

New Origins of Heavy Tails with Applications to Information Networks

Jian Tan

Department of Electrical Engineering
Columbia University, New York, NY 10027
jiantan@ee.columbia.edu

July 2009

Abstract

Since the first discoveries in early 1990s of the presence of heavy-tailed statistical characteristics of traffic streams in modern computer networks, there has been a large amount of research on documenting the empirical evidence of these phenomena in information networks, including the Ethernet traffic, VBR/MPEG video, file/Web document sizes, Web access patterns, Web graph, physical Internet topology, etc. These wide spread observations of heavy tails, in particular power laws, in the context of communication networks are not surprising since, starting from the early studies of Pareto in 1897 and later of Zipf in 1949, heavy tails have been repeatedly observed for over a hundred years in a variety of biological, technological and socioeconomic areas.

Similarly to the fact that the wide appearance of Gaussian/Normal distributions can be attributed to the generality of the central limit theorem, in this research, we investigate new universal laws that under general conditions could result in heavy tails. First, we propose the modulated branching processes to study the power laws that arise in the proportional growth systems. Under general polynomial Gärtner-Ellis conditions, we show that these classes of models result almost invariably in power laws. Informally, the interpretation of our main results suggests that alternating periods of growth and reduction, e.g., economic expansions and recessions, are primarily responsible for the appearance of power law distributions.

Second, we study the effects of retransmissions, which are at the core of all layers in modern communication network architectures. Maybe unexpectedly, we analytically show that, under quite general conditions, retransmission-based protocols over channels with failures may result in heavy-tailed delays and possibly zero throughput even if the distribution of packets (data units) is very concentrated, e.g., exponential or Gaussian. This phenomenon occurs irrespective of whether the cause of retransmissions is due to channel failures in the data link layer, collisions in ALOHA-type/CSMA protocols in the MAC layer, or the end-to-end acknowledgements in the transport layer. These theoretical findings are in agreement with empirical measurements which show that the utilization of the 802.11 protocol is only 40%, basically due to retransmissions. This work provides a new

explanation of the widely observed heavy-tailed delays in communication networks since channel failures and retransmissions are inherent part of modern network architectures.

Understanding the general mechanisms that cause heavy tails could provide guidance for creating novel networking algorithms that either avoid causing or utilize beneficially the heavy-tailed characteristics. Because of the extremely varying and possibly non-stationary nature of the heavy-tailed network environment, our design focuses on developing highly dynamic, adaptive (self-organizing), low-complexity and scalable algorithms. To this purpose, by using the insights developed from the analysis, we propose a new dynamic packet fragmentation algorithm that can reduce the effects of heavy tails. Based on the previously recorded periods of time when the channel was continuously good/available, our fragmentation algorithm divides the original packets into smaller fragments. For this algorithm, both our analysis and the simulation shows that the delay distribution has additional/higher moments relative to the scheme without fragmentation, which in particular ensures a positive throughput. In addition, the adaptivity of our algorithm is essential for the extremely dynamic/time varying wireless environment.

Similarly, the heavy-tailed properties can be beneficially exploited to improve algorithm designs. Along this direction, we design a new scheduling algorithm that is especially suitable for the heavy-tailed environment. This algorithm, termed comparison scheduling, can be shown in a certain sense to be close to the optimal. It is based on relative comparison of a newly arriving job to the sizes of the previous few arrivals. Thus, the adaptive and scalable nature of this comparison mechanism enables the scheduling algorithm to work well even when the traffic characteristics are nonstationary, highly variable and strongly correlated.

From a mathematical perspective, we develop and combine various techniques from probability theory that include exponential and subexponential sample path large deviations, branching processes and queueing theory during the course of our analysis. The new techniques developed in this thesis are likely to be useful in related areas of parallel computing, (average case) analysis of algorithms, probability and statistics, financial engineering and insurance risk theory. Since heavy-tailed phenomena are important for a broad range of disciplines, including socioeconomic (wealth distribution, stock prices, insurance, city population), complex biological networks (gene regulatory and protein-protein networks) and information technology, the new insights of this thesis might have a direct impact on these areas.

1 Introduction

Since the first discoveries in early 1990s [46] of the presence of heavy-tailed statistical characteristics of traffic streams in modern computer networks, there has been a large amount of research on documenting the empirical evidence of these phenomena, including Ethernet traffic [46], VBR video [22], long-tailed nature of file sizes [16], scene length distribution of MPEG video streams [27], etc. A rather comprehensive overview of the early work on the related phenomena in information networks can be found in [55]. More recently heavy-tailed distributions (power laws) have been found in the World Wide Web (Web) access patterns [17, 10], the Web graph [45], the physical Internet topology [18], etc.

These wide spread observations of heavy-tailed distributions, in particular power laws, in the context of communication networks are not surprising when put in perspective that these types of phenomena are found in a wide range of other domains, ranging from socioeconomic to biological and technological areas. Specifically, these types of distributions describe the city populations, species-area relationships, sizes of living organisms, value of companies, distributions of wealth, etc. Empirical observations of heavy tailed distributions and, in particular, power laws have a long history of over 100 years, starting from the discovery by Pareto [54] in 1897 that a plot of the logarithm of the number of incomes above a level against the logarithm of that level yields points close to a straight line, which is essentially equivalent to saying that the income distribution follows a power law. Hence, power law distributions are often called Pareto distributions. In a different context, early work by Arrhenius [4] in 1921 conjectured a power law relationship between the number of species and the census area, which was followed by Preston’s prediction in [58] that the slope on the log/log species-area plot has a canonical value equal to 0.262; for additional information and measurements on species-area relationships see [15, 57, 41]. Interestingly, there also exists a power law relationship between the rank of the cities and the population of the corresponding cities. This was proposed by Auerbach [7] in 1913 and later studied by Zipf [73] in 1949, after whom power law is also known as Zipf’s law. Similar observations have been made for firm sizes [2], and the gene family and protein statistics [28, 64].

Hence, these repeated empirical observations of heavy tails/power laws, over a period of more than a hundred years, strongly suggest that there exist general mathematical laws that govern these phenomena. Similarly to the fact that the wide appearance of Gaussian/Normal distributions can be attributed to the generality of the central limit theorem, in this paper, we investigate new universal laws, i.e., modulated branching processes and retransmissions over channel with failures, presented in Sections 2 and 3, respectively, that under very general conditions invariably result in heavy tails. Using the insight we obtained from the analysis, we discuss a new power law phenomena on finite population ALOHA with variable size packets in Section 4.

Now, going back to early 1990s, the empirical discoveries of heavy-tailed phenomena in communication networks were followed by the development of new models that were suitable for these new environments. Very early, it became clear that the traditional exponential/Poisson type models may not be suitable for these type of phenomena [56]. The primary reason is that network traffic characteristics exhibit an increased level of “burstiness” that spans over multiple time and space scales, which is often referred to as self-similarity. Typical approaches to capture this self-similar nature of network traffic include long range dependent Gaussian processes (e.g., see [22, 49]), multiple time scale (nearly decomposable) Markov models (e.g., see [31]) and heavy-tailed/subexponential fluid on-off/semi-Markov/(M/GI/ ∞) models (e.g., see [44]); for a rather complete list of references before 2000 see [55].

These models have motivated a line of analytical research that attempted to understand the performance of existing networking and queueing models, scheduling algorithms, etc, in the

presence of heavy-tailed traffic. However, creating new network architectures, protocols and algorithms that are specifically designed for operating in the heavy-tailed environment remains an area that is not fully exploited. In the second part of this paper, using the understanding of the general mechanisms that can cause heavy tails, we propose novel networking algorithms that either avoid causing or utilize beneficially the heavy-tailed characteristics. Because of the extremely varying and possibly non-stationary nature of the heavy-tailed network environment, our design focuses on developing highly dynamic, adaptive (self-organizing), low-complexity and scalable algorithms. To this purpose, we propose a new dynamic packet fragmentation algorithm in Section 5 that can reduce the effects of heavy tails. The adaptivity of this algorithm is essential for the extremely time varying wireless environment. In addition, to beneficially exploit the heavy-tailed properties, we design a new scheduling algorithm that is especially suitable for the heavy-tailed environment in Section 6. This scheduling mechanism works well even when the traffic characteristics are nonstationary, highly variable and strongly correlated.

2 Modulated branching processes and origins of power laws in proportional growth systems

As mentioned in the introduction, the repeated empirical observations of power laws have a long history over a period of more than a hundred years, which strongly suggest that there exist general mathematical laws that govern these phenomena. In this regard, after carefully examining the situations that result in power laws, we discover that most of them are characterized by the following three features. First, in the vast majority of these observations, e.g., city populations and sizes of living organisms, the objects of interest evolve due to the replication of their many independent components, e.g., birth-deaths of individuals and replications of cells. Secondly, the rate of replication of the many components is often controlled by exogenous parameters causing periods of baby booms and busts, economic growths and recessions, etc. Thirdly, the sizes of these objects often have lower boundaries, e.g., cities do not fall below a certain size, low income individuals are subsidized by the government, companies are protected by bankruptcy laws, etc.

In order to capture the preceding features, it is natural to propose *modulated branching processes* (MBP) with reflective or absorbing barriers as generic models for many of the observations of power laws. Indeed, one of our main results, presented in Theorem 2.2, shows that MBPs with reflective barriers almost invariably produce power law distributions under quite general *polynomial Gärtner-Ellis* conditions. The generality of our results could explain the ubiquitous nature of power law distributions. Furthermore, an informal interpretation of our main results, stated in Theorems 2.2 and 2.3 of Section 2.2, suggests that alternating periods of expansions and contractions, e.g., economic booms and recessions, are primarily responsible for the appearance of power law distributions. Actually, Theorem 2.3 shows that the distribution of the reflected MBP is exponentially bounded if the process has a tendency to contract.

Formal description of our reflected modulated branching process (RMBP) model is given in Section 2.1. In the singular case when the number of individuals born in each state of the modulating process is constant, our model reduces to a reflected multiplicative process. A rigorous connection (duality) between the reflected multiplicative processes (RMPs) and queueing theory was established in Section 5 of Goldie (1991) [24]; this duality was repeatedly observed and used later in, e.g., [69, 25]. In Subsection 2.1.1 we further emphasize this duality in the context of stationary and ergodic processes. We would like to point out that this duality makes a vast literature on queueing theory directly applicable to the analysis of RMPs. Informally, these results show that the role which exponential distributions play in queueing theory, and in additive reflected random walks in general, is represented by power law distributions in the framework of RMPs/RMBPs. For example, the power law distribution satisfies the *memoryless property* in the multiplicative world, playing an equivalent role to the memoryless exponential distribution in the additive world. Indeed, if $\mathbb{P}[M > x] = x^{-\alpha}$, $\alpha > 0$, $x \geq 1$, then, for $x, y \geq 1$, we obtain $\mathbb{P}[M > xy | M > x] = \mathbb{P}[M > y]$.

Furthermore, we would like to point out that the reflective nature of the barrier is not essential for producing power law distributions. Indeed, one only needs a positive lower barrier, e.g., porous, absorbing or reflective one, which is a natural condition since no physical object or socioeconomic one can approach zero arbitrarily close without repelling from it or simply disappearing. In many areas, objects of interest may not have a strictly reflecting barrier, but rather a porous one, e.g., cities may degenerate, bankruptcy protection may sometimes fail and a company can be liquidated. We discuss these situations, as well as some other related models including randomly stopped branching processes and truncated power laws in [40].

2.1 Reflected modulated branching processes

Let $\{J_n\}_{n > -\infty}$ be a stationary and ergodic modulating process that takes values in positive integers. Define a family of independent, non-negative, integer-valued random variables $\{B_n^i(j)\}$, $-\infty < i, j, n < \infty$, which are independent of the modulating process $\{J_n\}$. In addition, for fixed j , variables $\{B(j), B_n^i(j)\}$ are identically distributed with $\mu(j) \triangleq \mathbb{E}[B(j)] < \infty$.

Definition 2.1 A Modulated Branching Process (MBP) $\{Z_n\}_{n=0}^\infty$ is recursively defined by

$$Z_{n+1} \triangleq \sum_{i=1}^{Z_n} B_n^i(J_n), \quad (2.1)$$

where the initial value Z_0 is a positive integer. For increased clarity, we may explicitly write $\{Z_n^l\}$ when $Z_0 = l$.

Definition 2.2 For any $l \in \mathbb{N}$ and an integer valued Λ_0 , a Reflected Modulated Branching

Process (RMBP) $\{\Lambda_n\}_{n=0}^\infty$ is recursively defined as

$$\Lambda_{n+1} \triangleq \max \left(\sum_{i=1}^{\Lambda_n} B_n^i(J_n), l \right). \quad (2.2)$$

Remark 1 These types of modulated branching processes with a reflecting barrier appear to be new and, thus, the traditional methods for the analysis of branching processes [6] do not seem to directly apply.

Lemma 2.1 *Assume $\mathbb{E} \log \mu(J_0) < 0$, then, for any a.s. finite initial condition Λ_0 , Λ_n converges in distribution to*

$$\Lambda \stackrel{d}{=} \max_{n \geq 0} Z_{-n}.$$

2.1.1 Reflected multiplicative processes (RMP) and queueing duality

Note that in the special case $B_n^i(J_n) \equiv J_n$, reflected modulated branching processes reduce to reflected multiplicative processes with J_n being integer valued. In this subsection we assume that $\{J_n\}_{n \geq 0}$ is a positive, real valued process.

Definition 2.3 For $l > 0$ and $M_0 < \infty$, define a Reflected Multiplicative Process (RMP) as

$$M_{n+1} = \max(M_n \cdot J_n, l), \quad n \geq 0. \quad (2.3)$$

Without loss of generality we can assume $l = 1$. Now, let $X_n = \log J_n$ and $Q_n = \log M_n$ with the standard conventions $\log 0 = -\infty$ and $e^{-\infty} = 0$. Then, equation (2.3) is equivalent to

$$Q_{n+1} = \max(Q_n + X_n, 0), \quad (2.4)$$

which is the workload (waiting-time) recursion in a single server (FIFO) queue.

Lemma 2.2 *If $\mathbb{E} \log J_n < 0$, then M_n converges in distribution to an a.s. finite random variable M that satisfies*

$$M \stackrel{d}{=} \sup_{n \geq 0} \Pi_n, \quad (2.5)$$

where $\Pi_0 = 1$, $\Pi_n = \prod_{i=-n}^{-1} J_i$, $n \geq 1$.

The following theorem is a direct corollary of Theorem 1 in [23]; see also Theorem 3.8 in [14] and, for a more recent presentation, we refer the reader to [21].

Theorem 2.1 *Let $\{J_n\}_{n \geq 1}$ be a stationary and ergodic sequence of positive random variables. If there exists a function Ψ and positive constants α^* and ε^* such that*

- 1) $n^{-1} \log \mathbb{E}[(\Pi_n)^\alpha] \rightarrow \Psi(\alpha)$ as $n \rightarrow \infty$ for $|\alpha - \alpha^*| < \varepsilon^*$,
- 2) Ψ is finite and differentiable in a neighborhood of α^* with $\Psi(\alpha^*) = 0$, $\Psi'(\alpha^*) > 0$, and

3) $\mathbb{E} [(\Pi_n)^{\alpha^* + \varepsilon}] < \infty$, for $n \geq 1$ and some $\varepsilon > 0$,

then

$$\lim_{x \rightarrow \infty} \frac{\log \mathbb{P}[M > x]}{\log x} = -\alpha^*. \quad (2.6)$$

Remark 2 We refer to conditions 1) – 3) as the *polynomial Gärtner-Ellis conditions*.

2.2 Main results

This section presents our main results in Theorems 2.2 and 2.3. To avoid technical difficulties, we assume $\underline{\mu} \triangleq \inf_j \mu(j) > 0$. With a small abuse of notation, as compared to the preceding Subsection 2.1.1, we redefine here $\Pi_n = \prod_{i=-n}^{-1} \mu(J_i)$, $n \geq 1$, $\Pi_0 = l$ and $M = \sup_{n \geq 0} \Pi_n$.

Theorem 2.2 *Assume that the process $\{\Pi_n\}$ satisfies the polynomial Gärtner-Ellis conditions and $\sup_j \mathbb{E} [e^{\theta|B(j) - \mu(j)|}] < \infty$ for some $\theta > 0$, then,*

$$\lim_{x \rightarrow \infty} \frac{\log \mathbb{P}[\Lambda > x]}{\log x} = \lim_{x \rightarrow \infty} \frac{\log \mathbb{P}[M > x]}{\log x} = -\alpha^*. \quad (2.7)$$

Remark 3 Note that conditions 1) and 2) of Theorem 2.1 imply that there exists j such that $\mu(j) > 1$, since otherwise we have $\sup_\alpha \Psi(\alpha) \leq 0$, which would contradict $\Psi(\alpha^*) = 0$ and $\Psi'(\alpha^*) > 0$ in condition 2). The following theorem covers the opposite situation when the previous condition is not satisfied, i.e., $\sup_j \mu(j) < 1$.

Theorem 2.3 *If $\sup_j \mu(j) < 1$ and $\sup_j \mathbb{E} [e^{\theta|B(j) - \mu(j)|}] < \infty$ for some $\theta > 0$, then,*

$$\lim_{x \rightarrow \infty} \frac{\log \mathbb{P}[\Lambda > x]}{\log x} = -\infty. \quad (2.8)$$

Remark 4 Informally speaking, these two theorems show that the alternating periods of contractions and expansions, e.g., economic booms and recessions, are primarily responsible for the appearance of power law distributions; in other words, if there are no periods of expansions, i.e., the condition $\sup_j \mu(j) < 1$ of Theorem 2.3 is satisfied, then Λ has a tail that is lighter than any power law distribution.

2.3 Discussion of Related Models

Based on the study of reflected modulated branching processes, we discuss randomly stopped processes, modulated branching processes with absorbing barriers and truncated power laws in [40].

3 Retransmissions and induced heavy-tailed distributions

Retransmissions represent one of the most fundamental approaches in communication networks that guarantee data delivery in the presence of channel failures. These types of mechanisms

have been employed on all networking layers, including, for example, Automatic Repeat re-Quest (ARQ) protocol (e.g., see Section 2.4 of [11]) in the data link layer where a packet is resent automatically in case of an error; contention based ALOHA type protocols in the medium access control (MAC) layer that use random backoff and retransmission mechanism to recover data from collisions; end-to-end acknowledgement for multi-hop transmissions in the transport layer; HTTP downloading scheme in the application layer, etc.

We use the following generic channel with failures [39] to model the preceding situations. The channel dynamics is described as an on-off process $\{(A, U), (A_i, U_i)\}_{i \geq 1}$ with alternating periods when channel is available A_i and unavailable U_i , respectively; $(A, A_i)_{i \geq 1}$ and $(U, U_i)_{i \geq 1}$ are two independent sequences of i.i.d random variables. In each period of time that the channel becomes available, say A_i , we attempt to transmit the data unit of random size L . If $L \leq A_i$, we say that the transmission is successful; otherwise, we wait for the next period A_{i+1} when the channel is available and attempt to retransmit the data from the beginning.

It was first recognized in [20, 67] that this model results in power law distributions when the distributions of L and A have a matrix exponential representation. Under more general conditions, we discover in [39] that the distributions of N and T follow power laws with the same exponent α as long as $\log \mathbb{P}[L > x] \approx \alpha \log \mathbb{P}[A > x]$, which implies that power law distributions, possibly with infinite mean ($0 < \alpha < 1$) and variance ($0 < \alpha < 2$), may arise even when transmitting superexponential (e.g., Gaussian) documents/packets. In this paper, we further characterize this class of heavy-tailed distributions that are induced by retransmissions.

More precisely, we extend the results from [5, 39] under a more unified framework and study how the functional dependence between the data characteristics and channel dynamics in the form $(\mathbb{P}[L > x])^{-1} \approx \Phi(\mathbb{P}[A > x])^{-1}$ impacts the distribution of N , where the approximation \approx will be possibly differently defined according to the context. In the functional space of $\Phi(n)$, we identify several functional criticality points that define different classes of functional behavior of the distribution of N . Specifically, in Subsection 3.1.1, we show that if $\Phi(n)$ is dominantly varying then $\mathbb{P}[N > n] \approx \Phi(n)^{-1}$; see Proposition 3.3 and Theorem 3.2. The preceding tail equivalence between $\mathbb{P}[N > n]$ and $\Phi(n)^{-1}$ basically does not hold if $\Phi(x)$ is not dominantly varying. Furthermore, we show in a weaker form that if $\log(\Phi(n))$ is slowly varying, then $\log((\mathbb{P}[N > n])^{-1})$ is essentially slowly varying as well. Interestingly, if $\log(\Phi(n))$ grows slower than $e^{\sqrt{\log n}}$ then we have the asymptotic equivalence $\log(\mathbb{P}[N > n]) \approx -\log(\Phi(n))$ as shown in Theorem 3.3. However, if $\log(\Phi(n))$ grows faster than $e^{\sqrt{\log n}}$, this asymptotic equivalence does not hold and exhibits a different functional form.

Next, for lighter distributions of Weibull type, in Subsection 3.1.1, we show that if $\log(\Phi(n))$ is regularly varying with index $\beta > 0$, then basically one obtains Weibull distribution for N , i.e., $\log(\mathbb{P}[N > n]) \approx -(\log \Phi(n))^{1/(\beta+1)}$, as shown in Theorem 3.4, which we term moderately heavy (Weibull tail) asymptotics. Finally, in Subsection 3.1.1, we consider the situation when the separation between $\mathbb{P}[L > x]$ and $\mathbb{P}[A > x]$ is very large, i.e., their distributions are roughly separated by more than two exponential scales ($\log \log(\Phi(n)) \approx n^\gamma$). This separation

results in what we call the nearly exponential distribution for N in the form $\log(\mathbb{P}[N > n]) \approx -n/(\log n)^{1/\gamma}$.

After the preceding characterization of the different classes of distributional behavior for N , we study in Subsection 3.2 the total transmission time T . We use the large deviation results since T can be represented as the sum of L and $\{(A_i + U_i)\}_{1 \leq i < N}$. In this context, our primary results show that: (i) when $\Phi(\cdot)$ is regularly varying, we derive the exact asymptotics for T in Theorem 3.6. (ii) when $\log(\Phi(\cdot))$ is slowly varying, we obtain the logarithmic asymptotics for T in Theorem 3.7. (iii) when $\log(\Phi(\cdot))$ is regularly varying with positive index, we derive, in a different scale than in Theorem 3.7, the logarithmic asymptotics in Theorem 3.8. Interestingly, we want to point out that, unlike Theorems 3.6 and 3.7 requiring no conditions on A (Theorem 3.6 needs $\mathbb{E}[A] < \infty$), the minimum conditions needed for Theorem 3.8, as shown by Proposition 3.5, basically involve a balance between the tail decays of $\mathbb{P}[A > x]$ and $\mathbb{P}[L > x]$.

From a practical perspective, our results suggest that careful examination and possible redesign of retransmission based protocols in communication networks might be needed. This is especially the case for Ad Hoc and resource limited sensor networks, where frequent channel failures occur due to a variety of reasons, including signal fading, multipath effects, interference, contention with other nodes, obstructions, node mobility, and other changes in the environment [62]. In engineering applications, our main discovery is the matching between the statistical characteristics of the channel and transmitted data (packets). On the network application layer, most of us have been inconvenienced when the connections would brake while we are downloading a large file from the Internet. This issue has been already recognized in practice where software for downloading files was developed that would save the intermediate data (checkpoints) and resume the download from the point when the connection was broken. However, our results emphasize that, in the presence of frequently failing connections, the long delays may arise even when downloading relatively small documents. Hence, we argue that one might need to modify the application layer software, especially for the wireless environment, by introducing checkpoints even for small to moderate size documents. In this paper, we show that several well-known retransmission based protocols in different layers of networking architecture can lead to power law delays, e.g., ALOHA type protocols in MAC layer [37] (see Section 4) and end-to-end acknowledgements in transport layer [38]. These new findings suggest that special care should be taken when designing robust networking protocols, especially in the wireless environment where channel failures are frequent.

3.0.1 Description of the channel

Consider transmitting a generic data unit of random size L over a channel with failures. The channel dynamics is modeled as an on-off process $\{(A_i, U_i)\}_{i \geq 1}$ with alternating independent periods when channel is available A_i and unavailable U_i , respectively. In each period of time that the channel becomes available, say A_i , we attempt to transmit the data unit and, if $L \leq A_i$, we say that the transmission was successful; otherwise, we wait for the next period A_{i+1} when the channel is available and attempt to retransmit the data from the beginning.

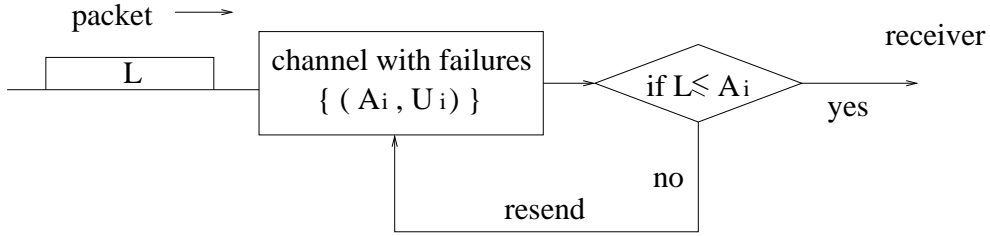


Figure 1: Packets sent over a channel with failures.

Assume that $\{U, U_i\}_{i \geq 1}$ and $\{A, A_i\}_{i \geq 1}$ are two mutually independent sequences of i.i.d. random variables.

Definition 3.1 The total number of (re)transmissions for a generic data unit of length L is defined as

$$N \triangleq \inf\{n : A_n \geq L\},$$

and, the total transmission time for the data unit is defined as $T \triangleq \sum_{i=1}^{N-1} (A_i + U_i) + L$.

The complementary cumulative distribution functions for A and L are denoted by $\bar{G}(x) \triangleq \mathbb{P}[A > x]$ and $\bar{F}(x) \triangleq \mathbb{P}[L > x]$.

We have identified in [39] that the transmission delay always has a power law tail if the hazard functions of A and L are asymptotically proportional. This result is quoted in the following theorem.

Theorem 3.1 *If there exists $\alpha > 0$, such that,*

$$\lim_{x \rightarrow \infty} \frac{\log \bar{F}(x)}{\log \bar{G}(x)} = \alpha, \quad (3.1)$$

then, we have

$$\lim_{n \rightarrow \infty} \frac{\log \mathbb{P}[N > n]}{\log n} = -\alpha.$$

The preceding result on power law distributions only covers a corner of the heavy-tailed distributions that can be induced by retransmissions. Indeed, when L has an infinite support, both N and T always decay slower than any exponential, as is shown below.

Proposition 3.1 *If $\bar{F}(x) > 0$ for all $x \geq 0$, then both N and T are subexponential in the following sense that, for any $\epsilon > 0$,*

$$e^{\epsilon n} \mathbb{P}[N > n] \rightarrow \infty \text{ as } n \rightarrow \infty \quad \text{and} \quad e^{\epsilon t} \mathbb{P}[T > t] \rightarrow \infty \text{ as } t \rightarrow \infty.$$

3.1 Main results

This section presents our main results. Here, we assume that $\bar{F}(x)$ is a continuous function with support on $[0, \infty)$. If $\bar{F}(x)$ is lattice valued, our results may still hold; see [39].

3.1.1 Asymptotics of the distribution of the number of retransmissions N

This subsection studies three scenarios: very heavy asymptotics (when $\log(\Phi(n))$ is slowly varying), medium heavy (Weibull) asymptotics (when $\log(\Phi(n))$ is regularly varying), and nearly exponential (when $\log \log(\Phi(n))$ is regularly varying), where within and between these subclasses we also identify critical functional points that define different distributional behavior of N .

Very heavy asymptotics

We term this subclass very heavy distributions since if $\log(\Phi(\cdot))$ is slowly varying, then the number of retransmissions N is always heavier than Weibull distribution, which is stated in the following Proposition 3.2.

Proposition 3.2 *If $\log(\Phi(\cdot))$ is slowly varying and*

$$\lim_{x \rightarrow \infty} \frac{\log(\bar{F}(x)^{-1})}{\log(\Phi(\bar{G}(x)^{-1}))} = 1, \quad (3.2)$$

then, for any $\epsilon > 0$, as $n \rightarrow \infty$,

$$\lim_{n \rightarrow \infty} \frac{\log(\mathbb{P}[N > n]^{-1})}{n^\epsilon} = 0.$$

In the remainder of this subsection we study the detailed structure of this class of distributions that have very heavy tails. The Weibull distribution will be studied in the next Subsection 3.1.1 on medium heavy asymptotics.

Proposition 3.3 *For $\Phi(\cdot)$ being dominantly regularly varying, i.e., it is non-decreasing in a neighborhood of infinity and $\overline{\lim}_{x \rightarrow \infty} \Phi(ex)/\Phi(x) < \infty$, if*

$$\bar{F}^{-1}(x) \sim \Phi(\bar{G}^{-1}(x)), \quad (3.3)$$

then, there is finite $c \geq 1$ such that

$$c^{-1} \leq \underline{\lim}_{n \rightarrow \infty} \mathbb{P}[N > n]\Phi(n) \leq \overline{\lim}_{n \rightarrow \infty} \mathbb{P}[N > n]\Phi(n) \leq c.$$

Proposition 3.4 *If (3.3) is satisfied and $\Phi(x)$ is eventually non-decreasing with*

$$\lim_{x \rightarrow \infty} \frac{\Phi(ex)}{\Phi(x)} = \infty,$$

then, $\lim_{n \rightarrow \infty} \mathbb{P}[N > n]\Phi(n) = \infty$.

When $\Phi(\cdot)$ is regularly varying, which is a subset of the dominantly regularly varying functions, we can compute the exact asymptotics of the distribution of N .

Theorem 3.2 Assuming $\bar{F}^{-1}(x) \sim \Phi(\bar{G}^{-1}(x))$ where $\Phi(\cdot)$ is regularly varying with index α , we obtain:

i) If $\alpha > 0$, then, as $n \rightarrow \infty$,

$$\mathbb{P}[N > n] \sim \frac{\Gamma(\alpha + 1)}{\Phi(n)}.$$

ii) If $\alpha = 0$ (meaning $\Phi(\cdot)$ is slowly varying) and $\Phi(x)$ is eventually non-decreasing, then, as $n \rightarrow \infty$,

$$\mathbb{P}[N > n] \sim \frac{1}{\Phi(n)}.$$

The condition of $\Phi(\cdot)$ being dominantly varying is basically necessary in order for $\mathbb{P}[N > n] \approx \Phi(n)^{-1}$ to hold. In the following theorem we extend the preceding logarithmic limit under a more general condition on $\Phi(\cdot)$.

Theorem 3.3 If an eventually non-decreasing function $\Phi(x) \triangleq e^{l(x)}$ satisfies (3.2) where $l(x)$ is slowly varying with

$$\lim_{x \rightarrow \infty} \frac{l\left(\frac{x}{l(x)}\right)}{l(x)} = 1, \quad (3.4)$$

then,

$$\lim_{n \rightarrow \infty} \frac{\log(\mathbb{P}[N > n]^{-1})}{\log \Phi(n)} = 1. \quad (3.5)$$

Medium heavy (Weibull) asymptotics

Now, we increase the separation such that $\Phi(x) = e^{R_\beta(x)}$ with $R_\beta(x)$ being regularly varying of index $\beta > 0$, and show that the distribution of N is of Weibull type.

Theorem 3.4 If an eventually non-decreasing function $\Phi(x) \triangleq e^{R_\beta(x)}$ satisfies (3.2) where $R_\beta(x) \equiv x^\beta l(x)$, $\beta > 0$ is regularly varying with $l(x)$ satisfying

$$\lim_{x \rightarrow \infty} \frac{l\left(\left(\frac{x}{l(x)}\right)^{\frac{1}{1+\beta}}\right)}{l(x)} = 1, \quad (3.6)$$

then,

$$\lim_{n \rightarrow \infty} \frac{\log(\mathbb{P}[N > n]^{-1})}{(\log \Phi(n))^{\frac{1}{\beta+1}}} = \beta^{\frac{1}{\beta+1}} + \beta^{-\frac{\beta}{\beta+1}}. \quad (3.7)$$

Nearly exponential asymptotics

Next, we investigate the situation when the separation $\Phi(x)$ is even larger than $e^{R_\beta(x)}$, which leads to the nearly exponential asymptotics for $\mathbb{P}[N > n]$.

Theorem 3.5 *If $\log(\bar{F}^{-1}(x)) \sim e^{R_\gamma(\bar{G}^{-1}(x))}$, where $R_\gamma(\cdot)$ is regularly varying with index $\gamma > 0$, then,*

$$\log \mathbb{P}[N > n]^{-1} \sim \frac{n}{R_\gamma^{\leftarrow}(\log n)}, \quad (3.8)$$

where $R_\gamma^{\leftarrow}(\cdot)$ is the asymptotic inverse of $R_\gamma(\cdot)$ as defined in Theorem 1.5.12 on p. 28 of [12].

Remark 5 In principle, one could study the situations when $\Phi(\cdot)$ grows faster than three exponential scales, which would make the distributions of N even closer to the exponential one. However, from a practical point of view, these cases will basically be indistinguishable from the exponential distribution and, thus, we omit these derivations.

3.2 Asymptotics of the total transmission time T

In this subsection, we compute the asymptotics of the total transmission time T based on the previous results on $\mathbb{P}[N > n]$. Our proving technique involves the relationship between N and T and the classical large deviation results. Theorem 3.6 and Theorem 3.7 characterize the exact asymptotics and logarithmic asymptotics for the very heavy case, respectively, and Theorem 3.8 derives the result for the moderate heavy (Weibull) case. Interestingly, we want to point out that, unlike Theorems 3.6 and 3.7 requiring no conditions on A (Theorem 3.6 needs $\mathbb{E}[A] < \infty$), the minimum conditions needed for Theorem 3.8, as shown by Proposition 3.5, basically involve a balance between the tail decays of $\mathbb{P}[A > x]$ and $\mathbb{P}[L > x]$.

Theorem 3.6 *If $\mathbb{E}[U^{(\alpha \vee 1) + \theta}] < \infty$, $\mathbb{E}[A^{1 + \theta}] < \infty$ and $\mathbb{E}[L^{\alpha + \theta}] < \infty$ for some $\theta > 0$, then, under the same conditions as in Theorem 3.2 i), we obtain, as $t \rightarrow \infty$,*

$$\mathbb{P}[T > t] \sim \frac{\Gamma(\alpha + 1)(\mathbb{E}[U + A])^\alpha}{\Phi(t)}.$$

Theorem 3.7 *Under the conditions of Theorem 3.3, if $\mathbb{P}[L > x] = O(\Phi(x)^{-(\delta+1)})$ and $\mathbb{P}[U > x] = O(\Phi(x)^{-(\delta+1)})$, $\delta > 0$, then, we obtain*

$$\lim_{t \rightarrow \infty} \frac{\log(\mathbb{P}[T > t]^{-1})}{\log(\Phi(t))} = 1.$$

Theorem 3.8 *Under the conditions of Theorem 3.4, if $\mathbb{P}[U > x] = O(e^{-(\log \Phi(x))^{(1+\delta)/(\beta+1)}})$ for some $\delta > 0$, $\mathbb{E}[A] < \infty$, and $\mathbb{P}[L > x] = O(e^{-x^\xi})$, $\mathbb{P}[A > x] = O(e^{-x^\zeta})$ with $\xi > \beta/(\beta + 1)$, $\zeta \geq 0$ satisfying $(1 - \zeta)\beta < \xi$, then, we obtain*

$$\lim_{t \rightarrow \infty} \frac{\log(\mathbb{P}[T > t]^{-1})}{(\log \Phi(t))^{\frac{1}{\beta+1}}} = \frac{\beta^{\frac{1}{\beta+1}} + \beta^{-\frac{\beta}{\beta+1}}}{(\mathbb{E}[A + U])^{\frac{\beta}{\beta+1}}}. \quad (3.9)$$

Basically, the condition $(1 - \zeta)\beta < \xi$ (or equivalently $\xi/(\xi + 1 - \zeta) > \beta/(\beta + 1)$) is needed since the following proposition shows that $\mathbb{P}[T > t]$ could have a heavier tail than predicted by (3.9) if $(1 - \zeta)\beta > \xi$.

Proposition 3.5 *If $\mathbb{P}[L > x] = e^{-x^\xi}$ and $\mathbb{P}[A > x] = e^{-x^\zeta}$ with $0 < \xi, \zeta < 1$, then, as $t \rightarrow \infty$,*

$$\mathbb{P}[T > t] \gtrsim e^{-2t^{\xi/(\xi+1-\zeta)}}.$$

4 Stability of finite population ALOHA with variable packets

ALOHA represents one of the first and most basic distributed Medium Access Control (MAC) protocols [1]. It is easy to implement since it does not require any user coordination or complicated controls and, thus, represents a basis for many modern MAC protocols, e.g., Carrier Sense Multiple Access (CSMA). The desirable properties of ALOHA, including its low complexity and distributed/asynchronous nature, make it a basis for many more sophisticated MAC protocols, e.g., CSMA.

Traditionally, the performance evaluation of ALOHA has focused on mean value (throughput) analysis, the examples of which can be found in every standard textbook on networking, e.g., see [48] and the references therein. However, it appears that there are no explicit and general studies (more than two users) of the distributional properties of ALOHA, e.g., delay distributions. In this regard, we consider a standard finite population ALOHA model with variable length packets [19] that have an asymptotically exponential tail. Surprisingly, we discover a new phenomenon that the distribution of the number of retransmissions (collisions) and time between two successful transmissions follow power law distributions, as stated in Theorem 4.1 on starting behavior and Theorem 4.2 on steady state behavior. Based on this observation, we derive new stability conditions for finite population ALOHA with variable packets in Theorem 4.3. Informally, our theorem shows that when the exponential decay rate of the packet distribution is smaller than the parameter of the exponential backoff distribution and the arrival rate, even the finite population ALOHA may have zero throughput. This is contrary to the common belief that the finite population ALOHA system always has a positive, albeit possibly small, throughput.

4.1 Model description

Consider $M \geq 2$ users sharing a common communication link (channel) of unit capacity. Each user can hold at most one packet in its queue and, when the queue is empty, a new packet is generated after an independent (from all other variables) exponential time with mean $1/\lambda$. Each packet has an independent length that is equal in distribution to a generic random variable L . A user with a newly generated packet attempts its transmission immediately and, if there are no other users transmitting during the same time, the packet is considered successfully transmitted. Otherwise, if the transmissions of more than one packet overlap, we say that there is a collision and the colliding packets need to be retransmitted; for a visual representation of the system see Figure 2. After a collision, each participating user waits (backoffs) for an independent exponential period of time with mean $1/\nu$ and then attempts to retransmit its packet. Each such user continues this procedure until its packet is successfully

transmitted and then it generates a new packet after an independent exponential time of mean $1/\lambda$.

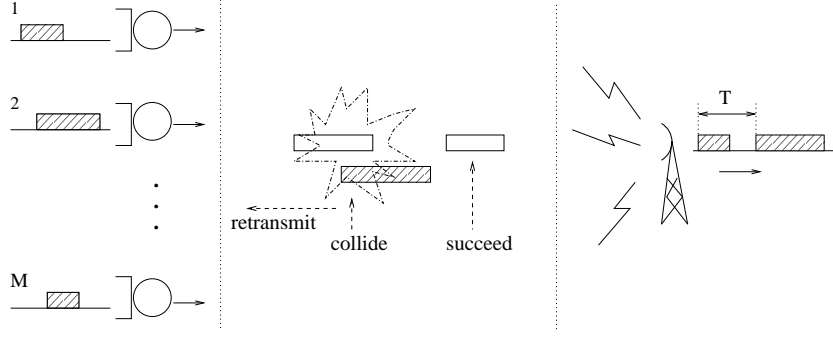


Figure 2: Finite population ALOHA model with variable size packets.

Let N_m be the number of (re)transmissions and T_m be the transmission time for the m th successfully received packet, respectively.

4.2 Main results

First, we study the starting behavior of our ALOHA model when the system starts from an empty state.

Theorem 4.1 *Assume that, for $\mu > 0$,*

$$\lim_{x \rightarrow \infty} \frac{\log \mathbb{P}[L > x]}{x} = -\mu, \quad (4.1)$$

and that at time $t = 0$ the system is empty, then, for any fixed $m \geq M$, the number of transmissions N_m and the transmission time T_m satisfy

$$\lim_{n \rightarrow \infty} \frac{\mathbb{P}[N_m > n]}{\log n} = \lim_{t \rightarrow \infty} \frac{\mathbb{P}[T_m > t]}{\log t} = -\frac{M\mu}{(M-1)\nu}.$$

When the system keeps running for a long period of time, we can show that the distributions of N_m and T_m are different from the preceding results. Let $\overline{\lim}$ denote both $\overline{\lim}$ and $\underline{\lim}$.

Theorem 4.2 *Assume that $\lambda = \nu > \mu/(M-1)$ and*

$$\overline{\lim}_{y \rightarrow \infty} \sup_{\delta y < x < y} \frac{1}{y-x} \log \left(\frac{\mathbb{P}[L > x]}{\mathbb{P}[L > y]} \right) \leq \mu \quad (4.2)$$

for $0 < \delta < 1$. If $\lambda = \nu > \mu/(M-1)$, we obtain

$$\lim_{n \rightarrow \infty} \overline{\lim}_{m \rightarrow \infty} \frac{\mathbb{P}[N_m > n]}{\log n} = \lim_{t \rightarrow \infty} \overline{\lim}_{m \rightarrow \infty} \frac{\mathbb{P}[T_m > t]}{\log t} = -\frac{\mu}{(M-1)\nu}. \quad (4.3)$$

4.3 Stability

We derive the stability condition of finite population ALOHA with variable packets in Theorem 4.3.

Theorem 4.3 *Under condition (4.1), if $\mu > (M - 1)\nu$, the ALOHA system has a positive throughput. Conversely, if $\lambda \geq \nu > 0$ and $\mu < (M - 1)\nu$, then, the system has a zero throughput.*

5 Dynamic packet fragmentation for wireless channels with failures

We show in Section 3 that retransmission based protocols can lead to power law delays and possibly zero throughput. Conventionally, in order to improve the network performance, the transmitting data unit is fragmented into shorter packets and sent out separately. The fragments are either reassembled at the next node, that is called intra-network fragmentation; or are reassembled at the destination, that is called inter-network fragmentation [68]. In general, we prefer to send packets as large as possible, rather than to use many small ones because much of the communication cost is counted per-packet instead of per-byte [42]. This desire for large packets, however, is restricted by the channel characteristics since, for example, large packets may cause many retransmissions and large delays that deteriorate the throughput performance [39]. Essentially this tradeoff between the communication cost/overhead and the channel utility/throughput involves an optimization problem over the constraints that are tightly related to the channel dynamics; the related work in this framework can be found in [53, 47].

In this section, motivated by the results from Section 3 and the preceding discussion of the time varying nature of the wireless environment, we study a dynamic packet fragmentation mechanism using the same channel model as in Section 3, except that upon the arrival of a packet of size L , we fragment it into several smaller packets of suitable sizes L_f and transmit them instead.

The algorithm that dynamically determines L_f is parameterized by two integers (k, m) and a constant $c > 0$; the algorithm keeps $k + m$ records of the lengths of the previous availability periods $\{A_i : -(k + m) \leq i \leq -1\}$. Upon the arrival of a data unit L , we set the maximum of c and the k th largest record \tilde{A}_k among $\{A_i : -(k + m) \leq i \leq -1\}$ to be the maximum length of the fragmented packet size L_m . If $L \leq L_m$, we do not fragment the data unit and send it out directly. If $L > L_m$, we fragment the data unit into $\lceil L/L_m \rceil$ packets with the last one possibly being shorter than L_m and all the other $\lceil L/L_m \rceil - 1$ ones being of equal size L_m .

The number of (re)transmissions for a packet of length L_f is defined as $N_f \triangleq \inf\{n : A_n \geq L_f\}$ and the transmission time for this packet is defined as $T_f \triangleq \sum_{i=1}^{N_f} (A_i + U_i)$. Thus, by taking the summation over the number of (re)transmissions of all the fragments, we obtain the total number of (re)transmissions \hat{N}_f and the corresponding total transmission time \hat{T}_f for

transmitting the data unit of length L

$$\hat{T}_f \triangleq \sum_{i=1}^{\hat{N}_f} (A_i + U_i). \quad (5.1)$$

Note that only for purposes of the analysis, we have assumed that all transmissions occur at the beginnings of available channel periods to avoid the possible technical difficulties caused by the lack of memoryless property. However, in actual implementation, one will be sending a new packet out immediately after the successful transmission of the previous packet.

5.1 Main results

Our algorithm consists of two components: measuring channel availability periods and fragmenting packets. In this section, to simplify the analysis, we decouple the channel measurements from packet fragmentation. In other words, we first measure channel availability periods after a successful transmission of the previous packet, and then fragment a new packet based on these measurements. Therefore, the $k + m$ records of lengths of the previous availability periods $\{A_i : -(k + m) \leq i \leq -1\}$ are i.i.d. random variables. In practice, this would be very inefficient and, instead, one could perform these two component functions concurrently and run the fragmentation algorithm in a more compact way.

Theorem 5.1 *Under the condition (3.1) and $\mathbb{E}[L^\beta] < \infty$ for all $\beta > 0$, we have*

$$\lim_{n \rightarrow \infty} \frac{\log \mathbb{P}[\hat{N}_f > n]}{\log n} = -(\alpha + k). \quad (5.2)$$

Furthermore, if $\mathbb{E}[(U + A)^{k+\alpha+\theta}] < \infty$ for some $\theta > 0$, then,

$$\lim_{t \rightarrow \infty} \frac{\log \mathbb{P}[\hat{T}_f > t]}{\log t} = -(\alpha + k). \quad (5.3)$$

Remark 6 As shown in Section 3, without using the fragmentation technique we would have

$$\lim_{n \rightarrow \infty} \frac{\log \mathbb{P}[N > n]}{\log n} = -\alpha,$$

where N is the number of (re)transmissions. Furthermore, Theorem 5.1 indicates that, even with our dynamic fragmentation, the distribution tails of the number of transmissions and total transmission time follow power laws. However, we can guarantee the increment of k additional moments to both N and T beyond α , and thus, put the tail behaviors under a desired control. Also, the parameter m can be tuned to yield a good throughput without deteriorating the tail performance.

6 Comparison scheduling algorithms for the heavy-tailed environment

It has been widely recognized that heavy-tailed distributions are suitable for modeling job sizes in information service networks, e.g., see [34] and the references therein. For heavy-tailed distributions, large jobs appear much more frequently than for the light-tailed ones, which imposes very different constraints in terms of optimizing the scheduling process as compared to the light-tailed scenarios. In particular, schedulers that may assign the server exclusively to a very large job, e.g., first come first serve (FIFO) discipline, can cause very large delays and, in general, suboptimal performance [3].

Hence, most of the practical schedulers utilize either the processor sharing (PS) and foreground background processor sharing (FBPS) disciplines because of their inherent fairness, or the shortest remaining processing time first (SRPT) discipline because of its known optimality under quite general conditions. In particular, it was shown in [66] that SRPT minimizes the number of customers in the G/G/1 queue over all work-conserving disciplines. For a recent survey on the performance of these disciplines in the context of heavy tails see [13].

It is well known that the sojourn time distributions under PS, FBPS and SRPT scheduling disciplines are asymptotically equivalent for power law distributions. This was originally proved in [50] and then later studied for regularly varying distributions in [13]; see also [36]. In other words, for large jobs, the waiting time does not depend on the choice of a specific scheduling discipline among PS, FBPS and SRPT.

In order to distinguish the performance of PS/FBPS and SRPT schedulers, we introduce a new notion of conditional waiting time distribution in [29, 30] that allows us to refine the result for medium size jobs. More formally, we show that even the relatively smaller jobs receive asymptotically the same residual capacity $1 - \rho$ as the larger ones for SRPT discipline, while, for PS/FBPS schedulers, these smaller jobs share the residual capacity equally with the larger jobs in the system. Hence, it appears that SRPT provides better and more uniform performance over a wide range of time scales.

However, as discussed in one of the very first papers on SRPT [65], this discipline may be quite difficult to implement. Clearly, its complicated preemptive nature requires keeping track of the remaining processing times for all jobs in the queue which may be prohibitive for systems with large job volumes, e.g., Web servers. In addition, it was shown in [65] that the expected number of preemptions per job is proportional to the load of the system, which can be quite large. Hence, even as early as 1966, it was recognized in [65] that one should try to approximate SRPT with less complex schedulers. The most apparent option, as suggested in [65], is to design a threshold-based static priority approximation to SRPT. Basically, the idea is to select a fixed number of thresholds m and then group jobs into $m + 1$ classes depending on which pair of thresholds a job size happens to fall between. Then, these classes are served according to the static priority discipline with higher priorities assigned to classes with smaller jobs. Since then, there has been a lot of work on threshold-based scheduling policies. For

example, it was shown in [8] that even with a single threshold, one can obtain the performance comparable to SRPT up to a constant factor in terms of the mean sojourn time for M/M/1 queue as well as for M/G/1 queue with finite variance Pareto service distribution.

Although it is encouraging that one can achieve a provably very good approximation of M/G/1/SRPT queue even with a very small number of static thresholds (only one in [8]), these solutions are likely not to perform well in practice since the traffic characteristics are often nonstationary, highly correlated (long range dependent) and very bursty (e.g., batch arrivals, etc); see [55, 70]. In order to overcome these difficulties, we propose a novel adaptive job classification (grouping) mechanism that is based on relative size comparison of a newly arriving job to the previous m arrivals; this scheduler is inspired by our conditional limit results. Specifically, if an arriving job is smaller than k and larger than $m - k$ of the previous m jobs, it is routed into class k . We also discuss refinements of the comparison grouping mechanism that improve the accuracy of the classification for both light-tailed and correlated job arrivals at the expense of a small (fixed) additional complexity in [30].

6.1 Heavy-tailed limits for medium size jobs with popular schedulers

6.1.1 Definitions and preliminary results

In this section we introduce the necessary notation and describe the existing and preliminary results. Let B_i and V_i denote the job size and the waiting time of the customer arriving at time T_i , respectively, where $\{B_i\}_{i>-\infty}$ are i.i.d. random variables. The arrival points $\{T_i\}_{i>-\infty}$ are assumed to be Poisson with rate λ and independent of job requirements $\{B_i\}_{i>-\infty}$. Hence, without loss of generality, in view of the PASTA property, we set $T_0 = 0$. The waiting time of a customer is defined as the amount of time between its arrival and departure, also referred to as sojourn time in the queuing literature. To present our main results, we need the following definitions.

Definition 6.1 A nonnegative random variable X or its distribution function (d.f.) F is called intermediately regularly varying, $X \in \mathcal{IR}$, if

$$\lim_{\eta \uparrow 1} \overline{\lim}_{x \rightarrow \infty} \frac{\mathbb{P}[X > \eta x]}{\mathbb{P}[X > x]} = 1.$$

Regularly varying distributions \mathcal{R}_α are the best-known examples from \mathcal{IR} . F is called regularly varying with index α , $X \in \mathcal{R}_\alpha$ ($F \in \mathcal{R}_\alpha$), if $F(x) = 1 - l(x)/x^\alpha$, $\alpha \geq 0$, where $l(x): \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is slowly varying, i.e., $\lim_{x \rightarrow \infty} l(\eta x)/l(x) = 1$, $\eta > 1$. We call F heavy-tailed ($X \in \mathcal{L}$ or $F \in \mathcal{L}$) if, for any fixed $y \in \mathbb{R}$, $\lim_{x \rightarrow \infty} \frac{\mathbb{P}[X > x - y]}{\mathbb{P}[X > y]} = 1$.

Let $\tilde{B}_i, 1 \leq i \leq m$ be the order statistics of $B_{-i}, 1 \leq i \leq m$ with the convention $\tilde{B}_0 = \infty$ and $\tilde{B}_{m+1} = 0$. To make the notation uniform, we assume that $\tilde{B}_0 = \infty, \tilde{B}_1 = 0$ for $m = 0$, and when it is necessary to emphasize the total number of random variables, we write explicitly $\tilde{B}_i^{(m)} \equiv \tilde{B}_i$.

Definition 6.2 Let $\mathcal{A}_k^{(m)} \triangleq \{\tilde{B}_{k+1}^{(m)} \leq B_0 < \tilde{B}_k^{(m)}\}$ for $m \geq k \geq 0$.

The asymptotic behavior of the sojourn time distribution for PS, FBPS and SRPT has been extensively studied under heavy-tails, e.g., see [74, 50, 13, 35, 36] and the references therein. We summarize these results for intermediately regularly varying distributions in the following theorem. In order to ease the notation we simply write $B \equiv B_0$ and $V \equiv V_0$.

For the rest of this section, we assume that the system has reached stationarity.

Theorem 6.1 *If $B \in \mathcal{IR}$ and $\mathbb{E}B^\alpha < \infty$ for some $\alpha > 1$, then, under the PS, FBPS or SRPT discipline, we have, as $x \rightarrow \infty$,*

$$\mathbb{P}[V > x] \sim \mathbb{P}[B > (1 - \rho)x].$$

6.1.2 Conditional limits

Using events $\mathcal{A}_k^{(m)}$, we can differentiate the performance of SRPT from PS and FBPS, as shown in the following theorem.

Theorem 6.2 *If $B \in \mathcal{IR}$ and $\mathbb{E}B^\alpha < \infty$ for some $\alpha > 1$, then, under either PS or FBPS discipline, we have for fixed k , as $x \rightarrow \infty$,*

$$\mathbb{P}\left[V > x, \mathcal{A}_k^{(m)}\right] \sim \mathbb{P}\left[B > \frac{(1 - \rho)x}{(1 + k)}, \mathcal{A}_k^{(m)}\right] \sim \frac{1}{k + 1} \binom{m}{k} \mathbb{P}\left[B > \frac{(1 - \rho)x}{k + 1}\right]^{k+1},$$

and under the SRPT discipline,

$$\mathbb{P}\left[V > x, \mathcal{A}_k^{(m)}\right] \sim \mathbb{P}\left[B > (1 - \rho)x, \mathcal{A}_k^{(m)}\right] \sim \frac{1}{k + 1} \binom{m}{k} \mathbb{P}[B > (1 - \rho)x]^{k+1}.$$

Remark 7 Note that $\mathcal{A}_k^{(m)}$ partitions the probability space into jobs of decreasing sizes as k increases. Interestingly, the result shows that, for the SRPT discipline, even the relatively much smaller job receives the entire long-term residual capacity $1 - \rho$, while, for PS/FBPS, this smaller job shares equally the residual capacity with the k larger ones. Hence, SRPT outperforms PS/FBPS for medium size jobs.

6.2 Adaptive and scalable comparison scheduling

Motivated by the preceding conditional limits presented in Section 6.1, we propose a novel adaptive and scalable comparison scheduling scheme.

6.2.1 Comparison splitting

The algorithm is based on relative size comparison of the arriving job to the previous m arrivals, $m \geq 1$. Specifically, if an arriving job is smaller than k and larger than $m - k$ of the previous m jobs, it is routed into class k , $0 \leq k \leq m$.

More formally, upon the arrival of job $i \geq 0$, we define $\tilde{B}_{i1} \geq \tilde{B}_{i2} \geq \dots \tilde{B}_{im}$ to be the order statistics of $\{B_{i-m}, B_{i-m+1}, \dots, B_{i-1}\}$ with $\tilde{B}_{i0} = \infty$ and $\tilde{B}_{i(m+1)} = 0$. Then, if $\tilde{B}_{i(k+1)} \leq B_i < \tilde{B}_{ik}$, the new arrival B_i is routed to class k , $0 \leq k \leq m$ and the i th arrival in class k is denoted as $B_i^{(k)}$. In order to initiate the comparison splitting process, assume that $B_i, -m \leq i \leq -1$ are already known; otherwise, one can simply set $B_i \equiv 0, -m \leq i \leq -1$.

Now, we argue that our comparison splitting actually does order jobs into classes that contain smaller jobs for larger class indexes. Indeed, when $B \in \mathcal{L}$, we obtain

$$\mathbb{P} \left[B_1^{(k)} > x \right] \sim \frac{1}{k+1} \binom{m}{k} \mathbb{P}[B > x]^{k+1} \quad \text{as } x \rightarrow \infty,$$

which implies a decreasing distribution tail when k increases.

From the description of the comparison splitter, we can see that its adaptive thresholds are determined by the order statistics of the previous m arrivals. Thus, it is reasonable to expect that, at least for a stationary input, the accuracy of the classification will increase if we obtain these thresholds using a longer history (than the preceding m arrivals). However, the increase of history may reduce the adaptability and add to the complexity of the algorithm. We discuss one refined splitting algorithm in [30].

6.2.2 Queueing analysis

Assume that jobs arrive according to a stationary renewal process $\{T_n\}$, $T_{-1} < 0 \leq T_0$ with finite mean $\mathbb{E}[T] < \infty$, where $T \stackrel{d}{=} T_1 - T_0$. The job sizes $\{B_n\}$ before the splitting are i.i.d and independent of $\{T_n\}$. To simplify the notation and analysis in this section, we say that the i th arrival to class k is equal to $B_i^{(k)} = B_i \mathbf{1}\{\tilde{B}_{i(k+1)} \leq B_i < \tilde{B}_{ik}\}$.

Furthermore, we assume that there is only one server with capacity c and that the $m+1$ classes are served jointly with a preemptive static priority discipline between classes. Suppose that the priorities of the classes are assigned in a decreasing order of the class index k , $0 \leq k \leq m$, i.e., class k receives service only if classes $i, k+1 \leq i \leq m$ are empty. Denote by $W_0^{(k)}$ the stationary workload of class k observed at arrival point T_0 . Let $\mu^{(k)} \triangleq \sum_{i=k}^m \mathbb{E}[B^{(i)}]$ and note that $\mu^{(0)} = \mathbb{E}[B]$.

Theorem 6.3 *If $\mathbb{P}[B > x] = l(x)/x^\alpha \in \mathcal{R}_\alpha$, $\alpha > 1$ and $\mathbb{E}[B] < c\mathbb{E}[T]$, then, as $x \rightarrow \infty$,*

$$\begin{aligned} \mathbb{P} \left[W_0^{(k)} > x \right] &\sim \frac{1}{c\mathbb{E}[T] - \mu^{(k)}} \int_x^\infty \mathbb{P} \left[B^{(k)} > u \right] du \\ &\sim \frac{1}{(k+1)(c\mathbb{E}[T] - \mu^{(k)})} \binom{m}{k} \frac{l(x)^{k+1}}{x^{\alpha k + \alpha - 1}}. \end{aligned}$$

7 Concluding remarks and further extensions

This paper investigates the origins of heavy-tailed distributions and design of new adaptive algorithms for the heavy-tailed environment. The presented materials can be organized in three general topics:

1. Origins of heavy-tailed (e.g., power law) distributions;
2. Redesign of network protocols to prevent the cause (or reduce the effects) of heavy tails;
3. Design of new adaptive (self-organizing) scheduling algorithms suitable for the heavy-tailed environment.

From a mathematical perspective, we develop and combine various techniques from probability theory that include exponential and subexponential sample path large deviations, branching processes and queueing theory during the course of our analysis. The new techniques developed in this thesis are likely to be useful in related areas of parallel computing, (average case) analysis of algorithms, probability and statistics, financial engineering and insurance risk theory. Since heavy-tailed phenomena are important for a broad range of disciplines, including socioeconomic (wealth distribution, stock prices, insurance, city population), complex biological networks (gene regulatory and protein-protein networks) and information technology, the new insights of this thesis might have a direct impact on these areas.

References

- [1] N. Abramson. The Aloha system - another alternative for computer communications. In *Proceedings of the Fall Joint Computer Conference*, pages 281–285, 1970.
- [2] L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, H. E. Stanley, and M. H. Stanley. Scaling behavior in economics: Empirical results for company growth. *Journal de Physique I*, 7:621–633, 1997.
- [3] V. Anantharam. Scheduling strategies and long-range dependence. *Queueing systems*, 33:73–89, 1999.
- [4] O. Arrhenius. *Journal of Ecology*, 9(95-99), 1921.
- [5] S. Asmussen, P. Fiorini, L. Lipsky, T. Rolski, and R. Sheahan. Asymptotic behavior of total times for jobs that must start over if a failure occurs. *Mathematics of Operations Research*, 33(4):932–944, November 2008.
- [6] K. Athreya and A. N. Vidyashankar. Branching processes. *Springer-Verlag*, 1972.
- [7] F. Auerbach. Das gesetz der belvolkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59:74–76, 1913.
- [8] N. Bansal and D. Gamarnik. Handling load with less stress. *Queueing Syst. Theory Appl.*, 54(1):45–54, 2006.
- [9] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: investigating unfairness. In *SIGMETRICS/Performance*, pages 279–290, 2001.
- [10] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns. *World Wide Web J.*, 2:15–28, 1999.
- [11] D. P. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, 2 edition, 1992.
- [12] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27. Cambridge University Press, 1987.
- [13] S. C. Borst, O. J. Boxma, R. Núñez-Queija, and A. P. Zwart. The impact of the service discipline on delay asymptotics. *Performance Evaluation*, 54(2):175–206, 2003.

- [14] C.-S. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.
- [15] E. F. Connor and E. D. McCoy. The statistics and biology of the species-area relationship. *The American Naturalist*, 113(6):791–833, Jun., 1979.
- [16] M. Crovella. The relationship between heavy-tailed file sizes and self-similar network traffic. In *9th INFORMS Applied Probability Conference*, Cambridge, Massachusetts, June 1997.
- [17] C. Cunha, A. Bestavros, and M. Crovella. Characteristics of World Wide Web client-based traces. *Technical Report TR-95010, Boston University*, pages 126–134, 1995.
- [18] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. In *Proceedings of the ACM SIGCOMM 1999 Conference*, pages 251–262, 1999.
- [19] M. J. Ferguson. An approximate analysis of delay for fixed and variable length packets in an unslotted ALOHA channel. *IEEE Transactions on Communications*, 25:644–654, July 1977.
- [20] P. M. Fiorini, R. Sheahan, and L. Lipsky. On unreliable computing systems when heavy-tails appear as a result of the recovery procedure. In *MAMA 2005 Workshop*, Banff, AB, Canada, June 2005; *ACM SIGMETRICS Performance Evaluation Review*, Volume 33, Issue 2, pp. 15–17, September 2005.
- [21] A. Ganesh, N. O’Connell, and D. Wischik. *Big Queues*. Springer-Verlag, 2004.
- [22] M. W. Garrett and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proceedings of ACM SIGCOMM*, pages 269–280, 1994.
- [23] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Appl. Prob.*, 31:131–156, 1994.
- [24] C. M. Goldie. Implicite renewal theory and tails of solutions of random equations. *The Annals of Applied Probability*, 1(1):126–166, February 1991.
- [25] W.-B. Gong, Y. Liu, V. Misra, and D. Towsley. Self-similarity and long range dependence on the Internet: a second look at the evidence, origins and implications. *Computer Networks*, 48:377–399, June 2005.
- [26] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems (TOCS)*, 21(2):207–233, 2003.
- [27] D. P. Heyman and T. V. Lakshman. Source models for VBR broadcast-video traffic. *IEEE/ACM Transactions on Networking*, 4(1):40–48, 1996.
- [28] M. A. Huynen and E. van Nimwegen. The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution*, 15(5):583–9, May 1998.
- [29] P. R. Jelenković, X. Kang, and J. Tan. Adaptive and scalable comparison scheduling. In *Proceedings of ACM SIGMETRICS’07*, volume 35, No.1, pages 215–226, San Diego, CA, USA, June 2007.
- [30] P. R. Jelenkovic, X. Kang, and J. Tan. Heavy-tailed limits for medium size jobs and comparison scheduling. *Annals of Operations Research*, Special Issue on Stochastic Performance Models for Resource Allocation in Communication Systems, 2008, to appear.
- [31] P. R. Jelenković and A. Lazar. On the dependence of the queue tail distribution on multiple time scales of ATM multiplexers. In *Proceedings of the Conference on Information Sciences and Systems*, pages 435–440, Baltimore, MD, March 1995.
- [32] P. R. Jelenković and A. A. Lazar. Evaluating the queue length distribution of an ATM multiplexer with multiple time scale arrivals. In *Proceedings IEEE INFOCOM*, San Francisco, CA, March 1996.
- [33] P. R. Jelenković, A. A. Lazar, and N. Semret. Multiple time scales and subexponentiality in MPEG video streams. In *International IFIP-IEEE Conference on Broadband Communications*, April 1996.

- [34] P. R. Jelenković and P. Momčilović. Capacity regions for network multiplexers with heavy-tailed fluid on-off sources. In *INFOCOM'01*, Anchorage, AK, April 2001.
- [35] P. R. Jelenković and P. Momčilović. Resource sharing with subexponential distributions. In *Proceedings of IEEE INFOCOM*, volume 3, pages 1316–1325. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, June 2002.
- [36] P. R. Jelenković and P. Momčilović. Large deviation analysis of subexponential waiting times in a processor-sharing queue. *Mathematics of Operational Research*, 28(3):587–608, August 2003.
- [37] P. R. Jelenković and J. Tan. Is ALOHA causing power law delays? In *Proceedings of the 20th International Teletraffic Congress*, Ottawa, Canada, June 2007; *Lecture Notes in Computer Science*, No 4516, pp. 1149–1160, Springer-Verlag, 2007. (Best Student Paper Award).
- [38] P. R. Jelenković and J. Tan. Are end-to-end acknowledgements causing power law delays in large multi-hop networks? In *14th Inform Applied Probability Conference*, Eindhoven, July 9–11 2007.
- [39] P. R. Jelenković and J. Tan. Can retransmissions of superexponential documents cause subexponential delays? In *Proceedings of IEEE INFOCOM'07*, pages 892–900, Anchorage, Alaska, USA, (submitted August 2006, accepted November 2006), May 2007.
- [40] P. R. Jelenković and J. Tan. Modulated branching processes, origins of power laws and queueing duality. Technical Report EE2007-09-25, Department of Electrical Engineering, Columbia University, New York, NY, September 2007. Eprint arXiv:0709.4297v1.
- [41] J. E. Keeley. Relating species abundance distributions to species-area curves in two mediterranean-type shrublands. *Diversity and Distributions*, (9):253–259, 2003.
- [42] C. A. Kent and J. C. Mogul. Fragmentation considered harmful. In *Proceedings of ACM SIGCOMM*, August 1987.
- [43] L. Kleinrock. *Queueing Systems, vol. II*. John Wiley and Sons, 1976.
- [44] M. Krunz and A. M. Makowski. A source model for VBR video traffic based on M/G/ ∞ input processes. In *Proceedings IEEE INFOCOM*, San Francisco, California, Apr. 1998.
- [45] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57–65, 2000.
- [46] W. E. Leland, W. Willinger, M. S. Taqqu, and D. V. Wilson. On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Computer Communication Review*, 25(1):202–213, 1995.
- [47] E. Modiano. An adaptive algorithm for optimizing the packet size used in wireless ARQ protocols. *Wireless Networks*, 5(4):279–286, 1999.
- [48] V. Naware, G. Mergen, and L. Tong. Stability and delay of finite user slotted ALOHA with multipacket reception. 51(7):2636–2656, July 2005.
- [49] I. Norros. A storage model with self-similar input. *Queueing Systems: Theory and Applications*, 16:387–396, 1994.
- [50] R. Núñez-Queija. *Processor-Sharing Models for Integrated-Service Networks*. PhD thesis, Eindhoven University of Technology, 2000.
- [51] M. Nuyens, A. Wierman, and B. Zwart. Preventing large sojourn times using smart scheduling. *Preprint*, 2007.
- [52] M. Nuyens and B. Zwart. A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing systems: Theory and Applications*, 54(2):85–97, 2006.
- [53] A. Orda and R. Rom. Optimal packet fragmentation and routing in computer networks. *Networks*, 29(1):11–28, December 1998.
- [54] V. Pareto. *Cours d'Economie Politique*, 2, Pichou, Paris, 1897.

- [55] K. Park and W. Willinger, editors. *Self-similar Network Traffic and Performance Evaluation*. Wiley, New York, 2000.
- [56] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1993.
- [57] J. B. Plotkin, M. D. Potts, D. W. Yu, S. Bunyavechewin, R. Condit, R. Foster, S. Hubbell, J. LaFrankie, N. Manokaran, H.-S. Lee, R. Sukumar, M. A. Nowak, and P. S. Ashton. Predicting species diversity in tropical forests. *PNAS*, 97(20):10850–10854, September 26, 2000.
- [58] F. Preston. The canonical distribution of commonness and rarity. *Ecology*, (43):185–215, 410–430, 1962.
- [59] I. A. Rai, E. W. Biersack, and G. Urvoy-Keller. Size-based scheduling to improve the performance of short TCP flows. *IEEE Network*, 19(1):12–17, January/February 2005.
- [60] I. A. Rai, G. Urvoy-Keller, M. K. Vernon, and E. W. Biersack. Performance analysis of LAS-based scheduling disciplines in a packet switched network. In *SIGMETRICS/Performance '04*, pages 106–117, New York, NY, USA, 2004.
- [61] K. Ramanan and A. L. Stolyar. Largest weighted delay first scheduling: Large deviations and optimality. *Annals of Applied Probability*, 11(1):1–48, February 2001.
- [62] T. S. Rappaport. *Wireless communications: principles and practise*. Prentice Hall, 2 edition, January 2002.
- [63] M. Rawat and A. Kshemkalyani. SWIFT: Scheduling in web servers for fast response time. In *Proceedings of the Second IEEE International Symposium on Network Computing and Applications*, page 15, Los Alamitos, CA, USA, April 2003.
- [64] A. Rzhetsky and S. M. Gomez. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, 17:988–996, 2001.
- [65] L. E. Schrage and L. W. Miller. The queue M/G/1 with the shortest sharing queue. *Operations Research*, 14:670–684, 1966.
- [66] P. J. Schweitzer. Perturbation theory and finite markov chains. *Journal of Applied Probability*, 5:401–413, 1968.
- [67] R. Sheahan, L. Lipsky, P. M. Fiorini, and S. Asmussen. On the completion time distribution for tasks that must restart from the beginning if a failure occurs. In *MAMA 2006 Workshop*, Saint-Malo, France, June 2006; *ACM SIGMETRICS Performance Evaluation Review*, Volume 34, Issue 3, pp. 24-26, December 2006.
- [68] J. Shoch. Packet fragmentation in Inter-network protocols. *Computer networks*, 3(1):3–8, 1979.
- [69] D. Sornette and R. Cont. Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *Journal de Physique I*, 7:3:431–444, March 1997.
- [70] M. Squillante, D. Yao, and L. Zhang. Web traffic modeling and web server performance analysis. In *Proceedings of the 38th Conference on Decision and Control*, pages 4432–4437, Phoenix, AZ, 1999.
- [71] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. In *ACM SIGMETRICS'03*, pages 238–249, San Diego, CA, USA, 2003.
- [72] R. W. Wolff. *Stochastic Modeling and Theory of Queues*. Prentice Hall, 1989.
- [73] G. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA, 1949.
- [74] A. Zwart and O. Boxma. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems: Theory and Applications*, 35(1/4):141–166, 2000.