1 Self-Similar Network Traffic: An Overview

Kihong Park

Network Systems Lab, Department of Computer Sciences, Purdue University, West Lafayette, IN 47907

Walter Willinger

Information Sciences Research Center, AT&T Labs-Research, Florham Park, NJ 07932

1.1 INTRODUCTION

1.1.1 Background

Since the seminal study of Leland, Taqqu, Willinger and Wilson [41] which set the groundwork for considering self-similarity an important notion in the understanding of network traffic including the modeling and analysis of network performance, an explosion of work has ensued investigating the multifaceted nature of this phenomenon.[1] The long held paradigm in the communication and performance communities has been that voice traffic and, by extension, data traffic are adequately described by certain Markovian models (e.g., Poisson) which are amenable to accurate analysis and efficient control. The first property stems from the well-developed field of Markovian analysis which allows tight equilibrium bounds on performance variables such as the waiting time in various queueing systems to be found. This also

[1] For a non-technical account of the discovery of the self-similar nature of network traffic, including parallel efforts and important follow-up work, we refer the reader to [71].

forms a pillar of performance analysis from the queueing theory side [38]. The second feature is, in part, due to the simple correlation structure generated by Markovian sources whose performance impact—e.g., as affected by the likelihood of prolonged occurrence of "bad events" such as concentrated packet arrivals—is fundamentally well-behaved. Specifically, if such processes are appropriately rescaled in time, the resulting coarsified processes rapidly lose dependence, taking on the properties of an independent and identically distributed (i.i.d.) sequence of random variables with its associated niceties. Principal among them is the exponential smallness of rare events, a key observation at the center of large deviations theory [70].

The behavior of a process under rescaling is an important consideration in performance analysis and control since buffering and, to some extent, bandwidth provisioning can be viewed as operating on the rescaled process. The fact that Markovian systems admit to this avenue of taming variability has helped shape the optimism permeating the late 1980s and early 1990s regarding the feasibility of achieving efficient traffic control for quality of service (QoS) provisioning. The discovery and, more importantly, succinct formulation and recognition that data traffic may not exhibit the hereto accustomed scaling properties [41] has significantly influenced the networking landscape, necessitating a reexamination of some of its fundamental premises.

1.1.2 What is self-similarity?

Self-similarity and fractals are notions pioneered by Benoit B. Mandelbrot [47]. They describe the phenomenon where a certain property of an object—e.g., a natural image, the convergent subdomain of certain dynamical systems, a time series (the mathematical object of our interest)—is preserved with respect to scaling in space and/or time. If an object is self-similar or fractal, its parts, when magnified, resemble—in a suitable sense—the shape of the whole. For example, the 2-dimensional Cantor set living on $A = [0, 1] \times [0, 1]$ is obtained by starting with a solid or black unit square, scaling its size by 1/3, then placing four copies of the scaled solid square at the four corners of A. If the same process of scaling followed by translation is applied recursively to the resulting objects ad infinitum, the limit set thus reached defines the 2-D Cantor set. This constructive process is illustrated in Figure 1.1.1. The limiting object—defined as the infinite intersection of the iterates—has the property that if any of its corners are "blown up" suitably, then the shape of the zoomed-in part is similar to the shape of the whole, i.e., it is *self-similar*. Of course, this is not too surprising since the constructive process—by its recursive action—endows the limiting object with the scale-invariance property.

The 1-dimensional Cantor set, e.g., as obtained by projecting the 2-D Cantor set

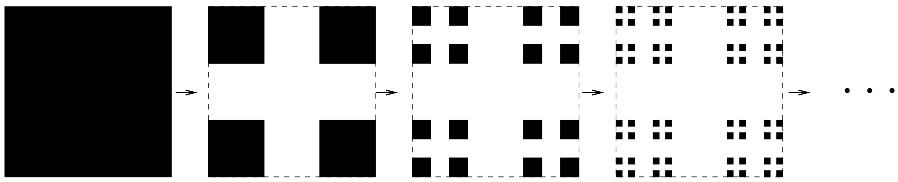

Fig. 1.1.1 2-dimensional Cantor set.

onto the line, can be given an interpretation as a traffic series $X(t) \in \{0, 1\}$—call it "Cantor traffic"—where $X(t) = 1$ means that there is a packet transmission at time t. This is depicted in Figure 1.1.2 (left). If the constructive process is terminated at iteration $n \geq 0$, then the contiguous line segments of length $1/3^n$ may be interpreted as *on-periods* or packet trains of duration $1/3^n$, and the segments between successive on-periods as *off-periods* or absence of traffic activity. Nonuniform traffic intensities may be imparted by generalizing the constructive framework via the use of probability measures. For example, for the 1-dimensional Cantor set, instead of letting the left and right components after scaling have identical "mass," they may be assigned different mass, subject to the constraint that the total mass be preserved at each stage of the iterative construction. This modification corresponds to defining a probability measure μ on the Borel subsets of $[0, 1]$ and distributing the measure at each iteration nonuniformly left and right. Note that the classical Cantor set construction—viewed as a map—is not measure-preserving. Figure 1.1.2 (middle) shows such a construction with weights $\alpha_L = 2/3$, $\alpha_R = 1/3$ for the left and right components, respectively. The probability measure is represented by "height"; we observe that scale-invariance is exactly preserved. In general, the traffic patterns producible with fixed weights α_L, α_R are limited, but one can extend the framework by allowing possibly different weights associated with every edge in the weighted binary tree induced by the 1-dimensional Cantor set construction. Such constructions arise in a more refined characterization of network traffic—called multiplicative processes or cascades—and are discussed in Chapter 20. Further generalizations can be obtained by defining different affine transformations with variable scale factors and translations at every level in the "traffic tree." The corresponding traffic pattern is self-similar if, and only if, the infinite tree can be compactly represented as a finite directed cyclic graph [8].

Whereas the previous constructions are given interpretations as traffic activity *per unit time*, we will find it useful to consider their corresponding *cumulative* processes which are nondecreasing processes whose differences—also called increment process—constitute the original process. For example, for the on/off

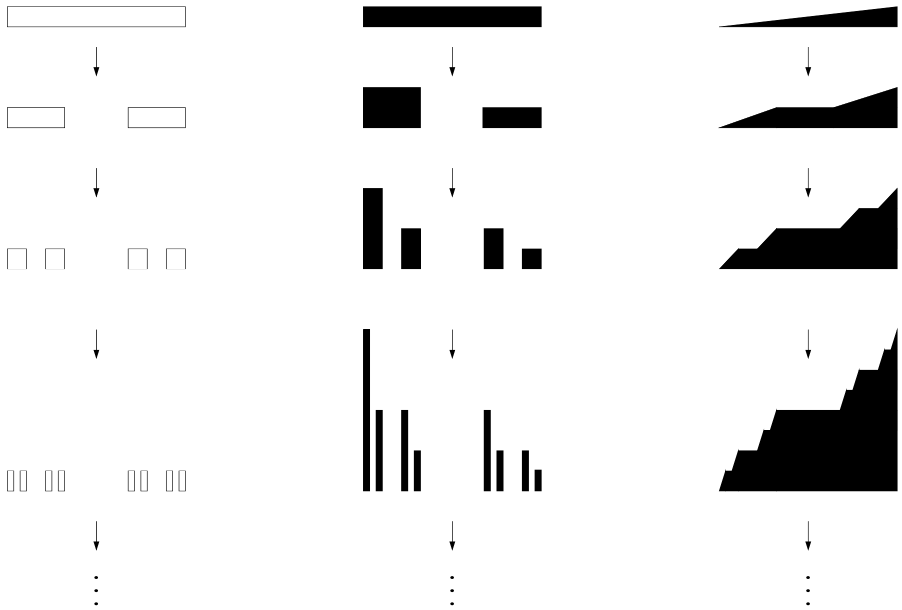

Fig. 1.1.2 Left: 1-dimensional Cantor set interpreted as on/off traffic. Middle: 1-dimensional nonuniform Cantor set with weights $\alpha_L = 2/3$, $\alpha_R = 1/3$. Right: Cumulative process corresponding to 1-dimensional on/off Cantor traffic.

Cantor traffic construction (cf. Figure 1.1.2 (left)), let us assign the interpretation that time is discrete such that at step $n \geq 0$, it ranges over the values $t = 0, 1/3^n, 2/3^n, \ldots, (3^n - 1)/3^n, 1$. Thus we can equivalently index the discrete time steps by $i = 0, 1, 2, \ldots, 3^n$. With a slight abuse of notation, let us redefine $X(\cdot)$ as $X(i) = 1$ if, and only if, in the original process $X(i/3^n) = 1$ and $X(i/3^n - \varepsilon) = 1$ for all $0 < \varepsilon < 1/3^n$. That is, for i values for which an on-period in the original process $X(t)$ begins at $t = i/3^n$, $X(i)$ is defined to be zero. Thus, in the case of $n = 2$, we have

$$X(0) = 0,\ X(1) = 1,\ X(2) = 0,\ X(3) = 1,\ X(4) = 0,$$
$$X(5) = 0,\ X(6) = 0,\ X(7) = 1,\ X(8) = 0,\ X(9) = 1.$$

Now, consider the continuous time process $Y(t)$ shown in Figure 1.1.2 (right) defined over $[0, 3^n]$ for iteration n. $Y(t)$ is nondecreasing, continuous, and it can be checked by visual inspection that

$$X(i) = Y(i) - Y(i-1), \qquad i = 1, 2, \ldots, 3^n,$$

and $X(0) = Y(0) = 0$. Thus $Y(t)$ represents the total traffic volume *up to* time t whereas $X(i)$ represents the traffic intensity during the i'th interval. Most importantly, we observe that exact self-similarity is preserved even in the cumulative process. This points toward the fact that self-similarity may be defined with respect to a cumulative process with its increment process—which is of more relevance for traffic modeling—"inheriting" some of its properties including self-similarity.

An important drawback of our constructions thus far is that they admit only a strong form of recursive regularity—that of *deterministic* self-similarity—and needs to be further generalized for traffic modeling purposes where stochastic variability is an essential component.

1.1.3 Stochastic self-similarity and network traffic

Stochastic self-similarity admits the infusion of nondeterminism as necessitated by measured traffic traces but, nonetheless, is a property that can be illustrated visually. Figure 1.1.3 (top-left) shows a traffic trace, where we plot throughput, in bytes, against time where time granularity is 100s. That is, a single data point is the aggregated traffic volume over a 100 second interval. Figure 1.1.3 (top-right) is the same traffic series whose first 1000 second interval is "blown up" by a factor of ten. Thus the truncated time series has a time granularity of 10s. The remaining two plots zoom in further on the initial segment by rescaling successively by factors of 10.

Unlike deterministic fractals, the objects corresponding to Figure 1.1.3 do not possess exact resemblance of their parts with the whole at finer details. Here, we assume that the measure of "resemblance" is the shape of a graph with the magnitude suitably normalized. Indeed, for measured traffic traces, it would be too much to expect to observe exact, deterministic self-similarity given the stochastic nature of many network events (e.g., source arrival behavior) that collectively influence actual network traffic. If we adopt the view that traffic series are sample paths of stochastic processes and relax the measure of resemblance, say, by focusing on certain statistics of the rescaled time series, then it may be possible to expect exact similarity of the mathematical objects and approximate similarity of their specific realizations with respect to these relaxed measures. Second-order statistics are statistical properties that capture burstiness or variability, and the autocorrelation function is a yardstick with respect to which scale-invariance can be fruitfully defined. The shape of the autocorrelation function—above and beyond its preservation across rescaled time series—will play an important role. In particular, correlation, as a function of time lag, is assumed to decrease polynomially as opposed to exponentially. The existence of nontrivial correlation "at a distance" is referred to as *long-range dependence*. A

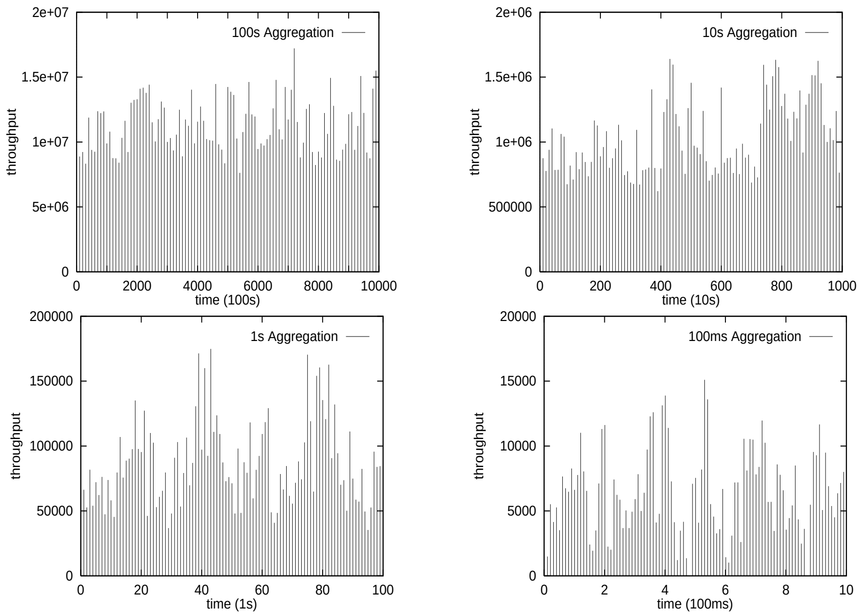

Fig. 1.1.3 Stochastic self-similarity—in the "burstiness preservation sense"—across time scales 100s, 10s, 1s, 100ms (top-left, top-right, bottom-left, bottom-right).

formal definition is given in Section 1.4.1.

1.2 PREVIOUS RESEARCH

1.2.1 Measurement-based traffic modeling

The research avenues relating to traffic self-similarity may be broadly classified into four categories. In the first category are works pertaining to *measurement-based traffic modeling* [13, 26, 34, 42, 56, 74] where traffic traces from physical networks are collected and analyzed to detect, identify as well as quantify pertinent characteristics. They have shown that scale-invariant burstiness or self-similarity is an ubiquitous phenomenon found in diverse contexts, from local area and wide area networks to IP and ATM protocol stacks to copper and fiber optic transmission media. In particular, [41] demonstrated self-similarity in a LAN environment (Ethernet), [56] showed self-similar burstiness manifesting itself in pre-World Wide Web WAN IP traffic, and [13] showed self-similarity for WWW traffic. Collectively, these measurement works constituted strong evidence that scale-invariant burstiness was not an isolated, spurious phenomenon but rather a persistent trait existing across a range of network environments.

Accompanying the traffic characterization efforts has been works in the area of statistical and scientific inference that has been essential to the detection and quantification of self-similarity or long-range dependence[2]. This work has been specifically geared toward network traffic self-similarity [28, 64], and has focused on exploiting the immense volume, high quality, and diversity of available traffic measurements; for a detailed discussion of these and related issues, see [72, 73]. At a formal level, the validity of an inference or estimation technique is tied to an underlying process that presumably generated the data in the first place. Put differently, correctness of system identification only holds when the data or sample paths are known to originate from specific models. Thus, in general, a sample path of unknown origin cannot be uniquely attributed to a specific model, and the main (and only) purpose of statistical or scientific inference is to deal with this intrinsically ill-posed problem by concluding whether or not the given data or sample paths are consistent with an assumed model structure. Clearly, being consistent with an assumed model does not rule out the existence of other models that may conform to the data equally well. In this sense, the aforementioned works on measurement-based traffic modeling have demonstrated that self-similarity is consistent with measured network traffic, and have resulted in adding yet another class of models—i.e., self-similar processes—to an already long list of models for network traffic. At a practical level, many of the commonly-used inference techniques for quantifying the degree of self-similarity or long-range dependence (e.g., Hurst parameter estimation) have been known to exhibit different idiosyncrasies and robustness properties. Due to their predominantly heuristic nature, these techniques have been generally easy to use and apply, but the ensuing results have often been difficult to interpret [64]. The recent introduction of wavelet-based techniques to the analysis of traffic traces [1, 23] represented a significant step toward the development of more accurate inference techniques that have been shown to possess increased sensitivity to different types of scaling phenomena with the ability to discriminate against certain alternative modeling assumptions, in particular, nonstationary effects [1]. Due to their ability to localize a given signal in scale and time, wavelets have made it possible to detect, identify, and describe *multifractal* scaling behavior in measured network traffic over fine time scales [23]: a nonuniform (in time) scaling behavior that emerges when

[2]The relationship between self-similarity and long-range dependence—they need not be one and the same—is explicated in Section 1.4.1.

studying measured TCP traffic over fine time scales, one that allows for more general scaling phenomena than the ubiquitous self-similar scaling property which holds for a range of sufficiently large time scales.

1.2.2 Physical modeling

In the second category are works on *physical modeling* that try to explicate the physical causes of self-similarity in network traffic based on network mechanisms and empirically established properties of distributed systems that, collectively, collude to induce self-similar burstiness at multiplexing points in the network layer. In view of traditional time series analysis, physical modeling affects model selection by picking among competing and—in a statistical sense—equally well fitting models those that are most congruent to the physical networking environment where the data arose in the first place. Put differently, physical modeling aims for models of network traffic that relate to the physics of how traffic is generated in an actual network, is capable of explaining empirically observed phenomena such as self-similarity in more elementary terms, and provides new insights into the dynamic nature of the traffic. The first type of causality—also the most mundane—is attributable to the arrival pattern of a single data source as exemplified by variable bit rate (VBR) video [10, 26]. MPEG video, for example, exhibits variability at multiple time scales which, in turn, is hypothesized to be related to the variability found in the time duration between successive scene changes [25]. This "single-source causality," however, is peripheral to our discussions for two reasons: one, self-similarity observed in the original Bellcore data stems from traffic measurements collected during 1989–1991, a period during which VBR video payload was minimal—if not nonexistent—to be considered an influencing factor[3], and two, it is well-known that VBR video can be approximated by short-range dependent traffic models which, in turn, makes it possible to investigate certain aspects of the impact on performance of long-range correlation structure within the confines of traditional Markovian analysis [32, 37].

The second type of causality—also called *structural causality* [50]—is more subtle in nature, and its roots can be attributed to an empirical property of distributed systems: the heavy-tailed distribution of file or object sizes. For the moment, a random variable obeying a *heavy-tailed* distribution can be viewed as giving rise to a very wide range of different values, including—as its trademark— "very large" values with non-negligible probability. This intuition is made more precise in Section 1.4.1.

[3]The same holds true for the LBL WAN data considered by Paxson and Floyd [56] and the BU WWW data analyzed by Crovella and Bestavros [13].

Returning to the causality description, in a nutshell, if end hosts exchange files whose size is heavy-tailed, then the resulting network traffic at multiplexing points in the network layer is self-similar [50]. This causal phenomenon was shown to be robust in the sense of holding for a variety of transport layer protocols such as TCP—e.g., Tahoe, Reno, and Vegas—and flow-controlled UDP, which make up the bulk of deployed transport protocols, and a range of network configurations. [50] also showed that research in UNIX file systems carried out during the 1980s give strong empirical evidence based on file system measurements that UNIX file systems are heavy-tailed. This is, perhaps, the most simple, distilled, yet high-level physical explanation of network traffic self-similarity. Corresponding evidence for Web objects, which are of more recent relevance due to the explosion of WWW and its impact on Internet traffic, can be found in [13].

Of course, structural causality would be meaningless unless there were explanations which showed why heavy-tailed objects transported via TCP- and UDP-based protocols would induce self-similar burstiness at multiplexing points. As hinted at in the original Leland *et al.* paper [41] and formally introduced in [74], the *on/off model* of Willinger *et al.* establishes that the superposition of a large number of independent on/off sources with heavy-tailed on- and/or off-periods leads to self-similarity in the aggregated process—a fractional Gaussian noise process—whose long-range dependence is determined by the heavy-tailedness of on or off-periods. Space aggregation is inessential to inducing long-range dependence—it is responsible for the Gaussian property of aggregated traffic by an application of the Central Limit Theorem—however, it is relevant to describing multiplexed network traffic. The on/off model has its roots in a certain renewal reward process introduced by Mandelbrot [46] (and further studied in [63]) and provides the theoretical underpinning for much of the recent works on physical modeling of network traffic. This theoretical foundation together with the empirical evidence of heavy-tailed on/off durations (as, for example, given for IP flow measurements [74]) represents a more low-level, direct explanation of physical causality of self-similarity, and form the principal factors that distinguish the on/off model from other mathematical models of self-similar traffic. The linkage between high-level and low-level descriptions of causality is further facilitated by [50] where it is shown that the application layer property of heavy-tailed file sizes is preserved by the protocol stack and mapped to approximate heavy-tailed busy periods at the network layer. The inter-packet spacing within a single session (or equivalently transfer/connection/flow), however, has been observed to exhibit its own distinguishing variability. This refined short time scale structure and its possible causal attribution to the feedback control mechanisms of TCP are investigated in [22, 23], and is the topic of on-going work.

1.2.3 Queueing analysis

In the third category are works that provide mathematical models of long-range dependent traffic with a view toward facilitating performance analysis in the queueing theory sense [2, 3, 17, 43, 49, 53, 66]. These works are important in that they establish basic performance boundaries by investigating queueing behavior with long-range dependent input which exhibit performance characteristics fundamentally different from corresponding systems with Markovian input. In particular, the queue length distribution in infinite buffer systems has a *slower-than-exponentially* (or *subexponentially*) decreasing tail, in stark contrast with short-range dependent input for which the decay is exponential. In fact, depending on the queueing model under consideration, long-range dependent input can give rise to *Weibullian* [49] or *polynomial* [66] tail behavior of the underlying queue length distributions. The analysis of such non-Markovian queueing systems is highly nontrivial and provides fundamental insight into the performance impact question. Of course, these works, in addition to providing valuable information into network performance issues, advance the state-of-the-art in performance analysis and are of independent interest. The queue length distribution result implies that buffering—as a resource provisioning strategy—is rendered ineffective when input traffic is self-similar in the sense of incurring a disproportionate penalty in queueing delay vis-à-vis the gain in reduced packet loss rate. This has led to proposals advocating a *small buffer capacity/large bandwidth* resource provisioning strategy due to its simplistic, yet curtailing influence on queueing: if buffer capacity is small, then the ability to queue or remember is accordingly diminished. Moreover, the smaller the buffer capacity, the more relevant short-range correlations become in determining buffer occupancy. Indeed, with respect to first-order performance measures such as packet loss rate, they may become the dominant factor. The effect of small buffer sizes and finite time horizons in terms of their potential role in delimiting the scope of influence of long-range dependence on network performance has been studied in [29, 58].

A major weakness of many of the queueing-based results [2, 3, 17, 43, 49, 53, 66] is that they are *asymptotic*, in one form or another. For example, in infinite buffer systems, upper and lower bounds are derived for the tail of the queue length distribution as the queue length variable approaches infinity. The same holds true for "finite buffer" results where bounds on buffer overflow probability are proved as buffer capacity becomes unbounded. There exist interesting results for zero buffer capacity systems [18, 19] which are discussed in Chapter 17. Empirically oriented studies [20, 33, 51] seek to bridge the gap between asymptotic results and observed behavior in finite buffer systems. A further drawback of current performance results is that they concentrate on first-order performance measures that relate to (long-term) packet loss rate but less so on second-order measures—e.g., variance of packet loss or delay, generically referred to as *jitter*—which are of import in multimedia communication. For example, two loss processes may have the same first-order statistic but if one has higher variance than the other in the form of concentrated periods of packet loss—as is the case in self-similar traffic—then this can adversely impact the efficacy of packet-level forward error correction used in the QoS-sensitive transport of real-time traffic [11, 52, 68]. Even less is known about transient performance measures which are more relevant in practice when convergence to long-term steady-state behavior is too slow to be of much value for engineering purposes. Lastly, most queueing results obtained for long-range dependent input are for open-loop systems that ignore feedback control issues present in actual networking environments (e.g., TCP). Since feedback can shape and influence the very traffic arriving at a queue [22, 50], incorporating their effect in feedback controlled closed queueing systems looms as an important challenge.

1.2.4 Traffic control and resource provisioning

The fourth category deals with works relating to the control of self-similar network traffic which, in turn, has two sub-categories: resource provisioning and dimensioning which can be viewed as a form of open-loop control, and closed-loop or feedback traffic control. Due to their feedback-free nature, the works on queueing analysis with self-similar input have direct bearing on the resource dimensioning problem. The question of quantitatively estimating the marginal utility of a unit of additional resource such as bandwidth or buffer capacity is answered, in part, with the help of these techniques. Of importance are also works on statistical multiplexing using the notion of effective bandwidth which point toward how efficiently resources can be utilized when shared across multiple flows [27]. A principal lesson learned in the resource provisioning side is the ineffectiveness of allocating buffer space vis-à-vis bandwidth for self-similar traffic, and the consequent role of short-range correlations in affecting first-order performance characteristics when buffer capacity is indeed provisioned to be “small” [29, 58].

On the feedback control side is the work on *multiple time scale congestion control* [67, 68] which tries to exploit correlation structure that exists across multiple time scales in self-similar traffic for congestion control purposes. In spite of the negative performance impact of self-similarity, on the positive side, long-range dependence admits the possibility of utilizing correlation at large time scales, transforming the latter to harness predictability structure which, in turn, can be affected to guide congestion control actions at smaller time scales to yield significant performance gains. The problem of designing control mechanisms that allow correlation structure at large time scales to be effectively engaged is a nontrivial technical challenge for two principal reasons: one, the correlation structure in question exists at time scales typically an order of magnitude or more above that of the feedback loop, and two, the information extracted is necessarily imprecise due to its probabilistic nature[4]. [67, 68] show that large time scale correlation structure can be employed to yield significant performance gains both for throughput maximization—using TCP and rate-based control—and end-to-end QoS control within the framework of adaptive redundancy control [52, 68]. An important by-product of this work is that the *delay-bandwidth product problem* of broadband networks, which renders reactive or feedback traffic controls ineffective when subject to long round-trip times (RTT), is mitigated by exercising control across multiple time scales. Multiple time scale congestion control allows uncertainty stemming from outdated feedback information to be compensated or "bridged" by predictability structure present at time scales exceeding the RTT or feedback loop (i.e., seconds vs. milliseconds). Thus even though traffic control in the 1990s has been occupied by the dual theme of large delay-bandwidth product and self-similar traffic burstiness, when combined, they lend themselves to a form of attack which imparts proactivity transcending the limitation imposed by RTT, thereby facilitating the metaphor of "catching two birds with one stone."

A related, but more straightforward, traffic control dimension is *connection duration prediction*. The works from physical modeling tell us that connections or flows tend to obey a heavy-tailed distribution with respect to their time duration or lifetime, and this information may be exploitable for traffic control purposes. In particular, heavy-tailedness implies that most connections are short-lived, but the bulk of traffic is contributed by a few long-lived flows [50]. By Amdahl's Law [4], it becomes relevant to carefully manage the impact exerted by the long-lived flows even if they are few in number[5]. The idea of employing "connection" duration was first advanced in the context of load balancing in distributed systems where UNIX processes have been observed to possess heavy-tailed lifetimes [30, 31, 40]. In contrast to the exponential distribution whose memoryless property renders prediction obsolete, heavy-tailedness implies *predictability*—a connection whose measured time duration

[4]We remark that understanding the correlation structure of network traffic at time scales below the feedback loop may be of relevance but remains, at this time, largely unexplored [22].

[5]A form of Amdahl's Law states that to improve a system's performance, its functioning with respect to its most frequently encountered states must be improved. Conversely, performance gain is delimited by the latter.

exceeds a certain threshold is more likely to persist into the future. This information can be used, e.g., in the case of load balancing, to decide whether it is worthwhile to migrate a process given the fixed, high overhead cost of process migration [31]. The ensuing opportunities have numerous applications in traffic control, one recent example being the discrimination of long-lived flows from short-lived flows such that routing table updates can be biased toward long-lived flows which, in turn, can enhance system stability by desensitizing against "transient" effects of short-lived flows [61]. In general, the connection duration information can also come from directly available information in the application layer—e.g., a Web server, when servicing a HTTP request, can discern the size of the object in question—and if this information is made available to lower layers, decisions such as whether to engage in open-loop (for short-lived flows) or closed-loop control (for long-lived flows) can be made to enhance traffic control [67].

1.3 ISSUES AND REMARKS

1.3.1 Traffic measurement and estimation

The area of traffic measurements—since the collection and analysis of the original Bellcore data [41]—has been tremendously active yielding a wealth of traffic measurements across a wide spectrum of different contexts supporting the view that network traffic exhibits self-similar scaling properties over a wide range of time scales. This finding is noteworthy given the fact that networks, over the past decade, have undergone significant changes in their constituent traffic flows, user base, transmission technologies, and scale with respect to system size. The observed robustness property or insensitivity to changing networking conditions justified calling self-similarity a *traffic invariant* and motivated focusing on underlying physical explanations that are mathematically rigorous as well as empirically verifiable. Robustness, in part, is explained by the fact that the majority of Internet traffic has been TCP traffic, and while in the pre-WWW days the bulk of TCP traffic stemmed from FTP traffic, in today's Internet, it is attributable to HTTP-based Web traffic. Both types of traffic have been shown to transport files whose size distribution is heavy-tailed [13, 56]. Physical modeling carried out in [50] showed that the transport of heavy-tailed files mediated by TCP (as well as flow-controlled UDP) induces self-similarity at multiplexing points in the network layer; it also showed that this is a robust phenomenon insensitive to details in network configuration and control actions in the protocol stack[6]. Measurement work has culminated in refined workload characterization at the application layer, including the modeling of user behavior [6, 7, 24, 48]. At the network layer, measurement analyses of IP traffic over fine time scales have lead to the multifractal characterization of wide area network traffic which, in turn, has bearing on physical modeling raising new questions about the relationship between feedback congestion control and short-range correlation structure of network traffic [22, 23]. The tracking of Internet workload and its characterization is expected to remain a practically important activity of interest in its own right. Demonstrating the relevance of ever refined workload models to networking research, however, will loom as a nontrivial challenge.

As with experimental physics, the measurement- or data-driven approach to networking research—rejuvenated by [41]—provides a balance to the more theoretical aspects of networking research, in the ideal situation, facilitating a constructive interplay of "give-and-take." A somewhat less productive consequence has been the discourse on short-range vs. long-range dependent mathematical models to describe measured traffic traces starting with the original Bellcore Ethernet data. At one level, both short-range and long-range dependent traffic models are parameterized systems that are sufficiently powerful to give rise to sample paths in the form of measured traffic time series. Mathematical system identification, under these circumstances, therefore, is an intrinsically ill-posed problem. Viewed in this light, the fact that different works can assign disparate modeling interpretations to the same measurement data, with differing conclusions, is not surprising [26, 33]. Put differently, it is well known that with a sufficiently parameterized model class, it is always possible to find a model that fits a given data set. Thus, the real challenge lies less in in mathematical model fitting but in *physical modeling*, an approach that in addition to describing the given data provides insight into the causal and dynamic nature of the processes that generated the data in the first place. On the positive side, the discussions about short-range vs. long-range dependence have brought out into the open concerns about nonstationary effects [16]—3pm traffic cannot be expected to stem from the same source behavior conditions as 3am traffic—that can influence certain types of inference and estimation procedures for long-range dependent processes. These concerns have spurned the development and adoption of estimation techniques based on wavelets which are sensitive to various types of nonstationary variations in

[6]Not surprisingly, extremities in control actions and resource configurations do affect the property of induced network traffic, in some instances, diminishing self-similar burstiness altogether [50]. Moreover, refined structure in the form of multiplicative scaling over sub-RTT time scales has only recently been discovered [23].

the data [1]. What is not in dispute are computed sample statistics—e.g., autocorrelation functions of measured traffic series—which exhibit nontrivial correlations at time lags on the order of seconds and above. Whether to call these time scales "long-range" or "short-range" is a matter of subjective choice and/or mathematical convenience and abstraction. What impact these correlations exert on queueing behavior is a function of how large the buffer capacity, the level of traffic intensity, and link capacity—among other factors—are [29, 58]. As soon as one deviates from empirical evaluation based on measurement data and adopts a model of the data, one is faced with the same ill-posed identification problem.

1.3.2 Traffic modeling

There exist a wide range of mathematical models of self-similar or long-range dependent traffic each with its own idiosyncrasies [5, 21, 23, 35, 43, 49, 53, 59, 74]. Some facilitate queueing analysis [43, 49, 53], some are physically motivated [5, 23, 74], and yet others show that long-range dependence may be generated in diverse ways [21, 35]. The wealth of mathematical models—while, in general, an asset—can also distract from an important feature endowed of the networking domain: the physics and causal mechanisms underlying network phenomena including traffic characteristics. Since network architecture—either by implementation or simulation—is *configurable*, from a network engineering perspective physical traffic models that trace back the roots of self-similarity and long-range dependence to architectural properties such as network protocols and file size distribution at servers have a clear advantage with respect to predictability and verifiability over "black box" models associated with traditional time series analysis. Contrast this with, say, economic systems where human behavior cannot be reprogrammed at will to test the consequences of different assumptions and hypotheses on system behavior. Physical models, therefore, are in a unique position to exploit this "reconfigurability trait" afforded by the networking domain, and use it to facilitate an intimate, mechanistic understanding of the system.

The on/off model [74] is a mathematical abstraction which provides a foundation for physical traffic modeling by advancing an explicit causal chain of verifiable network properties or events which can be tested against empirical data. For example, the factual basis of heavy-tailed on-periods in network traffic has been shown in [74], a corresponding empirical basis for heavy-tailed file sizes in UNIX file systems of the past whose transport may be the cause of heavy-tailed on-periods in packet trains has been shown in [50], and a more modern interpretation for the World Wide Web has been demonstrated in [13]. One weakness of the on/off model is its assumption of *independence* of on/off sources. This has been empirically addressed in [50] by studying the influence of dependence arising from multiple sources coupled at bottleneck routers sharing resources when the flows are governed by feedback congestion control protocols such as TCP in the transport layer. It was found that coupling did not significantly impact long-range dependence. A more recent study [22] shows that dependence due to feedback and inter-flow interaction may be the cause for multiplicative scaling phenomena observed in the short-range correlation structure, a refined physical characterization that may complement the previous findings which focused on coarser structure at larger time scales. We remark that the on/off model is able to induce both fractional Gaussian noise—upon aggregation over multiple flows and normalization—and a form of self-similarity and long-range dependence called asymptotic second-order self-similarity—a single process with heavy-tailed on/off periods—which constitute two of the most commonly used self-similar traffic models in performance analysis.

Finally, physical models, because of their grounding in empirical facts, influence the general argument advanced in Section 1.3.1 on the ill-posed nature of the identification problem. They can be viewed as tilting the scale in favor of long-range dependent traffic models. That is, since file sizes in various network related contexts have been shown to be heavy-tailed and the physical modeling works show that resulting traffic is long-range dependent, other things being equal, empirical evidence afforded by physical models biases toward a more consistent and parsimonious interpretation of network traffic as being long-range dependent as opposed to the mathematically equally viable short-range dependence hypothesis. Thus physical models, by virtue of their causal attribution, can also influence the choice of mathematical modeling and performance analysis.

1.3.3 Performance analysis and traffic control

The works on queueing analysis with self-similar input have yielded fundamental insights into the performance impact of long-range dependence, establishing the basic fact that queue length distribution decays slower-than-exponentially vis-à-vis the exponential decay associated with Markovian input [2, 3, 17, 43, 49, 53, 66]. In conjunction with observations advanced in [29, 58] on ways to curtail some of the effect of long-range dependence, a very practical impact of the queueing based performance analysis works has been the growing adoption of the resource dimensioning paradigm which states that buffer capacity at routers should be kept small while link bandwidth is to be increased. That is, the marginal utility of buffer capacity has diminished significantly vis-à-vis that of bandwidth. This is illustrated in Figure 1.3.1 which shows mean queue length as a function of buffer capacity at a bottleneck router when fed with self-similar input with varying degrees of long-range dependence but equal traffic intensity (roughly, α-values close to 1 imply strong long-range dependence whereas α-values close to 2 correspond to weak long-range dependence). In other words, when long-range correlation structure is weak, a buffer

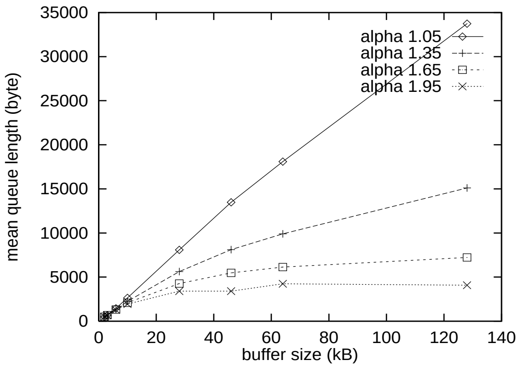

Fig. 1.3.1 Mean queue length as a function of buffer capacity for input traffic with varying long-range dependence ($\alpha = 1.05, 1.35, 1.65, 1.95$).

capacity of about 60kB suffices to contain the input's variability and, moreover, the average buffer occupancy remains below 5kB. However, when the long-range correlation structure is strong, an increase in buffer capacity is accompanied by a corresponding increase in buffer occupancy with the buffer capacity horizon at which the mean queue length saturates pushed out significantly.

In spite of the fundamental contribution and insight afforded by queueing analysis, as a practical matter, all the known results suffer under the limitation that the analysis is asymptotic in the buffer capacity: either the queue is assumed be infinite and asymptotic bounds on the tail of the queue length distribution are derived, or the queue is assumed to be finite but its overflow probability is computed as the buffer capacity is taken to infinity. There is, as yet, a chasm between these asymptotic results and their finitistic brethren which have alluded tractability. It is unclear whether the asymptotic formulae—beyond their qualitative relevance—are also practically useful as resource provisioning and traffic engineering tools. Further work is needed in this direction to narrow the gap. Another significant drawback of the performance analysis results—also related to the asymptotic nature of queueing results—is the focus on first-order performance indicators such as packet loss rate and mean queue length, which is even true in experimental studies. Second-order performance measures such as packet loss variance or delay variance—generically denoted jitter—play an important role in multimedia payload transport with real-time constraints. Even when a small buffer capacity resource provisioning policy is adopted to delimit the queueing aspect of self-similar traffic, if time-sensitive traffic flows are subject to concentrated periods of packet loss or severe inter-packet delay variation (even though packet loss rate may be small), then performance—as reflected by QoS—has degraded. The effectiveness of real-time QoS control techniques such as packet-level forward error correction are directly impacted by burstiness structure [11, 52, 68] and explicit incorporation of second-order performance measures must be affected to yield a balanced account of the performance impact question.

On the traffic control front, self-similarity—in spite of its detrimental performance aspect—implies the existence of correlation structure at a distance which may be exploitable for traffic control purposes. The framework of multiple time scale traffic control [67, 68, 69] exercises control actions across multiple time scales, using the information extracted at large time scales to modulate the output behavior of feedback congestion controls acting at the time scale of RTT. An important by-product of multiple time scale congestion control is the mitigation of the delay-bandwidth product problem which has been a pariah of reactive controls due to the outdatedness of feedback information in WAN environments which diminishes the effectiveness of reactive control actions. Figure 1.3.2 shows the performance gain of imparting multiple time scale capabilities on top of TCP Reno, Vegas, and Rate (a rate-based version of TCP) as a function of RTT. We observe that as RTT increases, performance enhancement vis-à-vis ordinary TCP due to multiple scale congestion control is amplified accordingly.

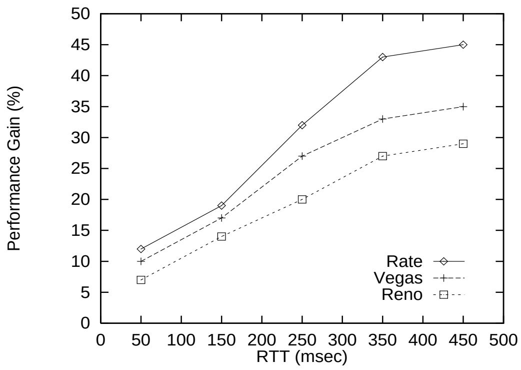

Fig. 1.3.2 Performance gain of TCP Reno, Vegas, Rate, when endowed with multiple time scale capabilities as a function of RTT.

The area of self-similar traffic control faces a number of challenges. First, self-similar traffic control, in the past, has received less attention than measurement/estimation, traffic modeling, and queueing analysis which is not too surprising since the problem of control is, in some sense, a natural continuation of research into "what is" type of questions that is followed by "what if" questions. Research into utilizing predictability stemming from long-range dependence and heavy-tailed connection durations is far from exhaustive, and further work is needed to explore the wide array of traffic control possibilities. Second, whereas long-lived connections—although few in number but contributing the bulk of traffic—constitute the primary target of traffic control, the effective management of short-lived connections—due to their sheer number—looms as an important problem. Maintenance of *persistent* state at end systems that is shared across multiple flows is a promising avenue that would allow open-loop traffic control to be sensitive to network state, thus imparting a measure of proactivity. Last but not least, analysis of feedback loop systems with respect to their stability and optimality including those arising in multiple time scale traffic control for self-similar traffic remains a challenge. New ideas and approaches are needed to succeed in our attempts to tractably analyze and understand large-scale, coupled, interacting complex systems such as the Internet.

1.4 TECHNICAL BACKGROUND

1.4.1 Self-similar processes and long-range dependence

1.4.1.1 Second-order self-similarity and stationarity Consider a discrete time stochastic process or time series $X(t)$, $t \in \mathbb{Z}$, where $X(t)$ is interpreted as the traffic volume—measured in packets, bytes, or bits—*at* time instance t. Of interest is also the interpretation that $X(t)$ is the total traffic volume *up to* time t, say, from time 0. To minimize confusion, when a "cumulative" view is taken, we will denote the process by $Y(t)$. We will then reserve $X(t)$ to be the *increment process* corresponding to $Y(t)$, i.e., $X(t) = Y(t) - Y(t-1)$.

For traffic modeling purposes, we would like $X(t)$ to be "stationary" in the sense that its behavior or structure is invariant with respect to shifts in time. In other words, t's responsibility as an *absolute* reference frame is relieved. Without some form of stationarity, "anything" is allowed and a model loses much of its usefulness as a compact description of (assumed) tractable phenomena. $X(t)$ is *strictly stationary* if $(X(t_1), X(t_2), \ldots, X(t_n))$ and $(X(t_1+k), X(t_2+k), \ldots, X(t_n+k))$ possess the same joint distribution for all $n \in \mathbb{Z}_+$, $t_1, \ldots, t_n, k \in \mathbb{Z}$. Denoting the k-shifted process or time series X_k, X and X_k are said to be equivalent in the sense of *finite-*

dimensional distributions, $X =_d X_k$. Imposing strict stationarity, it turns out, is too restrictive and we will be interested in a weaker form of stationarity—*second-order stationarity*[7]—which requires that the autocovariance function $\gamma(r,s) = E[(X(r) - \mu)(X(s) - \mu)]$ satisfies translation invariance, i.e., $\gamma(r,s) = \gamma(r+k, s+k)$ for all $r, s, k \in \mathbb{Z}$. The first two moments are assumed to exist and be finite, and we set $\mu = E[X(t)]$, $\sigma^2 = E[(X(t) - \mu)^2]$ for all $t \in \mathbb{Z}$. We will also assume $\mu = 0$. Since, by stationarity, $\gamma(r,s) = \gamma(r-s, 0)$, we denote the autocovariance by $\gamma(k)$.

To formulate scale-invariance, first define the *aggregated process* $X^{(m)}$ of X at aggregation level m,

$$X^{(m)}(i) = \frac{1}{m} \sum_{t=m(i-1)+1}^{mi} X(t).$$

That is, $X(t)$ is partitioned into non-overlapping blocks of size m, their values are averaged, and i is used to index these blocks. Let $\gamma^{(m)}(k)$ denote the autocovariance function of $X^{(m)}$. Under the assumption of second-order stationarity we arrive at the following definitions of second-order self-similarity.

Definition 1.4.1 **(Second-order Self-similarity)** $X(t)$ is *exactly second-order self-similar* with Hurst parameter H $(1/2 < H < 1)$ if

$$\gamma(k) = \frac{\sigma^2}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}) \tag{1.4.2}$$

for all $k \geq 1$. $X(t)$ is *asymptotically second-order self-similar* if

$$\lim_{m \to \infty} \gamma^{(m)}(k) = \frac{\sigma^2}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}). \tag{1.4.3}$$

It can be checked that (1.4.2) implies $\gamma(k) = \gamma^{(m)}(k)$ for all $m \geq 1$. Thus, second-order self-similarity captures the property that correlation structure is exactly—condition (1.4.2)—or asymptotically—the weaker condition (1.4.3)—preserved under time aggregation. The form of $\gamma(k) = ((k+1)^{2H} - 2k^{2H} + (k-1)^{2H})\sigma^2/2$ is not accidental and implies further structure—long-range dependence—to which we will return later. Second-order self-similarity (in the exact or asymptotic sense) has been a dominant framework for modeling network traffic and this is also reflected in the chapters of this book.

[7]Equivalent names are *weak*, *covariance*, and *wide sense* stationarity.

1.4.1.2 An allegory into distributional self-similarity To understand the particular form of $\gamma(k)$ in the definition of second-order self-similarity, we will make a short detour and discuss self-similar processes in slightly more generality. Further extensions and detailed treatments can be found in [9, 60].

Consider the cumulative process $Y(t)$, albeit in continuous time $t \in \mathbb{R}$. Following is a definition of self-similarity for continuous time processes in the sense of finite dimensional distributions.

Definition 1.4.4 **(H-ss)** $Y(t)$ is *self-similar with self-similarity parameter, i.e., Hurst parameter,* H $(0 < H < 1)$, denoted *H-ss*, if for all $a > 0$ and $t \geq 0$,

$$Y(t) =_d a^{-H} Y(at). \tag{1.4.5}$$

Thus $Y(t)$ and its time scaled version $Y(at)$—after normalizing by a^{-H}—must follow the same distribution. In the traffic modeling context, it is convenient to think of $Y(t)$ as the cumulative or total traffic up to time t. For $a > 1$—time is stretched or dilated—a contraction factor a^{-H} is applied to make the magnitude of $Y(at)$ comparable to that of $Y(t)$. For $a < 1$, the opposite holds true. As a varies, the scaling exponent H remains invariant. This is a most natural definition, however, it has an important drawback: unless $Y(t)$ is degenerate, i.e., $Y(t) = 0$ for all $t \in \mathbb{R}$, $Y(t)$ cannot be stationary due to the normalization factor a^{-H}. Its increment process $X(t) = Y(t) - Y(t-1)$, however, is another matter. In particular, consider the case where $Y(t)$ is H-ss and has *stationary increments*; in this case we say $Y(t)$ is *H-sssi*. Let us further assume that $Y(t)$ has finite variance. It can be checked that $E[Y(t)] = 0$, $E[Y^2(t)] = \sigma^2|t|^{2H}$, and

$$\gamma(k) = \frac{\sigma^2}{2}(|t|^{2H} - |t-s|^{2H} + |s|^{2H}). \tag{1.4.6}$$

This is achieved by noting that[8]

$$Y(t) =_d t^H Y(1)$$

from which it follows $E[Y^2(t)] = \sigma^2 t^{2H}$. The latter, then, can be used in the derivation of the autocovariance function (1.4.6). The increment process $X(t)$ has mean 0 and autocovariance $\gamma(k)$ as given in (1.4.2). The derivation is similar to that of $Y(t)$.

How does distributional self-similarity (of a continuous time process) tie in with

[8]From $a^H Y(t) =_d Y(at)$, substitute $t = 1$ and $a = t$.

second-order self-similarity (of a discrete time process) which requires exact or asymptotic invariance with respect to second-order statistical structure of the aggregated time series $X^{(m)}$? A key observation lies in noting that $X^{(m)}$ can be viewed as computing a sample mean

$$X^{(m)} = \frac{1}{m}\sum_{t=1}^{m} X(t) = m^{-1}(Y(m) - Y(0))$$
$$=_d m^{-1}m^H(Y(1) - Y(0)) = m^{H-1}X.$$

Thus, if $Y(t)$ is a H-sssi process then its increment process $X(t)$ satisfies

$$X =_d m^{1-H}X^{(m)} \tag{1.4.7}$$

which shows how $X^{(m)}$ is related to X via a simple scaling relationship involving H in the sense of finite dimensional distributions. (1.4.2) and (1.4.3), then, express the fact that, X and $m^{1-H}X^{(m)}$ are required to have exactly or asymptotically the same second-order structure. As a result, depending on whether a discrete time process $X(t)$ satisfies (1.4.7) for all $m \geq 0$ or only in the limit as $m \to \infty$, $X(t)$ is said to be *exactly self-similar* or *asymptotically self-similar*. Note that in the Gaussian case, this definition coincides with second-order self-similarity.

As a lead-in to the role of the parameter H, recall that the variance of the sample mean $\bar{Z}$ of a random variable Z satisfies var$(\bar{Z}) = \sigma_Z^2/m$ where m is the sample size. From (1.4.7) it follows that var$(X^{(m)}) = \sigma^2 m^{2H-2}$. When viewed as a sample mean where the samples are drawn *independently*, var$(X^{(m)})$ reduces to $\sigma^2 m^{-1}$ if $H = 1/2$. If $H \neq 1/2$, in particular, $1/2 < H < 1$, then

$$\text{var}(X^{(m)}) = \sigma^2 m^{-\beta}$$

with $0 < \beta < 1$ (and $H = 1 - \beta/2$) which hints at certain—and not just any—*dependency structure* in the "samples" (i.e., time series in our case) which causes var$(X^{(m)})$ to converge to zero slower than the rate m^{-1}.

1.4.1.3 Long-range dependence Thus far we have focused on explicating the role of self-similarity in the second-order stationary and distributional senses with little regard to the role of H and its range of values. Let us return to the definition of second-order self-similarity and its autocovariance $\gamma(k)$. Let $r(k) = \gamma(k)/\sigma^2$ denote the *autocorrelation function*. For $0 < H < 1$, $H \neq 1/2$, it holds

$$r(k) \sim H(2H-1)k^{2H-2}, \qquad k \to \infty. \tag{1.4.8}$$

In particular, if $1/2 < H < 1$, $r(k)$ asymptotically behaves as $c\,k^{-\beta}$ for $0 < \beta < 1$ where $c > 0$ is a constant, $\beta = 2 - 2H$, and we have

$$\sum_{k=-\infty}^{\infty} r(k) = \infty. \tag{1.4.9}$$

That is, the autocorrelation function decays slowly—i.e., hyperbolically—which is the essential property that causes it to be not summable. When $r(k)$ decays hyperbolically such that condition (1.4.9) holds, we call the corresponding stationary process $X(t)$ *long-range dependent*. $X(t)$ is *short-range dependent* if the autocorrelation function is summable[9]. An essentially equivalent definition can be given in the frequency domain where the *spectral density* $\Gamma(\nu) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} r(k)e^{ik\nu}$ is required to satisfy the property

$$\Gamma(\nu) \sim c|\nu|^{-\alpha}, \qquad \nu \to 0.$$

Here $c > 0$ is a constant and $0 < \alpha = 2H - 1 < 1$. Thus $\Gamma(\nu)$ diverges around the origin implying ever larger contributions by low frequency components.

Following are some simple facts regarding the value of H and its impact on $r(k)$. First, if $H = 1/2$, then $r(k) = 0$, and $X(t)$ is trivially short-range dependent by virtue of being completely uncorrelated. In the case where $0 < H < 1/2$, we have $\sum_{k=-\infty}^{\infty} r(k) = 0$, an artificial condition rarely encountered in applications. $H = 1$ is uninteresting since it leads to the degenerate situation $r(k) = 1$ for all $k \geq 1$. Finally, H-values bigger than 1 are prohibited due to the stationarity condition on $X(t)$.

1.4.1.4 Self-similarity vs. long-range dependence The preceding discussion indicates that there are self-similar processes that are not long-range dependent, and vice versa. For example, Brownian motion is 1/2-sssi with white Gaussian noise as its increment process, but the latter is not long-range dependent. Conversely, certain fractional ARIMA time series generate long-range dependence but they are not self-similar in the distributional sense. In the case of asymptotic second-order self-similarity, however, by the restriction $1/2 < H < 1$ in the definition, self-similarity implies long-range dependence, and vice versa. It is for this reason and the fact that asymptotic second-order self-similar processes are employed as "canonical" traffic models, that we sometimes use *self-similarity* and *long-range dependence*

[9]Technically more subtle definitions of long-range dependence are possible, but in this book, we will mainly rely on our working definition involving condition (1.4.9).

interchangeably when the context does not lead to confusion.

1.4.2 Impact of heavy tails

1.4.2.1 Heavy-tailed distribution There is an intimate relationship between heavy-tailed distributions and long-range dependence which we will discuss in the next sections. First, a few definitions and basic facts. A random variable Z has a *heavy-tailed distribution* if

$$\Pr\{Z > x\} \sim c x^{-\alpha}, \qquad x \to \infty \tag{1.4.10}$$

where $0 < \alpha < 2$ is called the *tail index* or *shape parameter* and c is a positive constant[10]. That is, the tail of the distribution, asymptotically, decays hyperbolically. This is in contrast to *light-tailed distributions*—e.g., exponential and Gaussian—which possess an exponentially decreasing tail. A distinguishing mark of heavy-tailed distributions is that they have infinite variance for $0 < \alpha < 2$, and if $0 < \alpha \leq 1$, they also have an unbounded mean. In the networking context, we will be primarily interested in the case $1 < \alpha < 2$. A frequently used heavy-tailed distribution is the *Pareto distribution* whose distribution function is given by

$$\Pr\{Z \leq x\} = 1 - \left(\frac{b}{x}\right)^{\alpha}, \qquad b \leq x,$$

where $0 < \alpha < 2$ is the shape parameter and b is called the *location parameter*. The mean is given by $\alpha k/(\alpha - 1)$. We remark that there are distributions—e.g., Weibull and log-normal—that have *subexponentially* decreasing tails but possess finite variance.

The main characteristic of a random variable obeying a heavy-tailed distribution is that it exhibits extreme variability. Practically speaking, a heavy-tailed distribution gives rise to very large values with nonnegligible probability so that sampling from such a distribution results in the bulk of values being "small" but a few samples having "very" large values. Not surprisingly, heavy-tailedness impacts sampling by slowing down the convergence rate of the sample mean to the population mean, dilating it as the tail index α approaches 1. For example, pending on the sample size m, the sample mean $\bar{Z}_m$ of a Pareto distributed random variable Z may significantly deviate from the population mean $\alpha k/(\alpha - 1)$, oftentimes underestimating it. In fact, the absolute

[10]Technically more subtle definitions involving slowly varying functions are possible and can be found in some chapters of this book. However, for practical purposes and to convey the main ideas, our working definition, centered around condition (1.4.10), will suffice.

estimation error $|\bar{Z}_m - E(Z)|$ asymptotically behaves as $m^{(1/\alpha)-1}$ (see, e.g., [15]), and thus for α-values close to 1, care must be given when sampling from heavy-tailed distributions such that conclusions about network behavior and performance attributable to sampling error are not advanced. A more detailed discussion of sampling issues is given in Chapter 3.

1.4.2.2 Heavy tails and predictability Heavy-tailedness of certain network related variables—e.g., file sizes and connection durations—can be shown to underlie the root cause of long-range dependence and self-similarity in network traffic. First, a simple fact on the intrinsic predictability associated with heavy-tailed random variables. Let Z be a heavy-tailed random variable interpreted as the *duration* or *lifetime* of a network connection (e.g., TCP connection, IP-flow, or session). Since connection durations are physically measurable events, assume that we observe—in time—that a connection has been active for $\tau > 0$ seconds. To simplify the discussion, assume time is discrete ($t \in \mathbb{Z}_+$) and $A : \mathbb{Z}_+ \to \{0, 1\}$ is an indicator function such that $A(t) = 1$ iff $Z \geq t$. We are interested in the probability that the connection will persist into the future given that it has been active for τ seconds. That is, we would like to estimate the conditional probability

$$\mathcal{L}(\tau) = \Pr\{\, A(\tau + 1) = 1 \mid A(t) = 1,\ 1 \leq t \leq \tau\,\}. \tag{1.4.11}$$

$\mathcal{L}(\tau)$ can be expressed as

$$\mathcal{L}(\tau) = 1 - \frac{\Pr\{Z = \tau\}}{\Pr\{Z \geq \tau\}}. \tag{1.4.12}$$

Let us first compute $\mathcal{L}(\tau)$ for light tails, in particular, distributions with asymptotically exponential tails $\Pr\{Z > x\} \sim c_1\, e^{-c_2 x}$ where $c_1, c_2 > 0$ are constants. The second term in (1.4.12) is computed by

$$\frac{\Pr\{Z = \tau\}}{\Pr\{Z \geq \tau\}} \sim \frac{c_1\, e^{-c_2 \tau} - c_1\, e^{-c_2(\tau+1)}}{c_1\, e^{-c_2 \tau}} = 1 - e^{-c_2}$$

for large τ, and we get $\mathcal{L}(\tau) \sim e^{-c_2}$. Thus for exponentially light tails, prediction is not enhanced by conditioning on ever longer periods of observed activity. For heavy tails, the corresponding derivations are

$$\frac{\Pr\{Z = \tau\}}{\Pr\{Z \geq \tau\}} \sim \frac{c\,\tau^{-\alpha} - c\,(\tau + 1)^{-\alpha}}{c\,\tau^{-\alpha}} = 1 - \left(\frac{\tau}{\tau + 1}\right)^{\alpha},$$

which yields

$$\mathcal{L}(\tau) \nearrow 1, \qquad \tau \to \infty. \tag{1.4.13}$$

Thus, the longer the period of observed activity, the more certain that it will persist into the future. In fact, it is straightforward to generalize (1.4.11) so that we can measure the *persistence* of activity $\delta \geq 1$ time units into the future, i.e.,

$$\mathcal{L}(\tau) = \Pr\{\, A(\tau + s) = 1,\ 1 \leq s \leq \delta \mid A(t) = 1,\ 1 \leq t \leq \tau \,\}.$$

This does not change the qualitative results: for the light-tailed case, $\mathcal{L}(\tau) \sim e^{-c_2\delta}$; for the heavy-tailed case, $\mathcal{L}(\tau)$'s asymptotic behavior follows $(1 + \delta/\tau)^{-\alpha} \nearrow 1$. Since $(1 + \delta/\tau)^{-\alpha} \leq e^{-\alpha\delta/\tau}$, we observe that in both cases predictability is exponentially sensitive to the prediction interval δ. However, in the heavy-tailed case, for any desired δ time unit "peek into the future," by conditioning the prediction on a sufficiently long past observation of activity, the prediction error can be reduced to an arbitrarily small level.

We remark that the mathematical implications of asymptotic analysis need not deter from the practical relevance of its conclusions, even considering the fact that tails are always finite in a physical network environment. First, if heavy-tails are modeled using the Pareto distribution, then its shape is hyperbolic across its *entire* range—not just asymptotically—and accurate finitary computations can be carried out. Second, given an empirical distribution with finite support, the fact that it has a finite cut-off point will not significantly influence the predictability computations carried out in practice as long as the tail is "sufficiently"—e.g., several orders of magnitude beyond the mean—long. As with time series, the identification problem of whether an empirical distribution is best modeled by heavy-tailed or light-tailed distributions is intrinsically ill-posed and secondary to the fact that the predictability structure as computed by (1.4.12) from *empirical distributions* is significant.

1.4.2.3 Heavy tails and long-range dependence As we saw in the previous section, heavy tails lead to predictability, and for a related reason, they lead to long-range dependence in network traffic. First, we give a definition of fractional Brownian motion (FBM) and its increment process—fractional Gaussian noise (FGN)—which are Gaussian self-similar processes with, and without, long-range dependence, first introduced by Mandelbrot [45]. Their Gaussian structure renders them especially useful as *aggregate* traffic models where aggregation of independent traffic sources—by the central limit theorem—leads to the Gaussian property. In practice, of course, traffic flows need not be independent if they engage in feedback control and share common resources at bottleneck routers. The definitions of FBM and FGN are couched in the framework of distributional self-similarity given in Section 1.4.1.2.

Definition 1.4.14 **(FBM)** $Y(t)$, $t \in \mathbb{R}$, is called *fractional Brownian motion* with parameter H, $0 < H < 1$, if $Y(t)$ is Gaussian and H-sssi.

Definition 1.4.15 **(FGN)** $X(t)$, $t \in \mathbb{Z}_+$, is called *fractional Gaussian noise* with parameter H if $X(t)$ is the increment process of FBM with parameter H.

By the definition of H-sssi, FBM reduces to Brownian motion—and FGN to white Gaussian noise—when $H = 1/2$. Thus $X(t)$, $t \in \mathbb{Z}_+$, becomes completely uncorrelated. Since Gaussian processes are characterized by their second-order structure, for each H, $0 < H < 1$, there is a unique Gaussian process that is the stationary increment of a H-sssi process. FBM is the corresponding unique Gaussian H-sssi process. By the same token, for Gaussian processes, distributional self-similarity and second-order self-similarity yield equivalent definitions.

Now to why heavy tails are considered the root cause of long-range dependence in network traffic. We take a constructive approach by presenting input processes—in various guises—with probabilistic activity times which, then, are shown to lead to long-range dependence if, and only if, they are heavy-tailed. We first present the on/off model by Willinger *et al.* [74] followed by a related model used by Likhanov *et al.* [43] which has a slightly different, but complementary, source arrival perspective.

The on/off model considers N independent traffic sources $X_i(t)$, $i \in [1, N]$, where each is a 0/1 *reward renewal process* with i.i.d. on-periods and i.i.d. off-periods. This just means that $X_i(t)$ takes on the values 1 ("on") and 0 ("off") on alternating, non-overlapping time intervals called on- and off-periods, respectively. $X_i(t) = 1$ is interpreted as there being a packet transmission. Thus, an on-period can be viewed as constituting a "packet train" [36]. Three such on/off sources and their aggregation is depicted in Figure 1.4.1. Let $S_N(t) = \sum_{i=1}^{N} X_i(t)$ denote the aggregate traffic at time t. Consider the *cumulative* process $Y_N(Tt)$ defined as

$$Y_N(Tt) = \int_0^{Tt} \left(\sum_{i=1}^{N} X_i(s) \right) ds, \tag{1.4.16}$$

where $T > 0$ is a scale factor that is explicitly incorporated. Thus $Y_N(Tt)$ measures the total traffic up to time Tt. What is the behavior of $Y_N(Tt)$ for large T and N? We will simplify the discussion so as to concentrate on the single salient feature of how heavy-tailedness influences long-range dependence. Let τ_{on} be the random variable describing the duration of the on-periods and let τ_{off} be the random variable

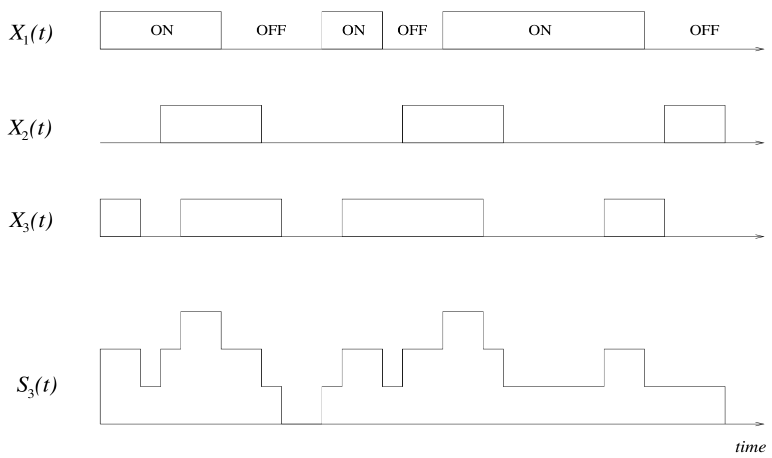

Fig. 1.4.1 $N = 3$ on/off sources $X_1(t), X_2(t), X_3(t)$, and their aggregation $S_3(t) = X_1(t) + X_2(t) + X_3(t)$.

associated with the durations of the off-periods. Let

$$\Pr\{\tau_{\text{on}} > x\} \sim c\,x^{-\alpha}, \qquad x \to \infty$$

where $1 < \alpha < 2$ and $c > 0$ is a constant. As to τ_{off}, it can be either heavy-tailed or light-tailed with finite variance. It can be shown [62, 74] that $Y_N(Tt)$ behaves like FBM in the following sense.

Theorem 1.4.17 (On-Off Model & FBM) *$Y_N(Tt)$ behaves statistically as*

$$\frac{E(\tau_{on})}{E(\tau_{on}) + E(\tau_{off})} NTt + CN^{1/2}T^H B_H(t) \tag{1.4.18}$$

for large T, N, where $H = (3-\alpha)/2$, $B_H(t)$ is FBM with parameter H, and $C > 0$ is a quantity depending only on the distributions of τ_{on} and τ_{off}.

Thus $Y_N(Tt)$ asymptotically behaves as fractional Brownian motion fluctuating around $NTtE(\tau_{\text{on}})/(E(\tau_{\text{on}}) + E(\tau_{\text{off}}))$ when suitably normalized. It is long-range dependent ($1/2 < H < 1$) iff $1 < \alpha < 2$, i.e., τ_{on}'s distribution is heavy-tailed. If neither τ_{on} nor τ_{off} is heavy-tailed, then $Y_N(Tt)$ is short-range dependent. It is in this sense that heavy-tailedness (in this case, of the on- or off-periods) is an essential component to inducing long-range dependence in the aggregated time series. Of less practical import in the networking context is the case when the off-period is heavy-tailed but the on-period is not, which nonetheless also yields long-range dependence.

A related but slightly different source model is obtained when viewing each source $i \in \mathbb{Z}_+$ as emitting a *singular* packet train but being otherwise silent [43]. Thus, a single on/off source in the on/off model can be construed to be the output behavior of a network host which may service multiple TCP connections, whereas in the singular packet train case, the source corresponds to a single TCP connection transporting a byte stream such as a file. To each source $i \in \mathbb{Z}_+$, we associate a time interval $[t_i, t_i + \tau_i)$, $t_i, \tau_i \in \mathbb{Z}_+$, where $X_i(t) = 1$ if $t \in [t_i, t_i + \tau_i)$, and 0 otherwise. We assume that the τ_i, $i \in \mathbb{Z}_+$, are i.i.d. and t_i is determined by a Poisson process $\xi(t)$ which indicates how many *new* connections arrive at time t.

$$X(t) = \sum_{i \in \mathbb{Z}_+} X_i(t) < \infty$$

then counts how many connections are active at time t. Alternatively, $X(t)$ can be viewed as the aggregate (over flows) traffic rate emitted at time instance t. The behavior of $X(t)$ and its generalized brethren can be analyzed directly [43, 65, 66], but a more succinct and elegant approach that reveals the influence of heavy-tailedness on long-range dependence can be found in a result due to Cox involving the M/G/∞ queueing system [12]. An M/G/∞ queue is defined to be the *busy server process* where connection arrivals are Poisson and each connection is serviced by a server—there are infinitely many—with a general service time. Thus, at any instance of time, we count how many servers are busy servicing requests. If the i.i.d. service times are given by $\tau = \tau_i$, $i \in \mathbb{Z}_+$, then it is easy to see that the busy server process in the M/G/∞ queue corresponds to the aggregate traffic rate $X(t)$ in the Poisson source model with a single on-period. Let τ be heavy-tailed with tail index $1 < \alpha < 2$.

Theorem 1.4.19 (M/G/∞ and LRD) *$X(t)$, $t \in \mathbb{Z}_+$, is asymptotically second-order self-similar with parameter $H = (3 - \alpha)/2$.*

Thus $1 < \alpha < 2$, via $1/2 < H = (3 - \alpha)/2 < 1$, is directly tied to long-range dependence. Theorem 1.4.19, in turn, implies by the previous correspondence that when connections with a single heavy-tailed on-period arrive in a Poisson manner, then the resulting aggregate traffic is long-range dependent. In its raw form, $X(t)$ has Poisson marginals [57], but it can be shown that FBM arises naturally as a limiting process by appropriately scaling the Poisson arrival rate and service times [39]. The M/G/∞ approach to modeling network traffic has proved useful in analyzing queueing behavior fed by long-range dependent input [54, 55].

We remark that from a purely *mathematical modeling* point-of-view, heavy-tailedness is not necessary to generate long-range dependence in aggregate traffic. As pointed out in [9] (and further explored in Chapter 11), an *infinite* aggregation of short-range dependent sources—in particular, heterogeneous on/off sources with exponential on/off times—can produce long-range dependence when suitably calibrated. *Finite* aggregations of short-range dependent sources, however, cannot induce long-range dependence, hence the assumption of infinite aggregation is crucial. Empirical traffic measurements provide strong evidence that file sizes and connection durations are heavy-tailed, and hence the *heavy-tailedness causes long-range dependence* rule-of-thumb is supported by physical modeling. The practical implications—if any—of the short-range dependent flows can produce long-range dependence observation are not clear, and we include them for completeness.

1.5 ORGANIZATION OF THE BOOK

This book is a collection of chapter contributions which brings together relevant works spanning a cross-section of topics covering traffic measurement, modeling, performance analysis, and traffic control for self-similar network traffic.

The first part of the book deals with traffic characterization, estimation, and simulation issues. Wavelet analysis is introduced as a powerful technique for both modeling and estimation of self-similar traffic. The wavelet based approach naturally lends itself to a multifractal view of network traffic where a shift in traffic properties at long and short time scales is captured using cascade constructions superimposed on heavy-tailed renewal processes. This is further discussed in Chapter 20. Complementing the theme of traffic modeling is the issue surrounding simulation, such as in the generation of synthetic workloads and self-similar traffic, which entails, in many instances, sampling from heavy-tailed distributions. Due to the slow convergence of sample statistics to population statistics, special care needs to be exercised when performing simulations that involve sampling from heavy-tailed distributions so as to not advance erroneous conclusions attributable to sampling effects including underestimation.

The second part of the book focuses on performance evaluation issues, in particular, queueing behavior of finite and infinite buffer systems when fed with long-range dependent input. Due to the breakdown of Markovian assumptions which are key to tractability in traditional queueing analysis, the technical challenges encountered with self-similar input are great. This part of the book gives an exposition of what is known about queueing with self-similar input, starting with the trademark phenomenon that queue length distribution decays polynomially—as opposed to exponentially—and advancing to packet scheduling, transient analysis, tight buffer asymptotics, and impact of resource boundedness and finite time horizons. Queueing based performance analysis also forms the foundation of traffic control based on resource provisioning and dimensioning. The traffic models considered can be viewed as variants of on/off renewal reward processes where session arrivals are allowed to be Poisson, however, on- or off-periods are assumed to be heavy-tailed. Some of the input processes are intimately related to fractional Brownian motion and its increment process, fractional Gaussian noise, which in turn can be analyzed by various techniques including large deviations theory.

The third part of the book covers traffic control issues that arise under self-similar traffic conditions. There are two main facets to the question, one centered on the problem of resource provisioning and dimensioning—a form of open-loop control—and ensuing trade-off relations, and the other based on the traditional traffic control framework of feedback control and its realization in network protocols including TCP. With respect to resource provisioning, due to the amplified queueing delay incurred when employing buffer dimensioning, an alternative resource provisioning strategy based on bandwidth dimensioning as the principal control variable has been advanced. In this "bufferless" traffic engineering regime, by reserving sufficient resources to meet the peak rate of multiplexed input traffic—i.e., over-provisioning—a desired level of quality of service in the form of statistical guarantees can be achieved. Feedback traffic control, on the other hand, represents a more subtle challenge where the central idea revolves around exploiting correlation structure at multiple time scales—in particular, "large" time scales exceeding the round-trip time associated with the feedback loop—as afforded by long-range dependence and self-similarity, to affect traffic control decisions executed at smaller time scales. When effectively facilitated, this can result in significant performance improvements including mitigation of the delay-bandwidth product problem in broadband wide area networks due to proactivity.

The last part of the book takes a bird's eye view of the manifold accomplishments and projects into the future promising research directions including those based on most recent developments. Chapter 20 focuses on traffic characterization and modeling issues with emphasis on a program for achieving a comprehensive understanding of network traffic and workloads, spanning both large- and small-time scale behaviors. Chapter 21 provides a complementary view concentrating on traffic control and performance evaluation issues which are expected to be of relevance in the design and management of the future Internet.

1.6 CHAPTER CONTRIBUTIONS

In the following, we give a brief outline of the various chapter contributions organized into three parts: (i) estimation and simulation, (ii) queueing with self-similar input, and (iii) traffic control and resource provisioning. We describe how each chapter fits into the overall picture and comment on the potential role and relevance of each chapter for future advances in these areas. The threefold categorization is not strict in the sense that some chapters encompass subject matters that cross the set boundaries. Also, part (ii) may be more generally characterized as performance evaluation with self-similar input as queueing is the predominant, but not exclusive, theme contained therein.

Chapter 1 by Park and Willinger serves as an introductory chapter that provides the necessary technical background including definitions for following the rest of the book. The chapter is self-contained and thus can also be read as a modern introduction to the topic of self-similar network traffic. It gives an overview of the various research activities surrounding self-similar traffic and outlines the principal issues in the areas of traffic modeling, statistical and scientific inference, performance analysis, and traffic control. Chapter 1 concludes with an overview of the book including the present section describing each chapter contribution.

1.6.1 Estimation and simulation

The chapter by Abry, Flandrin, Taqqu, and Veitch (Chapter 2) discusses the state-of-the-art in identification of scaling phenomena in traffic series—the crucial component of self-similarity—using the framework of wavelets. Due to their ability to localize a given signal or time series both in time and scale (or frequency), wavelets provide a powerful and refined technique for detecting and quantifying scaling behavior in measured traffic. Since wavelets are, in part, parameterized by scaling parameters, this lends itself naturally to a multi-scale representation and analysis of time series which, in turn, allows a qualitatively more informative and quantitatively more accurate estimation of underlying scaling properties. Abry *et al.* present a comprehensive overview of the fundamentals of wavelet analysis and its application to estimating scaling behavior in self-similar traffic, focusing on properties related to self-similar scaling. The chapter concludes with a discussion of the "inverse" operation, i.e., that of *generating* synthetic self-similar time series using wavelet expansions.

Chapter 20 by Riedi and Willinger describes an even further refined modeling based, in part, on the wavelet framework where a notion of large time scale and small time scale behavior and their observed empirical differences are captured. The novel aspect here is that the resulting representation combines in a natural manner self-similar and *multifractal* scaling behaviors. In one interpretation, multifractals can viewed as being composed of heavy-tailed renewal processes which collectively determine the large time scale behavior leading to long-range dependence, and a more fine-granular "within-connection" structure that reveals itself in a locally highly irregular scaling behavior and conforms to multifractality over fine time scales. The fact that long-range dependence of such processes does not depend on the finer details underlying variations within on-periods or connection times is shown by a result of Kurtz [39]. Riedi and Willinger, however, show that the fine granular structure at smaller time scales deviates significantly from the self-similar scaling behavior over large time scales and may therefore be relevant for traffic modeling purposes. They argue that multifractals in the form of certain cascade models or multiplicative processes provide a natural modeling framework, and also give initial evidence to suggest that the deviation from self-similar scaling observed in the short-correlation structure may be due to protocol stack effects such as those stemming from TCP's feedback congestion control.

Chapter 20 also provides a high-level overview of recent developments in traffic modeling with foundations in physical modeling. In this capacity, it serves to delineate a future program for traffic characterization and modeling with emphasis on achieving a comprehensive understanding of network traffic and workloads and their scaling behavior across multiple time scales.

Chapter 3 by Crovella and Lipsky serves to draw attention to issues surrounding simulation under self-similar traffic conditions which, in many instances, involve sampling from heavy-tailed distributions such as those arising in the context of generating long-range dependent traffic series as well as generating heavy-tailed workloads in related contexts (e.g., WWW). Slowly decaying tails lead to slow convergence of sample statistics to their corresponding population statistics, in fact, leading to biasedness in terms of underestimation. In other words, the sample mean is consistent but biased which has ramifications when performing simulations whose sampling frequency may not be sufficiently large. Crovella and Lipsky discuss various issues and possible remedies associated with this important practical problem.

1.6.2 Queueing with self-similar input

The chapter by Norros (Chapter 4) gives an updated overview of the fundamental queueing results associated with fractional Brownian motion (FBM) input processes, also called fractional Brownian storage. Due to the fundamental importance played by FBM and its increment process fractional Gaussian noise (FGN) as a model of *aggregate* traffic, understanding the queueing behavior under self-similar input as captured by FBM is of import to many other results, serving as a reference point. Norros derives the Weibullian tail behavior of the queue length distribution arising in fractional Brownian storage models with Hurst parameter in the range $1/2 < H < 1$, and discusses the importance of the Gaussian property of the input process and the role that it plays in the analysis.

In Chapter 5, Brichet, Massoulié, Simonian, and Veitch give a more refined extension of Gaussian input processes stemming from superposition of on-off processes in the heavy-load case where the arrival rate is close to the service rate. As the number of i.i.d. on-off sources increases, they are able to derive limit results that characterize the queue length distribution asymptotics and show it to be Weibullian, consistent with the result in the FBM case.

The chapter by Boxma and Cohen (Chapter 6) discusses queueing behavior as a function of the service discipline, in particular, first-come-first-served (FCFS), processor sharing (PS), and last-come-first-served preemptive resume (LCFS-PR), in the context of the M/G/1 queueing model. Due to the close connection between heavy-tailed on-off processes and the M/G/1 system when the service time is heavy-tailed, the latter represents a natural queueing model in which to incorporate the impact of long-range dependence. Their main contribution is toward deriving heavy-traffic—utilization approaches 1—limit theorems for the M/G/1 queue under the aforementioned service disciplines and showing that the tail of the queue length distribution can indeed be significantly less heavy when using PS and LCFC-PR in place of FCFS. The conclusions advanced have potential applications to packet and workload scheduling in router and Web server design.

Chapter 7 by Resnick and Samorodnitsky investigates performance issues for three classes of input models where heavy tails induce long-range dependence on the input side and a single server works at constant rate. In particular, they consider a single on-off renewal process with heavy-tailed on-periods, a finite number of i.i.d. heavy-tailed renewal processes, and lastly, an infinite number of sources whose transmission times or on-periods are heavy-tailed. In all three cases, the input process is long-range dependent and queueing affected performance leads to polynomial (vs. exponential) dependence which agrees with related queueing results.

The chapter by Likhanov (Chapter 8) derives results for the queueing behavior for a broad class of asymptotic second-order self-similar processes—also related to the M/G/∞ model—which can be viewed as superposition of sessions that arrive in a Poisson fashion and whose session durations are heavy-tailed. Each session or connection, however, can be viewed as having a finite lifetime—after a burst of activity, it is silent forever thereafter—which distinguishes it from the on/off model where connections alternate between active and idle periods ad infinitum. The author establishes asymptotic bounds for the queue length distribution which relate the various parameters of the asymptotic second-order self-similar arrival process model to performance, giving rise to polynomially decaying tails of the queue length distribution.

In Chapter 9, Makowski and Parulekar present a comprehensive treatment of the large buffer asymptotics for the M/G/∞ input model—a constant rate server fed with M/G/∞ inputs—using large deviation techniques to approximate the tail behavior. They show that there exist compact relationships between buffer occupancy asymptotics and the service time distribution of M/G/∞, but the tightness of buffer asymptotics is influenced by the shape of the service time distribution—exponential or subexponential—the latter leading to upper and lower bounds that collapse only under certain restrictions. In addition to its technical content, the chapter provides a detailed discussion on why the M/G/∞ model is a useful queueing model under which to study performance issues relating to long-range dependence, and explicate their performance results with respect to other well-known results in buffer asymptotics for long-range dependent input. In particular, the authors point out that the same buffer asymptotics can be induced by vastly different input streams—long-range and short-range dependent.

Chapter 10 by Jelenković investigates the subexponential queueing behavior of very general queueing systems when subject to subexponential arrival processes. In particular, he considers the class of GI/GI/1 queueing systems and some of its variations, including finite buffers and truncated heavy-tailed arrivals. It is shown that the asymptotic approximation for the loss rate in a finite buffer GI/GI/1 queue is independent of the service process, and buffer capacity dimensioning—in certain buffer size regimes—can exert a significant impact on performance. The chapter concludes with a study of multiplexing behavior under long-range dependent input for fluid queues. It is shown that dominance is at play which allows simplified reasoning of multiplexing effects by suitable replacement of dominated input processes by more simple ones (e.g., constant bit rate).

The chapter by Jacquet (Chapter 11) is concerned with long-range dependent processes that are the superposition of an infinite number of suitably calibrated on-off sources with *exponential* on-off times. While in the networking context such constructions are artificial and have little in common with empirical information gained from measured traffic, it serves to show that mathematically—under *infinite* aggregation—light-tailed on- and off-periods can induce long-range dependence, and due to the inherently Markovian nature of the individual components they may be useful when studying the behavior of queueing systems. For example, Jacquet uses

Mellin transforms, a transform technique used in traditional teletraffic theory, to track the polynomial tails of the ensuing queue length distribution in a simple queueing system.

Chapter 12 by Heyman and Lakshman discusses the relative import of short-range and long-range dependence for performance analysis in certain networking contexts—e.g., where buffer capacities are sufficiently "small," for certain types of applications such as variable bit rate video, and for certain first-order performance measures such as long-term packet loss rate—where short-range dependence can dominate queueing. The existence of such regimes is not surprising but worthwhile remarking as with other manifold qualifications on the impact and role of self-similarity and heavy-tailedness discussed in the book. In these cases, given the choice between equally well-fitting short-range dependent "black-box" processes and long-range dependent "physical" models, the short-range dependent one can be effectively employed for performance analysis purposes. However, a change in the underlying assumptions—different networking context, application, or performance measures—renders this approach inflexible in contrast with physical models which are more robust and parsimonious. The limitation of short-range dependent black box models is most pronounced when studying "what if" questions, in particular, those involving traffic control with feedback.

Chapter 13 by Li and Li is the last chapter of the second part of the book and presents a transient analysis of queueing behavior under long-range dependent input. The analysis is "transient" in the sense that loss probabilities are derived *conditioned* on the state of the system at a previous time instance which facilitates tractability while utilizing the correlation structure present in the stationary i.i.d. increments of the long-range dependent input considered. This work can be viewed as an effort toward understanding genuinely transient phenomena in queueing systems with self-similar input, noting that complete transient analysis is a difficult task even for simple Markovian systems such as M/M/1 which involves modified Bessel functions. The practical relevance of transient results becomes obvious in the presence of heavy-tailed distributions which can slow down convergence to steady-state to the point where its value for engineering purposes is greatly diminished. Chapter 17—in the third part of the book—gives a complementary and more sophisticated form of transient analysis geared toward an application to traffic control.

1.6.3 Traffic control and resource provisioning

Chapter 14 by Park, Kim, and Crovella discusses how the causality of self-similar network traffic can be traced back to a high-level structural property of the underlying networked system, namely, the heavy-tailed nature of file size or Web document distributions at the application layer. The authors show that when objects sampled from such distributions are exchanged via the mediation of a "typical" protocol stack—application layer (e.g., FTP, HTTP), transport layer (e.g., TCP, flow-controlled UDP), network layer (e.g., IP)—with focus on the transport layer which governs congestion control and reliability (if so desired), then the transfer and manifestation of the application layer causal seed of self-similarity at multiplexing points in the network layer in the form of self-similar traffic is robust, being largely impervious to the details of the actions carried out in the protocol stack and network configuration. In conjunction with evidence of heavy-tailed file size distribution observed in UNIX file systems of the past—i.e., in the 1980's well before the onset of the World Wide Web and its constituent traffic—the chapter provides a causal, physical explanation for why self-similar traffic may be so ubiquitous, and is a phenomenon likely to persist in the future Internet. The chapter also shows that if the transport protocol behaves in "extreme" ways—e.g., minimal or no congestion control—it is possible for the protocol stack to exert sufficient influence such that the transfer mechanism of causality is significantly impeded.

Chapter 20 and its discussion of multifractal IP traffic with different scaling behavior at small and large time scales where the multiplicative scaling at small time scales—at roughly sub-RTT times—is taken to stem from the actions associated with TCP's feedback congestion control, can be viewed as a refined characterization of the influence of the protocol stack, in this case, for short-term correlation structure of network traffic.

In Chapter 15, Feldmann presents an empirical study of the characteristics of TCP connection arrivals and shows that in today's Internet, in addition to self-similarity at the packet level, self-similar scaling is already encountered at the session or application layer when analyzing time series of the number of TCP connections per time unit. To this end, Feldmann relies on wavelet-based inference techniques and uncovers that various facets of TCP connection arrival characteristics conform to Weibullian-type distributions. This detailed workload characterization is relevant from both traffic modeling and control perspectives since knowing the structure of TCP connection arrivals and their durations can help in devising improved traffic control mechanisms.

The chapter by Roberts (Chapter 16) gives a high-level discussion of traffic control and resource provisioning issues under long-range dependent traffic conditions. The basic premise is predicated on segmenting traffic into two broad classes—stream and elastic traffic—where the former are subjected to open-loop control, i.e., resource reservation, and the latter are handled using closed-loop control. Due to the heavy queueing cost associated with provisioning resources using buffer sizing, it is explicitly proposed that for stream traffic, bandwidth allocation with small buffer capacity be the default resource allocation policy employed. Roberts sketches the components of a multiservice network architecture advocating measurement-based admission control for stream traffic considered effective for self-similar traffic, and end-to-end feedback control for elastic traffic, with pricing applicable to both. The influence of self-similarity and heavy-tailedness on architectural considerations and traditional traffic control are discussed throughout.

In Chapter 17, Duffield and Whitt adopt a bufferless model for performance analysis and traffic control where instantaneous offered load is given by a long-term level process (i.e., "DC" component shifts across a range of traffic levels) with "within-level" fluctuations. They investigate the problem of approximating the conditional mean of aggregate traffic—conditioned on past traffic profile or demand parameterized by level and age (or duration)—using numerical transform inversion. They show that the age variable plays an important role in facilitating prediction. Duffield and Whitt show applications of the "transient analysis framework" by estimating the probability of high levels of congestion in steady state using a large deviation principle approximation. They also analyze the converse situation captured by the time to recover—i.e., return to a traffic level corresponding to a given resource capacity—after the excursion. The approach advanced in Chapter 17 is interesting due to its focus on the long-range dependent aggregate input process, dispensing with its impact on queueing, and directly analyzing the transient or dynamic variability structure based on the predictability inherent to long-range dependent processes. A similar "on-line" framework is adopted in Chapter 18 where long-term predictability structure is exploited for feedback congestion control.

Tuan and Park (Chapter 18) show that in spite of the "bad news" associated with scale-invariant burstiness, there is "good news" in the sense of there being the potential of exploiting long-term correlation structure present in long-range dependent traffic for traffic control purposes. They advance the *multiple time scale congestion control* framework and show that nonnegligible correlations at large time scales can be effectively detected on-line and engaged to improve the performance of feedback congestion controls in rate-based settings. The central idea underlying the technology is *selective aggressiveness control* which allows explicit prediction of large time scale network state to be used to modulate the aggressiveness of bandwidth consumption behavior exhibited by feedback congestion control acting at small time scales (i.e., time scale of RTT). An important consequence is the mitigation of the delay-bandwidth product problem of reactive controls in broadband wide area networks.

Finally, Chapter 19 by Adas and Mukherjee addresses the problem of how resource reservation in a time-division multiplexing set-up—per-VC framing—can be used to facilitate end-to-end quality of service (QoS) under long-range dependent traffic conditions. The asynchronous framing approach described follows the resource provisioning paradigm espoused for long-range dependent traffic, namely, that of bufferless queueing, which then allows computation of QoS guarantees by appealing to the Central Limit Theorem and equivalent bandwidth computations.

References

1. P. Abry and D. Veitch. Wavelet analysis of long-range dependent traffic. *IEEE Trans. Information Theory*, 44(1):2–15, 1998.

2. A. Adas and A. Mukherjee. On resource management and QoS guarantees for long range dependent traffic. In *Proc. IEEE INFOCOM '95*, pages 779–787, 1995.

3. R. Addie, M. Zukerman, and T. Neame. Fractal traffic: measurements, modelling and performance evaluation. In *Proc. IEEE INFOCOM '95*, pages 977–984, 1995.

4. Gene Amdahl. Validity of the single-processor approach to achieving large scale computing capabilities. In *AFIPS Conference Proc.*, pages 483–485, 1967.

5. V. Anantharam. Queueing analysis with traffic models based on deterministic dynamical systems. In *Proc. 25th Allerton Conference on Communication, Control and Computing*, pages 233–241, 1996.

6. M. Arlitt and C. Williamson. Internet Web servers: workload characterization and performance implications. *IEEE/ACM Trans. Networking*, 5(5):631–645, 1997.

7. P. Barford and M. Crovella. Generating representative workloads for network and server performance evaluation. In *Proc. ACM SIGMETRICS '98*, pages 151–160, 1998.

8. Michael Barnsley. *Fractals Everywhere*. Academic Press, 1988.

9. Jan Beran. *Statistics for Long-Memory Processes*. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, NY, 1994.

10. J. Beran, R. Sherman, M. S. Taqqu and W. Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications* **43**, pp. 1566–1579, 1995.

11. Ernst Biersack. Performance evaluation of forward error correction in ATM networks. In *Proc. ACM SIGCOMM '92*, pages 248–257, 1992.

12. D. R. Cox. Long-range dependence: a review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. Iowa State Univ. Press, 1984.

13. M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. In *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, May 1996.

14. M. E. Crovella and A. Bestavros Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking* **5**, pp. 835–846, 1997.

15. M. Crovella and L. Lipsky. Long-lasting transient conditions in simulations with heavy-tailed workloads. In *Proc. 1997 Winter Simulation Conference*, 1997.

16. N. G. Duffield, J. T. Lewis, N. O'Connel, R. Russell, and F. Toomey. Statistical issues raised by the bellcore data. In *Proc. 11th IEE Teletraffic Symposium*, 1994.

17. N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. *Mathematical Proc. of the Cambridge Phil. Soc.* **118**, pp. 363–374, 1995.

18. N. Duffield and W. Whitt. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems*, 26:69–104, 1997.

19. N. Duffield and W. Whitt. A source traffic model and its transient analysis for network control. *Stochastic Models*, 14:51–78, 1998.

20. A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. Networking*, 4:209–223, 1996.

21. A. Erramilli, R. Singh, and P. Pruthi. An application of deterministic chaotic maps to model packet traffic. *Queueing Systems*, 20:171–206, 1995.

22. A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger. Dynamics of IP traffic: A study of the role of variability and the impact of control. In *Proc. ACM SIGCOMM '99*, pp. 301–313, 1999.

23. A. Feldmann, A. C. Gilbert, and W. Willinger. Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic. In *Proc. ACM SIGCOMM '98*, pages 42–55, 1998.

24. A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz. The changing nature of network traffic: Scaling phenomena. *Computer Communication Review* **28**, pp. 5–29, 1998.

25. M. R. Frater, P. Tan and J. F. Arnold. Variable bit rate video traffic on the broadband ISDN: Modelling and verification. In J. Labetoulle, J. W. Roberts, editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, pp. 1351–1360, Elsevier, Amsterdam, The Netherlands, 1994.

26. M. Garret and W. Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proc. ACM SIGCOMM '94*, pages 269–280, 1994.

27. R. J. Gibbens. Traffic characterization and effective bandwidths for broadband network traces. In F. P. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications*, pages 169–179. Clarendon Press, Oxford, 1996.

28. A. C. Gilbert, W. Willinger, and A. Feldmann. Scaling analysis of conservative cascades, with applications to network traffic. *IEEE Trans. Information Theory*, 45(3):971–991, 1999.

29. M. Grossglauser and J-C. Bolot. On the relevance of long-range dependence in network traffic. In *Proc. ACM SIGCOMM '96*, pages 15–24, 1996.

30. M. Harchol-Balter. Process lifetimes are not exponential, more like $1/t$: implications on dynamic load balancing. Technical report, EECS, University of California, Berkeley, 1996. CSD-94-826.

31. M. Harchol-Balter and A. Downey. Exploiting process lifetime distributions for dynamic load balancing. In *Proceedings of SIGMETRICS '96*, pages 13–24, 1996.

32. D. Heyman and T. Lakshman. Source models for VBR broadcast video traffic. *IEEE/ACM Transactions on Networking*, 4(1), June 1996.

33. D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Transactions on Networking*, 4(3):301–317, June 1996.

34. C. Huang, M. Devetsikiotis, I. Lambadaris, and A. Kaye. Modeling and simulation of self-similar variable bit rate compressed video: a unified approach. In *Proc. ACM SIGCOMM '95*, pages 114–125, 1995.

35. Philippe Jacquet. Analytic information theory in service of queueing with aggregated exponential on/off arrivals. In *Proc. 25th Allerton Conference on Communication, Control and Computing*, pages 242–251, 1996.

36. R. Jain and S. Routhier. Packet trains—measurements and a new model for computer network traffic. *IEEE J. Select. Areas Commun.*, 4(6):986–995, 1986.

37. P. Jelenkovic and B. Melamed. Automated TES modeling of compressed video. In *Proc. IEEE INFOCOM '95*, pages 746–752, 1995.

38. Leonard Kleinrock. *Queueing Systems, Volume 1: Theory*. Wiley-Interscience, New York, 1975.

39. T. G. Kurtz. Limit theorems for workload input models. In F. P. Kelly, S. Zachary, and I. Ziedins, editors, *Stochastic Networks: Theory and Applications*. Clarendon Press, Oxford, 1996.

40. W. E. Leland and T. J. Ott. UNIX process behavior and load balancing among loosely-coupled computers. In O. J. Boxma, J. W. Cohen and H. C. Tijms, editors, *Teletraffic Analysis and Computer Performance Evaluation*, pp. 191–208, Elsevier, Amsterdam, The Netherlands, 1986.

41. W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of ethernet traffic. In *Proc. ACM SIGCOMM '93*, pages 183–193, 1993.

42. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

43. N. Likhanov, B. Tsybakov, and N. Georganas. Analysis of an ATM buffer with self-similar ("fractal") input traffic. In *Proc. IEEE INFOCOM '95*, pages 985–992, 1995.

44. R. Lukose and B. Huberman. Surfing as a real option. In *Proc. 1st International Conference on Information and Computation Economies*, pages 45–51, 1998.

45. B. Mandelbrot and J. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Rev.*, 10:422–437, 1968.

46. B. B. Mandelbrot. Long-run linearity, locally gaussian processes, h-spectra and infinite variances. *Intern. Econom. Rev.*, 10:82–113, 1969.

47. B. B. Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman and Company, New York, 1982.

48. D. Menasce and V. Almeida. *Capacity Planning for Web Performance: Metrics, Models, and Methods*. Prentice Hall, 1998.

49. I. Norros. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.

50. K. Park, G. Kim, and M. Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Proc. IEEE International Conference on Network Protocols*, pages 171–180, 1996.

51. K. Park, G. Kim, and M. Crovella. On the effect of traffic self-similarity on network performance. In *Proc. SPIE International Conference on Performance and Control of Network Systems*, pages 296–310, 1997.

52. K. Park and W. Wang. QoS-sensitive transport of real-time MPEG video using adaptive forward error correction. In *Proc. IEEE Multimedia Systems '99*, pages 426–432, 1999.

53. M. Parulekar and A. Makowski. Tail probabilities for a multiplexer with self-similar traffic. In *Proc. IEEE INFOCOM '96*, pages 1452–1459, 1996.

54. M. Parulekar and A. Makowski. M/G/∞ input processes: A versatile class of models for traffic network. In *Proc. IEEE INFOCOM '97*, 1997.

55. M. Parulekar and A. Makowski. Tail probabilities for a multiplexer driven by M/G/∞ input processes (i): preliminary asymptotics. *Queueing Systems*, 27:271–296, 1997.

56. V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. In *Proc. ACM SIGCOMM '94*, pages 257–268, 1994.

57. V. Paxson and S. Floyd. Wide-area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking* **3**, pp. 226–244, 1995.

58. B. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. In *Proc. ACM SIGCOMM '96*, pages 3–14, 1996.

59. Bo Ryu. Fractal network traffic modeling: past, present, and future. In *Proc. 25th Allerton Conference on Communication, Control and Computing*, pages 252–260, 1996.

60. G. Samorodnitsky and M. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman and Hall, New York, London, 1994.

61. A. Shaikh, J. Rexford, and K. Shin. Load-sensitive routing of long-lived IP flows. In *Proc. ACM SIGCOMM '99*, pages 215–226, 1999.

62. M. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review*, 26:5–23, 1997.

63. M. S. Taqqu and J. B. Levy. Using renewal processes to generate long-range dependence and high variability. In E. Eberlein and M. S. Taqqu, editors, *Progress in Prob. and Stat. Vol. 11*. Birkhauser, Boston, 1996.

64. M. S. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: an empirical study, 1995. Preprint.

65. B. Tsybakov and N. D. Georganas. On self-similar traffic in ATM queues: Definitions, overflow probability bound and cell delay distribution. *IEEE/ACM Trans. Networking*, 5(3):379–409, 1997.

66. B. Tsybakov and N. D. Georganas. Self-similar traffic and upper bounds to buffer overflow in an ATM queue. *Performance Evaluation*, 36(1):57–80, 1998.

67. T. Tuan and K. Park. Multiple time scale congestion control for self-similar network traffic. *Performance Evaluation*, 36:359–386, 1999.

68. T. Tuan and K. Park. Multiple time scale redundancy control for QoS-sensitive transport of real-time traffic. To appear in *Proc. IEEE INFOCOM '00*, 2000.

69. T. Tuan and K. Park. Performance evaluation of multiple time scale TCP under self-similar traffic conditions. Technical report, Dept. of Computer Scieces, Purdue University, 1999. CSD-TR-99-040.

70. A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis*. Chapman&Hall, London, 1995.

71. W. Willinger. The discovery of self-similar traffic. In G. Haring, C. Lindemann and M. Reiser, editors, *Performance Evaluation: Origins and Directions*, LNCS, Springer-Verlag, New York (to appear).

72. W. Willinger and V. Paxson. Discussion of ‘Heavy tail modeling and teletraffic data’ by S. I. Resnick. *The Annals of Statistics* **25**, pp. 1856–1866, 1998.

73. W. Willinger and V. Paxson. Where mathematics meets the Internet. *Notices of the AMS* **45**, pp. 961–970, 1998.

74. W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. In *Proc. ACM SIGCOMM ’95*, pages 100–113, 1995.