networks: LAN, MAN and BISDN

Local area networks (LAN's) were introduced in the 70's to interconnect data processing equipment. The Ethernet was an early LAN technology and it remains one of the most popular in use today.

From today's perspective Ethernet has two disadvantages. Firstly the relatively low speed (10 Mbit/s) and secondly the limited range (the physical span is limited to a few kilometers). So people kept on looking for something else and they came up with broadband integrated services digital network (BISDN).

In absence of practical experience with gigabit networks (like BISDN), the challenge for today's traffic engineers is to gain an understanding of the likely characteristics of future BISDN traffic. Due to the already existing need, LAN interconnection services are expected to become an immediate and major BISDN application. Therefore, understanding the characteristics of LAN traffic such as Ethernet will be very useful. So we want to model Ethernet traffic.

The paper [1] uses statistical techniques to arrive at an Ethernet traffic model. In doing so, it diverges from the usual approach in telecommunication systems, where traffic models are typically judged by how well they predict the performance of the queueing model alone, and almost never by how well the model fits actual traffic.

A detailed analysis of the actual traffic data leads to the conclusion that Ethernet traffic is statistically *self-similar*. Self-similar processes were introduced by Kolmogorov. Intuitively, they can be recognized by the fact that they look the same in different scales. In terms of stochastic modeling, self-similar processes or their increments processes are almost exclusively used in situations where the modeler tries to account for the presence of long-term correlations in a parsimonious manner.

It is easy to see that self-similar models fit Ethernet LAN better than conventional traffic models, all of which ignore the presence of long-term correlations in the data. It is possible to arrive at this conclusion through simple plots of the traffic over a range of different time scales, as we will show now.

2.2 Traffic Measurements

The writers of [1] concentrate exclusively on the data analysis and modeling aspects of the Ethernet traffic measurements. For a detailed description of how the data is collected the writers refer to other literature. There is no limitation on the amount of traffic that can be collected. The traffic measurements are of unusual quality. The traffic measurements in this paper were collected at the Bellcore Morris Research and Engineering Center. With these data sets, they were able to investigate features of the observed traffic, e.g. self-similarity.

With the graphics of figure 1 (p 6) it is easy to see that Ethernet traffic is self-similar. On the left-side of the picture we can see the data that was collected. Each subsequent plot is obtained from the previous one by increasing the time resolution by a factor 10 and by concentrating on a randomly chosen subinterval. We can observe that all the plots are very similar to one another. So Ethernet traffic seems to look the same in the large time scales as in the small ones. So it is now (intuitively) clear that Ethernet traffic is self-similar. This scale-invariant or self-similar feature of Ethernet traffic is drastically different from stochastic models for packet traffic currently considered in the literature. On can easily conclude this by looking at the right-side of the picture. This sequence was obtained in the same way as the left sequence, except that it depicts synthetic traffic generated from a comparable compound Poisson process.

So now we know that using a Poisson process to model traffic is not a wise course of action. We have to look at distributions that are self-similar e.g. *heavy-tailed distributions*. In the next section some basic notions about these distributions are given.

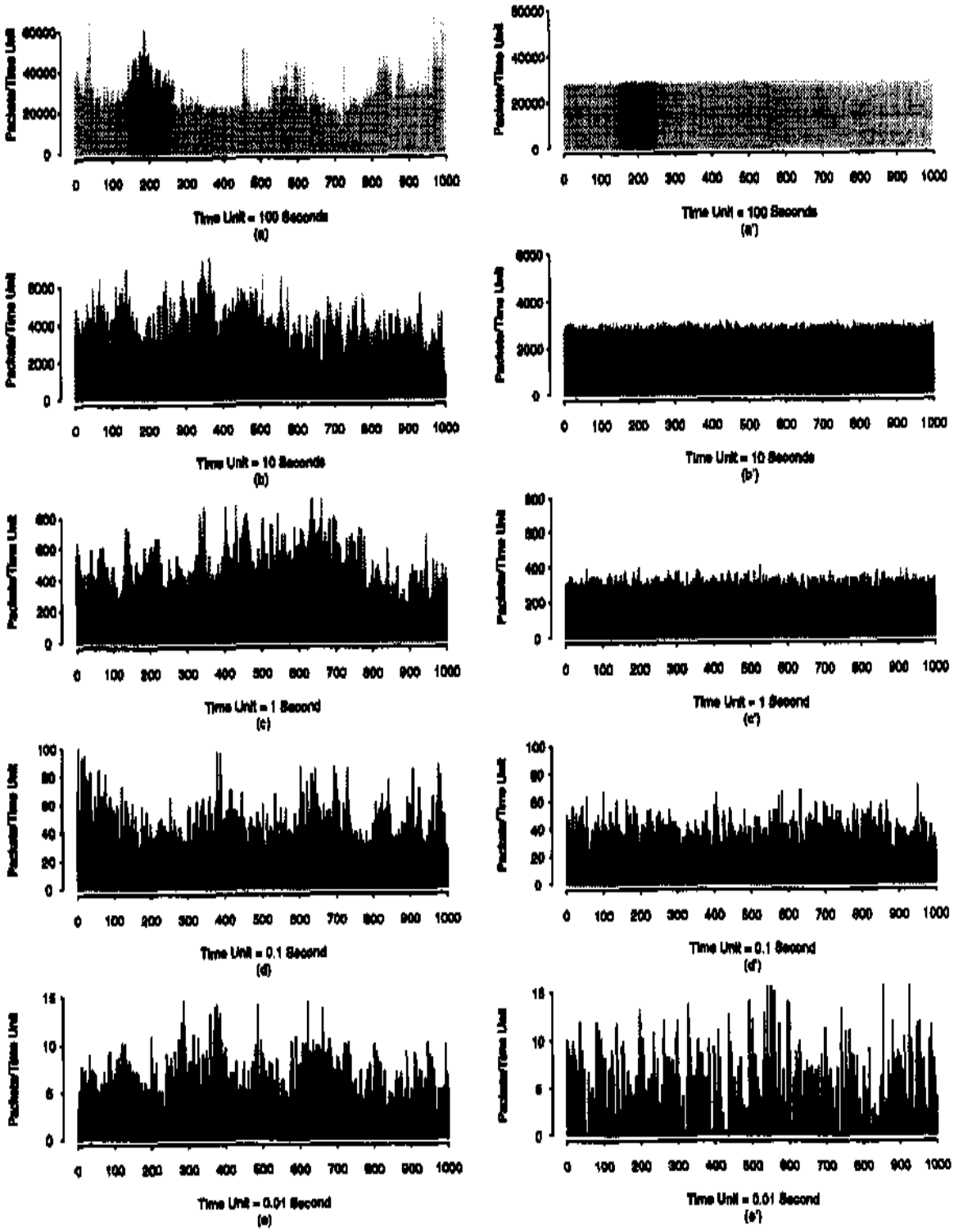

Figure 1: Indication of self-similarity: Ethernet traffic on five different time scales (a)-(e): for comparison, synthetic traffic from an appropriately chosen Poisson model on the same five different time scales (a')-(e')

3 A primer on heavy-tailed distributions

This section contains a brief overview of the theory of *heavy-tailed distributions*. It is based on the survey [2].

3.1 Gives some definitions and notations about these distributions. Next subexponential and equilibrium distributions will be discussed. 3.4 Gives an introduction on the interesting class $\mathcal{S}^*$. In 3.5 we can find some theory about the GI/GI/1 queue. Finally in 3.6 and 3.7 self-similarity and long-range dependence will be explained.

3.1 Heavy-tailed distributions

First some definitions and notation.
Given a non-negative random variable X, its distribution function is denoted by

$$F(x) = P(X \leq x) \tag{1}$$

and its tail is denoted by

$$\overline{F}(x) = 1 - F(x) = P(X > x) \tag{2}$$

F is *heavy tailed* if, for all $\epsilon > 0$,

$$E(e^{\epsilon X}) = \infty$$

A major subclass of heavy-tailed distributions is the class of *long-tailed* distributions.
A distribution function F (or the random variable X) is said to be *long-tailed* if $\overline{F}(x) > 0, x \geq 0$ and

$$\lim_{x\to\infty} P(X > x + y | X > x) = \lim_{x\to\infty} \frac{\overline{F}(x+y)}{\overline{F}(x)} = 1, \qquad \forall y \geq 0. \tag{3}$$

Letting $a(x) \sim b(x)$ means that $\frac{a(x)}{b(x)} \to 1$ as $x \to \infty$, we can express (3) as

$$\overline{F}(x+y) \sim \overline{F}(x), \qquad \forall y \geq 0 \tag{4}$$

A standard example of a heavy-tailed distribution is the *Pareto distribution*. This distribution has tail $\overline{F}(x) = x^{-\alpha}$, $x \geq 0$, where $\alpha > 0$ is a parameter.

The class of long-tailed distributions is denoted by $\mathcal{L}$ (so $F \in \mathcal{L}$ or $X \in \mathcal{L}$).

3.2 Subexponential distributions

Now we define a subclass $\mathcal{S}$ of $\mathcal{L}$, the subexponential distributions.
First some notation.

Let $F^{*n}(x)$ denote the n-fold convolution of $F(x)$ then

$$F^{*n}(x) = \int_0^x F^{*n-1}(x-y)dF(y) \tag{5}$$

with

$$F^{*2}(x) = \int_0^x F(x-y)dF(y) \tag{6}$$

The distribution function F (or the random variable X) is called *subexponential* if $\overline{F}(x) > 0, x \geq 0$ and for all $n \geq 2$,

$$\lim_{x\to\infty} \frac{\overline{F^{*n}}(x)}{\overline{F}(x)} = n. \tag{7}$$

The above definition can be restated as

$$P(X_1 + ... + X_n > x) \sim nP(X > x) \tag{8}$$

and this can equivalently be stated as

$$P(X_1 + ... + X_n > x) \sim P(\max(X_1, ..., X_n) > x). \tag{9}$$

Equation (9) is also known as the *catastrophe principle*, a sum is large most likely due to *one* large term.

An interesting example of subexponential distributions is presented by *regularly varying distributions*.
With $\alpha > 0$, $\overline{F}$ is said to be regularly varying with index $-\alpha$ if it is a regularly varying function, that is, if

$$\lim_{x\to\infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = t^{-\alpha}, \qquad t > 0. \tag{10}$$

Such tails can be equivalently represented in the form $\overline{F}(x) = L(x)x^{-\alpha}$, where $L(x)$ is a slowly varying function (that is, regularly varying with $\alpha = 0$; $\frac{L(tx)}{L(x)} \to 1$).

3.2.1 Convolutions and tail equivalence

$\mathcal{S}$ is not closed under convolution. Here follow some results about convolution.

Proposition

If $F \in \mathcal{S}$ and G is any distribution function with a "lighter" tail than F, that is, G satisfies

$$\lim_{x\to\infty} \frac{\overline{G}(x)}{\overline{F}(x)} = 0, \tag{11}$$

then $F * G \in \mathcal{S}$ with $\overline{F * G}(x) \sim \overline{F}(x)$

Proposition

If $F \in \mathcal{S}$ and G is any distribution function such that

$$\overline{G}(x) \sim c\overline{F}(x), \qquad c > 0 \tag{12}$$

then $G \in \mathcal{S}$ and $F * G \in \mathcal{S}$ and $\overline{F * G}(x) \sim (1 + c)\overline{F}(x)$.

3.2.2 Random sums

If $\{X_n\}$ is an i.i.d. sequence of random variables, and $N \geq 0$ is an integer-valued random variable, it is of interest to determine when a random sum of the form

$$Y = \sum_{n=1}^{N} X_n \tag{13}$$

is subexponential. A useful special case and related results are provided by the next proposition.

Proposition

Suppose Y is of the form (13) with i.i.d. $\{X_n\}$ distributed as F, and $N \geq 0$ is an integer-valued random variable independent of $\{X_n\}$ such that $0 < E(N) < \infty$.

- If $F \in \mathcal{S}$, and if $E(e^{\epsilon N}) < \infty$, for some $\epsilon > 0$, then $Y \in \mathcal{S}$ and in fact

$$P(Y > x) \sim E(N)\overline{F}(x) \tag{14}$$

- If (14) holds and if $P(N = n) > 0$ for some $n \geq 2$, then $F \in \mathcal{S}$ and thus $Y \in \mathcal{S}$ also.
- If N has a geometric distribution, $P(N = n) = (1 - a)a^n$, $n \geq 0$, with $0 < a < 1$, then $F \in \mathcal{S}$ if and only if $Y \in \mathcal{S}$ if and only if (14) holds.

3.3 Equilibrium distributions

For any non-negative random variable X with distribution F and finite mean $\frac{1}{\mu}$, the *equilibrium distribution* F_e is defined by

$$F_e(x) = \mu \int_0^x \overline{F}(y)dy, \qquad x > 0. \tag{15}$$

Here follow some results about the equilibrium distribution.

Lemma

If $F \in \mathcal{L}$, then $F_e \in \mathcal{L}$ and $\overline{F_e}$ is heavier than $\overline{F}$;

$$\lim_{x\to\infty} \frac{\overline{F}(x)}{\overline{F_e}(x)} = 0 \tag{16}$$

Lemma

For any distribution function F (of a non-negative random variable) with finite mean μ^{-1}

$$\overline{F}(x) \leq \mu^{-1}(\overline{F_e}(x-1) - \overline{F_e}(x)), \qquad x \geq 0. \tag{17}$$

Corollary

If $F_e \in \mathcal{L}$, then $\overline{F_e}$ is heavier than $\overline{F}$; previous lemma holds.

3.4 The class $\mathcal{S}^* \subset \mathcal{S}$

We can restrict the class $\mathcal{S}$ to the class $\mathcal{S}^* \subset \mathcal{S}$.
First we define $\mathcal{S}^*$.

Let F be a distribution function on $[0, \infty)$ such that $\overline{F}(x) > 0$, $x \geq 0$. We say that $F \in \mathcal{S}^*$ if F has finite first moment $\frac{1}{\mu}$ and

$$\lim_{x\to\infty} \mu \int_0^x \frac{\overline{F}(x-y)}{\overline{F}(x)} \overline{F}(y)dy = 2. \tag{18}$$

Class $\mathcal{S}^*$ is naturally interesting because of the following property.

Proposition

If $F \in \mathcal{S}^*$, then both F and F_e are subexponential.
In particular $\overline{F * F_e}(x) \sim \overline{F_e}(x)$ holds if $F \in \mathcal{S}^*$.

3.5 Delay asymptotics for the FIFO GI/GI/1 queue

We consider here a single-server queueing model in which customer interarrival times $\{T_n\}$ are i.i.d., as $A(x) = P(T \leq x)$ with finite non-zero mean $E(T) = \frac{1}{\lambda}$, and service times $\{S_n\}$ are i.i.d., as $F(x) = P(S \leq x)$ with finite non-zero mean $E(S) = \frac{1}{\mu}$. The two sequences are assumed independent. $\rho = \frac{\lambda}{\mu} < 1$ (stability). Customers join the queue and are served in the order they arrive. The delay of the n^{th} customer is denoted by D_n and satisfies the recursion

$$D_{n+1} = (D_n + S_n - T_n)_+, \qquad n \geq 0. \tag{19}$$

D denotes steady-state delay: $P(D \leq x) = \lim_{n\to\infty} P(D_n \leq x)$, and has the same distribution as the maximum of the negative drift random walk

$$R_n = \sum_{j=1}^{n}(S_j - T_j), \qquad n \geq 1, \qquad R_0 = 0; \tag{20}$$

$$D = \max_{n\geq 0} R_n. \tag{21}$$

The following theorem shows how fundamental $\mathcal{S}$ is in the context of queues.

Theorem

- If S_e is subexponential, then
$$P(D > x) \sim \frac{\rho}{1-\rho} P(S_e > x). \tag{22}$$
- If the arrival process is a homogenous Poisson process at rate λ (M/G/1), then D is subexponential if and only if (22) holds.

We don't give a proof of this theorem. But with the following intuition from [3] we can believe that this theorem is valid.

Let us focus on the workload in the system at time $t = 0$. The assumption is that a large workload level is most likely due to the prior arrival of one customer with a large service requirement S, let us say at time $t = -y$ (this assumption is nothing but an educated guess at this stage). Note that from $t = -y$ onward, the workload decreases in a roughly linear fashion at rate $1 - \rho$. So in order for the workload at time $t = 0$ to exceed the level x, the service requirement S must be larger than $x + y(1 - \rho)$. Observing that customers arrive according to a Poisson process of rate λ, integrating with respect to y and making the substitution $z = x + y(1 - \rho)$, we obtain, for large x,

$$\begin{aligned} P(D > x) &\approx \int_{y=0}^{\infty} P(S > x + y(1-\rho))\lambda dy \\ &= \frac{\lambda}{1-\rho}\int_{z=x}^{\infty} P(S > z)dz \\ &= \frac{\rho}{1-\rho} P(S_e > x). \end{aligned}$$

And this gives exactly the result of the theorem.

3.6 Definition of Self-Similar Processes

In this section some definitions and properties about *self-similar* processes are stated.

A stochastic process $\mathcal{X} = \{X(t), t \geq 0\}$ is (strictly) *self-similar* with parameter H if $\{X(t), t \geq 0\}$ and $\{\gamma^{-H} X(\gamma t), t \geq 0\}$ have the same finite-dimensional distributions for any $\gamma > 0$.

For some purposes it suffices to give an intuitive explanation.
If $T(.)$ is self-similar with parameter H, then

$$Var(T(t)) \sim c_2 t^{2H} \tag{23}$$

A stochastic process that satisfies relation (23) is called *long-range dependent*.

Before we can define long-range dependent we first have to define a special function.

Let $\mathcal{X} = \{X(t), t \geq 0\}$ be some strictly stationary stochastic process.
Define the autocorrelation function

$$c(t) = \frac{Cov(X(s), X(s+t))}{Var(X(s))} \tag{24}$$

The following definition is standard.

$\mathcal{X}$ is *short-range dependent* if $\int_0^\infty |c(t)|dt < \infty$.
If $\int_0^\infty |c(t)|dt = \infty$, then $\mathcal{X}$ is *long-range dependent*.

3.7 Explaining long-range dependence via heavy tails

Several studies at the application level[1] indicate that long-range dependence may be caused by heavy-tailedness of certain traffic characteristics.
Let Y be a generic file size or transmission time. Then, typically,

$$P(Y > t) \sim c_3 t^{-\alpha}, \qquad 0 < \alpha < 2. \tag{25}$$

This result is confirmed by a number of measurements.

Now we want to refer to [5] to explain a bit more about the relation between long-range dependence and heavy tails.

[1] A number of studies that tried to explain results by examining quantities related to network traffic at a much higher level of aggregation.

A way to introduce long-range dependence in an input process is to take a fluid queue fed by a single on/off source, and to assume that a typical on-period A has the following tail:

$$P(A > t) \sim h_a t^{-a}, \qquad t \to \infty, \tag{26}$$

and/or that a typical off-period S has the following tail:

$$P(S > t) \sim h_s t^{-s}, \qquad t \to \infty, \tag{27}$$

with $1 < a, s < 2$ and h_a, h_s positive constants.

So if $1 < a < 2$ the distributions are heavy tailed. We can allow more sources, some of them having a heavy tailed on- and/or off-period distribution.
If at least one of the tails of the on- or off-period is heavy tailed, then the input process is long-range dependent (cf. the following theorem).

Theorem

> Assume that for all i, the distribution of A_{i1} or S_{i1} is non-lattice. Then the stationary process $(r^*(t))_{t\geq 0}$ is long-range dependent if and only if $E[A_{i1}^2] = \infty$ or $E[S_{i1}^2] = \infty$ for some i.

This theorem explains the long-range dependence of a fluid queue with an infinite buffer and outflow rate equal to one, fed by $N > 1$ independent on/off sources. For $1 \leq i \leq N$, we assume that source i, in his active period (A_i), has an input rate $r_i \geq 1$. S_i characterizes the silent period, so the period in which the source generates no input. The global traffic is characterized by the input rate process $r^*(t) := \sum_{i=1}^{N} r_i^*(t)$.

4 The impact of the service discipline on delay asymptotics

In this section we give a brief overview of [3]. In this article a survey is given concerning the tail behavior of the waiting (or sojourn) time distribution in the M/G/1 queue with regularly varying [2] service time distribution.

The key result appears to be the following. The service discipline plays a crucial role with respect to this tail behavior. In particular the tail of the waiting time distribution for the FCFS, LCFS-NP queues is regularly varying of one degree heavier than the tail of the service time distribution. And in contrast to this, the tail of the sojourn time distribution for the PS, FBPS, SRPTF and LCFS-PR queue is just as heavy as the tail of the service requirement time distribution.

In 4.1 we state Theorems 1-6, which give details about the above mentioned results. In the third section we elaborate on the transform approach which can be used to prove the Theorems. Subsection 4.3 tells about tail equivalence via conditional moments and in section 4.4 there will be some heuristics about sample-path techniques.

4.1 Model description and main results

In this section the model is formally described, some notations and concepts are introduced and an overview of the main results is given.

We focus on the M/G/1 queue. Customers arrive according to a Poisson process, with rate λ, at a single server. The service requirements $B_1, B_2, ...$ of the customers are identically distributed, with distribution function $B(.)$ and mean β. There is no restriction on the number of customers in the system. The stability condition equals $\rho = \lambda\beta < 1$. We study the steady-state sojourn time S of a customer, and in some cases also the steady-state waiting time W until service begins.

First we introduce some useful terminology.
For any stochastic variable X with distribution function $F(.)$, with $E(X) < \infty$, denote by $F^r(.)$ the distribution function of the residual lifetime of X, i.e., $F^r(x) = \frac{1}{E(X)} \int_0^x (1 - F(y))dy$, and by X^r a stochastic variable with distribution $F^r(.)$.
We focus on the class $\mathcal{R}$ of regularly varying distributions. $\mathcal{R}_{-\nu}$ contains all regularly varying functions of index $-\nu$. We already know that this class is a subset of the class of subexponential distributions.

In the remainder of 4.1, an overview of the tail asymptotics of the waiting-and/or sojourn time distributions in the M/G/1 queue for six key disciplines is given.

[2] Regularly varying functions are an example of subexponential distributions, so heavy-tailed.

4.1.1 The M/G/1 FCFS queue

Customers who come first are the first ones to get service (First-Come-First-Served). The next theorem characterizes the tail asymptotics of the distribution of the steady-state waiting time W for FCFS service discipline.

Theorem 1 *In the case of regular variation, i.e.,* $P\{B > .\} \in \mathcal{R}_{-\nu}$,

$$P\{W > x\} \sim \frac{\rho}{1-\rho} P\{B^r > x\}, \qquad x \to \infty. \tag{28}$$

4.1.2 The M/G/1 PS queue

The Processor Sharing (PS) service discipline operates as follows. If there are $n \geq 1$ customers present, then they are all served simultaneously, each at rate $\frac{1}{n}$.
In the M/G/1 PS queue the sojourn time tail is just as heavy as the service requirement tail as stated in the following theorem. S_{PS} denotes the steady-state sojourn time.

Theorem 2 *If* $P\{B > .\} \in \mathcal{R}_{-\nu}$,

$$P\{S_{PS} > x\} \sim P\{B > (1-\rho)x\}, \qquad x \to \infty. \tag{29}$$

4.1.3 The M/G/1 LCFS-PR queue

In the Last-Come-First-Served Preemptive-Resume (LCFS-PR) discipline, an arriving customer K is immediately taken into service. However, this service is interrupted when another customer arrives, it is only resumed when all customers who have arrived after K have left the system.
The fact that no customer has to wait for the completion of a residual service requirement, suggests that the tail of the sojourn time distribution is just as heavy as the tail of the service requirement distribution. S_{LPR} denotes the steady-state sojourn time.

Theorem 3 *If* $P\{B > .\} \in \mathcal{R}_{-\nu}$,

$$P\{S_{LPR} > x\} \sim \frac{1}{1-\rho} P\{B > (1-\rho)x\}, \qquad x \to \infty. \tag{30}$$

4.1.4 The M/G/1 LCFS-NP queue

Let W_{LNP} denote the steady-state waiting time in the M/G/1 Last-Come-First-Served Non-Preemptive (LCFS-NP) queue. The impossibility of preemption suggests that the tail of W_{LNP} will be determined by the tail of a residual service requirement. Indeed,

Theorem 4 *If* $P\{B > .\} \in \mathcal{R}_{-\nu}$,

$$P\{W_{LNP} > x\} \sim \rho P\{B^r > (1-\rho)x\}, \qquad x \to \infty. \tag{31}$$

4.1.5 The M/G/1 FBPS queue

The Foreground-Background Processor Sharing (FBPS) discipline allocates an equal share of the service capacity to the customers which so far have received the least amount of service. It is proven that the tail of the distribution of the sojourn time S_{FB} is the same as that for the ordinary PS discipline.

Theorem 5 *If* $P\{B > .\} \in \mathcal{R}_{-\nu}$ *with* $1 < \nu < 2$,

$$P\{S_{FB} > x\} \sim P\{B > (1-\rho)x\}, \qquad x \to \infty. \tag{32}$$

4.1.6 The M/G/1 SRPTF queue

With this service discipline (Shortest-Remaining-Processing-Time-First) the total service capacity is always allocated to the customer(s) with the shortest remaining processing time. The service of a customer is preempted when a new customer arrives with a service requirement smaller than the remaining service requirement of the customer being served. The service of the customer that is preempted is resumed as soon as there are no other customers with a smaller amount of work in the system. For the sojourn time S_{SR} there is the following theorem.

Theorem 6 *If* $P\{B > .\} \in \mathcal{R}_{-\nu}$ *with* $1 < \nu < 2$,

$$P\{S_{SR} > x\} \sim P\{B > (1-\rho)x\}, \qquad x \to \infty. \tag{33}$$

Note that the tail of the service requirement distribution behaves as those of the PS and FBPS disciplines.

4.2 Transform approach

There exists a very useful relation between the tail behavior of a regularly varying probability distribution and the behavior of its Laplace Stieltjes Transform (LST) near the origin. That relation often enables us to conclude from the form of the LST of the waiting- and/or sojourn time distribution, that the distribution itself is regularly varying at infinity. In the following lemma this relation is presented.

Lemma 1

Let $n < \nu < n+1$, $C \geq 0$. The following statements are equivalent:

$$\phi_n\{s\} = (C + o(1))s^{\nu} L(\frac{1}{s}), \qquad s \downarrow 0, s \in \mathbb{R}, \tag{34}$$

$$1 - F(x) = (C + o(1))\frac{(-1)^n}{\Gamma(1-\nu)} x^{-\nu} L(x), \qquad x \to \infty, \tag{35}$$

where $\phi_n\{s\} := (-1)^{n+1}[\phi\{s\} - \sum_{j=0}^{n} \mu_j \frac{(-s)^j}{j!}]$, and where $F(.)$ is the distribution of a non-negative random variable, with LST $\phi\{s\}$ and finite first n moments $\mu_1, ..., \mu_n$ (and $\mu_0 = 1$).

The theorems that were discussed in 4.1 can be proved by using Lemma 1. We will only look at the proof of Theorem 1. For the other proofs we refer to [3].

4.2.1 Proof of Theorem 1 using the transform approach

Firstly we repeat Theorem 1 which concerns the M/G/1 FCFS queue:

In the case of regular variation, i.e., $P\{B > .\} \in \mathcal{R}_{-\nu}$,

$$P\{W > x\} \sim \frac{\rho}{1-\rho} P\{B^r > x\}, \qquad x \to \infty. \tag{36}$$

In this queue, the LST of the steady-state waiting-time distribution is given by the Pollaczek-Khintchine formula:

$$E\{e^{-sW}\} = \frac{1-\rho}{1-\rho\beta^r\{s\}}, \qquad s \geq 0, \tag{37}$$

where $\beta^r\{s\} = \frac{1-\beta\{s\}}{\beta s}$ is the LST of the residual service requirement distribution $B^r(.)$. A Karamata theorem implies that if $P\{B > .\} \in \mathcal{R}_{-\nu}$, the integrated tail $P\{B^r > .\} \in \mathcal{R}_{1-\nu}$. More precisely, if

$$P\{B > x\} \sim x^{-\nu} L(x), \qquad \nu > 1, \qquad x \to \infty, \tag{38}$$

then

$$P\{B^r > x\} = \frac{1}{\beta} \int_x^\infty P\{B > y\} dy \sim \frac{1}{(\nu-1)\beta} x^{1-\nu} L(x), \qquad x \to \infty. \tag{39}$$

We now demonstrate how the following statement, which implies Theorem 1, is obtained from the LST expression (37) and Lemma 1. For $1 < \nu < 2, x \to \infty$,

$$P\{B > x\} \sim x^{-\nu} L(x) \Leftrightarrow P\{W > x\} \sim \frac{\rho}{1-\rho} \frac{1}{(\nu-1)\beta} x^{1-\nu} L(x). \tag{40}$$

It follows from (38) and Lemma 1 that

$$1 - \beta^r\{s\} = 1 - \frac{1-\beta\{s\}}{\beta s} = -(\frac{\Gamma(1-\nu)}{\beta} + o(1)) s^{\nu-1} L(1/s), \qquad s \downarrow 0. \tag{41}$$

Combining this result with (37) yields:

$$1 - E\{e^{-sW}\} = \frac{\rho(1-\beta^r\{s\})}{1-\rho\beta^r\{s\}} \sim -\frac{\rho}{1-\rho} \frac{\Gamma(1-\nu)}{\beta} s^{\nu-1} L(1/s), \qquad s \downarrow 0. \tag{42}$$

Another application of Lemma 1 gives the $\Rightarrow$ part of (40). The reverse part is obtained in a similar way.

4.3 Tail Equivalence via Conditional Moments

With heavy-tailed distributions, it is often the case that large occurrences of the variables of interest (e.g., a customer's waiting time or sojourn time) are essentially caused by a single large occurrence of one input variable (e.g., a service requirement). In this section a generic approach will be described that may be used to prove that the tails of the distributions of the causal variable and the resultant variable are equally heavy.

In the queues we already mentioned, a customer's own service requirement (denoted by B) or the residual service requirement of some other customer (denoted by B^r) will play the role of the *causal* variable X and the customer's sojourn time (denoted by S) that of the *resultant* Y. We use $S(\tau)$ to denote a customer's sojourn time given that the residual service time equals τ. Consequently we may write $S(B)$ or $S(B^r)$ for the unconditional sojourn time S.

Theorem 7 below relates the tails of the distributions of S and the causal variable X. Firstly we have to make two assumptions: one regarding the distribution of the causal variable and one regarding $S(\tau)$.

Assumption A

$$P\{X > .\} \in \mathcal{R}_{-\alpha}, \qquad \alpha > 0. \tag{43}$$

Assumption B

1. $E\{S(\tau)\} \sim \bar{g}\tau$, for some $\bar{g} > 0$;
2. With α as in Assumption A, there exists $\kappa > \alpha$ such that

$$P\{S(\tau) - E\{S(\tau)\} > t\} \leq \frac{h(\tau)}{t^{\kappa}} \tag{44}$$

   with $h(\tau) = o(\tau^{\kappa-\delta}), \tau \to \infty$, for some $\delta > 0$;
3. $S(\tau)$ is stochastically increasing in $\tau \geq 0$, i.e., for all $t \geq 0$, the probability $P\{S(\tau) > t\}$ is non-decreasing in $\tau \geq 0$.

We can now state Theorem 7:

Theorem 7 *Suppose Assumptions A and B are satisfied. Then the tails of the distributions of the random variables X and $S(X)$ are equally heavy in the sense that:*

$$P\{S(X) > \bar{g}x\} \sim P\{X > x\}. \tag{45}$$

In particular, the distribution of $S(X)$ is also regularly varying with the same index $-\alpha$ as that of X.

For the proof of this theorem we refer to [3].

Theorem 7 can be used to show for several queueing models that the tail of the sojourn time distribution is as heavy as that of the (residual) service requirement distribution. Assuming that the service requirement distribution is regularly varying, it suffices to verify that $S(\tau)$, the sojourn time conditioned on the (residual) service requirements, satisfies Assumption B.

We know that for the PS, FBPS, SRPTF and LCFS-PR queues the tail of the sojourn time distribution is just as heavy as the tail of the service requirement time distribution. So Theorem 2,3,5 and 6 can be proved using tail equivalence via conditional moments. For the proofs we also refer to [3].

4.4 Sample-Path Techniques

In this section it is described how sample-path techniques may be used to determine the tail asymptotics of the delay distribution in the M/G/1 queue for various disciplines.

In 3.5 we already sketched a heuristic derivation of the tail asymptotics of the workload D.

We look at the tail asymptotics of the delay distribution in the M/G/1 queue. The delay distribution strongly depends on the service discipline that is used. We now look at the FCFS queue; for the others we refer to [3].

4.4.1 Proof of Theorem 1 using sample-path techniques

For the FCFS, the waiting time is simply equal to the workload at the time of arrival. Because of PASTA, it then follows ,using the theorem of subsection 3.5, that

$$P\{W_{FCFS} > x\} \sim \frac{\rho}{1-\rho} P\{B^r > x\}, \tag{46}$$

which agrees with Theorem 1.
For a detailed and rigorous proof, one should derive a lower and upper bound for $P\{W_{FCFS} > x\}$ and show that they both approach $\frac{\rho}{1-\rho} P\{B^r > x\}$ in the limit (cf. [4]).

In this section we have looked at the tail behavior of the waiting- and/or sojourn time distributions for several M/G/1 queues. It turns out that, if the service time distribution is regularly varying with index $-\nu$, the waiting time distribution in FCFS and LCFS-NP is heavier. This is in contrast with the service disciplines like PS, FBPS, SRPTF and LCFS-PR. These disciplines all yield a sojourn-time tail of index $-\nu$.

References

[1] W. Willinger, M. S. Taqqu, W. E. Leland and D. V. Wilson, Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements, Statistical Science, Vol. 10, No. 1, 67-85, 1995.

[2] K. Sigman, A Primer on Heavy-tailed Distributions, Advances in Computational Mathematics, Queueing Systems 33, 1999.

[3] S.C. Borst, O.J. Boxma, R. Nunez-Queija, A.P. Zwart, The impact of the service discipline on delay asymptotics, Performance Evaluation, 2003.

[4] A.P. Zwart, Queueing Systems with Heavy Tails, PHD Thesis, Eindhoven University of Technology, 2001.

[5] O.J.Boxma, V.Dumas, Fluid queues with long-tailed activity period distributions, Computer Communications 21, 1509-1529, 1998.