The Wiley Encyclopedia of Telecommunications

Network Traffic Modeling

Michael Devetsikiotis
ECE Department, NC State University, Raleigh, NC 27695-7911, USA

Nelson L. S. da Fonseca
Institute of Computing, State University of Campinas, Brazil

Abstract

Telecommunication networks are built to carry "traffic" in the form of data being transmitted. Efforts in network design, control or management require decisions and optimization actions that in turn require traffic *models*. This is why the science and "art" of traffic modeling has been playing a crucial role in the area of communication network design and operation. Traffic modeling has an analytical aspect, whereby suitable stochastic models are devised, and attributed to different types of data sources and network types; and a computational aspect where actual parameter values are estimated from measured data.

In this article, the most common models used in network traffic modeling are presented and basic concepts on traffic characterization are provided. Models for individual traffic sources as well as for aggregates of sources are also introduced. Finally, a brief discussion on the state-of-the art and on future trends is presented.

Keywords

Network traffic, stochastic processes, renewal models, regression, long-range dependence, self-similarity, envelope processes, effective bandwidth.

Cross-references

Network congestion control, traffic characterization, network simulation, network performance

1 Introduction

Traffic, that is, data being transmitted, is what telecommunication networks are built to carry. The fascinating and unprecedented "Internet revolution" has led to an ever increasing need for larger amounts of data to be transmitted, as well as fast increasing expectations in terms of the diversity and quality of the transmitted data. Modern networks are expected to accommodate a very heterogeneous traffic mix including traditional telephone calls, data services, world-wide web browsing, and video or other multimedia information. In this context, network designers and telecommunication engineers are called upon to design, control and manage networks of increasing transmission speed ("bandwidth"), size and complexity. Any effort in network design, control or management requires decisions and optimization actions that in turn require accurate prediction of the performance of the system under design or control. This is why the science and "art" of traffic modeling has been playing a crucial role in the area of communication network design and operation [7].

The amount of traffic per unit time arriving at a network access point, the number of Internet access requests in an hour or the traffic *workload* through an Internet provider's nodes ("routers" or "switches") is a real physical quantity, despite the fact that it consists of bits and bytes and *not* of atoms or molecules. This physical quantity is highly variable with time and space, and appears irregular or, *random.* Furthermore, network traffic usually exhibits visual clusters of activity separated by less active intervals, what is described in the telecommunications lingo, as *bursty* behavior. In order to predict the performance of networks carrying this variable and diverse traffic, researchers and telecommunication engineers utilize analysis (closed-form mathematics), numerical approximations, computer simulation, experimentation with real systems (in the laboratory or in the field), and heuristic or *ad hoc* projections based on past experience. All of the above require, to a larger or smaller degree, some representation or abstraction of *real-life* network traffic, that is, traffic "models".

Traffic modeling, has a theoretical/analytical aspect, whereby suitable stochastic models are devised in the mathematical sense, and attributed to different types of data sources and network types. Each model has a number of parameters that determine specific aspects such as mean value, higher moments, autocorrelation function, marginal density, etc. Such models include [1]:

- Renewal models
- Markov and semi-Markov processes
- Autoregressive processes (AR, ARMA, and ARIMA)
- Specially invented processes like Transform-Expand-Sample (TES), SRP, DAR and other
- Long-range dependent, self-similar and multifractal processes

There are also key *computational* and statistical aspects to traffic modeling: After deciding on or hypothesizing about a model (or model family) in the abstract, particular values have to be chosen for the parameters of the model. This usually means performing *matching* or *fitting* where parameter values are estimated statistically from the measured traffic data. Depending on the number of parameters involved, the type of model, and the nature of the data, this task may be far from straightforward and quite time-consuming. The moments to be estimated also depend on the type of network and traffic source, and represent an assumption in themselves. Typical traffic sources include:

- Voice, very important for its dominant presence in telephone networks
- Video, especially digital, compressed video (e.g., MPEG)
- Data applications such as FTP, TELNET, SMTP, HTTP
- Traffic in local area and campus networks (LAN and MAN)
- Aggregated traffic on network "trunks" over wide area networks (WAN)

In this article, we present the most common stochastic processes used for traffic modeling, in Section 2. Such processes can be used either to model the aggregate traffic of several sources (flows, connections, calls) on a network link or can be used to model individual sources, such as the stream generated by a phone call. Models for specific sources are introduced in Section 3. Some special aspects of traffic modeling related to network performance, namely the concepts of *effective bandwidths* and *envelope processes* are discussed briefly in Section 4. Finally, conclusions and some current open and challenging issues are discussed in Section 5.

2 Traffic Models

2.1 General Background

Traffic modeling starts usually by a researcher or telecom engineer collecting samples of traffic during a period of time ("traffic traces") from a specific source and/or at a specific point in the network (e.g., access point, router port or transmission link). Before stochastic modeling is applied, care must be taken to remove *determinism* and identifying the "residual uncertainty" [16] so that what remains to be modeled is truly stochastic in nature, and *stationary* (i.e., does not have fundamental properties that change with time of the day or month). At a second step, the data is analyzed and a stochastic model is proposed so that a realization of the stochastic process matches the data trace. A theoretical traffic model has to be checked against several data traces before one can be confident of its accuracy. In what follows, we present stochastic processes commonly used to describe traffic streams.

Network traffic can be *simple* or *compound*. Simple traffic corresponds to single arrivals of discrete data entities (e.g., "packets") and is typically described as a *point process* [9], that is, a sequence of arrival instants $T_1, T_2, \ldots, T_n, \ldots...$, with $T_0 = 0$. Point processes can be described equivalently by counting processes and interarrival time processes. A counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, non-negative integer-valued stochastic process, where $N(t)$ is the number of traffic arrivals in the interval $(0, t]$. An interarrival time process is a real-valued random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$ is the length of the time interval separating the n-th arrival from the previous one.

Compound traffic consists of *batch arrivals*, that is, multiple units possibly arriving simultaneously at an instant T_n. In the case of compound traffic, we also need to know the real-valued random sequence $\{B_n\}_{n=1}^{\infty}$, where B_n is the (random) number of units in the batch.

In some cases, it is more appropriate or convenient to assume that time is *slotted*, which leads to *discrete-time* traffic models. This means that arrivals may take place only at integer times T_n and inter-arrival periods are also integer valued. Furthermore, there are cases where the natural structure of the traffic is such that interarrival times are deterministic or periodic, with only the amount of arriving *workload* changing from arrival to arrival (e.g., compressed video "frames", arriving every 1/30th of a second).

A simple way to represent a stochastic process is to give the moments of the process. More particularly, the first and the second moments which are called the mean, variance and autocovariance functions. The mean function of the process is defined by $\mu_t = E(X_t)$. The variance function of the process is defined by $\sigma_t^2 = E[(x_t - \mu_t)^2]$ and the autcovariance function between X_{t_1} and X_{t_2} is defined by $\gamma(t_1, t_2) = E[(X_{t_1} - \mu_{t_1})(X_{t_2} - \mu_{t_2})]$.

Another topic that is very relevant in traffic modeling is that of traffic "burstiness". Burstiness is present in a traffic process if the interarrival times process $\{A_n\}$ tends to give rise to runs for several short interarrival times followed by relatively long ones. With typical network traffic exhibiting patterns and bursts that co-exist over many magnitudes of time scales (from minutes to hours to days) come the notion of time scale invariance. Time scale refers to the change or immunity to change of the process structure upon scaling of the time axis. A process $\{X_t\}$ can be defined as scaling invariant if for some $\alpha \in [a, b]$ the process is equal in distribution to its scaled version $\{X_{\alpha t}\}$. If a traffic is not scale invariant then when studying its behavior as time scales increase, it will show that the bursts and random fluctuations degenerate toward a white noise, non-bursty type of traffic.

The marginal distribution of a process $\{X_t\}$ captures the steady state first order distribution of X and is considered the primary characteristic in describing network traffic. Assuming that the process is Wide-Sense Stationary (WSS), the marginal distribution becomes invariant to time and is then defined by the one dimensional probability density function (PDF): $f_X(x) = f_{X_t}(x) = \frac{d}{dx} Pr[X_t \leq x]$. The PDF describes the probability that the data will assume a value within some given range at any instant of time.

The autocorrelation function of a process $\{X_t\}$ captures the second order measurement of the process and it is used as a supplement to the marginal distribution. The autocorrelation function for network traffic describes the general dependence of the values at another time. Assuming the process is WSS, then the autocorrelation between the data values at times t and $t + k$ is defined as follows:

$$\rho(k) = \frac{E[X_t X_{t+k}] - (E[X_t])^2}{E[(X_t - E[X_t])^2]}$$

where k is called the “lag”, the difference or distance between time points under consideration. If the autocorrelation function $\rho(k)$ of $\{X_k\}$ is equal to zero for all values of $k \neq 0$, then $\{X_k\}$ is of the *renewal* type. Markov and other *short-range dependent* (SRD) models have a correlation structure that is characterized by an *exponential* decay, which leads to $\sum_k \rho(k) < \infty$.

On the other hand, many real traffic traces exhibit *long-range dependence* (LRD) and can be modeled by self-similar and multifractal models later in this article. For these processes, the autocorrelation function decays slowly (say, polynomially instead of exponentially) in a way that makes the autocorrelation non-summable, i.e., $\sum_k \rho(k) \to \infty$ [23].

2.2 Short Range Dependent Models

2.2.1 Renewal Models

Renewal models have been used for a long time due to their simplicity and tractability. For this type of traffic, the inter-arrival times are independent and identically distributed (IID), with an arbitrary distribution. The major modeling drawback of renewal processes is that the autocorrelation function of A_n is *zero* except for lag $n = 0$. Hence, renewal models do not generally capture the behavior of high speed network traffic in an accurate manner.

Within the renewal family, *Poisson* models are the oldest and most widely used, having been historically closely linked to traditional telephony and the work of A. K. Erlang. A Poisson process is a renewal process with *exponentially* distributed interarrival times with rate λ, i.e., $P[A_n \leq t] = 1 - e^{-\lambda t}$. It is also a counting process with $P[N(t) = n] = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$, and independent numbers of arrivals in disjoint intervals. Poisson processes are very appealing due to their attractive memory and aggregation properties.

2.2.2 Markov Models

Unlike renewal traffic models, Markov and Markov-renewal traffic models [7, 1] introduce dependence into the random sequence A_n. Consequently, they can potentially capture traffic burstiness, due to non-zero autocorrelations of A_n. Consider a Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space. M behaves as follows: it stays in state i for an exponentially distributed holding time which depends on i alone; it then jumps to state j with probability p_{ij}, such that the matrix $P = [p_{ij}]$ is a probability matrix. In a simple Markov traffic model, each jump of the Markov process corresponds to an arrival, so interarrival times are exponentially distributed, and their rate parameter depends on the state from which the jump occured. Arrivals may be single, a batch of units or a continuous quantity.

Markov-modulated models constitute another important class of traffic models. Let $M = \{M(t)\}_{t=0}^{\infty}$ be a continuous-time Markov process, with state space of $1, 2, \ldots, m$. Now assume that while M is in state k, the probability law of traffic arrivals is completely determined by k. Thus, the probability law for arrivals is *modulated* by the state of M. The modulating process can be more complicated than a Markov process (so the holding times need not be restricted to exponential random variables), but such models are far less analytically tractable.

The most commonly used Markov-modulated model is the MMPP (Markov-Modulated Poisson Process) model, which combines a modulating (Markov) process with a modulated Poisson process. In this case, while in state k of M, arrivals occur according to a Poisson process of rate k. As a simple example, consider a two-state MMPP model, where one state is an "on" state with a positive Poisson rate, and the other is an "off" state with a rate of zero. Such models have been widely used to model voice traffic sources.

A semi-Markov process is a generalization of Markov processes, that allows the holding time to follow an arbitrary probability distribution. This destroys the Markov property since times are not exponentially distributed, however it allows for more general models of traffic. When values from a semi-Markov chain are generated, the next state is chosen first, followed by a value for the holding time. If the holding times are ignored, then the sequence of states will be a discrete time Markov chain, referred to as an *embedded* Markov chain.

2.2.3 Autoregressive Models

The autoregressive model of order p, AR(p), is a process $\{X_t\}$ whose current value is expressed as a finite linear combination of previous values of the process plus a white noise process ϵ_t: $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t$ where ϕ_i are constants and X_{t-i} are past values of the process at time $t - i$. The recursive form of this model makes it a popular modeling candidate as it makes it straightforward to *generate* an autocorrelated traffic sequence, e.g., variable-bit-rate (VBR) video traffic [1, 7]. However, autoregressive models cannot simultaneously match the empirical marginal distribution of arbitrary traffic such as video.

Another model of the same family is the Autoregressive Moving Average model of order (p, q), denoted by ARMA(p, q): $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$. Due to the larger number of parameters, ARMA models are more flexible than AR models and can be used in more cases. However, estimation of its parameters is more involved.

2.2.4 TES: Transform-Expand-Sample

Transform-Expand-Sample (TES) models represent another important class of models appropriate for modeling autocorrelated traffic streams. This family of models aims to capture *both* autocorrelation and *marginal distribution* of the empirical traffic trace, in fact it was historically the first traffic model explicitly devised to accomplish exactly this dual purpose and specifically for network traffic data. TES models capture stationary, correlated time series and also allow one to generate synthetic streams of real-looking traffic streams to drive simulations of networks [7].

TES models include two types of TES processes: TES$^+$ and TES$^-$. TES$^+$ produces sequences with positive autocorrelation at lag 1, while TES$^-$ produces negative autocorrelation at lag 1. The TES$^+$ process is more suitable for modeling network traffic. To define the TES$^+$ process, we first introduce a *modulo-1* operation. The *modulo-1* of a real number x, denoted by $< x >$, is defined as $< x >= x - \lfloor x \rfloor$, where $\lfloor x \rfloor$ is the maximum integer less than x. The recursive construction of the background TES$^+$ process is defined by:

$$U_n^+ \quad = \quad \begin{cases} U_0^+ & n = 0 \\ < U_{n-1}^+ + V_n > & n > 0 \end{cases}$$

where $\{V_n\}$ is a sequence of IID random variables referred to as *innovations* and U_0^+ is uniformly distributed on $[0, 1)$ and independent of $\{V_n\}$. The resulting sequence $\{U_n^+\}$ has a $[0, 1)$ uniform marginal distribution, and autocorrelation function determined by the probability density function $f_V(t)$ of V_n. The choice of $f_V(t)$ determines the correlation structure of the resulting process. From this background sequence the output process of the model referred to as the foreground sequence, $\{X_n^+\}$ is created by "distorting" each U_n^+ by $X_n^+ = F^{-1}(U_n^+)$, where F is the marginal distribution of the empirical data [9].

2.2.5 Other Short Range Dependent Models

Another interesting model is the Spatial Renewal Process (SRP). SRP efficiently models processes exhibiting arbitrary marginal distribution and aperiodically decaying autocorrelation (see [20] and references therein).

A Discrete Autoregressive model of order p, denoted as DAR(p), generates a stationary sequence of discrete random variables with an arbitrary probability distribution and with an auto-correlation structure similar to that of an AR(p). DAR(1) is a special case of DAR(p) process: it has a smaller number of parameters than general Markov chains, simpler parameter estimation, and can match arbitrary distributions. Moreover, the analytical queuing performance is tractable ([1] and references therein).

2.3 Long Range Dependent and Self-Similar Traffic Models

Measurements and statistical analysis of real traces performed during the 1990's revealed that traffic exhibits large irregularities (*burstiness*) both in terms of extreme variability of traffic intensities as well as persistent autocorrelation. Network traffic often looks extremely irregular at different time scales [12, 17] and such extreme behavior is not exhibited by the traditional Poisson traffic which smoothes out when aggregated at coarser time scales. If traffic were to follow a Poisson or Markov arrival process, it would have a characteristic burst length which would tend to be smoothed by averaging over a long enough time scale. Instead, measurements of real traffic indicate consistently that significant traffic burstiness is present on a wide range of time scales.

This behavior is reminiscent of and has been modeled according to *self-similar* processes. Self-similar or *fractal* modeling has been used in a number of research areas such as hydrology, financial mathematics, telecommunications, and chaotic dynamics [4, 24]. Internet traffic and more generally broadband network traffic, is an area where fractal modeling has become popular more recently. Such modeling has also been related to the observation of on-off traffic with "heavy tailed" distribution [23].

2.3.1 Heavy-Tailed ON-OFF Models

The fractal nature of network traffic is consistent with and predicted by the behavior of the individual connections that produce the aggregate traffic stream. In WAN traffic, individual connections correspond to "sessions", where a session starts at a random point in time, generates packets or bytes for some time and then stops transmitting. On the other hand, in LAN traffic, individual connections correspond to an individual source-destination pair. Individual connections are generally described using simple traffic models such as ON-OFF sources.

Traditional ON-OFF models assume finite variance distributions for the duration of the ON and the OFF periods. The aggregation of a large number of such processes results in processes with very small correlations. On the other hand, a positive random variable Y is called heavy-tailed with tail index α, if it satisfies: $P[Y > y] = 1 - F(y) \approx cy^{-a}$, $y \to \infty$, $0 < \alpha < 2$, where $C > 0$ is a finite constant independent of y. This distribution has infinite variance. Furthermore, if $1 < \alpha < 2$, then it has a finite mean. The superposition of many such sources was shown to produce aggregate traffic that exhibits long-range dependence and even self-similarity [12, 22].

2.3.2 Monofractal Models

Self-similarity in a process indicates that some aspect of the process is *invariant* under scale-changing transformations, such as "zooming" in or out. In network traffic, this is observed when traffic becomes bursty, exhibiting significant variability, on many or all time scales. The appeal and modeling convenience of self-similar processes lies in the fact that the degree of self-similarity of a series can be expressed using only one parameter. The *Hurst* parameter, H, describes the speed of decay of the series autocorrelation function. For self-similar series the value of H is between 0.5 and 1. The degree of self-similarity increases as the Hurst parameter approaches unity.

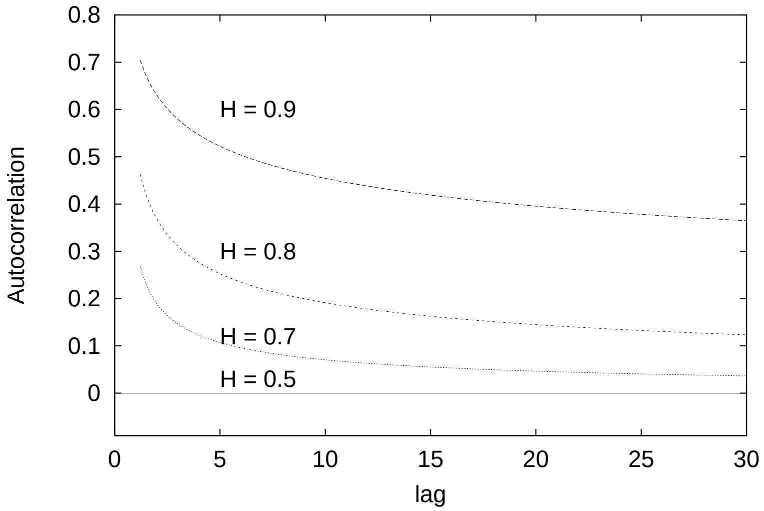

Figure 1: The autocorrelation as a function of time for different value of H.

A process $\{X_k\}$ whose autocorrelation function, $\rho(k)$, takes the form $\rho(k) \approx ck^{-\beta}$, $0 < \beta < 1$, for large k and a constant $c > 0$, is said to be *long-range dependent.* This implies that the autocorrelation function decays slowly and is not summable, thus $\sum_k \rho(k) \to \infty$. Figure 1 shows the autocorrelation as a function of time for streams with different H value. Note that for streams with greater H the autocorrelation decays slower as a function of time.

In the case of traffic traces, self-similarity is used in the distributional sense: when viewed at varying scales, the object's distribution remains unchanged. Equivalence in distribution between X and Y is denoted by $X \overset{d}{=} Y$. We provide in the following certain common definitions of self-similar traffic processes, following [22]: Let $X = \{X_t : t = 1, 2, 3, \ldots\}$ be a second-order stationary sequence with mean $\mu = E[X_t]$, variance $\sigma^2 = var(X_t)$, and autocorrelation function $r(k) = \frac{E[(X_{t+k}-\mu)(X_t-\mu)]}{\sigma^2}$. Let $X^{(m)}(t) = \frac{1}{m}(X_{tm-m+1} + \ldots + X_{tm})$, $m = 1, 2, 3, \ldots$, be the corresponding aggregated sequence with level of aggregation m, obtained by dividing the original sequence X into non-overlapping blocks of size m and averaging over each block. The index t labels the block. For each $m = 1, 2, 3, \ldots$ let $\mathbf{X}^{(m)} = \{X^{(m)}(k) : k = 1, 2, 3, \ldots\}$ denote the averaged process with autocorrelation function $r^{(m)}(k)$.

• A process X is called exactly second-order self-similar with parameter $H = 1 - (\frac{\beta}{2}),\ 0 < \beta < 1$ if its correlation coefficient is $r(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}]$, $k = 1, 2, 3, \ldots$.

• A strict-sense stationary process X is called strictly self-similar with parameter $H = 1 - (\frac{\beta}{2}),\ 0 < \beta < 1$, if $X \overset{d}{=} m^{1-H} X^{(m)}$. If X is strictly self-similar, then it is also exactly second order self-similar. The opposite is not true, except for Gaussian processes.

- A process X is called asymptotically second-order self-similar with parameter $H = 1 - (\frac{\beta}{2})$, $0 < \beta < 1$, if $lim_{m \to \infty} r^{(m)}(k) = \frac{1}{2}\delta^2(k^{2-\beta})$, $k = 1, 2, 3, \ldots$ where $\delta^2(f(x)) = f(x + \frac{1}{2}) - f(x - \frac{1}{2})$.
- A strict sense stationary process X is called strictly asymptotically self-similar if $X^{(m)} \overset{d}{=} X$, $m \to \infty$. Note that a strictly asymptotically self-similar process is not necessarily asymptotically second-order self-similar.

2.3.3 Fractal Gaussian Noise and Fractal Brownian Motion

The Fractal Brownian Motion (FBM) is a self-similar process with Gaussian stationary increment [14]. The increment process is called Fractal Gaussian Noise and its autocorrelation function is invariant under aggregation and is given by: $r(k) = 1/2[|k + 1|^{2H} - 2|k|^{2H} + |k - 1|^{2H}]$. The FBM process accurately models Ethernet, ATM and FDDI traffic, as well as video sources. The aggregate of on-off sources with heavy tails tends to an FBM.

The analysis of a queuing system with FBM input is quite challenging. However, it becomes manageable if the fractal Brownian traffic [15] process is used instead. The fractal Brownian traffic is defined as the fluid input in time interval $(s, t]$, and is given by $A(s, t) = m(t - s) + \sigma(Z_t - Z_s)$ where m is the mean input rate, σ^2 is the variance of traffic in a time unit and Z_x is a normalized fractal Brownian Motion, defined as a centered Gaussian process with stationary increments and variance $E[Z_t^2] = t^{2H}$.

2.3.4 Distorted Gaussian

The Distorted Gaussian (DGauss) model begins with a Gaussian process with a given autocorrelation structure and maps it into an appropriate marginal distribution. Examples of this popular traffic generation technique include the Autoregressive-To-Anything Process [20] and the self-similar traffic model in [8].

Many techniques exist to generate Gaussian time series (Gaussian in the marginal distribution) with a wide range of autocorrelation decay characteristics. A background Gaussian process Z_k is imparted with an autocorrelation structure $\rho'(t)$ and is run through a fitting function $X_k = F_X^{-1}(F_N(Z_k))$ to map its values into an appropriate distribution. Due to the background-foreground transformations, pre-compensation is applied to the background autocorrelation ρ' such that the resulting output autocorrelation ρ matches the desired specification [8].

2.3.5 Fractal Lévy Motion

In [11] the authors introduced, a teletraffic model which takes into account, in addition to the Hurst parameter $H \in [1/2, 1)$, the Lévy parameter $\alpha \in (1, 2]$. This was the so-called *fractional Lévy motion* (fLm), mentioned by Mandelbrot in [14]. Two important subclasses of Lévy motion exist: (i) the well-known ordinary Lévy motion (oLm), an α-stable process (distributed in the sense of P. Lévy) with independent increments, which is a generalization of the ordinary Brownian motion (the Wiener process), and (ii) the fractional Lévy motion, a self-similar and stable distributed process, which generalizes the fractional Brownian motion (fBm), has stationary increments and an infinite “span of interdependence”.

Several self-similar stable motions have been proposed for traffic modeling. These processes combine, in a natural way, both scaling behavior and extreme local irregularity.

2.4 Multifractal Models

Historically following self-similar models, researchers have been studying also the possibility of modeling network traffic with *multi fractal* processes (see [16] and references within). It appears that even though measured network traffic is consistent with asymptotic self-similarity, it also exhibits small time scaling features that differ from those observed over larger time scale. This small-time scaling behavior has been related to communication protocol-specific mechanisms and end-to-end congestion control algorithms that operate at those small time scales (less than a few hundred milliseconds). Modeling network traffic with multifractals has the potential of capturing the observed scaling phenomena at large as well as small time scales and thus to naturally extend and improve the original self-similar models of measured traffic.

To quantify the local variations of traffic at a particular point in time t_0, let $Y = \{Y(t), 0 < t < 1\}$ denote the traffic rate process representing the total number of packets or bytes sent over a link in an interval $[t_0, t_0 + t]$. The traffic has a *local scaling component* $\alpha(t_0)$ at time t_0 if the traffic rate process behaves like $t^{\alpha(t_0)}$ as $t \to 0$. In this context, $\alpha(t_0) > 1$ relates to instants with low intensity levels or small local variations, and $\alpha(t_0) < 1$ is found in regions with high level of burstiness or local irregularities.

If $\alpha(t_0)$ is constant for all t_0, then the traffic is *monofractal.* Equivalently, if $\alpha(t_0) = H$ for all t_0, then the traffic is exactly self similar, with Hurst parameter H. On the other hand, if $\alpha(t_0)$ is not constant and varies with time, the traffic is *multifractal.*

The multifractal appearance of WAN traffic is attributed to the existence of certain multiplicative mechanisms in the background. Multifractal processes are well modeled using multiplicative processes or "conservative cascades". The latter are a fragmentation mechanism, which preserves the mass of the initial set (or does so in the expected value sense). The generator of the cascade is called the fragmentation rule and the mathematical construct that describes the way mass is being redistributed is called the limiting object or multifractal. Modern data networks together with their protocols and controls can be viewed as specifying the mechanisms and rules of a process that fragments units of information at one layer in the networking hierarchy into smaller units at the next layer, and so on.

Multifractal processes are a generalization of self-similar processes. Hence, self-similar processes are also multifractal, but the reverse is not always true. This leads to the important modeling question: which of the two type of models is more appropriate in a given case? A method for distinguishing between the two models was proposed in [19]. Their conclusion was that traffic traces from environments were well modeled using self-similar models and that more sophisticated models such as multifractals were not needed. On the other hand, in WAN environments, there were cases where self-similar models were not deemed adequate and where multifractal models appeared to be more appropriate.

2.5 Fluid Traffic Models

In fluid traffic modeling, individual units such as packets, are not explicitly modeled. Instead, traffic is viewed as a "stream of fluid" arriving at a certain *rate* that may be changing. Fluid models can simplify analysis due to their lower "resolution" or level of detail. More importantly, fluid models can make network simulation much more efficient, since the computer representation of the fluid traffic requires much fewer "events" (e.g., rate changes) that need to be tracked.

In modern high speed networks such as Asynchronous Transfer Mode (ATM) networks, the size of individual packets is often fixed and very small (e.g., 53 bytes), relative to the total transmission speed and aggregate volume of information being transmitted (e.g., hundreds of megabits or gigabits per second). Therefore fluid modeling may be appropriate in such cases and, in general, whenever individual packets can be thought of as effectively insignificant with respect to the total traffic. The validity of this approximation depends very much on the time scale involved as well as the point of interest inside the network (e.g., access points versus large routers in the middle of the network).

Fluid models [9] typically assume that sources are bursty, commonly of the "on-off" type. In the "off" state, there is *no* traffic arriving, while in the on state traffic arrives at a constant rate. To maintain analytical tractability, the durations of "on" an off periods are assumed exponentially distributed and mutually independent (that is,they form an alternating renewal process).

3 Source Models

In this section, the modeling of different type of traffic sources is discussed. The flow generated by some network sources are regulated by the stack of protocols used in the network. Such type of sources is called elastic sources. Sources whose flow do not depend on network protocol are called streaming sources. First, streaming multimedia sources are introduced, followed by the modeling of elastic sources. At last, a discussion on a general characterization of traffic streams, called effective bandwidth, is given.

3.1 Data

Data streams were traditionally modeled by Poisson processes. The rationality behind it was that the superposition of several independent renewal processes tends to a Poisson process.

The nature of traffic changes as new applications becomes a significant part of the network traffic. SMTP, e-mail, Telnet and FTP were responsible for most of the traffic in pre-web time. As the use of web services became predominant, Internet traffic presents new patterns. Most of today network traffic is based on TCP. Internet traffic observed at long time scale exhibits self-similarity. However, at short time scales, typically shorter than a round trip time, Internet traffic presents high variability. At short time scales, Internet traffic marginal distribution is non-Gaussian and the scaling exponent of the variance is smaller than the asymptotic exponent. In other words, at short time scales, Internet traffic exhibits multi-fractal scaling, with different moments of the traffic showing scaling described by distinct exponents. However, at high frequencies it can be modeled as self-similar, level [5] [4]. Such behavior is originated by the complex interaction between network protocols which governs the network flow and TCP sources.

3.2 Voice

The packet stream from a voice source can be characterized by an *on/off* model, i.e. during silent periods no packet is generated and during "talkspurts" periods packets are generated either at exponentially distributed intervals or at constant intervals depending whether compression algorithms are used or not. The residence time in each state is exponentially distributed.

A popular approach to analyze a multiplexer fed by several *on/off* sources is to use Markov modulated processes to mimic the superposition process. The arrival rates and the transition probabilities of the underlying Markov chain are defined in a way that certain statistics of the Markov modulated process have the same numerical value of the corresponding statistics of the superposition process. The advantage of adopting a two-state process is to keep the complexity of both the matching procedure and the queuing solution low. In a two state MMPP there are only four parameters to be determined: the arrival rate and the sojourn time in each state. Several procedures are available to set these four parameters. Most of procedures consider two super-states: the underloaded and the overloaded states [18]. In the overload state, the packet generation rate (due to the number of source in state *on*) exceeds the server capacity, whereas in the underload state it is below the server capacity.

3.3 Video

The bit rate of a video stream depends not only on the coding algorithm, but also on the level of activity of a scene. Whenever there is a scene change, a new scene has to be encoded, generating a high number of bits to be transmitted, and, consequently high bit rates.

The MPEG coding scheme is widely used for several types of applications. MPEG streams consist of a series of frames. In MPEG-2, there are three types of frames: Intracoded (I), Predictive (P) and Bidirectional (B). A periodic sequence of I, B, P frames is called Group of Pictures (GOP). An MPEG transmission consists of one GOP after the other. Typically I frames will have more bits than P and B frames, and B frames will have the least number of bits. The size of I frames can be approximated by a Normal distribution, whereas the size of B and P frames can be approximated either by a Gamma or by a lognormal distribution. Figure 3 illustrates the bit rate

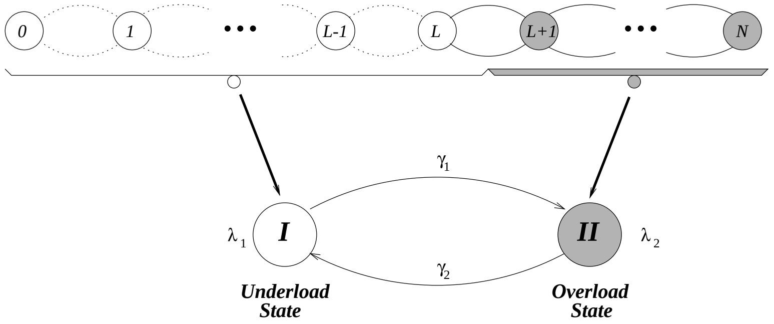

Figure 2: Modeling the Superposition of Voice Sources as a two-state MMPP. The States of the Original Markov Chain Represent the Number of Sources in State On and the Arrival Rate in the *nth* State is n Times the Arrival Rate in State On. The Superstates of the 2-State MMPP Correspond to Underload/Overload Periods Depending Whether or not the Aggregated Arrival Rate Surpasses the Channel Capacity

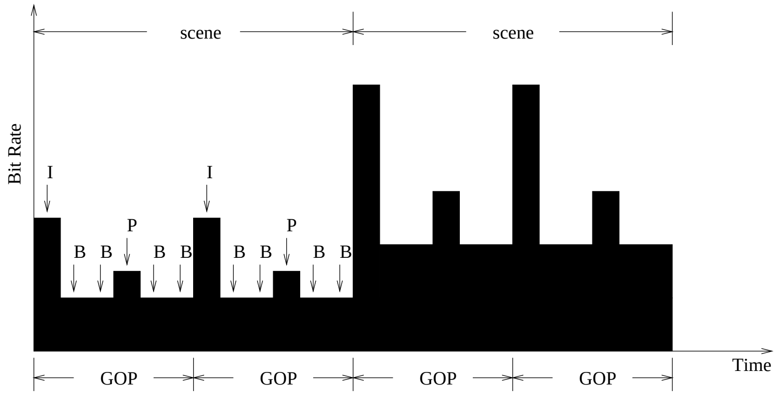

Figure 3: Bit Rate of a Video Transmission

profile of a typical video stream. The high peaks correspond to scene changes, whereas the low peaks correspond to the activity within a scene.

Video traffic exhibits long range dependencies [2] [8]. The repetitive pattern of GOPs introduces strong periodic components in the autocorrelation function (ACF). Video streams are usually modeled either by a fractal Brownian motion process or by a fractal ARIMA (0,d, 0) process, which are LRD processes. However, some researchers advocate that, for finite buffer, long-term correlation have minor impact on queuing performance, and, therefore, Markovian models should be used, since only short term correlations impact the performance. The discrete first-order autoregressive model, DAR(1), is a popular Markovian process used for video modeling. Actually, Markovian models give rise to ACF of the form $\rho(k) \sim e^{-\beta k}(\beta > 0)$, whereas an LRD process exhibits ACF of the form $\rho(k) \sim k^{-\beta} = e^{-\beta logk}(\beta > 0)$ [10]. In fact, the performance of fractal models may be overly sensitive to the buffer size, and, consequently, may underestimate the actual performance. On the other hand, Markovian models provide good performance under heavy loads; however they perform poorly under light loads

[10].

3.4 Elastic Sources

The amount of data an application can pump into the network is often regulated by the network protocols and their congestion control mechanisms which probe the available bandwidth to determine the amount of data which can be transmitted. Traffic sources whose transmission rate depend on network congestion status are called elastic sources. Examples of elastic sources are the Available Bit Rate service (ABR) in ATM networks and the Transmission Control Protocol (TCP), largely deployed in the Internet.

TCP congestion control mechanism is a window based one. Segments, packets in TCP language, are transmitted and acknowledgments from the receiver are expected. Each segment has a sequence number, set at the receiving end. Acknowledgments specify the sequence number of the acknowledged segment. Acknowldegements are cumulative, i.e., an acknowledgment notifies the transmitter that all the segments with a lower sequence number were properly received. The time from sending a packet to receiving its acknowledgment is called Round Trip Time (RTT). TCP controls a connection rate by limiting the number of transmitted-but-yet-to-be-acknowledge segments. In the beginning, the window size is set to one. Every time an acknowledgment is received, i.e., at every RTT, the window size is doubled, and the window grows up to a threshold. After this threshold, the window is incremented by one segment.

Whenever an acknowledgment does not arrive after a pre-defined interval, a timeout event occurs and the threshold is set to one-half the current congestion window and the congestion window is set to one. If the transmitter receives three consecutive acknowledgments for the same segment, it is assumed that the next segment was lost and the window is set to one-half its current value.

The evolution of the window size between loss events can be analyzed in order to determine a TCP connection throughput, i.e., the amount of data a TCP connection pumps into the network per unit of time by observing the window size evolution between loss events. The distribution of interloss periods as well as the distribution of the type of the loss event should be taken into consideration in this computation.

4 Effective Bandwidths and Envelope Processes

Most communications services are subject to performance constraints designed to guarantee a minimal quality of service (QoS). Consider a general traffic stream offered to a deterministic server, and assume that some prescribed parameterized performance constraints are required to hold. The effective bandwidth of the traffic stream corresponds to the minimal deterministic service rate, required to meet these constraints. Queuing-oriented performance constraints include bounds on such statistics as queuing delay quantiles or averages, server utilization, and overflow probabilities. The effective bandwidth concept serves as a compromise between two alternative bandwidth allocation schemes, representing a pessimistic and an optimistic outlook. The strict one allocates bandwidth based on the stream peak rate, seeking to eliminate losses, whereas the lenient one allocates bandwidth based on the stream average rate, merely seeking to guarantee stability.

Let us formally define effective bandwidth: Let $X[0,t]$ be the workload that arrived during time interval $[0,t]$ for a traffic stream. Effective bandwidth of the traffic stream is defined as $\alpha(s) = \lim_{t\to\infty} \frac{1}{st} \log E(e^{sX[0,t]})$, which is a function of $s > 0$, the so-called space parameter. In the effective bandwidth theory, s is the asymptotic exponential decay rate of queue size distribution tail probability with respect to queue size. That is, when the service rate is $\alpha(s)$, the queue size distribution tail probability with respect to queue size is $P(Q > B) \approx \exp(-sB)$, where Q denotes the queue size. This is why s is called the space parameter.

The notion of effective bandwidth provides a useful tool for studying resource requirements of telecommunications services and the impact of different management schemes on network performance. Estimates of effective bandwidths are called empirical effective bandwidths, while the analytical form are called analytical effective bandwidths.

Envelope Processes

An envelope process is a function which provides a bound for the amount of work generated in a traffic stream during a certain time interval. If $A(t_2 - t_1)$ is the amount of bits generated during the interval $t_2 - t_1$, $\hat{A}(t)$ is an envelope process for $A(t)$ if and only if $\hat{A}(t_2 - t_1) > A(t_2 - t_1)$, for any $t_2 > t_1$. For any traffic stream, $A(t)$, there is a whole family of possible envelope processes, however the lowest bound is the one of interest. Envelope processes are useful tools since, in general, they require a small number of parameters, i.e., they are a *parsimonious* way of representing a stochastic process. However, dimensioning based on envelope processes may overestimate the required resources. Moreover, envelope processes are not appropriate for the study of phenomena at the cell time scale, such as cell discarding.

Network services can be either deterministic or statistical. Accordingly, deterministic and stochastic envelope processes are defined. A deterministic envelope process is a strict upper bound on the amount of work arriving during an interval for a traffic stream. A commonly used envelope process is $\int_0^t A(t) < \rho t + \sigma$, where ρ is the source mean arrival rate and σ is the maximum amount of work allowed in a burst [3]. $\rho t + \sigma$ is a model for the output of a leaky bucket regulator where ρ is the leaky rate and σ the bucket size.

In a stochastic envelope process, the amount of work generated in a certain interval may surpass a deterministic bound with a certain probability value. An accurate stochastic envelope process for a fractal Brownian motion process is $\rho t + k\sigma t^H$ where ρ is the mean arrival rate, σ is the standard deviation and H is the Hurst parameter [6]. Note that the amount of work is not a linear function of time, it has a t^H which takes into account long periods of arrivals.

5 Conclusions

The aim of traffic modeling is to provide network designers with simple means to predict the network load, and consequently, the network performance. Since the early days of telephony network engineers have been engaged in understanding the nature of network traffic, and its impact on Quality of Service provisioning. Traffic models mimic the traffic patterns observed in real networks. The suitability of a traffic model is related to the degree of accuracy of the conclusions that can be drawn from studies using such a model. Therefore, there is no unique model for a certain type of traffic, but models with different degrees of accuracy.

With the advent of integrated networks, Poisson models for traffic streams were replaced by more sophisticated short range dependent models which considered the correlation pattern besides the mean arrival rate. By 1993, the seminal work of Leland et al. [12] demonstrated the fractal nature of LAN traffic. Several other works followed showing that other types of traffic such as video traffic were also fractal. Recent studies have shown that Internet traffic is not precisely fractal at small time scales, but can be represented well as fractal at larger time scales. The understanding of the impact of multifractality on network performance is still an open problem.

Traffic patterns are influenced by several factors such as the nature of file size, human think time, protocol fragmentation and congestion control mechanisms. New challenging problems in traffic modeling will certainly exist when multimedia applications become a significant part of the whole network traffic.

References

[1] A. Adas. Traffic Models in Broadband Networks. IEEE Communications Magazine, July 1997.

[2] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger. Variable-bit-rate video traffic and long range dependence. *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 1566-1579, 1995.

[3] R. L. Cruz. A calculus for network delay, part I: Network elements in isolation. In *IEEE Transactions on Information Theory*, volume 37, pages 114 – 131. IEEE, January 1991.

[4] A. Erramilli, O. Narayan, A. Neidhardt and I. Sanjee. Performance Impacts of Multi-Scaling in Wide Area TCP/IP traffic. in *Proc. INFOCOM'00* 2000.

[5] A.Feldmann, A.C. Gilbert, W.Willinger and T.G.Kurtz. Looking behind and beyond self-similarity: On scaling phenomena in measured WAN traffic. *in Proc. of the 35th Annual Allerton Conference on Communications, Control and Computing*, pp. 269-280, 1997

[6] N.L.S. Fonseca, G.S. Mayor and C.A.V. Neto. On The Equivalent Bandwidth of Self Similar Sources. In *ACM Transactions on Modeling and Computer Simulation*,vol 10, no 3, pp 104-124, 2000.

[7] V. Frost and B. Melamed. Traffic Modeling for Telecommunications Networks. IEEE Communications Magazine, March 1994.

[8] 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. SIGCOMM'95 Conf.*, pp. 114-125, 1995.

[9] D. Jagerman, B. Melamed and W. Willinger. Stochastic modeling of traffic processes. In *Frontiers in Queuing: Models, Methods and Problems, J. Dshalalow, Ed.*, CRC Press, 1996.

[10] M. M. Krunz and A. M. Makowski. Modeling Video Traffic using $M/G/\infty$ Input Process: A Comparison Between Markovian and LRD Models. *IEEE Journal on Selected Areas in Communications* vol. 16, no. 5, pp. 733-745, June 1998.

[11] N. Laskin, I. Lambadaris, F. Harmantzis and M. Devetsikiotis. Fractional Lévy Motion and its Application to Traffic Modeling. Computer Networks, Special Issue on Long-Range Dependent Traffic Engineering, 2002.

[12] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic (extended version). IEEE/ACM Trans. on Networking, vol. 2, no. 1, pp. 1–15, 1994.

[13] B. Liu, D.R. Figueiredo, Y. Guo, J. Kurose, D. Towsley. A Study of Networks Simulation Efficiency: Fluid Simulation vs. Packet-level Simulation. In *Proc. IEEE Infocom, Alaska,* Apr. 2001.

[14] B.B. Mandelbrot and J. W. Van Ness. Fractal Brownian motions, fractional noises and applications. In *SIAM Rev.*, 10:422-437, 1968.

[15] I. Norros. A Storage Model with Self-similar Input. In *Queuing Systems*, 16:387-396, 1994.

[16] K. Park and W. Willinger (Editors). Self Similar Network Traffic and Performance Evaluation. Wiley Interscience 2000.

[17] V. Paxson and S. Floyd. Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transaction on Networking*, vol. 3, no 3, pp. 226-244, 1995.

[18] J. A. Silvester, N. L. S. Fonseca, and S. S. Wang. D-Bmap models for the performance analysis of ATM networks. In Performance Modeling of ATM Networks. D. Kouvatsos editor, pp. 325-346, Chapman and Hall Publishers, 1995.

[19] M. S. Taqqu, V. Teverovsky and W. Willinger. Is Network Traffic Self-Similar or Multifractal? *Fractals*, pp. 63-73, 1997.

[20] T. Taralp, M. Devetsikiotis, and I. Lambadaris. In Search of Better Statistics for Traffic Characterization. Journal of the Brazilian Computer Society, special issue on Traffic Modeling and Control of Wired and Wireless Networks. Vol. 5, No. 3, pp. 5-13, April 1999.

[21] S. Tartarelli, M. Falkner, M. Devetsikiotis, I. Lambadaris and S. Giordano. Empirical Effective Bandwidths. Proceedings of IEEE GLOBECOM 2000, Vol 1 672-678.

[22] B. Tsybakov and N. D. Georganas. Self-Similar Processes in Communication Networks. IEEE Transactions on Information Theory, 44(5), September 1998.

[23] W. Willinger and V. Paxson. Where Mathematics Meets the Internet. Notices of the American Mathematical Society, 45(8), pp. 961-970, Sept. 1998.

[24] W. Willinger and M. S. Taqqu and R. Sherman and D. V. Wilson. Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Transactions on Networking*, vol 5, no 1, pp. 71-86, Feb 1997.