Statistical Modeling and Performance Analysis of Multi-Scale Traffic

Nelson X. Liu, *Member, IEEE,* and John S. Baras, *Fellow, IEEE*

***Abstract* — In this paper we propose a new statistical model for multi-scale traffic, and present an exact queueing analysis for the model. The model is based on the central moments and the marginal distributions of the cumulative traffic loads in different time scales. Only the first two moments are needed to characterize the traffic process, which greatly simplifies the representation and estimation. The queueing analysis uses a very general approach and can evaluate not only the steady state performance but also the transient queueing behavior. The analysis reveals that there exist two classes of packet losses, the absolute loss and the opportunistic loss, both of which can be examined exactly with the method. Based on the statistical model and the queueing analysis method, a compound model is constructed for the practical multi-scale traffic, and its performance is evaluated from various aspects. This work provides a good basis for practical application of the mutli-scale traffic characterizations in network dimensioning and resource management.**

***Index Terms*— Traffic modeling, multi-scale traffic, queueing analysis, log-normal distribution, loss probability**

I. Introduction

SCALING and multi-scale behaviors, such as the long-range dependence, the self-similarity, and the multi-fractality, have been commonly viewed as the most significant characteristics of the Internet traffic today [11] [13] [18] [19] [20]. They are found not only in the wired networks but also in the wireless networks [21] [29], the Ad hoc networks [12], and the satellite networks [26]. These behaviors generally mean that the traffic is bursty in many time scales and among many orders of statistics. They make the network performance much worse than that in traditional Gaussian and short-range dependent traffic environment. Modeling of scaling and multi-scale traffic is thus of great importance for planning, dimensioning, control, and performance guarantee of various types of networks.

Significant effort has been put on the research of scaling and multi-scale traffic in the last decade. Many important models have been proposed, such as the heavy-tailed on-off model [10] [14] and the factional Brownian motion [16] [27] for the scaling phenomenon, and the random cascade [8] and the multi-fractal wavelet model [24] for the multi-scale behavior. However, dealing with multi-scale traffic is a difficult task. Generally speaking, the following difficulties are facing many models. First, few models can address the full range of behaviors of practical traffic. In the examples above, the former two are used for the long-range behavior while the latter two emphasize the small scale behavior. Second, multi-scale traffic models often involve a number of parameters, which make it difficult to measure or estimate them. Third, exact queueing analysis is hardly possible. The asymptotic queueing analysis has become the dominant approach, and sometimes the only approach available. These significantly limit the application of the traffic models for network performance optimization in practice. Especially, the third problem is most serious because the queueing analysis relating the traffic parameters with the network performance is a key part in performance engineering. The asymptotic approach often gives an oversimplified picture of the relations for a complex traffic process like the multi-scale traffic. Some parameters of practical importance may be ignored in the analysis.

In this paper we propose a new framework for formulating scaling and multi-scale behaviors, and present a corresponding queueing analysis method. The work is intended to avoid difficulties mentioned above, and pave the way for fully taking advantage of scaling and multi-scale behaviors in traffic management, control, and network design. Two basic requirements are borne in mind in developing the model. First, it should keep key characteristics of scaling and multi-scale processes that are important for the network performance. Second, it should allow exact queueing analysis rather than just asymptotic analysis, though specific analytical technique may be needed. These two requirements being fulfilled, the model should also involve as few parameters as possible. Thus the traffic measurement and parameter estimation are made easy. In short, it is intended to be an engineering-oriented, performance evaluation friendly model.

For scaling and multi-scale traffic, what are most important for the network performance are actually the power-law behaviors of different orders of statistics at many time scales. Based on this insight, we model the traffic by solely characterizing the different orders of moments. To make the model more practical, the marginal distributions at the time scales of interest are taken into account. Then the model can be completely characterized with only the first two orders of moments. It can be shown that higher-order statistics, though not explicitly defined, have the power-law like behaviors in nature, which resemble those of the real traffic very well.

An exact queueing analysis method is presented for the

This work was supported by DARPA under Contract No. N66001-00-C-8063.

The authors are with the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mails: {nliu, baras} @ isr.umd. edu).

statistical multi-scale traffic model. In fact, it uses a fundamental approach and applies to an even more general class of traffic. For a stationary traffic process, if only the distribution of the cumulative traffic load is known, the queue content distribution at any time can be obtained with our method. It is especially suitable for analyzing the multi-scale traffic specified by the distribution-plus-two-moments model. The queueing result is given in integral form, which can be computed numerically. Our method can evaluate not only the steady state loss probability, but also the transient queueing behavior. It reveals that the packet losses of scaling and multi-scale traffic actually are of two types based on their origins: the absolute loss and the opportunistic loss. Behaviors of both can be examined with our method of analysis. As far as we know, this is the first time that an exact queueing analysis is presented for multi-scale traffic.

The paper is organized as follows. In Section II, the statistical model for scaling and multi-scale traffic is developed, and important properties of the model are studied. In Section III, queueing analysis of the traffic model is presented. Loss behaviors of building blocks of the model, persistent scaling and multi-scaling processes, are examined. Practical multi-scale traffic often combines scaling and multi-scaling behaviors in one process. Section IV presents a compound multi-scale traffic model and evaluates its performance from various aspects, including steady state loss probability, transient queueing behavior, heavy-traffic performance, and effects of buffer size. Section V concludes the paper.

II. Statistical Modeling of the Multi-scale Traffic

A. Scaling and Multi-Scaling Processes

A critical feature of the multi-scale scaling phenomena is that many orders of statistics of the traffic have power-law like behaviors that span many time scales. On one hand, this means the traffic has complex and strong dependence structures inherently. On the other hand, the traffic appears very bursty, and the burstiness looks similar at many time scales. These properties cause the network performance to be much worse than that in traditional Gaussian and short-dependent traffic environment. From the point of view of the stochastic analysis, the existence of the power-law like behaviors of the statistics is the key reason for this. Following this insight, we re-define the scaling and multi-scaling processes, and use them as components to construct the multi-scale process. We put the power-law like behaviors of statistics at the center of the formulation, and thus keeps the effects of the traffic on the network performance.

The definition is in terms of the central moments of the cumulative traffic load (CTL) process. Let $X(t)$ be the traffic rate at t. Then $W(t) = \int_0^t X(t)dt$ will be the arriving load up to t. Obviously, $W(t)$ is a non-decreasing process. Denote by $V(t, \Delta t) = W(t+\Delta t) - W(t)$. Assume the increment process is stationary, i.e., $V(t, \Delta t) = V(\Delta t)$. The average traffic rate is $\lambda = \lim_{\Delta t\to\infty}(V(\Delta t) / \Delta t)$. In the following description, $y(p) \sim z(p)$ means $\lim_{u\to p}(y(u) / z(u)) = c$, where $0 < c < \infty$ is a constant.

Definition 1: Given $T > 0$, $W(t)$ is said to be a *scaling* process up to order L at time scale T if there exists an integer $L > 0$, a constant $0 < \alpha(T) < 1$, and a small constant $\varepsilon > 0$ such that for any $\tau \in (T-\varepsilon, T+\varepsilon) \cap \tau > 0$ and any $0 < l \le L$

$$E[|V(\tau) - \lambda\tau|^l] \sim \tau^{l\alpha(T)} \tag{2.1}$$

For an applicable T, the scaling property of this process can be characterized with a vector $(\alpha(T), L)$. $\alpha(T)$ is the scaling exponent of the process.

Definition 2: Given $T > 0$, $W(t)$ is said to be a *multi-scaling* process up to order L at time scale T if there exist integers L, $M > 0$, a set $A = \{\alpha_i(T): 0 < \alpha_i(T) < 1, i \le M\}$, a set $\Phi = \{\phi_i(T): 0 < \phi_i(T) < 1, i \le M, \sum_{i=1-M}\phi_i(T) = 1\}$, and a small constant $\varepsilon > 0$ such that for any $\tau \in \{\tau: T-\varepsilon < \tau < T+\varepsilon, \ \tau > 0\}$ and any $0 < l \le L$

$$E[|V(\tau) - \lambda\tau|^l] \sim \sum_{i=1}^{M}\phi_i(T)\tau^{l\alpha_i(T)} \tag{2.2}$$

For a given T, a scaling property of this process can be characterized with (A, Φ). Φ is a set of scaling exponents of the process. Obviously, a scaling process is a multi-scaling process with a single scaling exponent $\alpha(T)$.

Let $c^{(l)}$ be a coefficient on the right hand side of (2.2) (or (2.1)) for it to become an equality. $\boldsymbol{c} = (c^{(1)}, c^{(2)}, \ldots, c^{(L)})$ is the vector of coefficients for all L orders of moments. Assume $W(t)$ has multi-scaling properties (A_1, Φ_1) and (A_2, Φ_2) at two time scales $T_1 > 0$ and $T_2 > 0$ ($T_1 \ne T_2$), and the coefficient vectors are $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, respectively. $W(t)$ is said to be *consistent* at T_1 and T_2 if $\boldsymbol{c}_1 = \boldsymbol{c}_2$ and $(A_1, \Phi_1) = (A_2, \Phi_2)$.

Definition 3: Given a time scale section S: (t_1, t_2), $W(t)$ is said to be a *persistent multi-scaling* process in S if it is consistent at any at two time scales $T_1 \in S$ and $T_2 \in S$ ($T_1 \ne T_2$). If only one scaling exponent exists for all time scales in S, it is called a *persistent scaling* process in S. A *persistent multi-scaling* (include *scaling*) process in $(0, \infty)$ is simply called the *persistent multi-scaling* (include *scaling*) process.

Obviously, to specify the scaling property of a persistent multi-scaling process in a given section S also needs a vector (A, Φ), the same as that for a time scale $T \in S$.

Using the scaling and the multli-scaling processes as components, we can construct the multi-scale process. It is always safe to say that a multi-scale process is a process that is multi-scaling at any valid time scale for it. But this does not help simplify the problem. Instead, we may construct the multi-scale process in the following way, though other constructions may also be possible.

Definition 4: $W(t)$ is said to be a *multi-scale* process if there exist a series of non-overlapped time scale sections S_i: $(t_{i,1}, t_{i,2})$, $i = 1, 2, 3, \ldots, N$, $N > 0$, such that $W(t)$ is a persistent multi-scaling (include scaling) process in section S_i. To cover the whole time line, let $t_{0,1} = 0$ and $t_{N,2} = \infty$. The scaling properties at separate points $t_{i,1}$ and $t_{i,2}$ are defined as being consistent

with the sections on their immediate right sides except at $t_{N,2}$, which is defined as being consistent with the section on its left side.

To define scaling and multi-scaling processes in terms of the central moment is the first important decision we made in this model. Based on previous expositions about scaling and multi-scale phenomena [9] [23] [24], one may be tempted to use the moments about the origin instead of the central moments. However, this would cause some problem. Suppose we change the definition (2.1) of the scaling process to the following

$$E[|V(T)|^l] \sim T^{l\alpha(T)} \quad (2.1')$$

Then for $l = 1$, $E[V(T)] \sim T^{\alpha(T)}$. When $T \to \infty$, $E[V(T)] / T \sim T^{\alpha(T)-1} \to 0$. But the practical traffic rate is a positive process. This conflict can be solved by using definition (2.1), from which we can get $E[V(T)] / T \to \lambda$.

The fractional Brownian motion (fBm) is an important model for self-similar traffic. As an example, the following theorem characterizes its scaling property in terms of above definitions.

Theorem 1: The fBm is a persistent scaling process.

Proof: Write the fBm with Hurst parameter $0 < H < 1$ as $W_H(at) = a^H W_H(t)$. Since its increment process is a stationary Gaussian process, we have $V(\Delta t) = W_H(\Delta t) = \Delta t^H W_H(1)$ and the mean is $\lambda = 0$. Therefore, $\forall T > 0$,

$$E[|V(T) - \lambda T|^l] = E[|V(T)|^l]$$
$$= E[|W_H(T)|^l] = E[|T^H W_H(1)|^l]$$
$$= T^{lH} E[|W_H(1)|^l] \quad (2.3)$$

This completes the proof.

□

The multifactal process used in [9] [23] [24] to model the small scale behavior can be viewed as a special case of the multi-scaling process defined here, i.e., a multi-scaling process up to infinite order at time scale $T \to 0$.

B. Model Restriction

To constrain the model with marginal CTL distributions is another important decision we made for the model. Analyses of real traffic traces have shown that the Internet traffic behaviors are time scale dependent. Two important time scales that are generally emphasized: the large or slow time scale from 100 ms up to several minutes, during which the traffic shows long-range dependence or self-similarity, and the small or fast time scale less than 100 ms, during which the traffic shows clear multi-scaling property. There may be a short transition phase around 100 ms, which is approximately the time scale of the roundtrip time. It is known that the increment process of the CTL is strongly Gaussian at the large scale and strongly non-Gaussian at the small scale [1] [7] [9] [15] [23]. This property is robust and can be found in almost every trace. Based on the multiplicative formulation of the multi-scale process [9] [23], it can be deduced that the non-Gaussian distribution in the small scale is roughly log-normal. With this knowledge, the modeling of the multi-scale behaviors can be significantly simplified. This is because that knowing all orders of moments does not necessarily mean we know the distribution. Knowledge about the distribution can greatly confine the model. This would make the model more specific, more practical, and easier for performance analysis.

Given the CTL having normal or log-normal distribution, we need only know the first and the second-order moments to completely decide its moments of all orders. If only we define the first two orders of moments according to the scaling or the multi-scaling process, the traffic behavior of such processes are fully specified. They have at least second order scaling property. Because both the distributions and the first two orders of moments can be validated in practice, this model gives us confidence that it is sufficient for practical traffic. The simple specification also has other advantages. First, it reduces the complexity of traffic estimation. Second, it avoids over-modeling. Generally speaking, high-order statistics of practical traffic may not strictly conform to a scaling or multi-scaling process, though there is likeness. Specifying the higher-order moments may create the problem of over-modeling. The distribution-plus-two-moments model need not explicitly define the higher-order moments and thus avoids the problem. In the following we will give the formal definition of the model. We will still use the terms "scaling process" and "multi-scaling process" but with restricted meanings. They will be used in this sense throughout the remaining part of this paper. Let μ and σ^2 represent the mean and the variance of $V(\Delta t)$.

Definition 5: Given $T > 0$, a cumulative traffic process $W(t)$ is said to be a *scaling* process at time scale T if all of the following are satisfied:

i) $W(t)$ has a stationary increment at time scale T, i.e., $V(t, T) = V(T)$.

ii) $V(T)$ has a normal distribution $N(\mu, \sigma^2)$:

$$f_{V(t)}(v) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(v-\mu)^2}{2\sigma^2}] \quad (2.4)$$

where μ and σ^2 satisfy the conditions iii) and iv).

iii) $\mu = \lambda T$ (2.5)

iv) There exist a constant $0 < \alpha(T) < 1$ and a small constant $\varepsilon > 0$ such that for any $\tau \in \{\tau: T-\varepsilon < \tau < T+\varepsilon, \ \tau > 0\}$

$$\sigma^2 \sim \tau^{2\alpha(T)} \quad (2.6)$$

Definition 6: Given $T > 0$, a cumulative traffic process $W(t)$ is said to be a *multi-scaling* process at time scale T if all of the following are satisfied:

i) $W(t)$ has a stationary increment at time scale T, i.e., $V(t, T) = V(T)$.

ii) $V(T)$ has a log-normal distribution $L(\varpi, \theta^2)$:

$$f_{V(t)}(v) = \frac{1}{\sqrt{2\pi}\theta v} \exp[-\frac{(\ln v - \varpi)^2}{2\theta^2}] \quad (2.7)$$

where ϖ and θ are constants related with T.

iii) μ and σ^2 satisfy the following conditions:

iii-a) $\mu = \lambda T$ (2.8)

iii-b) There exists an integer $M > 0$, a set $A = \{\alpha_i(T): 0 < \alpha_i(T) < 1, \ i \le M\}$, a set $\Phi = \{\phi_i(T): 0 < \phi_i(T) < 1, \ i \le M, \ \sum_{i=1-M}\phi_i(T) = 1\}$, and a small constant $\varepsilon > 0$ such that for any $\tau \in \{\tau: T-\varepsilon < \tau < T+\varepsilon, \ \tau > 0\}$ such that

$$\sigma^2 \sim \sum_{i=1}^{M} \phi_i(T)\tau^{2\alpha_i(T)} \tag{2.9}$$

(2.9) means there exists a probability measure for A, and $\alpha_i(T)$ occurs with the probability $\phi_i(T)$. The continuous version of (2.9) is

$$\sigma^2 \sim \int_{-\infty}^{\infty} f_{A(T)}(\alpha)\tau^{2\alpha} d\alpha \tag{2.10}$$

where $f_{A(T)}(\alpha)$ denotes the probability density function of the scaling exponent. This applies when infinitely many scaling exponents exist.

The parameters ϖ and θ of the multi-scaling process can not be directly measured. They can be decided through μ and σ^2. We can write down μ and σ^2 of the log-normal distribution as

$$\mu = \exp(\varpi + \theta^2/2) \tag{2.11}$$

$$\sigma^2 = \exp(2\varpi + \theta^2)[\exp(\theta^2) - 1] \tag{2.12}$$

Therefore,

$$\varpi = \ln \mu - \frac{1}{2}\ln(\frac{\sigma^2}{\mu^2} + 1) \tag{2.13}$$

$$\theta = \sqrt{\ln(\frac{\sigma^2}{\mu^2} + 1)} \tag{2.14}$$

With definition 5 and 6, new definitions of the persistent multi-scaling process and the multi-scale process follow, which are the same with definition 3 and 4 in words, but use restricted meanings of the scaling and the multi-scaling processes.

C. *Second-Order Moment of the Multi-scaling Process*

Behaviors of scaling and multi-scaling processes depend on their first- and second-order moments. Their first-order moments, as given in (2.5) and (2.8), behave in the same manner. Their second-order moments, however, display significant difference. The second-order moment of the scaling process, as given in (2.6), is simple. We will examine the behavior of that of the multi-scaling process. The continuous version of the scaling exponent distribution in (2.10) will be used. The discrete version in (2.9) can be analyzed in a similar way.

Many traffic analyses [8] [15] [23] reveal that the scaling exponent set A of a multi-scaling process spans a wide scope in [0, 1] and even goes beyond 1. There usually exists a central scaling exponent that holds the highest probability. The probability density function decreases rapidly on both sides symmetrically with the increase of the distance from the center. For simplicity, we assume the scaling exponents follow a normal distribution $N(\tilde{\alpha}, \tilde{\sigma}^2)$ at a time scale T, where $\tilde{\alpha}$ and $\tilde{\sigma}^2$ are the mean and the variance of the scaling exponents. We omit the subscript T for $\tilde{\alpha}$ and $\tilde{\sigma}^2$ here. Then for time scale T the integral (2.10) is

$$\sigma^2 \sim \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp[-\frac{(\alpha - \tilde{\alpha})^2}{2\tilde{\sigma}^2}] T^{2\alpha} d\alpha \tag{2.15}$$

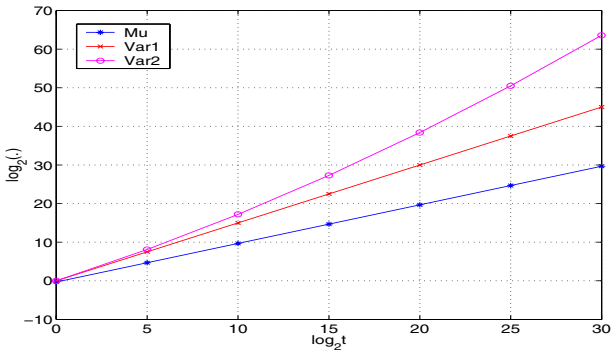

Fig.1. Increases of the mean (Mu) and the variances of the persistent scaling traffic (Var1) and the multi-scaling traffic (Var2) with time scale.

Let $z = T^{2\alpha}$. Then $\alpha = \ln(z) / (2\ln(T))$, and $d\alpha / dz = dz / (2\ln(T)z)$. Then (2.15) becomes

$$\sigma^2 \sim \int_0^{\infty} z \frac{1}{\sqrt{2\pi}(2\ln(T)\tilde{\sigma})z} \exp[-\frac{(\ln(z) - (2\ln(T)\tilde{\alpha})^2}{2(2\ln(T)\tilde{\sigma})^2}] dz \tag{2.16}$$

The right hand side of (2.16) is nothing but the expectation of a log-normal distribution with parameter $2\ln(T)\tilde{\alpha}$ and $(2\ln(T)\tilde{\sigma})^2$. In fact, if we view $z = T^{2\alpha}$ as a random variable that is a function of the random variable α, it is easy to deduce that z has a log-normal distribution. With the property (2.11) of the log-normal distribution, we can immediately write down

$$\sigma^2 \sim \exp[2\ln(T)\tilde{\alpha} + 2(\ln(T)\tilde{\sigma})^2] = \exp[2\ln(T)(\tilde{\alpha} + \ln(T)\tilde{\sigma}^2)]$$
$$= T^{2\tilde{\alpha}} T^{2\tilde{\sigma}^2 \ln(T)} \tag{2.17}$$

Comparing (2.17) with (2.6), we see if a persistent multi-scaling process and a scaling process have the same mean and the same average scaling exponent, the variance of the former is generally greater than that of the latter, especially when T is big. While the variance of the latter keeps a strict, constant scaling property, that of the former has a quasi, nonlinear scaling property, as can be seen clearly in figure 1. In the log-log graph, the variance of the scaling process ("Var1") is a straight line and that of the multi-scaling process ("Var2") is a nonlinearly increasing curve. The latter deviates from the mean ("Mu") faster than the former. Based on these differences, we may also expect that the multi-scaling process has worse network performance than the scaling process, as will be confirmed in Section III.

With (2.17) we can also decide the relations between the parameters ϖ and θ of the log-normal distribution and the scaling exponents and T. Referring to (2.13) and (2.14), we can write down

$$\varpi = \ln[\frac{\lambda T}{\sqrt{(c/\lambda^2)T^{-2(1-\tilde{\alpha})}T^{2\tilde{\sigma}^2\ln(T)} + 1}}] \tag{2.18}$$

$$\theta = \sqrt{\ln[(c/\lambda^2)T^{-2(1-\tilde{\alpha})}T^{2\tilde{\sigma}^2\ln(T)} + 1]} \tag{2.19}$$

where c is some finite constant.

D. *High-Order Moments*

An important feature of practical scaling or multi-scale

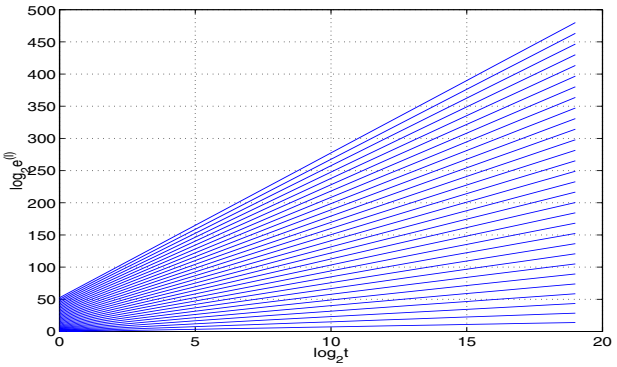

Fig.2. Log-log graph of different orders of central moments vs. time scale for the persistent scaling process (bottom-up: orders 1 ~ 30).

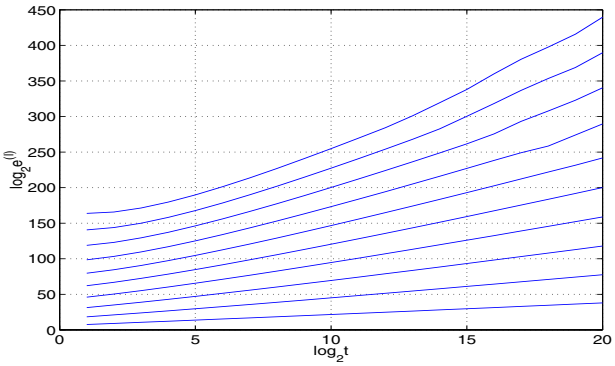

Fig.3. Log-log graph of different orders of central moments vs. time scale for the persistent multi-scaling process (bottom-up: orders 2, 4, …, 20).

traffic processes is that their high-order moments also show power-law like behaviors. Though not specifying it explicitly, the statistical multi-scale model holds this property inherently. Consider a normal random variable $Z \sim N(\mu, \sigma^2)$. Its absolute central moments are

$$E[|Z-\mu|^l] = \int_{-\infty}^{\infty} \frac{|z-\mu|^l}{\sqrt{2\pi}\sigma} \exp[-\frac{(z-\mu)^2}{2\sigma^2}]dz \quad (2.20)$$

Let $s = (z-\mu)/\sigma$. We get

$$E[|Z-\mu|^l] = \sigma^l \int_0^{\infty} \frac{2s^l}{\sqrt{2\pi}} \exp(-\frac{s^2}{2})ds \quad (2.21)$$

The integral on the right hand side is a Gaussian integral which we denote as $K(l)$. Then

$$E[|Z-\mu|^l] = K(l)\sigma^l \quad (2.22)$$

Developing the Guassian integral we get

$$K(l) = \begin{cases} (l-1)!!, & l = even \\ \frac{2^{l/2}[(l-1)/2]!}{\sqrt{\pi}}, & l = odd \end{cases} \quad (2.23)$$

Therefore, for a scaling process at time scale T, using (2.5) and (2.6) we can write down

$$E[|\Delta W(T) - \lambda T|^l] \sim K(l) T^{l\alpha(T)} \quad (2.24)$$

So the scaling process still keeps the scaling property strictly in high order moments.

Now let us see the multi-scaling process. Consider a log-normal random $Z \sim L(\varpi, \theta^2)$. We still use μ and σ^2 to represent its mean and variance. Its high order non-central moments are

$$E[Z^l] = \exp(l\varpi + l^2\theta^2/2) \quad (2.25)$$

With the relation between the central moments and the non-central moments, we can write

$$E[(Z-\mu)^l] = \sum_{k=0}^{l} C_l^k (-1)^k \mu^k E[Z^{l-k}] \quad (2.26)$$

where $C_l^k = l!/(k!(l-k)!)$. For simplicity, we test only even order central moments, which equal to the absolute central moments of same orders. Inserting (2.11) and (2.25) to (2.26), for an even l, we get

$$E[|Z-\mu|^l] = \sum_{k=0}^{l} C_l^k (-1)^k \exp[l\varpi + \frac{k}{2}\nu^2 + \frac{(l-k)^2}{2}\nu^2] \quad (2.27)$$

Replacing ϖ and θ with (2.13) and (2.14), we have

$$E[|Z-\mu|^l] = \sum_{k=0}^{l} C_l^k (-1)^k \mu^l (\frac{\sigma^2}{\mu^2}+1)^{\frac{(l-k)^2-(l-k)}{2}} \quad (2.28)$$

Therefore, for a multi-scaling process with μ and σ^2 given in (2.8) and (2.9), we can write

$$E[|\Delta W(T) - \lambda T|^l]$$

$$= \lambda^l T^l \sum_{k=0}^{l} C_l^k (-1)^k (\frac{c}{\lambda^2} \sum_{i=1}^{M} \phi_i(T) T^{-2(1-\alpha_i(T))} + 1)^{\frac{(l-k)^2-(l-k)}{2}} \quad (2.29)$$

where c is some constant. Obviously, when fully developed, (2.29) is the sum of a power series of T, which has a similar form as (2.9), though the set of exponents here is richer. This suggests that the high-order moments preserve the quasi multi-scaling property.

To see the high-order moment behaviors visually, we calculate them numerically for the persistent scaling and multi-scaling processes, and draw their changes with t on a log-log graph, as shown in figure 2 and 3, respectively. Comparing them with those from real traffic data [8] [9] [23], which have been viewed as the main evidence of the multi-scale behavior, we see they are extremely similar. In figure 2, every moment vs. time is a straight line with a slope proportional to its order. They share the same scaling exponent i.e., the Hurst parameter of the process. In figure 3, the curves are not completely straight but all have scaling trends, which demonstrates the existence of the multi-scaling property.

E. Parameter Characterization

Now the multi-scale traffic is modeled with scaling or multi-scaling processes in a series of time scale sections. Once the sections are known, the characterization of the traffic can be done efficiently. For a scaling section, what we need to know are three parameters (λ, α, c_1). Here λ and α are the average traffic rate and the scaling exponent for that section. c_1 is the coefficient on the right hand side of (2.6) for it to become an equality. As can be seen from (2.5) and (2.6), these three parameters completely decide μ, σ^2, and all high order moments of the scaling process. For a multi-scaling section, one more dimension is needed, i.e., the scaling structure (A, Φ) or $f_{A(T)}(\alpha)$. With the Gaussian assumption of the scaling

exponents, this is to require $\tilde{\alpha}$ and $\tilde{\sigma}^2$. So a compact representation of the multi-scaling section is (λ, c_2, $\tilde{\alpha}$, $\tilde{\sigma}^2$). c_2 is the coefficient on the right hand side of (2.9) for it to become an equality. For performance evaluation purposes, the real traffic may be simply approximated with only two time scale sections, i.e., the large scale section and the small scale section. Given the stationarity of the process, λ can be estimated without distinguishing the sections and shared by both. Considering the complexity of the multi-scale behavior, this characterization is very compact and convenient for performance modeling. It makes the multi-scale traffic model quite accessible for engineering network design.

F. Model Validation

From previous arguments, two conditions are sufficient to decide the model. Both the normal or the log-normal CTL distribution and the behaviors of the first two orders of moments can be validated in practical traffic. There exists a lot of work related with this in literature [7] [13] [15] [23] [24] [28]. So we will not repeat them here. In Section IV we will demonstrate that the queueing analysis results based on this model match the performance of real traffic very well. This further validates the model.

III. A General Approach for Performance Analysis

A. General Analysis of the Fluid Queue

Exact queueing analyses of the scaling and the mutli-scaling processes are very difficult. Common queueing analysis techniques like *M*/*M*/1, *M*/*G*/1, even *G*/*G*/1 does not help much here. The main approach used so far is asymptotic analysis, especially the technique of large deviations [5] [6]. However, the limitation of the asymptotic approach is obvious: it generally does not apply to the finite buffer; it is usually only related with the scaling exponent and can not evaluate the effect of other parameters. In short, as a limiting bound, it is more useful for qualitative evaluation rather than as a quantitative means. Moreover, to our best knowledge, the queueing analysis taking into account the multi-scale property has only been considered in one paper [22], where a loose higher bound for a wavelet-based multi-scale traffic model is given. In this Section, we will present a general approach for exact analysis of the fluid queue fed by multi-scale processes. The basic idea dates back to Benes [2], and has been adopted in performance evaluation of broadband networks [17] [25]. We provide a new interpretation of it, and apply it to analyze the scaling and the multi-scaling processes. The method is especially suitable for the traffic model presented in Section II.

Consider that we want to evaluate the queueing performance of a general traffic process. To do so, we go back to the fundamental relations of the queueing system. For a single server fluid queue running in the stable region and having enough space to buffer the transient bursts, we have the following balance equation:

$$Q(t_0)+V(t-t_0)=Q(t)+O(t-t_0) \tag{3.1}$$

Here $Q(t)$ is the queue length at t, $V(t-t_0)=W(t)-W(t_0)$ is the CTL arriving in the period (t_0, t), and $O(t-t_0)$ denotes the traffic load leaving in (t_0, t). It simply states that for any period (t_0, t) the arriving CTL plus the queue content at time t_0 is equal to the amount of the traffic load leaving the queue in the same period plus the content remaining at time t. Assume $V(t) = 0$ and $Q(t) = 0$ at time $t = 0$. Let $t_0 = 0$. Then we get the following relation:

$$Q(t)=[V(t)-O(t)]^+ \tag{3.2}$$

where the operator $[\cdot]^+$ means $max(\cdot, 0)$. It is used to emphasize the fact that $Q(t)$ is non-negative at any time t. Let C be the constant service rate of the queue. It is easy to see that $O(t)$ is

$$O(t)=C(t-I(t)) \tag{3.3}$$

Where $I(t)$ is the total server idle time up to t. Let $\Lambda(t) = C \cdot I(t)$, which is the maximum amount of traffic the server could serve in time $I(t)$ if the traffic load were available. We call it the "virtual" traffic load (VTL) up to t. Let $Y(t)=V(t)-Ct$. This is the "net" traffic load (NTL) that $W(t)$ owns beyond Ct (it can be negative in general). With (3.2) and (3.3) we get

$$Q(t)=[V(t)-Ct+\Lambda(t)]^+=[Y(t)+\Lambda(t)]^+ \tag{3.4}$$

For a queue length $q > 0$, using the law of total probability we then get

$$\begin{aligned} P[Q(t)>q] &= P[Y(t)+\Lambda(t)>q, Y(t)>q] \\ &\quad + P[Y(t)+\Lambda(t)>q, Y(t)\le q] \\ &= P[Y(t)>q]+P[Y(t)\le q<Y(t)+\Lambda(t)] \end{aligned} \tag{3.5}$$

Physically, (3.5) means the overflow probability above q is the sum of the probability that NTL exceeds q and the probability that NTL does not exceeds q but NTL plus VTL exceeds q. This actually indicates two different origins of packet losses. Based on the difference, we call the first term on the right hand side of (3.5) the *absolute loss* probability, and the second term the *opportunistic loss* probability. They are denoted as $P_{abs}(t)$ and $P_{opp}(t)$, respectively.

The absolute loss probability is

$$\begin{aligned} P_{abs}(t) &= P[Y(t)>q] \\ &= P[V(t)>Ct+q]=\int_{Ct+q}^{\infty} f_{V(t)}(v)dv \end{aligned} \tag{3.6}$$

Denote the integral on the right hand side as

$$J_t(C,q)=\int_{Ct+q}^{\infty} f_{V(t)}(v)dv \tag{3.7}$$

The opportunistic loss probability is more complex. With the result by Benes [2, Chapter 2], we can write down

$$\begin{aligned} P_{opp}(t) &= P[Y(t)\le q<Y(t)+\Lambda(t)] \\ &= \frac{\partial}{\partial q}\int_0^t P[Y(t)-Y(u)\le q, Q(u)=0]du \\ &= \frac{\partial}{\partial q}\int_0^t P[Y(t)-Y(u)\le q \mid Q(u)=0]P[Q(u)=0]du \end{aligned} \tag{3.8}$$

Assume $Q(t)$ is stationary. Define the network utility as $\eta=\lambda/C$. Let

$$\rho=1-\eta=1-\frac{1}{C}\lim_{t\to\infty}\frac{V(t)}{t}=-\lim_{t\to\infty}\frac{Y(t)}{Ct} \tag{3.9}$$

Then the following relation holds almost surely [3].

$$P[Q(u)=0]=\rho \tag{3.10}$$

So (3.8) becomes

$$\begin{aligned} P_{opp}(t) &= \frac{\partial}{\partial q}\int_0^t P[Y(t-u)\le q]\rho\, du \\ &= \rho\int_0^t \frac{\partial}{\partial q} P[V(u)\le Cu+q]du \\ &= \rho\int_0^t f_{V(u)}(v)\,|_{v=Cu+q}\, du \end{aligned} \tag{3.11}$$

The integral in (3.11) is a function of C and q when t is fixed. Denote

$$G_t(C,q)=\int_0^t f_{V(u)}(v)\,|_{v=Cu+q}\, du \tag{3.12}$$

Use $P_{\text{loss}}(t)$ to represent the total loss probability. With (3.5), (3.6), (3.7), (3.11), and (3.12), we can write down

$$P_{loss}(t)=P_{abs}(t)+P_{opp}(t)=J_t(C,q)+\rho G_t(C,q) \tag{3.13}$$

This is a general formula for the exact queueing behavior of virtually any type of traffic. Given C and q, the loss probability of scaling and multi-scale processes at any time can be evaluated through numerical computation. Since (3.13) calculates the loss probability directly using the CTL distribution, it is almost customized for the multi-scale traffic model developed in Section II. In fact, that is the main motive that we use this approach. Finally, the loss probability of the persistent scaling process is

$$\begin{aligned} P_{loss}(t) &= \int_{Ct+q}^{\infty}\frac{1}{\sqrt{2\pi c}t^{\alpha}}\exp[-\frac{(v-\lambda t)^2}{2ct^{2\alpha}}]dv \\ &+(1-\frac{\lambda}{C})\int_0^t\frac{1}{\sqrt{2\pi c}u^{\alpha}}\exp[-\frac{(Cu-\lambda u+q)^2}{2cu^{2\alpha}}]du \end{aligned} \tag{3.14}$$

The loss probability of the persistent multi-scaling process is

$$\begin{aligned} &P_{loss}(t) \\ &=\int_{Ct+q}^{\infty}\frac{\exp[-\frac{[\ln(v\sqrt{(c/\lambda^2)t^{-2(1-\tilde{\alpha})}t^{2\tilde{\sigma}^2\ln(t)}+1})-\ln(\lambda t)]^2}{2\ln[(c/\lambda^2)t^{-2(1-\tilde{\alpha})}t^{2\tilde{\sigma}^2\ln(t)}+1]}]}{\sqrt{2\pi\ln[(c/\lambda^2)t^{-2(1-\tilde{\alpha})}t^{2\tilde{\sigma}^2\ln(t)}+1]}\cdot v}dv \\ &+ \\ &(1-\frac{\lambda}{C})\int_0^t\frac{\exp[-\frac{[\ln((Cu+q)\sqrt{(c/\lambda^2)u^{-2(1-\tilde{\alpha})}u^{2\tilde{\sigma}^2\ln(u)}+1})-\ln(\lambda u)]^2}{2\ln[(c/\lambda^2)u^{-2(1-\tilde{\alpha})}u^{2\tilde{\sigma}^2\ln(u)}+1]}]}{\sqrt{2\pi\ln[(c/\lambda^2)u^{-2(1-\tilde{\alpha})}u^{2\tilde{\sigma}^2\ln(u)}+1]}\cdot(Cu+q)}du \end{aligned} \tag{3.15}$$

B. Behavior of the Absolute Loss

The absolute loss probability in (3.6) has been widely used as an approximation of the overall loss probability or as the lower bound of it. Looking closely at it, we have the following theorem about its behavior.

Theorem 2: If $\lambda < C$, the absolute loss probability $P_{\text{abs}}(t)$ of a persistent scaling or multi-scaling process goes to zero when $t \to \infty$.

Proof: For a persistent scaling process, with (2.4) and (3.7) $J_t(C,q)$ can be written as

$$J_t(C,q)=\frac{1}{2}-\frac{1}{2}erf[\frac{Ct+q-\mu(t)}{\sqrt{2}\sigma(t)}] \tag{3.16}$$

where $\mu(t)$ and $\sigma(t)$ are μ and σ in(2.4), and $erf(z)$ is the error function

$$erf(z)=\frac{2}{\sqrt{\pi}}\int_0^z e^{-u^2}du \tag{3.17}$$

The term inside $erf(\cdot)$ in (3.16) is

$$z=\frac{Ct+q-\lambda t}{\sqrt{2}ct^{\alpha(t)}}=\frac{C-\lambda}{\sqrt{2}c}t^{1-\alpha(t)}+\frac{q}{\sqrt{2}c}t^{-\alpha(t)} \tag{3.18}$$

In the stable regime of the queueing system, i.e., $\lambda < C$, when $t \to \infty$ the first term on the right hand side of (3.18) $\to \infty$ and the second term $\to 0$, thus $z \to \infty$. This makes $erf(z) \to 1$, and $J_t(C,q)$ in (3.16) $\to 0$. This means that the absolute loss for the persistent scaling traffic will eventually be zero, and the long-term packet loss rate is governed by the opportunistic loss.

For a persistent multi-scaling process, with (2.7) $J_t(C,q)$ is

$$J_t(C,q)=\frac{1}{2}-\frac{1}{2}erf[\frac{\ln(Ct+q)-\varpi(t)}{\sqrt{2}\theta(t)}] \tag{3.19}$$

where $\varpi(t)$ and $\theta(t)$ are ϖ and θ in(2.7). This time the term inside $erf(\cdot)$ in (3.19) is

$$\begin{aligned} z &= \frac{\ln(Ct+q)-\ln\mu+\frac{1}{2}\ln(\sigma^2/\mu^2+1)}{\sqrt{2\ln(\sigma^2/\mu^2+1)}} \\ &= \frac{\ln(qt^{-1}+C/\lambda)}{\sqrt{2\ln[(c^2/\lambda^2)t^{-2(1-\alpha(t))}+1]}}+\frac{1}{2\sqrt{2}}\sqrt{\ln[(c^2/\lambda^2)t^{-2(1-\alpha(t))}+1]} \end{aligned} \tag{3.20}$$

With a similar analysis, we can see that when $t \to \infty$, the first term on the right hand side of (3.20) goes to ∞ and the second term goes to 0. This again results in $z \to \infty$, then $erf(z) \to 1$, and then $J_t(C,q) \to 0$. So the long-term loss behavior of the multi-scale traffic is also governed by the opportunistic loss. However, the convergence speed in this case is exponentially slower than that of the scaling traffic.

Because $P_{\text{abs}}(t)$ goes to zero when $t \to \infty$, it is generally improper to use it to approximate the overall loss probability. However, we will show in Section IV that it converges extremely slowly for heavy traffic. In that situation it is a good approximation before the steady state is reached.

C. Behavior of the Opportunistic Loss

For the opportunistic loss probability, we have the following theorem.

Theorem 3: If $\lambda < C$, the opportunistic loss probability of a persistent scaling or multi-scaling process increases monotonically with t and converges to $\rho \sup_{t>0} G_t(C,q)$ when t

$\to \infty$.

Proof: If $\lambda < C$, the opportunistic loss probability is obviously bounded above by 1. Because $f_{V(u)}(v) > 0$, from (3.11) the integral function $G_t(C,q)$ is a monotonically increasing function of t. So if only $f_{V(u)}(v)$ is a valid probability distribution function, $\sup_{t>0} G_t(C,q)$ exists and $\rho \sup_{t>0} G_t(C,q) \le 1$. Therefore, for a persistent scaling or multi-scaling traffic, $P_{\text{opp}}(t)$ increases monotonically converges to $\rho \sup_{t>0} G_t(C,q)$.

□

We immediately get the following theorem about the steady state performance.

Theorem 4: The total loss probability in the steady state, P_{steady}, is

$$P_{steady} = \lim_{t\to\infty} P_{loss}(t) = \rho \sup_{t>0} G_t(C,q) \qquad (3.21)$$

Proof: From theorem 2 and 3, $P_{\text{abs}}(t) \to 0$ and $P_{\text{opp}}(t) \to \rho \sup_{t>0} G_t(C,q)$ when t $\to \infty$. So $P_{steady} = \lim_{t\to\infty} P_{loss}(t) = \lim_{t\to\infty} P_{opp}(t) = \rho \sup_{t>0} G_t(C,q)$.

□

In particular, for the persistent scaling process the steady state loss probability is

$$P_{steady} = (1-\frac{\lambda}{C})\int_0^\infty \frac{1}{\sqrt{2\pi c u^{\alpha}}} \exp[-\frac{(Cu-\lambda u+q)^2}{2cu^{2\alpha}}]du \qquad (3.22)$$

For the persistent multi-scaling process, it is

$$P_{steady} = (1-\frac{\lambda}{C})\int_0^\infty \frac{\exp[-\frac{[\ln((Cu+q)\sqrt{(c/\lambda^2)u^{-2(1-\tilde{\alpha})}u^{2\tilde{\sigma}^2\ln(u)}+1)-\ln(\lambda u)]^2}{2\ln[(c/\lambda^2)u^{-2(1-\tilde{\alpha})}u^{2\tilde{\sigma}^2\ln(u)}+1]}]}{\sqrt{2\pi\ln[(c/\lambda^2)u^{-2(1-\tilde{\alpha})}u^{2\tilde{\sigma}^2\ln(u)}+1]}\cdot(Cu+q)}du \qquad (3.23)$$

D. Examples

Figure 4 to 6 shows the absolute, the opportunistic, and the total loss probabilities of a persistent scaling process (PS1) and a persistent multi-scaling process (PS2). The two processes have the same average arrival rate and share the same link speed and buffer size. PS2's average scaling exponent equals to the only scaling exponent of PS1. Just as we analyzed above, the absolute loss goes to zero eventually. But it first reaches to a peak value before going down. The opportunistic loss increases monotonically and converges to a limit. It dominates the total loss probability, and the latter follows the same trend and goes to the steady state. Most strikingly, the figures show the performance difference between the two processes is huge: P_{steady} of PS1 is about 0.5% while that of the PS2 is four times higher, about 2.5%. This means that if the traffic is a multi-scaling process in the large scale, the packet losses would be very high. Fortunately, it is just a theoretical case that has not been seen in real situations. The practical traffic generally

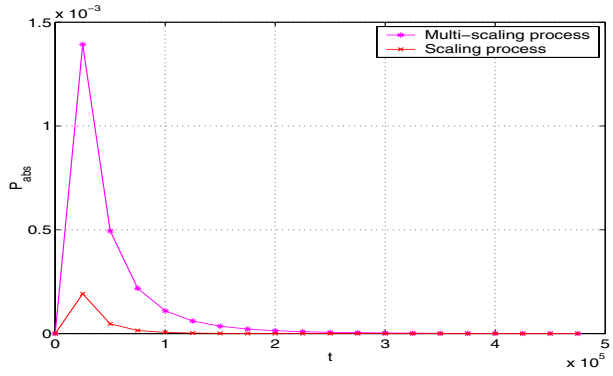

Fig.4. Behaviors of the absolute packet losses of persistent scaling and multi-scaling processes.

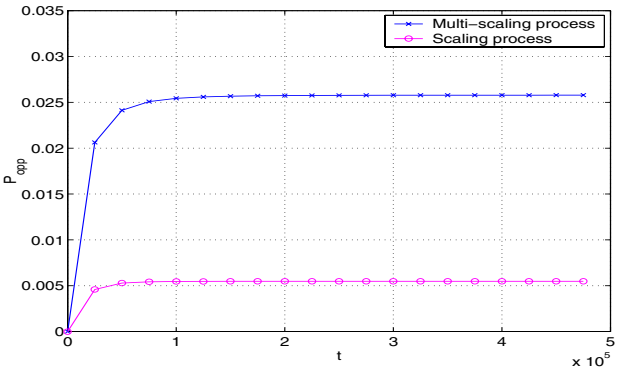

Fig.5. Behaviors of the opportunistic packet losses of persistent scaling and multi-scaling processes.

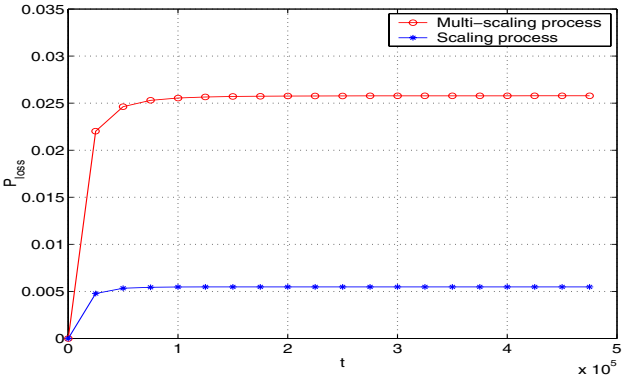

Fig.6. Behaviors of the total packet losses of persistent scaling and multi-scaling processes.

holds the multi-scaling property only in the small scale, and always turns into the scaling process in the large scale. Its performance will be specifically studied in next Section.

IV. Performance Evaluation of Practical Multi-scale Traffic

A. Loss Probability of the Multi-scale Traffic

The Internet traffic behaviors are different at the large time scale and the small scale. These two time scales are also of most interest as far as the queueing performance is concerned. In different hours of a day a clear drift of the level of the traffic load may be observed. However, it occurs in the administrative time sale and generally does not affect the assumptions and conclusions of the queueing analysis.

We separate two time scale sections accordingly in our analysis. Let $[0, t_I]$ and $[t_I, t_{II}]$ be the small time scale section and the large time scale sections, respectively. Here $t_{II} = \infty$. Define

$$U(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{4.1}$$

Let $f_{\mathrm{I}}(w)$ and $f_{\mathrm{II}}(w)$ denote the log-normal and the normal distributions given in (2.7) and (2.4), respectively. Then $V(t)$'s distribution at any time scale t is

$$f_{V(t)}(v) = f_I(v)[U(t) - U(t - t_I)] + f_{II}(v)[U(t - t_I) - U(t - t_{II})] \tag{4.2}$$

Represent the integrals in (3.7) and (3.12) for the multi-scaling traffic as $J_t^I(C,q)$ and $G_t^I(C,q)$, and those for the scaling traffic as $J_t^{II}(C,q)$ and $G_t^{II}(C,q)$. Based on (4.2) and (3.7) and (3.12), we can show that the integrals for the overall traffic are

$$J_t(C,q) = J_t^I(C,q)[U(t) - U(t - t_I)] + J_t^{II}(C,q)[U(t - t_I) - U(t - t_{II})] \tag{4.3}$$

and

$$G_t(C,q) = G_t^I(C,q)[U(t) - U(t - t_I)] + [G_t^{II}(C,q) - G_{t_I}^{II}(C,q) + G_{t_I}^I(C,q)][U(t - t_I) - U(t - t_{II})] \tag{4.4}$$

Denote the loss probabilities in (3.15) and (3.14) for the multi-scaling traffic and the scaling traffic as $P_{loss}^I(t)$ and $P_{loss}^{II}(t)$, respectively, and the opportunistic loss probabilities of them as $P_{opp}^I(t)$ and $P_{opp}^{II}(t)$. We finally get

$$P_{loss}(t) = P_{loss}^I(t)[U(t) - U(t - t_I)] + P_{loss}^{II}(t)[U(t - t_I) - U(t - t_{II})] - [P_{opp}^{II}(t_I) - P_{opp}^I(t_I)][U(t - t_I) - U(t - t_{II})] \tag{4.5}$$

This formula means that the multi-scaling property solely affects the transient queueing behavior in the small scale, and also contributes to the large scale loss probability. However, the contribution is rather limited if t_I is small. Then the steady state performance is governed by the scaling property in the large scale. With (4.5) and the results in Section III, we can compute the loss probability at any time for the multi-scale traffic.

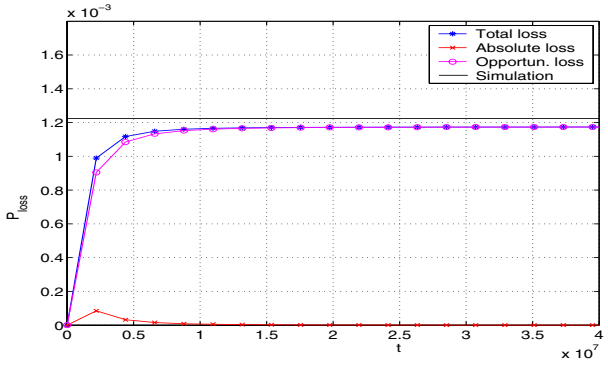

Fig.7. Comparison of the analytical and the simulation results for the steady state loss probability of the multi-scale traffic.

B. Steady state performance

To calculate the steady state loss probability based on the model, we estimate the model parameters for the large scale described in Section II-E from the real traffic trace, and apply them in formula (4.5). It is an approximate calculation because we ignore the multi-scaling property in the small scale, which will be examined specifically in Section IV-C. Generally, it is a fairly mature subject to estimate the scaling exponent and the parameters of normal and log-normal distributions, for which we will not get into details in this paper. The real traffic data we use is the well-known trace LBL-TCP-3 [30] from Lawrence Berkeley Laboratory, which has been used as a representative for scaling and multi-scale behaviors [23] [28]. With an average rate of 282.12Kbps, the traffic generates a network utility of 28.21% on the 1Mbps Ethernet link on which it was captured.

Figure 7 shows the total loss probability and its components based on the model for the real trace, and compares the steady state loss probability with the simulation result. The constant value indicated by the top plain line is the average loss probability measured in the simulation, in which the real trace is fed into a FIFO queue. The maximum queue length is set as 300KB. We see the absolute loss and the opportunistic loss behave like we described in Section III: the absolute loss eventually vanishes and the opportunistic loss ends up in a steady state. The convergence of the total loss probability is fairly quick and the final state is quite stable. The value in the steady state is so close to the result from the simulation that we can only see a difference of 0.002%. Thus the queueing analysis just taking into account the large scale scaling behavior can provide a very good prediction for the steady state loss performance.

C. Transient Behavior

Now that the scaling behavior alone determines the steady state queueing performance, what is the role of the multi-scaling behavior? With the performance analysis model we can show that it has effect on the transient behavior. The transient behavior indicates the change of the loss probability in the

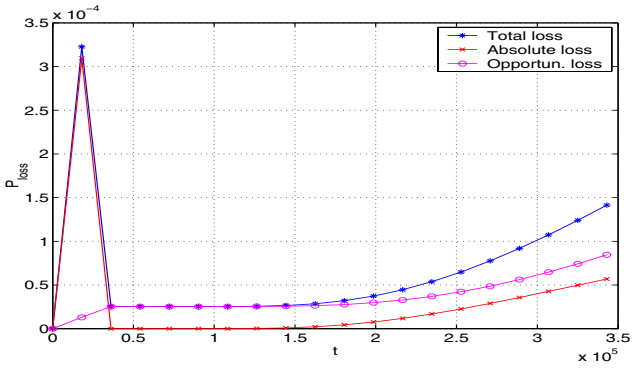

Fig.8. Transient loss behavior of the multi-scale traffic.

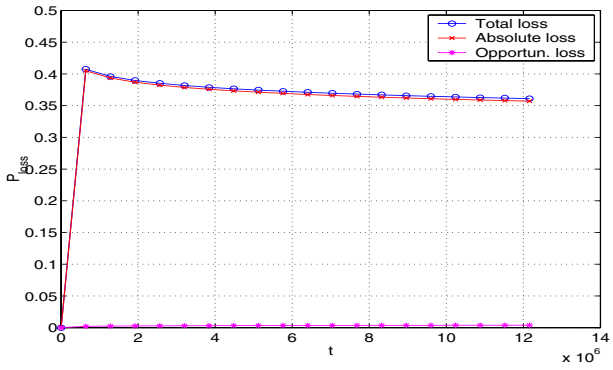

Fig.9. Loss behavior of the multi-scale traffic under heavy load ηz= 0.9.

transient period, i.e., the period before the steady state is reached. Figure 8 illustrates the early transient period of the loss probability for trace LBL-TCP-3. In this analysis, t_1 is identified[1] and parameters for both the large scale and the small scale are first estimated. Then the loss probabilities at different times are computed with (4.5). Different from that of the persistent scaling process in figure 6, the total loss probability goes down sharply at t_1 after an initial quick start, and then evolves like a normal scaling process. We claim that this initial jump of loss probability is exactly the consequence of the multi-scaling behavior. As we have known from Section III, a multi-scaling process performs much worse than a scaling process if they have similar (average) scaling exponents. So the multi-scaling phase builds up a higher loss probability before t_1 than if it were a scaling process in that phase. After t_1 the loss probability follows that of a scaling process so it drops down to a corresponding level and then develops on that basis. We see the dramatic change is mainly due to the absolute loss. The opportunistic loss keeps increasing across t_1. Clearly, this transient behavior does not affect the performance in the large scale much.

D. Heavy Traffic Behavior

From (3.9), when $\eta \to 1$, then $\rho \to 0$, and the opportunistic loss would be close to zero. We also know the absolute loss in the steady state is zero, too. Thus the total loss would go to zero for the heavy load traffic. This is intuitively incorrect. So we guess the loss behavior in this case must be different from what we observed so far. Figure 9 gives the answer to this seeming contradiction. It shows the absolute, the opportunistic, and the total loss probabilities under the load η = 0.9. We see that at such a high load, the value of the opportunistic loss indeed is very small. On the contrary, the value of the absolute loss is large. Mathematically, it is not difficult to verify these with the equations (3.6) and (3.11). So the thing is, although the opportunistic loss still increases with time and the absolute loss still decreases, both of them converge extremely slowly. In the practical measurement of a limited period, we are likely to find that the total loss is non-zero, and is approximately the absolute loss. So in the heavy load case, the absolute loss can be used as an approximation of the total loss. This is consistent with the analysis in Section III.

E. Loss Probability vs. Buffer Size

As we have mentioned in Section I, the queueing performance evaluation for the scaling and the mutli-scaling processes usually uses the asymptotic approach, i.e., pursues the limiting result when the buffer size $q \to \infty$. Then the result is applied to the finite buffer for an approximate analysis, which may give a very poor prediction. A well-known result for the scaling process is that $P_{steady} \sim q^{-2(1-H)}$ for $q \to \infty$ [4] [15], where $0 < H < 1$ is the Hurst parameter, i.e., the sole scaling exponent of the process. Our queueing model in this paper, however, is an exact analysis and can give numerical result for the finite buffer. We will evaluate the effect of the buffer size on the loss probability with the model, and compare the result with those from the simulation and the asymptotic approach[2]. Figure 10 gives the results for the trace LBL-TCP-3. The simulation data indicate that the loss probability decreases quickly with the increase of buffer size for moderate buffer sizes. The change can be predicted very well with our queueing analysis, while the asymptotic method performs poorly. The results also testify that using bigger buffer is actually more beneficial than expected with the power law or the asymptotic approach if the buffer size is in the moderate range.

V. Conclusion

Scaling and multi-scale phenomena of network traffic have important impacts on various aspects of network design, control, and management. However, existing models are often

[1] A simple approximate way to identify t_1 is through the log-log graph of σ^2 vs. time scale. This graph generally gives a good indication of the traffic behaviors at different time scales. t_1 is the time scale beyond which the graph is roughly a straight line and below which it is obviously nonlinear.

[2] The asymptotic formula used is $P_{steady} = \beta q^{-2(1-H)}$, where the coefficient β needs to be decided. Because the formula is supposed to be able to predict the loss probabilities accurately for very large buffer sizes, we collect sample loss probabilities from simulations using very large buffers, and estimate β based on the samples.

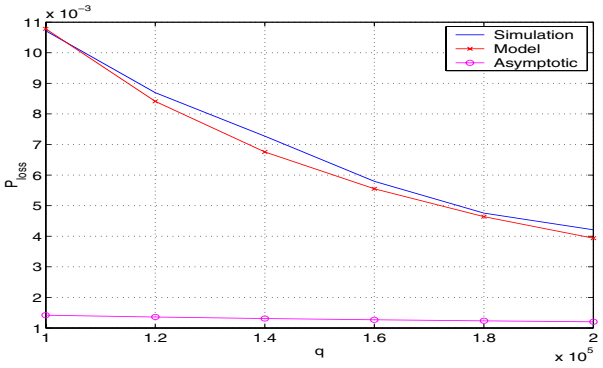

Fig.10. Steady state loss probabilities for different buffer sizes.

very complex and not suitable for performance analysis, which greatly limit their applications for network optimization. This paper proposes a new statistical model for scaling and multi-scale traffic, and presents a general queueing anlaysis technique for it. The statistical model employs the central moments and the marginal distributions of the cumulative traffic load process. Only the first two orders of moments are needed to define the process. At the same time, all high-order moments have the power-law like properties inherently. Thus the model provides a good approximation of the multi-scale behavior while maintaining the simplicity for measurement and estimation. Unlike most of the queueing techniques used for scaling and mutli-scale processes so far, the queueing analysis presented in this paper is neither an asymptotic approach nor a bounding approximation. It is a general fluid queueing analysis method that can give numerical result. Two types of packet losses, the absolute loss and the opportunistic loss, are identified with the method, and their behaviors are determined. The analysis can evaluate not only the steady state performance but also the transient queueing behavior. Various dimensions of performance of scaling and multi-scale traffic, such as the heavy-traffic performance and the loss probabilities for different buffer sizes, can be easily evaluated with the analysis method.

The traffic model and the queueing analysis method pave the way for taking full advantage of the scaling and multi-scale properties in practical network planning, dimensioning, and traffic control and management. Our future work will focus their applications in these issues.

ACKNOWLEDGMENT

We would like to thank Prof. Armand Makowski, Huigang Chen, and Majid Raissi-Dehkordi for helpful discussions.

REFERENCES

[1] P. Abry, P. Flandrin, M.S. Taqqu and D. Veitch, "Wavelets for the analysis, estimation and synthesis of scaling data," In *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. Wiley Interscience, 1999.

[2] V. Benes, *General Stochastic Processes in the Theory of Queues*, Reading, MA: Addison Wesley, 1963.

[3] A.A. Borovkov, *Stochastic Processes in Queueing Theory*, NY: Springer-Verlag, 1976.

[4] O.J. Boxma, "Fluid queues and regular variation," *Performance Evaluation*, 27&28: 699-712, 1996.

[5] J.D. Deuschel and D.W. Stoock, *Large Deviations*, Boston: Academic Press, 1989.

[6] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single-server queue, with applications," *Math. Proc. of the Camb. Phil. Society*, 118:363--374, 1995.

[7] A. Feldmann, A.C. Gilbert, W. Willinger, and T.G. Kurtz, "The changing nature of network traffic: scaling phenomena," *Computer Communication Review*, April 1998.

[8] A. Feldmann, A.C. Gilbert, and W. Willinger, "Data networks as cascades: investigating the multifractal nature of internet WAN traffic," *ACM SIGCOMM* 1998.

[9] A. C. Gilbert, W. Willinger, and A. Feldmann, "Scaling analysis of random cascades, with applications to network traffic," *IEEE Transactions on Infomation Theory*, April 1999.

[10] D. Heath, S. Resnick, and G. Samorodnitsky, "Heavy tails and long range dependence in on/off processes and associated fluid models," *Math. Oper. Res.*, 23(1):145-165, 1998.

[11] M. Krunz and A. Makowski, "Modeling video traffic using M/G/infinity input processes: a compromise between markovian and LRD models," *IEEE Journal on Selected Areas in Communications*, 16(5):733-748, June 1998.

[12] S. S. Kulkarni and G. R. Dattatreya, "Statistically multiplexed adaptive operation of ad hoc networks with self-similar traffic," *1999 IEEE Emerging Technologies Symposium on Wireless Communications and Systems*, Richardson, TX, 1999.

[13] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, "On the self-similar nature of Ethernet traffic," *IEEE/ACM Transactions on Networking*, 2:1-15, 1994.

[14] J.B. Levy and M.S. Taqqu, "Renewal reward processes with heavy-tailed interrenewal times and heavy-tailed rewards," *Bernoulli*, 6(1):23-44, 2000.

[15] X. Liu and J. Baras, "Understanding multi-scale network traffic: a structural TCP traffic model," *Submitted*.

[16] T. Mikosch, S. Resnick, H. Rootzen and A.W. Stegeman, "Is network traffic approximated by stable levy motion or fractional brownian motion?" Technical Report, Cornell University, 1999.

[17] I. Norros, J. Roberts, *et al*, "The superposition of variable bit rate sources in an ATM multiplexer," *IEEE Journal on Selected Areas of Communications*, 9(3):378-387, 1991.

[18] K. Park, G. Kim, and M. Crovella, "On the effect of traffic selfsimilarity on network performance," in Proc. *SPIE Intl. Conf. Perf. and Control of Network Systems*, 1997.

[19] K. Park and W. Willinger, "Self-similar network traffic: an overview," In *Self-Similar Network Traffic and Performance Evaluation*, Wiley-Interscience, 2000.

[20] V. Paxson and S. Floyd, "Wide-area traffic: the failure of poisson modeling," *IEEE/ACM Transactions on Networking*, 3(3):226-244, June 1995.

[21] J. Potemans, J. Theunis, *et al*, "Measuring Self-Similar Wireless Data Traffic for Multimedia Applications," *2001 International Conference on Third Generation Wireless and Beyond*, San Francisco, USA, 2001.

[22] V. J. Ribeiro, R. H. Riedi, *et al*, "Multiscale queuing analysis of long-range-dependent network traffic," *IEEE INFOCOM 2000*.

[23] R.H. Riedi, M.S. Crouse, V.J. Ribeiro, and R.G. Baraniuk, "Multifractal wavelet model with application to tcp network traffic," *IEEE Special Issue on Information Theory*, 45:992-1018, April 1999.

[24] R.H. Riedi and J.L. Vehel, "TCP traffic is multifractal: a numerical study," Preprint, 1997.

[25] J. Roberts, U. Mocci, and J. Virtamo, "Broadband network teletraffic: performance evaluation and design of broadband multiservice networks", Final Report of COST 242, 1996.

[26] B. Ryu, "Modeling and simulation of broadband satellite networks: part II -- traffic modeling," *IEEE Communication Magazine*, No. 7, 1999.

[27] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *ACM/SIGCOMM Computer Communica-tions Review*, 27:5-- 23, 1997.

[28] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, 5(1):71-86, January 1997.

[29] J. Zhang, M. Hu, and N. Shroff, "Bursty data over CDMA: MAI self-similarity, rate control, and admission control," *IEEE INFOCOM 2002*.

[30] http://ita.ee.lbl.gov/html/traces.html.