Applications of Fractals in Engineering for Realistic Traffic Processes

Ashok Erramilli, James Gordon and Walter Willinger

Bell Communications Research, 331 Newman Springs Rd., Red Bank, N.J., U.S.A.

Recent traffic measurement studies in Bellcore and elsewhere have shown that traffic characteristics and resource usage patterns in emerging networks and services are likely to be very complex and can be expected to include features that are more efficiently described in terms of fractal processes than conventional stochastic processes. We show that fractal characterizations are applicable to arrival, service and quality of service processes, and we relate them to several traffic engineering problems of current interest. Specifically, we illustrate that (i) operational measurements in existing packet networks are poor indicators of actual load, (ii) for heavy-tailed service time densities, convergence to Erlang B results may be too slow to be useful in practice, and (iii) long term loss probabilities are inadequate to describe packet loss processes.

1. INTRODUCTION

The rapid evolution of telecommunications technology over the last decade and, more specifically, the widespread deployment of packet based services from ISDN through ATM and beyond, have outpaced developments in teletraffic theory - at least to the extent it is practiced. This is in part due to the uncertainties in the traffic characteristics of emerging networks and services, and to the difficulties in characterizing the complexity of traffic arrival and resource usage patterns in emerging networks. Recent measurement studies in Bellcore and elsewhere [1-4] have indicated just how complex these traffic patterns can be. Analysis of traffic data from networks and services as diverse as ISDN packet networks, Ethernet LANS, CCSNs and Variable Bit Rate (VBR) video all convincingly demonstrate the presence of features such as self similarity, long range dependence, slowly decaying variances, heavy-tailed or power law distributions, and fractal dimensions. These features are more characteristic of fractal processes than of conventional stochastic processes. Conventional traffic processes are Markovian in nature and can be characterized by exponential decays. In contrast, measured traffic processes seem to be more accurately characterized by power law decays.

We demonstrate in this paper that ideas from fractals are in fact applicable to many aspects of teletraffic systems (e.g., arrival and service time processes, as well as quality of service) and allow for efficient and parsimonious modeling of actual packet traffic. Although classical Markovian traffic models can in principle always be used to accurately describe any finite set of traffic measurements, the resulting models needed to capture the fractal-like nature of measured traffic are bound to be very complex and highly parameterized. Ultimately, in the context of analyzing, designing and engineering communication networks, it is practical applicability that will determine the relative merits of fractal traffic models. To this end, we show in this paper that (i) existing packet network traffic measurements (which are typically conducted over 'coarse' time scales on the order of minutes) are inadequate, and we consider alternatives to fine time scale measurements that take into account fractal properties; (ii) heavy-tailed service time densities can lead to actual blocking levels that can be very different from those predicted by classical theory (despite the well known insensitivity of the Erlang B results), mainly because of very slow convergence to Erlang.
B results; and (iii) packet losses are bursty, and it may be insufficient to describe them solely in terms of long term loss probabilities.

The paper is organized as follows. Section 2 summarizes recent measurement results and provides some background on fractal properties; Section 3 considers several engineering impacts of these properties, and Section 4 provides a summary of our results with suggestions for further work.

2. PACKET TRAFFIC MEASUREMENTS AND THEIR FRACTAL PROPERTIES

In this section, we describe four network environments from which actual traffic measurements have been collected and summarize the major findings from recent extensive statistical analyses of these data sets.

2.1 Packet Traffic Measurements

Leland et al. [3] report on a data collection and analysis effort of high time-resolution traffic measurements on several Ethernet LANs (10 Mbps) at the Bellcore Morristown Research and Engineering Center (MRE) during a 4-year period, from March 1989 to March 1992. This traffic is representative of LAN traffic offered to a high-speed public network (e.g., SMDS or Frame Relay) supporting LAN interconnection services. A day-long trace from a moderately loaded network (about 10% utilization) typically yields tens of millions of packets; a number of uninterrupted week-long traces (with no loss of data) are available resulting in hundreds of millions of packets. For details about this data collection and analysis effort, see Leland et al. [3].

Another data collection and analysis project involving a collection of about 20 sequences of variable-bit-rate (VBR) video data (resulting in hundreds of thousands of video traffic measurements) is reported in Beran et al. [2]. VBR traffic is widely expected to be one of the major contributors to future ATM network traffic. The outstanding features of these traffic measurements are that the sequences were generated by different hardware or software codecs, that they represent a variety of different scene types (e.g. video conferencing, video phone, full motion broadcast or studio quality video) and that they typically consist of tens of thousands of observations (number of ATM cells per frame).

Meier-Hellstern et al. [1] present ISDN packet traffic measurements from a terminal-based office automation application. The entire data stream transmitted and received on a user’s ISDN D-channel (16 Kbps) was captured and accurately time-stamped. User applications were primarily mail or file retrieval and word processing. Users were secretaries, managers, and engineers, and each user was monitored for one week, 24 hours daily, using a commercially available monitor. Data were collected for a total of 8 users, resulting in a total of 44 calls and 116000 packets. For a more detailed description of the measurement configuration, data collection, and model fitting effort, see Meier-Hellstern et al. [1].

Duffy et al. [4] describe the collection and analysis of packet traffic measurements from the Common Channel Signaling Network (CCSN) which uses the Signaling System Number 7 (SS7) protocol for communication and operates at a speed of 56 Kbps. The data consist of individual time-stamped SS7 messages (to the millisecond accuracy) on links monitored at Signaling Transfer Points (STPs). Data were collected from four different working CCSN subnetworks and from a variety of different link-types, resulting (to date) in about 150 million SS7 messages. For further details about the data collection and an in-depth analysis of CCSN traffic, see Duffy et al. [4].
2.2 Fractal Properties of Packet Traffic

Extensive statistical analyses of the above-mentioned traffic data convincingly demonstrate the presence of self-similarity in our measured data (see for example Leland et al. [3]). Intuitively, self-similar packet traffic exhibits structural similarities across all (or at least a wide range) of time scales and manifests itself in the absence of a natural length of a “burst”; at every time scale ranging from milliseconds to minutes and hours, similar-looking traffic bursts are evident. More formally, for a covariance stationary stochastic process \( X = (X_1, X_2, X_3, \ldots) \) with autocorrelation function \( r(k) \), let \( X^{(m)} = (X_k^{(m)} : k = 1, 2, 3, \ldots) \) denote the new covariance stationary time series (with autocorrelation function \( r^{(m)}(k) \)) obtained by averaging the original series \( X \) over non-overlapping blocks of size \( m \). That is, for each \( m = 1, 2, 3, \ldots \) the aggregated process \( X^{(m)} \) is given by \( X_k^{(m)} = (X_{km-m+1} + \ldots + X_{km}) / m, k = 1, 2, 3, \ldots \). The process \( X \) is called exactly (second order) self-similar if \( r^{(m)}(k) = r(k) \) for all \( m = 1, 2, 3, \ldots \), \( k = 1, 2, 3, \ldots \). In other words, \( X \) is exactly self-similar if the aggregated processes \( X^{(m)} \) are indistinguishable from \( X \) at least with respect to their autocorrelation functions. We say that \( X \) is asymptotically (second order) self-similar if the autocorrelations of the aggregated processes \( X^{(m)} \) agree with the autocorrelations of \( X \) asymptotically (i.e., for large \( m \) and large \( k \)). The most striking feature of self-similar processes is that their aggregated processes \( X^{(m)} \) possess a non-degenerate correlation structure as \( m \to \infty \). This behavior manifests itself in a number of equivalent ways: (i) the variance of the sample mean decreases more slowly than the reciprocal of the sample size, i.e., \( \text{var} X^{(m)} = m^{-\beta} \), as \( m \to \infty \) and \( 0 < \beta < 1 \) (slowly decaying variances), (ii) the autocorrelations decay much more slowly than exponential, i.e., \( r(k) = k^{-\beta} \), as \( k \to \infty \), \( 0 < \beta < 1 \), implying a non-summable autocorrelation function \( \sum r(k) = \infty \) (long-range dependence), and (iii) the spectral density \( \tilde{S}(\tilde{f}) \) obeys a power-law behavior around the origin of the form \( S(f) = f^{-\alpha} \), as \( f \to 0 \), \( 0 < \alpha < 1 \) (1/f noise).

In stark contrast, typical packet traffic models currently considered in the literature can be characterized by the property that their aggregated processes \( X^{(m)} \) tend to second order pure noise, i.e., \( r^{(m)}(k) \to 0 \), as \( m \to \infty \), \( k = 1, 2, 3, \ldots \). As a result, the variance of their sample means decreases like the reciprocal of the sample means, their autocorrelations decay exponentially fast, and their spectral densities are finite at the origin.

Mathematically, self-similarity in measurements from aggregated traffic (e.g., Ethernet, ISDN, CCSN traffic) can be explained by a simple aggregation argument (see Leland et al. [3] for details); aggregating, for example, many elementary renewal reward processes (representing individual user traffic) produces self-similarity in the limit as the number of users increases. The crucial property that distinguishes the renewal reward process source model from commonly assumed source models is that the inter-renewal intervals (i.e., the lengths of the “active/inactive” periods) are heavy-tailed or, using Mandelbrot’s terminology, exhibit the infinite variance syndrome. More precisely, a random variable \( U \) is called “heavy-tailed” if \( P(U > u) = u^{-\alpha} \), as \( u \to \infty \), \( \alpha > 0 \). For example, the stable (Pareto) distribution with parameter \( 1 < \alpha < 2 \) satisfies this property and has infinite variance. Evidence in support of the infinite variance syndrome in packet traffic measurements already exists. For example, Meier-Hellstern et al. [1] observed that the extreme variability in the ISDN data mentioned above cannot be adequately captured using traditional packet traffic models but, instead, seems to be best described with the help of heavy-tailed distributions. Another - rather unexpected - source for heavy-tailed distributions is the CCSN traffic data mentioned above. In fact, Duffy et al. [4] show that the call holding time distribution for calls originating during high traffic periods is heavy-tailed with an estimated \( \alpha \)-value of about 2.0; for calls originating during light traffic periods, the estimated \( \alpha \)-
value drops down to about 1.0.

Finally, Erramilli et al. [5] propose using fractal dimensions (see Mandelbrot [6]) in order to characterize the fractal-like nature of the above traffic measurements. Intuitively, a dimension is an indication of the extent to which a set or an object (e.g., arrival times) fills the space in which it is embedded. For example, the widely used correlation dimension has been shown to be computationally feasible and it is associated with a measure known as the correlation integral \( C(g) \), defined as the number of pairs of points \( i, j \) whose distance is less than \( g \), i.e.,

\[
C(g) = \frac{1}{N^2} \sum_i \sum_j C_{ij} \]

where \( C_{ij} \) is 1 or 0 depending on whether the points indexed \( i, j \) in the set are separated by a distance less than \( g \), or greater than \( g \). The correlation dimension is then defined as

\[
D_c = \lim_{g \to 0} \frac{\log(C(g))}{\log(g)}
\]

if this limit exists. For self-similar sets, the correlation integral scales as \( C(g) \sim g^{D_c} \) over a range of time scales, and \( D_c \) is found by estimating the slope of a plot of \( \log(C(g)) \) against \( \log(g) \).

3. Applications

In this section, we will consider the impact of fractal properties on a number of different aspects of teletraffic systems: arrival streams, service times, and quality of service. In particular, we address below issues of: (i) measuring rates for bursty packet traffic; (ii) interpreting Erlang B-type results in the presence of heavy-tailed service times; and (iii) assessing QOS beyond long-term loss rates. (Additional potential impacts of fractal properties are discussed in [3]).

3.1. Measuring Packet Rates

Measuring traffic rates is a fundamental monitoring operation in all packet-based networks. In existing engineering practice, traffic levels are monitored by measuring either packet counts, or switch element occupancies. These measurements are typically made over intervals of the order of 15 to 30 minutes, even for high speed data networks. The collection and reporting of measurements on finer time scales is usually precluded by the limitations of Operational Support (OS) systems. Capacity exhaust of the monitored resource is indicated if busy period traffic levels consistently exceed a specified rate or utilization threshold, which is typically in the range 60% - 80% of the maximum throughput of the network element. This threshold is determined on the basis of prior testing or modeling. According to such measures, existing packet networks are often lightly loaded i.e., loaded well below recommended utilization thresholds. In spite of this, it is not unusual to observe traffic related problems such as occasional customer complaints of poor performance, and packet call blocking due to network congestion. The fact that traffic related problems occur at such low occupancy levels is one of the paradoxes of packet switching, and it can be related to the fractal nature of packet traffic.

Intuitively, the problem arises because packet traffic bursts occur over time scales much smaller than measurement intervals on the order of 30 minutes. As a result, traffic peaks are averaged out to much lower values. This is illustrated in figure 1, which shows successive one-second counts of a 15 minute segment of ISDN packet traffic, obtained by merging individual DTE traffic streams. The average rate over the 15 minute period is about 21 packets per second (pps), but it can be seen that the traffic consistently peaks well above this value, with a maximum rate of about 79 pps. All other Ethernet and ISDN data sets we have analyzed show the same qualitative behavior, and the data set in figure 1 is typical.
The performance / engineering consequences of this effect are illustrated by using the data stream of Figure 1 as input to a queueing simulation. An FCFS server with deterministic service times and unlimited buffers is assumed, and the utilization of the server is varied by changing the service time. The average waiting times in this system are plotted as a function of the 15 minute utilization in Figure 2. The simulated waiting times can be compared with the waiting times obtained using the Queueing Network Analyzer (QNA), which is based on standard two moment approximations for queueing systems. The arrival process is characterized by the mean arrival rate and the squared coefficient of variation $C_d^2$ (equal to 1.57 for the data set). Note that QNA uses more information than is typically available from operational measurements. The agreement with the waiting time obtained from the actual data is nevertheless very poor. The capacity of the server, as indicated by the knee of the curve, is as much as 80%, according to the two moment model, whereas delays start to increase sharply around 50% with the actual data. Qualitatively similar results were obtained with every other ISDN and Ethernet data set we analyzed. One can conclude that existing measurements are poor indicators of actual network load, at least as they are currently interpreted.

This leads to several fundamental questions: (i) over what time scales should operational rate and utilization measurements be made for a given packet network? (ii) What is the minimum set of operational measurements that have to be collected to accurately assess network load and performance? (iii) How should these measurements be interpreted?

Given that current OSs do not have the capacity to collect frequent finer time scale measurements, an alternative is to exploit self similarity to infer finer time scale behavior from coarse time scale measurements. As an example, consider the traffic rate in an interval, conditioned on the information that the interval is centered on an arrival. The conditional rates and utilizations over short time scales can be inferred from longer time scale measurements using the following relation, valid over a range of time scales known as the scaling region [5]:

$$\lambda_2 = \lambda_1 \cdot \left(\frac{t_1}{t_2}\right)^{1-D_c}$$

where $0 \leq D_c \leq 1$ is the correlation dimension of the arrival process discussed in Section 2. Thus the conditional rate can be inferred by multiplying existing operational measurements by a scale factor, which depends on $D_c$ and the lower and upper cut-offs of the scaling region over which it is estimated. Such conditional or “local” rates may be better indicators of queueing performance than the long term average rate.

The advantage of this approach is that one can compensate for the inadequacies of existing measurements without proposing additional operational measurements given estimates of $D_c$ and the scaling region. Disadvantages include the error introduced by any extrapolation procedure, and the difficulties of estimating fractal dimensions and scaling regions. Recent analyses have indicated the significance of other fractal parameters in predicting queueing performance (Norros [7]), and it remains to be seen if this can be exploited in overcoming limitations with existing operational measurements.

### 3.2. Heavy-Tailed Service Densities

Heavy-tailed densities, which are characteristic of fractal point processes, are appropriate for modeling service times in a number of applications, besides the call holding times indicated by
analysis of CCS data (see [4]), and of individual call records (see Bolotin [9]). In general, circuit switched data services (for example, Switched 56 services) can be expected to have heavy-tailed service time densities. Even in packet based networks, there are resources that are held for the duration of a call or a session, which can range from milliseconds through several hours, for e.g., with Constant Bit Rate (CBR) services in ATM, call buffers in connection oriented networks, and modem pools that provide dial up access. Our experience with circuit switched data services indicates that there are numerous difficulties in accurately engineering these services, which arise in part due to the heavy-tailed service time densities. This is contrary to expectation, given the well known insensitivity of the Erlang B results to service time characteristics besides the mean rate. While these results hold asymptotically as long as the mean rate exists, in practice convergence to these results may be so slow that there can be considerable deviations from the theory over time scales of engineering interest.

To illustrate some of these difficulties, we simulated a 10-server queueing system with no waiting room, Poisson arrivals with unit rate and two types of service time distribution: exponential, and Pareto, defined by the following cumulative distribution function

$$\Pr[t \leq T] = 1 - \frac{\beta^\alpha}{(\beta + T)^\alpha}$$  \(3\)

where \(\alpha\) and \(\beta\) are parameters. To illustrate the range of possible blocking behaviour, we used a mean service time of 9.685, and Pareto service time distributions having three different values of \(\alpha\):

- \(\alpha = 2.5\), resulting in a distribution with finite mean and variance. The blocking in the exponential and Pareto systems both converge to the expected Erlang B result (0.2) within the engineering period (which may typically involve as few as \(\sim 10^3\) events) (Figure 3a).
- \(\alpha = 1.5\) corresponding to a distribution with finite mean and infinite variance. The Pareto system converges so slowly (\(\sim 10^6\) arrivals) to the Erlang B result that one can significantly underestimate the true blocking probability in typical engineering intervals (Figure 3b)
- \(\alpha = 0.5\) in this pathological case \(\alpha = 0.5\), both the mean and variance of the service time distribution are infinite. The utilization is ill-defined, but \(\beta\) was chosen so that the 95th percentiles of the Pareto and exponentials matched. The Pareto blocking probability starts off very low, and then tends asymptotically to one (Figure 3c). In this case, one could obtain any value between 0 and 1 for the blocking probability, depending on the interval over which measurements were taken.

In principle, one could resolve the rate of convergence problem (assuming there is convergence) by extending the length of the engineering period. However, over longer observation intervals, assumptions about the stationarity of arrival processes do not hold, and the Erlang B results are once again inapplicable (see Smith [10] and Davis et al. [11]). Smith identifies another problem, which we have not modeled, arising from high reattempt rates (caused by auto dialers, for example). With heavy-tailed densities, reattempts are more likely to be correlated, inflating the observed blocking rate. The development of accurate engineering procedures in the presence of heavy-tailed service time distributions will have to account for these and related problems. It remains to be seen if fractal scaling of the service rate processes is applicable here. Note that the convergence from below observed in the Pareto system (Figure 3) arises because the service rate
over smaller time intervals, conditioned on a departure, can be much greater than the long term rate.

3.3 Packet Losses

Packet loss processes are well known to be very highly bursty, but packet losses are nevertheless almost always characterized by their long term rates. Ramaswami et. al. [12, 13] identify the problem of serial correlations in losses, and demonstrate the limitations of using long term rates to describe packet losses. In addition to this problem, which arises because of periodicities in the arrival process, all the limitations of long term average rates in describing bursty processes (as shown by Mandelbrot for transmission errors [6], and as argued here for packet arrival processes) also apply in the case of packet losses. Briefly, any packet loss rate measurement is likely to be arbitrary over a wide range of time scales, and the long term rate (for e.g., ATM cell loss rates of the order of 10^{-10}) may be too low to be physically meaningful. This is a conclusion also drawn by Cohen and Heyman [14].

As with the cases of transmission errors and packet arrivals, fractal characterizations are applicable in describing packet loss processes. To illustrate this point, we have analyzed the loss processes in queueing simulations driven by Ethernet traffic traces. Table 1 gives the results of a correlation dimension analysis for two data sets. The burstiness of the loss process is indicated by the fractional correlation dimension. Table 1 also demonstrates that when packet losses occur, they occur at much higher rates than the long term rates, and correspondingly the impact on applications will be considerably more than that indicated by the long term rate. Other fractal parameters (such as the Hurst parameter) may also be applicable to characterize the loss process. We are currently analyzing such issues in more depth, using more extensive data sets from simulations, as well as packet loss traces from the Internet.

<table>
<thead>
<tr>
<th></th>
<th>#pkts</th>
<th>#Losses</th>
<th>Loss Prob.</th>
<th>Loss rate, worst 1 sec interval</th>
<th>Correlation Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>2500000</td>
<td>4749</td>
<td>0.019</td>
<td>0.34</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>1000000</td>
<td>7315</td>
<td>0.0073</td>
<td>0.27</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

3.4 Fractal Queueing

The presence of fractal properties in actual arrival, service time and quality of service processes may serve as a motivation for the development of a 'fractal queueing' theory, to analyze, for example, the performance implications of processes with long range dependence. One possibility is that fractal properties impact performance indirectly, by biasing long term traffic measurements; they can then be accounted for by scaling or transforming inputs to conventional queueing models. Alternately, one could consider directly analyzing models that use fractal characterizations as inputs. Given the lack of a Markovian structure, such models will be extremely difficult to analyze. There are, however, two promising approaches, one based on Fractional Brownian Motion models (see Norros [7]), and the other based on a dynamical sys-
tems approach using chaos theory [15]. The development of newer analysis techniques is perhaps
necessary to fully assess the performance and traffic engineering impacts of all the features found
in actual traffic processes.

4. Summary

In order to keep pace with the rapid development of telecommunications technology, tele-
traffic theory must address the complexity of actual traffic processes in ways that can be applied
in practice. The complexity of actual traffic processes is indicated by the presence of the so-called
fractal properties in recent measurement studies of traffic from ISDN packet, Ethernet, Variable
Bit Rate (VBR) Video and CCS networks. In practical terms, actual traffic has features that span a
wide range of time scales and can be described in terms of power law densities and correlations;
in contrast, the range of time scales that can be efficiently (i.e., with a small number of parameters)
captured by conventional models is limited by the underlying Markovian structure. This complex-
ity is perhaps most naturally and efficiently addressed in terms of fractal models.

In this paper, we considered the practical impacts of such fractal properties. We illustrated
the inadequacy of existing operational measurements in packet networks, and considered alterna-
tives based on fractal measures. We also considered the impact of heavy-tailed service densities on
blocking in loss systems, and showed that convergence to Erlang B may be so slow that the results
may not be useful in practice. We discussed the inadequacy of describing packet losses in terms of
long term rates, and proposed alternate fractal characterizations. Understanding the impacts of
such fractal properties is essential to resolving a number of traffic engineering problems of practi-
cal interest, and one can speculate on the need for a "fractal queueing" theory to facilitate this.

ACKNOWLEDGEMENTS

We are indebted to Kathy Meier-Hellstern of AT&T Bell Labs for providing the ISDN data,
Jean Bolot of INRIA for providing Internet packet loss data, Will Leland and Dan Wilson of
Bellcore for the Ethernet data, Ali Tabatabai for the VBR data, and Diane Duffy and Allan McIn-
tosh of Bellcore for the CCSN traffic data.

REFERENCES


and Long Range Dependence,” accepted for publication in IEEE /ACM Trans. on Communica-
tions (subject to revisions), 1993.


CCSN/SS7 Traffic Data from Working CCS Subnetworks”, IEEE J. Select. Areas Commun., 1994
(to appear).


