Queueing Analysis of Multiscale Network Traffic

HONG Fei ¹ and BAI Yu ²

¹ School of Computer Science, BeiHang University
No.37, XueYuan Road, HaiDian District, Beijing, P.R.China (100083)
hongfei@buaa.edu.cn

² School of Mathematical Sciences, Peking University
HaiDian District, Beijing, P.R.China (100871),
pkuby@math.pku.edu.cn

Abstract: Recent research shows that one of remarkable characteristic is the fractal nature of the investigated network traffic, especially the multiscale phenomena over small time scale. In this paper, based on multifractal wavelet analysis, we perform the efficient queue tail analysis in an infinite capacity single server queue serviced at a constant rate driven by general multiscale input process. With the scale nature of wavelet, it is more convenient and accurate to perform the tail asymptotic analysis in wavelet domain than the analysis one. These results are validated by queuing simulation of the measured network traffic traces.

Keywords: Queueing Analysis, Multiscale, Wavelet, Network Traffic.

1. INTRODUCTION

Teletraffic research papers have reported the high variability and burstiness nature of network traffic in several LAN/WAN environments in the last decade. Moreover, it seems that most of the measured network traffic exhibits properties of scale invariance. It means that within a range of scales no characteristic dominant scale can be identified and some statistical properties within this range are not changing. This remarkable scaling phenomenon called for the fractal modeling of the investigated LAN/WAN traffic [6, 16, 17, 18].

In the fractal modeling framework long-range dependence (LRD) and self-similarity have been analyzed intensively, and a number of studies are focused on how to detect accurately the LRD property and how to estimate the Hurst parameter [2]. LRD is revealed by the power law decay of the autocorrelation function at large lags, i.e. \( r(k) \propto c \cdot k^{-2H} \), \( k \to \infty \), \( H \in (0.5, 1) \), where \( c \) is a constant. The degree of this slow decay is determined by the Hurst parameter \( H \).

A large group of traffic models (Fractional Brownian Motion (FBM) models, FARIMA models, Cox’s M/G/∞ models, ON/OFF models, etc.) to capture LRD and self-similar properties has also been developed [13]. Among these models the FBM [14] was found to be a popular parsimonious and tractable model of traffic aggregation [9]. The performance
implications of the fractal property are also addressed in a series of studies [7, 8]. After a number of new measurements and deeper analysis of network traffic it was discovered that the LAN/WAN traffic has a more complex scaling behavior which cannot be described by LRD and self-similarity [6, 18]. More precisely, it has been found that aggregated network traffic is asymptotically self-similar over time scales of the order of a few hundreds of milliseconds and above but it exhibits multifractal scaling below this time scale [6]. It has been also pointed out that the transition from the multifractal to self-similar scaling occurs around time scales of a typical packet round-trip time in the network. However, some studies showed that multifractal scaling can also be present even at large time scales [12]. Therefore the monofractal traffic models (e.g. FBM) are inadequate to characterize the network traffic and multifractal traffic models with a much more flexible rule for the scaling law seem to be needed, especially for some WAN environments. Multifractal models can allow a compact description of a complex scaling behavior and it can also capture the non-Gaussian character of network traffic. Multifractal models imply the non-redundant scaling behavior of moments of many orders. The physical explanations and engineering implications are also addressed in several papers, e.g. [6].

A stochastic process $X(t)$ is called multifractal [10] if it has stationary increments and satisfies

$$E[X(t)^q] = c(q)t^{\tau(q)+1},$$

(1)

For some positive $q$, where $\tau(q)$ is called the scaling function of multifractality and $c(q)$ is independent of $t$. An easy consequence of this definition is that $\tau(q)$ is a concave function [10].

If the scaling function $\tau(q)$ is a linear function of $q$ the process is called monofractal. Multifractality is thus defined as a global property of the process moments. The definition is very general and it covers a very large class of processes. Multifractal processes are also called processes with scaling property.

From a practical point of view queuing analysis of fractal traffic is a very important issue for network dimensioning and management. Therefore the study of queuing systems with fractal traffic input is a challenge in queuing theory. In the recent years the performance of queues with LRD or self-similar input has been deeply analyzed. A collection of studies has proven that the FBM based models have a tail queue distribution that decays asymptotically like a Weibullian law, i.e., $P[Q > b] \propto \exp(-\delta b^{2-2H})$, where $\delta$ is a positive constant that depends on the service rate of the queue [3, 14]. This important result shows that queues with FBM input ($H > 1/2$) have a much slower decay than that of the exponential.

However, there is a lack of queuing results available in the cases when the input traffic has a more complex scaling behavior. Especially, queuing systems with multifractal input are an undiscovered field and only a few results have been published in the literature. Vehel et al. [19] suggested a cascade model for TCP traffic based on the retransmission and congestion...
avoidance mechanisms with no performance analysis. Riedi et al. [16, 17] developed a multiscale queuing analysis in the case of tree-based multiscale input models. Gao et al. simulated queues fed by multiplicative multifractal processes in [7] but provided no analytical results. In contrast to these results we consider general multifractal in the wavelet domain. And derive analytical results for the queue tail asymptotic.

Our aim is to contribute to the queuing theory of multiscale queues and also to the traffic engineering implications. In this paper we present an efficient and accurate analysis of multifractal queues in the wavelet domain, including the tail asymptotic, special case, and practical applications. With the scaling nature of wavelet, the queue result is much better than the traditional analysis.

2. QUEUEING MODEL

All we consider a simple queuing model: a single server queue in continuous time, the serving principle for offered work is defined to be FIFO (First In, First Out), the queue has infinite buffer and constant service rate $c$. Denote by $K(t)$ the total size of work arriving to the queue from time instant $-t$ in the past up to this moment, time instant 0. The so called workload process $W(t)$ is the total amount of work stored in the buffer in time interval $(-t, 0)$, that is:

$$ W(t) = K(t) - ct $$

(2)

Our interest, however, is the current buffer length of the queue, denoted by $Q$. This is the queue length in the equilibrium state of the queue when the system has been running for a long time and the initial queue length has no influence. If this state of the system does exist, i.e., stationary and ergodicity of the workload process hold, and the stability condition for the system is also satisfied, i.e., $\lim_{s \to 0} E[K(t)]/t < s$ then:

$$ Q = \sup_{t \geq 0} W(t) $$

(3)

Where $W(0)$ is assumed to be 0. This equation is also referred to as Lindley’s equation.

The input process $X(t)$ is considered as a general multifractal process which is defined by (1), this definition, presented by Mandelbrot et al. in [10], describes multifractal process in terms of moments which leads to a more intuitive understanding of multifractality.

In [20], the main proposition for the analytical approximation for queue tail probabilities is presented as following:
Proposition 1: The probabilities for the queue tail asymptotic of a single queuing model with general multifractal input is accurately approximated by

\[
\log(P(Q > b)) \approx \min_{q \geq 0} \log(c(q)) \frac{\left[ b \tau_0(q) \right]^\tau(q)}{b q \tau(q)} \frac{\left[ c(q - \tau(q)) \right]}{c(q)} \text{by large } b, q \geq \tau(q)
\] (4)

where \( \tau_0(q) := \tau(q) + 1 \). The scaling functions \( \tau(q) \) and \( c(q) \) are the functions which define the multifractal input process.

3. QUEUEING ANALYSIS

In this section, we mainly focus on the queue tail in the wavelet domain. Here we suppose readers are familiar with the discrete wavelet transform (DWT). Denote by \( x(t) \) the measured network traffic, then through DWT by Haar wavelet, \( u_{j,k} \) and \( w_{j,k} \) represent the corresponding scale and wavelet coefficient.

Since the Haar scaling coefficients \( u_{j,k} \) represent the local mean of the signal at different scales and shifts, they are non-negative if and only if the signal itself is non-negative; that is, \( x(t) \geq 0 \Leftrightarrow u_{j,k} \geq 0, \forall j, k \). This condition leads us directly to a set of constraints on the Haar wavelet coefficients. With the constraint \( u_{j,k} \geq 0 \), we obtain the condition:

\[
x(t) \geq 0 \Leftrightarrow |w_{j,k}| \leq u_{j,k}, \forall j, k
\] (5)

The positive constraints inspire a very simple multiscale, multiplicative signal model for positive process. In the multifractal wavelet model we compute the wavelet coefficients recursively by:

\[
W_{j,k} = A_{j,k} U_{j,k}
\] (6)

With \( A_{j,k} \) be a random variable supported on the interval \([-1,1]\). We assume that the \( A_{j,k} \)'s are independent. Together with (6), we obtain:

\[
U_{j,2k} = 2^{-1/2} (1 + A_{j+1,k}) U_{j+1,k}, \quad U_{j,2k+1} = 2^{-1/2} (1 - A_{j+1,k}) U_{j+1,k}
\] (7)

Now, we consider the queuing problem in the discrete condition, that is let \( Q \) represent the
queue size at time instant $t$. Denote by $K_r$ the aggregate traffic arriving between time instants $-r+1$ and 0. In the following, we refer to $K_r$ as representing data at time-scale $r$. We set $K_0 := 0$. Considering the equation (3), we can compute as following:

$$Q_r = \max(Q_r + K_r - c, 0) = \max(Q_r + K_r - 2c, K_r - c, 0)$$  \hspace{1cm} (8)

Processing (5) iteratively, at last we can obtain the common equation:

$$Q_0 = \max(Q_{r+1} + K_r - rc, ..., K_0)$$  \hspace{1cm} (9)

Since $Q_r \geq 0$ for all $r$ we must have $Q_0 \geq \sup_{r \in \mathbb{Z}} (K_r - rc)$. Denoting by $-j$ the last instant the queue was empty before time instant 0, we obtain $Q_0 = K_j - jc \leq \sup_{r \in \mathbb{Z}} (K_r - rc)$. Thus if the queue was empty at some time in the past, $Q_0 = \sup_{r \in \mathbb{Z}} (K_r - rc)$  \hspace{1cm} (10)

We will study the quantity $\tilde{Q}_0$ obtained by taking the supermum in (7) over a finite subset of $Z_+$ corresponding to powers of 2 values for $n$—the dyadic time scale, which is easy for wavelet analysis, that is:

$$\tilde{Q}_0 := \sup_{m \in \{0, ..., n\}} (K_{2^m} - c2^m)$$  \hspace{1cm} (11)

for some fixed $n \in \mathbb{Z}_+$. Clearly, $\tilde{Q}_0 \leq Q_0$.

We will approximate the desired tail queue probability by $P(\tilde{Q}_0 > b) \approx P(Q_0 > b)$.

Here we will concentrate on the tail queue probability of the models through wavelet analysis. The Haar wavelet scaling coefficients on the branch are related to the quantities $K_{2^n}$, that is:

$$K_{2^n} = 2^{-i/2} U_{i,2^{n-1}} \quad \text{for} \quad i = 0, 1, ..., n$$  \hspace{1cm} (12)
This makes a queuing analysis feasible. Then introduce lemma 1 as following:

Lemma 1: Assume that the events \( E_i \) are of the form \( E_i = \{ J_i < b \} \), where \( J_i = J_{i-1} + L_{i-1} \) for \( 1 \leq i \leq n \) and where \( J_0, L_0, \ldots, L_n \) are independent. Then for \( 1 \leq i \leq n \)

\[
P(E_i \mid E_{i-1}, \ldots, E_0) \geq P(E_i)
\]

Let \( E_i \) denote the event \( \{ K_{2^{n-i}} < b + c2^{n-i} \} \) and assuming them to be independent, then we can get

\[
P(\hat{Q} > b) = 1 - P(\cap_{i=0}^n E_i) \leq 1 - \prod_{i=0}^n P(E_i) = P_{\text{app}}(\hat{Q} > b)
\]

(13)

Then we arrive at an upper bound approximation \( P_{\text{app}}(\hat{Q} > b) \) of \( P(\hat{Q} > b) \). That is

\[
P(\hat{Q} > b) \approx 1 - \prod_{i=0}^n P(E_i) = P_{\text{app}}(\hat{Q} > b)
\]

(14)

As to the specific multifractal wavelet analysis, we can get:

\[
K_{2^{n-i}} = U_{0,0} \prod_{j=0}^{i-1} (1 - A_j)/2
\]

(15)

So the event \( E_i \) is

\[
E_i = \{ K_{2^{n-i}} < b + c2^{n-i} \} = \{ \log(U_0) + \sum_{j=0}^{i-1} \log(1-A_j) < \log(b2^i + c2^n) \}
\]

(16)

For choosing the beta distribution as those \( A_j \), we use the cumulative value to compute \( P(E_i) \), then we can get the queue tail asymptotic.

4. SIMULATION RESULTS

We use the well-known real data traces in our experiments, that is AUCK[1] which contains the number of bytes per 2ms of recorded WAN traffic (mostly TCP packet). The mean rate of AUCK is 1.46Mbps, and AUCK contains \( 1.8 \times 10^6 \) data points. The Hurst parameter of AUCK
obtained from the variance-time plot using time scales 512ms to 262.144s is \( H = 0.86 \). We perform our simulation and analysis under a single queue with constant service rate and AUCK traffic input. We process the real traffic simulation, the analysis estimation from (4) and estimation based on wavelet separately to validate the queue analysis simulation and prove the wavelet method present better results.

To compute the (4), we choose the time unit as dyadic ones, that is:

\[
E[ X^{2^m \Delta t} | q ] = c(q)(2^m \Delta t)^{\tau_0(q)} \quad m = 1, 2, 3...
\]

Choose the \( \Delta t \) as the time unit, then:

\[
\log E[ X^{2^m \Delta t} | q ] = \tau_0(q) \log m + \log c(q) \quad q>0
\]

(17)

by simple linear regression, we can compute the discrete \( t(q) \) and \( c(q) \), then the analysis estimation can be done.

Through Discrete wavelet transform, we can get the needed parameter to perform our estimation (14).

In this paper, we present two simulation results in figure 1 and figure 2 separately. The Hurst parameter is 0.8 in figure 1, and 0.85 in figure 2. The X axis indicates the queue size in 100kb unit, Y axis indicates \( P(Q > b) \). The upper dash-dotted line is the real traffic simulation, the middle solid line is the estimation based on wavelet, and the lower dotted line is the analysis estimation. Intuitively in both simulation results, the estimation in wavelet presents a better estimation.

![Figure 1 Queue Analysis Results](image-url)
5. CONCLUSION

The remarkable characteristic in investigated network traffic have arose the interests of many researchers, which is convinced to be the scale nature, especially the multiscale phenomena over small time scale. In this paper, we present the efficient queue tail analysis with multiscale input based on multifractal wavelet analysis, with the scale nature of wavelet, it is more convenient and accurate to perform the tail asymptotic analysis. These results validated by queuing simulation of measured network traffic. Further research may be carried out the prediction and control based on the queue results in the wavelet domain.

REFERENCES

1. NLANR, Auckland-II trace archive. http://moat.nalnr.net/Traces