JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 25, 1783-1801 (2009)

Modeling and Analysis of Wireless LAN Traffic*

DASHDORJ YAMKHIN AND YOUJIP WON+
Department of Electronics and Computer Engineering
Hanyang University
Seoul, 133-791 Korea
E-mail: {dashdorj; yjwon}@ece.hanyang.ac.kr

In this work, we present the results of our empirical study on 802:11 wireless LAN network traffic. We collect the packet trace from existing campus wireless LAN infrastructure. We analyze four different data sets: aggregate traffic, upstream traffic, downstream traffic, tcp only packet trace from aggregate traffic. We analyze the time series aspects of underlying traffic (byte count process and packet count process), marginal distribution of time series, and packet size distribution. We found that in all four data sets there exist long-range dependent properties in terms of byte count and packet count process. Inter-arrival distribution is well fitted with Pareto distribution. Upstream traffic, *i.e.* from the user to Internet, exhibits significant difference in packet size distribution from the rests. Average packet size of upstream traffic is 151:7byte while average packet size of the rest of the data sets are all greater than 260bytes. Packets with full data payloads constitute 3% and 10% in upstream traffic and downstream traffic, respectively. Despite the significant difference in packet size distribution, all four data sets have similar Hurst values. The Hurst alone does not properly explain the stochastic characteristics of the underlying traffic. We model the underlying traffic using fractional-ARIMA (FARIMA) and fractional Gaussian Noise (FGN). While the fractional Gaussian Noise based method is computationally more efficient, FARIMA exhibits superior performance in accurately modeling the underlying traffic.

Keywords: network traffic, modeling, analysis, self-similarity, long-range dependence, fractional-ARIMA, fractional Gaussian noise

1. INTRODUCTION

Over the past several decades network and communication technology has been a significant and growing component of Internet traffic. Integrated broadband networks are expected to support various traffic types such as data, voice, image and video. Traffic generated from these services is substantially different in its statistical characteristics and networks are required to maintain a certain level of throughput during each session. For example, voice traffic has a bandwidth requirement of several Kbps and is delay sensitive, while high speed data traffic for file transfers or LAN/WAN interconnection requires hundreds of Mbps and is loss sensitive. Complex network traffic requires elaborate modeling and analysis which can be quite unconventional in an engineering sense. Traditional modeling tools and techniques, both theoretical and empirical, have been able to characterize and understand the behavior of network traffic to a rather limited extent. However, the discovery of the free-scaling nature of measured teletraffic has led to the creation of modeling solutions that can approximate data characteristics much better than

Received December 24, 2007; revised April 22 & July 16, 2008; accepted October 2, 2008.
Communicated by Ten-Hwang Lai.
* This work was funded by National Research Lab Grant (ROA02007-000-200114-0) by KOSEF.

previous techniques. As a result, self-similar processes have been used to successfully model data exhibiting long-range dependence in a variety of different scientific fields, including hydrology [2], geophysics [8], biology [4], telecommunication networks [16] and economics [7].

The objective of this study is to obtain a comprehensive understanding of the underlying packet level traffic including the stochastic characteristics of the underlying time series, marginal distribution, and packet size distribution. Our study is based upon the real network packet trace data collected from an existing public 802.11 network. We analyzed the underlying network traffic from an upstream/downstream point of view with regard to the aggregate traffic. We found that aggregate traffic, upstream, and downstream traffic all exhibit long memory properties. However, the packet size distributions of upstream and downstream traffics are different. In our study, it was found that the average packet size in downstream traffic is much larger (298byte) than in upstream traffic (151:7byte). This is because a large fraction of packets in downstream traffic carries a full data payload whereas upstream traffic usually carries signal packets, *e.g.* SYN, ACK, and *etc.*, which carry 8 byte signal information. In this study, our work is concentrated on characterizing traffic in wireless LAN environment. Traffic studies consist of either the study of aggregate traffic or the study of application dependent traffic characteristics. Our work focuses on aggregate traffic. Traffic characteristics based upon the application type is a very important topic. We like to address the application dependent traffic characterization study in a separate context. We develop an analytical model for underlying network traffic using fractional Gaussian Noise (FGN) and fractional ARIMA (FARIMA) to model the long memory nature of the underlying traffic. The FARIMA process is better than FGN in modeling the underlying traffic in that it can consider short range as well as long range dependence.

2. RELATED WORK

A few works have focused on identifying human behaviors in term of mobility patterns [15] in wireless environments. These studies reflect a variety of wireless environments, such as university campuses, auditoria and enterprise networks. Balachandran *et al.* [1] analyzed user behavior and network performance in public-area wireless networks. They developed a parameterized model for wireless users for use with analytic and simulation studies and for the application of workload analysis results to issues in wireless network deployment, such as capacity planning and potential network optimizations. Gurtov *et al.* [10] designed transport protocols that can be improved by providing easy models (consider the interplay between wireless links and transport protocols) of wireless links that strike a balance between realism, generality and detail. Park *et al.* [20] showed that the degree to which file sizes are heavy-tailed directly determines the degree of traffic self-similarity. By measuring self-similarity via the Hurst parameter H and file size distribution by its power-law exponent, they show that there is a nearly linear relationship between x and y over a wide range of network conditions and when the relationship is subject to the influence of the protocol stack. This mechanism gives a simple explanation of why self-similar network traffic may be observed in many diverse contexts. Mah *et al.* [14] developed an empirical model of network traffic produced by HTTP.

Instead of relying on server or client logs, their approach is based on packet traces of HTTP conversations. Through traffic analysis, they determined statistics and distributions for higher-level quantities such as the size of HTTP files, the number of files per web page, and user browsing behavior. Paxson *et al*. [21] presented a fast Fourier transform method for synthesizing approximate self-similar sample paths and, fractional Gaussian noise. Their method is as fast or faster than existing methods and appears to generate close approximations to true self-similar sample paths. Grossglauser *et al*. [9] argue that most recent modeling work has failed to consider the impact of two important parameters: (1) finite range of time scales of interest in performance evaluation and prediction problems; and (2) first-order statistics such as the marginal distribution of a process. Specifically, their model is a modulated fluid traffic model in which the correlation function of the fluid rate is asymptotically second-order self-similar with a given Hurst parameter, then drops to zero at a cutoff time lag. Tudjarov *et al*. [17] analyzed different protocols, *e.g*. TCP and UDP, and performed statistical analysis through the correlation coefficients, covariance, and self-similarity degree. Their experimental study captured traffic with a Hurst parameter around 0:7-0:75. They use the Maximum Likelihood approach to fit the obtained time series to existing distributions, such as Pareto and exponential distribution, where the first distribution is a self-similar process and the second is not. Leveraging the tree structure of the model, Vinay *et al*. [23] derived a multiscale queuing analysis that provides a simple closed form approximation to the tail queue probability. The analysis is applicable not only to the MWM (Multifractal Wavelet model) but to tree-based models in general, including fractional Gaussian noise. Their result clearly indicates that the marginal distribution of traffic at different time-resolutions affects queuing and that a Gaussian assumption can lead to over-optimistic predictions of tail queue probability even when taking LRD (Long Range Dependence) into account. Oliveira *et al*. [3] examined the long range dependent nature of application traffic in wireless environment. They analyzed three types of traffic in wireless LAN environment: http, ftp and video streaming, and examined the self-similarity of the network traffic. This in contrast to our study in which we examined the degree of self-similarity in wireless LAN traffic and the effectiveness of the existing stochastic model, fractional Gaussian Noise and fractional ARIMA in modeling the wireless LAN traffic. While our work touches upon the workload of a Fast Ethernet layer and TCP layer for one and two direction traffic data sets (for packet size and packet counts), we are mainly interested in the statistical characteristics and the estimated the Hurst parameter and modeled the Fractional Gaussian noise and FARIMA by using a self similar method in the field of wireless technology. We also modeled for inter arrival time by using Pareto, Weibull and Lognormal type distributions and got the results that is Pareto type distribution is more correlated to real data networks traffic data sets. Closer to our work, the authors in [18, 19] briefly describe the self similar network traffic in wired networks. These studies are however on self-similar based which has been shown to have a highly self similar nature in the autocorrelation function and inter arrival time figures whereas in our study we have calculated its statistical distributions. The rest of the paper is organized as follows. Section 3 presents the mathematical notion of self-similarity. Section 4 describes measurement setup. Section 5 calculated the traffic statistics. Section 6 describes the long range dependent. Section 7 carries the modeling of wireless LAN traffic. Section 8 concludes the paper.

3. SYNOPSIS: LONG RANGE DEPENDENT PROPERTY

The notion of a long memory process or self-similar process has been widely used to explain the nature of a various network traffic, *e.g.* WAN (Wide area network) [13], LAN (Local Area Network) [16], variable rate video [12, 28] *etc*. Determining whether a certain underlying process is self similar or not is quite a subjective issue. Still, it is true that the notion of self-similarity provides an effective explanation for the empirical phenomenon without which it is difficult to explain. Before we move along, we revisit the mathematical definition of self-similarity. The basic definition of a self-similar process is as follows:

A continuous-time stochastic process $\{X_t\}$ is strongly self-similar with a self-similarity parameter H ($0 < H < 1$), if for any positive stretching factor c, the rescaled process with time scale ct, $c^{-H}X_{ct}$, is equal in distribution to the original $\{X_t\}$ [25]. This means that, for any sequence of time points t_1, t_2, t_3, …, t_n, and for all $c > 0$, $\{c^{-H}X_{ct_1}, c^{-H}X_{ct_2}, \ldots, c^{-H}X_{ct_n}\}$, has the same distribution as $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ [25]. In discrete-time case, let $\{X_k\} = \{X_k: k = 0, 1, 2, \ldots\}$ be a (discrete-time) stationary process with mean μ, and variance σ^2. Let $\{X_k^m\} = \{X_1^m, X_2^m, \ldots\}$, $m = 1, 2, 3, \ldots$, be a sequence of batch means, *i.e.* $X_k^m = (X_{km-m+1} + \ldots + X_{km})/m$, $k \geq 1$. The process $\{X_t\}$ with autocorrelation function $\rho_k \to k^{-\beta}$, as $k \to \infty$, $0 < \beta < 1$, is called exactly self-similar with $H = 1 - (\frac{\beta}{2})$, if $\rho_k^m = \rho_k$, for any $m = 1, 2, 3, \ldots$. In other words, the process $\{X_k\}$ and the averaged processes $\{X_k^m\}$, $m \geq 1$, have identical correlation structure. The process $\{X_k\}$ is asymptotically self-similar with $H = 1 - (\frac{\beta}{2})$, if $\rho_k^{(m)} \to \rho_k$, as $m \to \infty$ [27]. There are many approaches for making a certain process self-similar. One of the most widely used approaches is to use Fractional Gaussian Noise [25]. The incremental process $\{Y_k\} = \{X_k - X_{k-1}\}$, $k > 0$, is called the fractional Gaussian Noise (FGN) process if $\{X_k\}$ designates a fractional Brownian motion (FBM) random process. Fractional Brownian is the model which is used widely for modeling self-similar processes as it provides tractable analysis techniques [18]. The main properties of self-similar processes include [2, 5, 13]: (i) *Slowly decaying variance*, *i.e.*, the variance of the sample mean decreases more slowly than the reciprocal of the sample size, that is, $Var[X_k^{(m)}] \to c\tau^{-\beta}$ as $\tau \to \infty$, where c is constant and $0 < \beta < 1$ [24]. In this process, the Hurst parameter corresponds to $H = 1 - (\frac{\beta}{2})$ [28]. (ii) Covariance structure of the self-similar process looks as $\gamma_x(t, s)\frac{\sigma^2}{2}\{t^{2H} - |t-s|^{2H} + s^{2H}\}$. (iii) Its auto-correlation function ρ_k is non-summable, *i.e.*, $\sum_{k=0}^{\infty}\rho_k = \infty$. The speed of decay of autocorrelations is more hyperbolic than exponential.

There are a number of ways to determine the degree of self-similarity, H. These include rescaled adjusted R/S statistics [25], variance time plots [13], wavelet analysis [28] *etc*. In this work, we use R/S plot in determining the H parameter of the underlying traffic. For a given set of numbers $\{X_1, X_2, \ldots, X_n\}$, a Hurst parameter H can be estimated from the rescaled adjusted range $\frac{R(n)}{S(n)}$ (or R/S statistics) where $R(n) = \max\{\sum_{i=1}^{k}(X_i - \mu), 1 \leq k \leq n\}$. An asymptotic slope on a loglog plot of R/S statistics represents the H parameter. Further interested users are referred to [18, 25].

4. MEASUREMENTS SETUP

Fig. 1 shows the network configuration and connection used in this study. To a cover wider geographical area, it is more cost effective to use wireless network than to use wired network technology. Particularly, in a sparsely populated country such as Mongolia, wireless network is the more preferred communication medium. In a wireless network there is one six-sector antenna system where each sector antenna approximately covers 60° degree angle and adjacent sector antennas slightly overlap with each other. 40 wireless clients are connected to 2 Access Points of the provider. Routing of all connections, and also the control and management of throughput (Traffic Shaping, QoS) are carried out with a router. Each wireless client has throughput ranging from 64 up to 512 kbps.

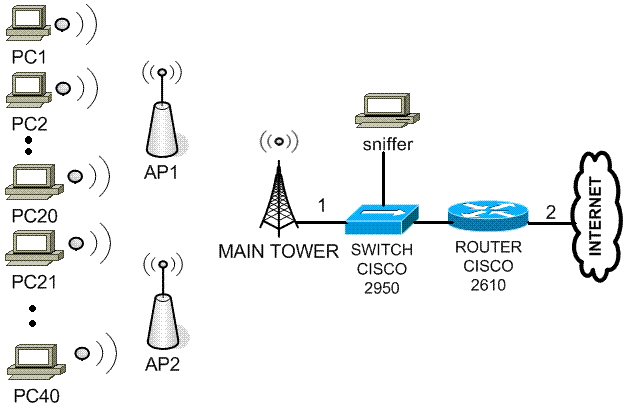
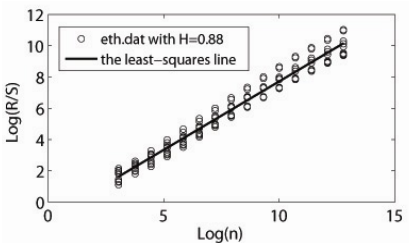
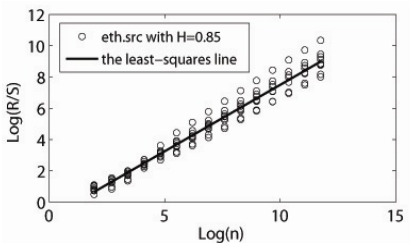


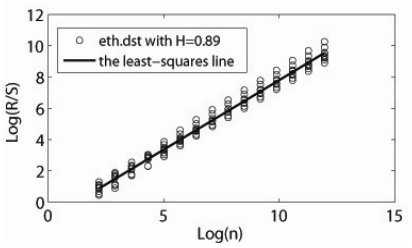
Fig. 1. The wireless network configuration.



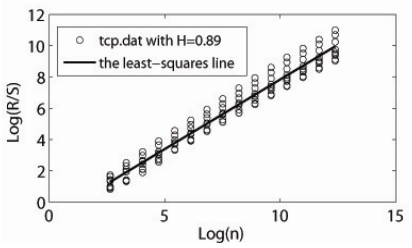
(a) eth.dat: $H = 0.88$.



(b) eth.src: $H = 0.85$.



(c) eth.dst: $H = 0.89$.



(d) tcp.dat: $H = 0.89$.

Fig. 2. Degree of self-similarity in datasets: *R*/*S* plots.

We use sniffer to collect the packet trace. A sniffer is program or a device that eavesdrops on the network traffic by grabbing information traveling over a network. These devices work in this manner because the Ethernet was built around a principle of sharing. Computers can be made to accept messages that are not meant for them. A computer connected to the LAN has 2 addresses. One is the MAC address that uniquely identifies each node in a network and is stored on the network card itself. It is the MAC address that is utilized by the Ethernet protocol while building 'frames' to transfer data to and from a machine. The other is the IP address which is used by applications. The data link layer uses an Ethernet header with the MAC address of the destination machine rather than the IP address. The network layer is responsible for mapping IP network addresses to the MAC address as required by the Data link Protocol. It initially look up the MAC address of the destination machine in a table, usually called the ARP cache. If no entry is found for the IP address, the address resolution protocol broadcasts a request packet (ARP request) to all machines on the network. The machine with that address responds to the source machine with its MAC address. This MAC address then gets added to the source machines ARP Cache. This MAC address is then used by the source machine in all its communications with the destination machine. Traffic used to collect personal computer running Windows 2000 Professional by its protocol analyzer Ethereal (sniffer), which ensures the accuracy of determining the time stamp of a package of 10^{-6} seconds. Sniffer is connected to the network so as to record traffic going through Point "1" and simultaneously through Point "2" in Fig. 1. Please note that the point of "1" receiving traffic information sharing among wireless customers, and with it the traffic flow of information between customers and Internet. After a point "2" is only the latest of them. All packages are recorded down to the file format tcpdump. More than 12.7 million packages were collected in our study. Of these, 70 percent were used to construct the TCP datagram.

5. TRAFFIC STATISTICS

5.1 Primitive Statistics

We collected the packet traffic from 10:00 to 17:00 on March 18, 2005. Our traffic data consists of four data sets shown in Table 1. Table 2 summarizes the trace data statistics. We obtain packet traces from two different layers of protocol stacks: Ethernet layer and TCP layer. Average packet size for "eth.dat", "eth.src", "eth.dst" and "tcp.dat" are 267.4, 151.7, 298.5 and 270.7byte, respectively. "eth.src" has different characteristics from the other three files. Average packet size of "eth.src" is 151.7byte, which is much smaller than the average packet size of the other files. For the other three files, average packet size is 267bytes or larger. Variance of packet size is also much smaller in "eth.src" than in the other files. Variance of packet size in "eth.src" is approximately 33% of the variance of packet size in the other files. Average inter-arrival times are 2msec, 5.5msec, 4.5msec, 3msec for "eth.dat", "eth.src", "eth.dst", and "tcp.dat", respectively. Fig. 3 shows the result of the aggregated traffic for 1 hour time intervals for the trace data. We collected four traffic data sets which there are haven in 2 direction "eth.dat", "tcp.dat" and 1 direction "eth.src", "eth.dst" data traffic sets with 1 microsecond time interval. The objective of this work is to examine the characteristics of wireless network traffic.

Table 1. File description of wireless network.

	File Name	File Description	Protocol layer
1	eth.dat	Aggregate traffic, captured at point 1	2 (Ethernet)
2	eth.src	Upstream traffic, captured at point 2	2 (Ethernet)
3	eth.dst	Downstream traffic, captured at point 2	2 (Ethernet)
4	tcp.dat	TCP traffic, captured at point 2	4 (TCP)

Table 2. Traffic of data sets (10:00-17:00, Mar-18-2005).

Data Set	Number of packets	Packet size (byte)		Inter arrival time (msec)	
		μ	σ^2	μ	σ^2
eth.dat	12 715 077	267.4	204760	2	10
eth.src	4 554 667	151.7	80103	5.5	67
eth.dst	5 586 555	298.5	239430	4.5	71
tcp.dat	8 468 547	270.7	279430	3	45

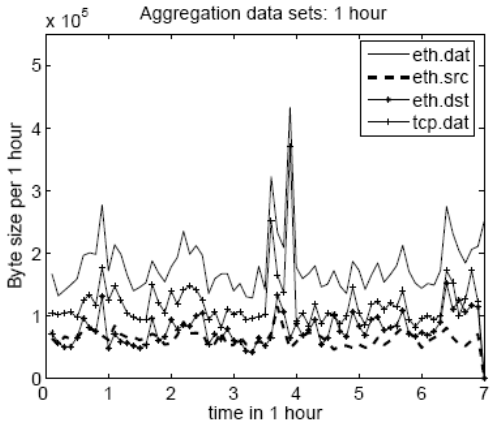


Fig. 3. Statistical analysis for data sets.

5.2 Application Layer Statistics

To understand the behavior of the application layer traffic and provide reference to study of scheduled traffic for the wireless network, this section gives detail statistics and understandings of TCP wireless traffic, specifically its port usages at point 1 at Fig. 1. The TCP traffic trace used in the paper has been manipulated and sanitized to hide the source and destination IP addresses, since it causes security issues. However, the port numbers in the traces are intact, which is meant to be and used for only scholarly purposes. Internet Assigned Numbers Authority (IANA) provides the list of well known ports in the range of 0 to 1023, and Tables 3 and 4 distinguishes well known ports and registered port numbers in the captured wireless traffic. Top ten of each region is listed respectively. Wireless traffic trace shows simple behavior where the majority services exploited by the users are World Wide Web, E-mail, FTP, and DNS services.

Table 3. Ranks of TCP traffic well known port and its applications.

Rank	Ports	Applications	Percentage	Packet Sum
1	80	WWW HTTP	37.80%	117Mbyte
2	443	HTTPS	22.59%	70Mbyte
3	135	EPMAP	8.35%	26Mbyte
4	25	SMTP	5.29%	16.5Mbyte
5	110	POP3	3.21%	10.0Mbyte
6	20	FTP-DATA	2.01%	6.2Mbyte
7	53	DNS	2.00%	6.2Mbyte
8	21	FTP-Control	0.26%	0.8Mbyte
9	119	NNTP	0.12%	0.3Mbyte
10	143	IMAP	0.07%	0.2Mbyte

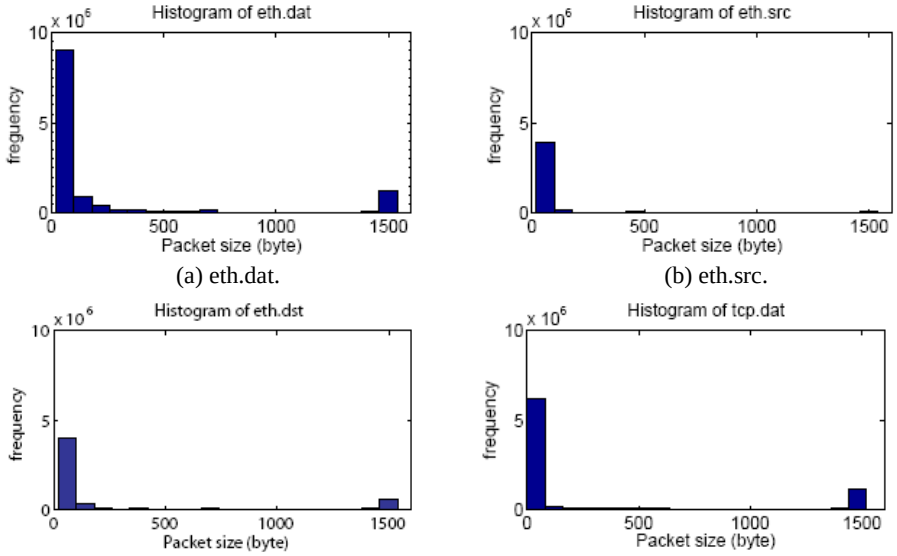
Table 4. Ranks of TCP traffic registered ports and its applications.

Rank	Ports	Applications	Percentage	Packet Sum
1	4899	RAdmin Port	2.57%	8.0Mbyte
2	1494	ICA	2.52%	7.8Mbyte
3	1790	NMSP	1.29%	4.0Mbyte
4	5190	AOL	0.92%	2.8Mbyte
5	4662	OMS	0.85%	2.6Mbyte
6	4000	ICQ	0.59%	1.8Mbyte
7	49753	Unknown	0.41%	1.2Mbyte
8	8080	HTTP Alternate	0.36%	1.1Mbyte
9	1297	SDPROXY	0.29%	0.9Mbyte
10	1934	IBM LM Appl. Agent	0.28%	0.8Mbyte

These top ten rank services holds about 80% of wireless TCP traffic. There were total of 5637 distinct port numbers found in the traffic trace. There are 23 well known ports, and the rest are used by registered ports. From the Table 3, it is comprehensible that wireless users tend to navigate the WWW, check E-mails, and uses it to download contents. Table 4 shows registered ports used in the captured traffic trace. Port number 4899 is used by remote control software called RAdmin, which allows mobile computer users to remotely connect to their desktop computers.

5.3 Packet Size Distribution

We examine the packet size distribution in the underlying traffic. As mentioned earlier, the packet size distribution of the upstream traffic, “eth.src” is smaller than the packet sizes of the others three data sets. Fig. 4 presents the histograms of the packet sizes. In “eth.dat”, “eth.dst” and “tcp.dat” more than 10% of the packets carry a full data payload (1500bytes). On the other hand, in ”eth.src”, the percentage of packets with a full data payload is 3% . Figs. 4 (a)-(d), show the respective packet size distributions for the wireless network traffic for eth.dat, eth.src, eth.dst, and tcp.dat. Packet sizes from 60bytes to 1514bytes were generated by the wireless traffic data. The histogram clearly demon-



(a) eth.dat. (b) eth.src.

(c) eth.dst. (d) tcp.dat.

Fig. 4. Histogram in datasets.

Table 5. Packet size statistics of wireless network.

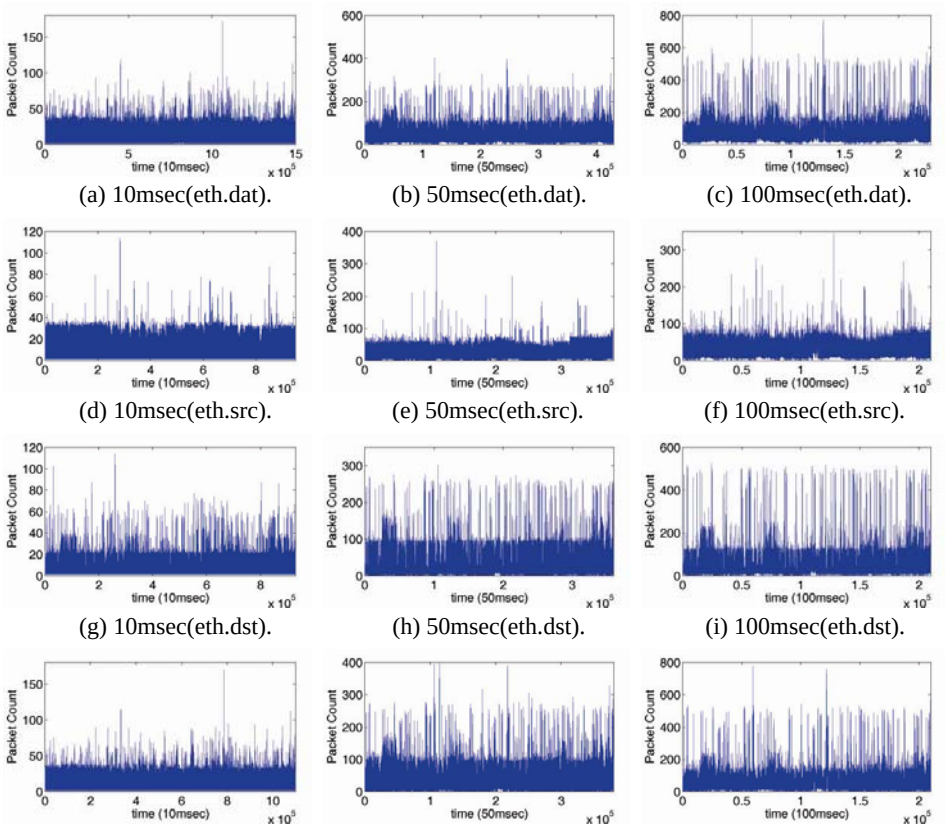
Rank	eth.dat		eth.src		eth.dst		tcp.dat	
	size	%	size	%	size	%	size	%
1	60	24.6%	60	33%	62	30%	40	66.6%
2	62	16.6%	74	25.2%	60	17.3%	1500	12.1%
3	74	13.0%	62	8.6%	1514	10.4%	66	2.0%
4	1514	9%	66	5.7%	74	8.2%	1488	1.4%
5	66	3%	80	2.7%	106	3.6%	52	0.6%
6	64	2.8%	1514	2.6%	78	3.2%	46	0.58%
7	214	2.3%	102	2%	64	2.6%	192	0.5%
8	150	2%	91	1.7%	80	2.3%	1420	0.47%
9	106	1.7%	64	1.4%	709	2%	628	0.38%
10	94	1.6%	87	1.2%	66	1.5%	54	0.37%
11	78	1.5%	82	0.7%	84	1.14%	1400	0.33%
12	80	1.4%	99	0.66%	1434	0.88%	1216	0.26%

strates the heavy tail presence in the traffic distribution (see a histograms for eth.dat, eth.src, eth.dst and tcp.dat in Figs. 4 (a)-(d)). Table 5 illustrates the percentages of most common packet sizes. Packets with less than 80bytes are usually used to carry control messages. As can be seen in Table 5, more than 85% of the packets are less than 100bytes. This characteristic persists through all four traces. In terms of the packet count, packets with full data payloads constitutes approximately 10% and 2.6% in downstream and up-stream traffic, respectively.

6. LONG RANGE DEPENDENCY

6.1 Empirical Observation

We first visually examine the burstiness of the traffic under different time scale. We plot the packet count and the byte count process in 10msec, 50msec, 100msec intervals for the four data sets. As can be seen, we do not observe a noticeable smooting effect even though we increase the time scale of interest by orders of magnitude in Figs. 5 and 6. The Hurst parameters are presented in Table 7. Table 6 presents the result of statistical parameters for synthetic data set of FGN and FARIMA. Our work focuses on analyzing the aggregate traffic in wireless LAN environment and on examining the effectiveness of the existing stochastic model for wireless LAN traffic. As currently it stands, we do not examine the application type of the packets and indeed our packet trace does not include application specific information. We plan to examine the application specific traffic characteristics in the separate context.



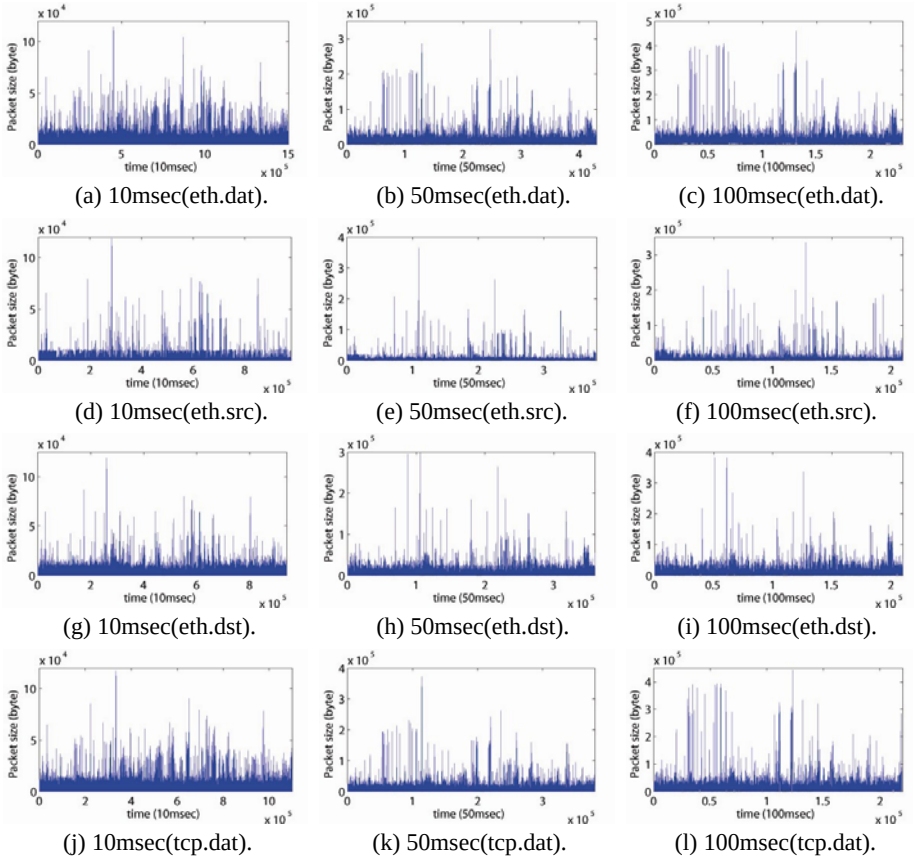
(a) 10msec(eth.dat). (b) 50msec(eth.dat). (c) 100msec(eth.dat).

(d) 10msec(eth.src). (e) 50msec(eth.src). (f) 100msec(eth.src).

(g) 10msec(eth.dst). (h) 50msec(eth.dst). (i) 100msec(eth.dst).

(j) 10msec(tcp.dat). (k) 50msec(tcp.dat). (l) 100msec(tcp.dat).

Fig. 5. Byte count for eth.dat, eth.src, eth.dst, tcp.dat.



(a) 10msec(eth.dat). (b) 50msec(eth.dat). (c) 100msec(eth.dat).

(d) 10msec(eth.src). (e) 50msec(eth.src). (f) 100msec(eth.src).

(g) 10msec(eth.dst). (h) 50msec(eth.dst). (i) 100msec(eth.dst).

(j) 10msec(tcp.dat). (k) 50msec(tcp.dat). (l) 100msec(tcp.dat).

Fig. 6. Packet count for eth.dat, eth.src, eth.dst, tcp.dat.

Table 6. Description of model parameters.

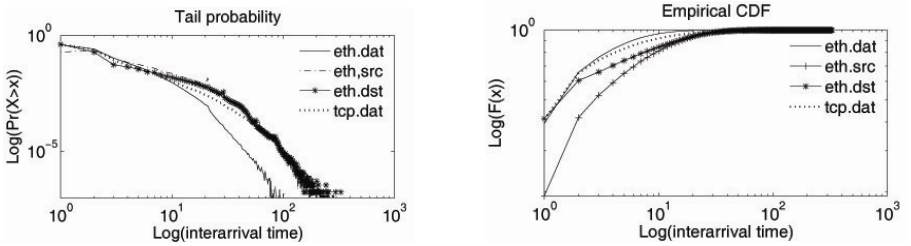
Data sets	μ (byte)	σ^2 (byte)^2	*H* parameter		
			Trace Data	FARIMA	FGN
eth.dat	267.4	204760	0.88	0.85	0.82
eth.src	151.7	80103	0.85	0.84	0.83
eth.dst	298.5	239430	0.89	0.84	0.83
tcp.dat	270.7	279430	0.89	0.83	0.83

Table 7. *H* parameter with 10msec, 100msec aggregation.

	H parameter							
	eth.dat		eth.src		eth.dst		tcp.dat	
	10ms	100ms	10ms	100ms	10ms	100ms	10ms	100ms
Byte count	0.82	0.78	0.74	0.85	0.81	0.75	0.82	0.79
Packet count	0.84	0.79	0.80	0.79	0.83	0.80	0.84	0.82

6.2 Sample Autocorrelations

Fig. 9 illustrates the sample autocorrelations for the underlying traffic. As can be seen, ACF decays very slowly with respect to the lag. There are variety of factors which cause the network traffic (Layer 3) or Ethernet traffic (Layer 2) to be appear self-similar. These include the heavy tailed nature of the file size distribution, TCP congestion algorithm [25], on-off distribution of a user's computer usage model, or a combination of these factors. The properties of self-similar processes lend themselves to the following methods for estimating H [6, 13, 25]. We obtain the Hurst parameter of each data set by an R/S plot. The R/S statistics of an aggregated process increases linearly (for large n) in log-log plots over n. The slope of the regression line for these R/S samples are an estimate for the Hurst parameter H for the 4 data sets (see in Figs. 2 (a)-(d)). We calculated the cumulative distribution function and tail probability function for interarrival time in Fig. 7.



(a) Cumulative distribution function for inter arrival time.

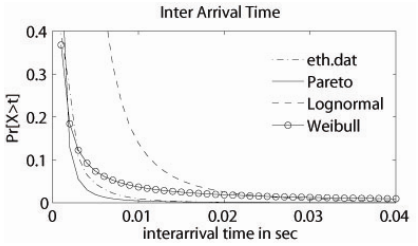
(b) Tail probability function for inter arrival time.

Fig. 7. Inter-arrival time statistics.

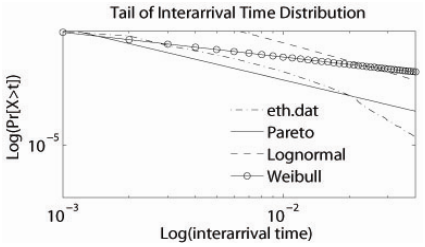
7. MODELING OF WIRELESS LAN TRAFFIC

7.1 Background

A number of analytical models have been proposed to model the long memory nature of the underlying traffic. In this paper, we use fractional Gaussian noise (FGN) and fractional ARIMA (FARIMA) to model the underlying traffic. FGN algorithm is based on synthesizing sample paths that have the same power spectrum as FGN. These sample paths can then be used in simulation as traces of self-similar network traffic. The algorithm is a fast approximation of the power spectrum of an FGN process; this approximation also has applications for fast estimation of the strength of long-range dependence (Hurst parameter) present in network arrival processes [21]. We also generated the actual traffic data by using an FFT algorithm for fractional Gaussian noise synthesis (see a more detailed algorithm in the next chapter). Recent real traffic measurements found that the co-existence of both long-range and short-range dependence in traffic traces [18]. Models are required to describe both long-range and short-range dependence simultaneously. We consider the F-ARIMA (fractional autoregressive integrated moving average) model as one of the better models with this capability. We provide a procedure to fit a FARIMA model to the actual traffic trace, as well as a method to generate a FARIMA process with

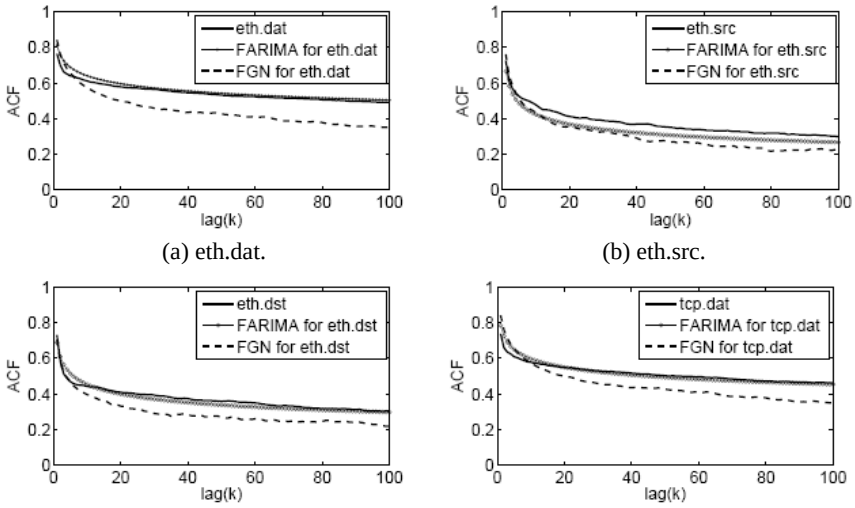


(a) Inter-arrival time fitting: eth.dat vs. Pareto, Weibull, Log-normal distribution.



(b) Tail probability function for inter-arrival time.

Fig. 8. Probability of inter-arrival time.



(a) eth.dat. (b) eth.src.

(c) eth.dst. (d) tcp.dat.

Fig. 9. Autocorrelations of sample data sets and ACF of synthetic traffic generated by FGN and FARIMA.

given parameters [10]. We generated the raw packets process using a Gaussian Motion noise generator in the same manner as F-ARIMA by using Hosking's algorithm. We estimate the Hurst parameter from the wireless LAN traffic and by using that the Hurst parameter modeled the synthetic traffic data for wireless LAN and calculated the auto-correlation function for the FGN data series and, the FARIMA data series and found that it is strongly correlated, similar to measurement of the a wireless LAN traffic data (in Fig. 9) and may have LRD structure. The obvious clustering of R/S plot points around a linear line suggests the presence of LRD of about H = 0:88 (in Figs. 2 (a)-(d)). We also simulated the interarrival time by using the Pareto, Weibull and Lognormal distributions in Figs. 8 (a) and (b). In Fig. 9 FGN, FARIMA simulation, also exhibits LRD and Self-similar properties in the LRD test. We also performed the simulation for Interarrival time with heavy-tailed distribution. LRD test applied to the reported traffic showed the presence of LRD with all most the near values for H [18, 19].

7.2 Self-similar Process from Fractional Gaussian Noise

Fractional Gaussian Noise and Fractional ARIMA are two of most widely studied self-similar processes [21]. In addition to these models, there are a number of stochastic models for synthesizing self-similar processes, *e.g. M*/*G*/∞ [22], Random Midpoint Displacement [18], Fractional Brownian Motion [25], Wavelet model [28]. In this study, we use FGN and FARIMA to model the empirical traffic data and examine their effectiveness in describing these self similar processes. Fractional Gaussian noise (FGN) is a process of FBM (Fractional Brown Motion) increments, *i.e.* $X_H(t) = 1/\delta[B_H(t + \delta) - B_H(t)]$, where δ is an increment and $B_H(t)$ is Fractional Brownian Motion. The $X_H(t)$ process is normally distributed *i.e. N* (0; σ^2), the normalized covariation function being of the form.

$$r(\tau) = (|\tau + 1|^{2H} - 2|\tau|^{2H} + |\tau - 1|^{2H})/2$$

All aggregated processes $X_H^{(m)}(t)$ have the same distribution for any $0 < H < 1$. FGN is exactly a self-similar process with the Hurst exponent H varying in the interval $1/2 < H < 1$. FGN is fully described by two parameters only, by variance and Hurst exponent H. The substantial argument in favour of FGN models in networks is that in many cases traffic can be considered as a superposition of a large number of separate independent ON/OFF sources having distributions with heavy tails for the ON-period duration [26]. For this case, after subtraction of the average arrival speed and necessary normalization in accordance with the central limit theorem, the aggregated ON/OFF sources (cumulative arrivals) converge to the Gaussian FBM. Therefore, the self-similar traffic (for the increment process) can be modeled as the model *FGN + mean* with a given variance and H. Unfortunately, FGN models have strict limitations when adapted to network traffic. First of all, it is not enough to have the H parameter only cover the complex correlation structure of real network processes. Moreover, other studies have proven the importance of short-term correlations for buffering and for discovery of significant timescales [16]. Secondly, the Gaussian features of FGN models may not correspond with reality, *e.g.* when the mean-square deviation is greater than the mean value. In this case, FGN outputs contain a large number of negative values. Thirdly, many real network application processes are not Gaussian, especially for small timescales. Due to the complex correlation structure of the underlying traffic, Fractional ARIMA model may be preferable to FGN despite its high computational complexity, $\vartheta(n^2)$ where n is the number of samples.

Paxson develop an efficient method to generate self-similar sample path using fast Fourier transform method [21]. This algorithm is based on a calculation of the power spectrum density with the use of a period gram (the power spectrum at the given frequency is represented by independent exponential random variables). In the first stage the complex numbers are constructed, their magnitudes are regulated by the normal distribution and after that the inverse FFT is fulfilled. Fig. 10 shows how self-similar sequences have been generated by means of the FFT. Fractional Gaussian Noise process is modeled as: $X(t) = \mu + \sigma * Z_H(t)$ where $Z_H(t)$ denotes the fractional Gaussian Noise process with Hurst parameter H. μ and σ is mean and standard deviation of $X(t)$ since $Z_H(t)$ is centralized normal process (see a the algorithm for FGN in Fig. 10). The illustration use of the model is shown by applying it to the Wireless LAN aggregate traffic presented

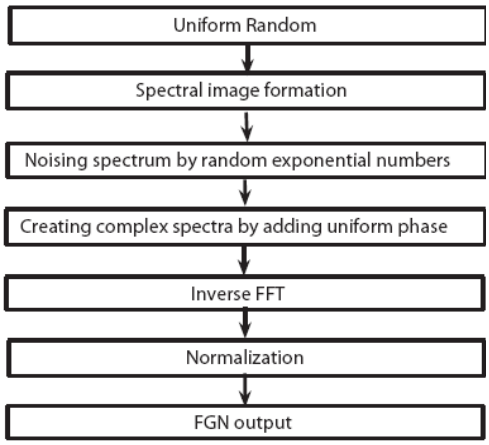


Fig. 10. Fast fourier transform algorithm for FGN.

above μ and σ, are estimated from the data sets of "eth.dat", "eth.src", "eth.dst" and "tcp.dat" frames for real wireless data traffic. The estimate of the Hurst parameter of the data sets are presented in Table 2 for "eth.dat", "eth.src", "eth.dst" and "tcp.dat" for our previous LRD test.

7.3 Self-similar Process from Fractional-ARIMA

We use the FARIMA (fractional autoregressive integrated moving average) process, which has some advantages over other models based on other fractal processes. FBM, the aggregation of ON/OFF sources with high changeability, *etc*., is an example of such processes. The fractional Brownian motion has only one parameter, which controls the correlation function, and therefore there is no flexibility for short-range dependence modeling. The aggregation of a large number of ON/OFF sources with infinite variance for ON and OFF periods allows for the formation of long-range dependence and can be used to cover the asymptotic behavior of long-range traffic. However, the possibility of short-range behavior simulation is not possible. *FARIMA* (*p*; *d*; *q*) models have three parameters, *p*, *d* and *q*, that control the correlation structure. Therefore, they can cover the short-range dependence as well as long-range dependence. It is necessary to have a model that will be able to cover the short-range dependence, the long range dependence and the arbitrary distribution. A *FRIMA*(*p*; *d*; *q*) process is a process where *d* is the level of differencing, *p* is the auto regression order, and *q* is the moving average order; *p* and *q* have non-negative integer values and *d* have non-integer value. We synthesize underlying packet traces using a fractional ARIMA process. We apply *FARIMA* (0; *d*; 0) to generate a synthetic sequence. *d* is the fractional differencing order, $0 < d < 1/2$, $H = d + 1/2$. We use Hosking's algorithm for generating self-similar processes using the FARIMA model [10, 11]. In practice, this method is very computationally intensive for generating long sequences. A number of Mathematical models have been proposed for Internet backbone traffic. They include fractional Gaussian Noise (FGN), fractional Brownian Motion, *M*/*G*/∞, MMPP (Markov Modulated Poisson Process) and *etc*. The contribution of this work is twofold. First, we carefully analyze the stochastic characteristics of the underlying

Wireless LAN traffic. Second, we examine the existing stochastic model for network traffic against our empirical data and verify the effectiveness of the respective models. It is found that FARIMA better models the underlying network traffic. This study delivers valuable and meaning result full which bridges the gap between empirical traffic study and analytical modeling. In this work, we found that FARIMA model more accurately represents the stochastic characteristics of the underlying traffic.

8. CONCLUSION

In this study, we perform comprehensive analysis on 802.11 network traffic. We capture packet traces from the existing wireless LAN environment. We analyze four data sets: aggregate traffic, upstream traffic, downstream traffic, and TCP only aggregate traffic. We examine the primitive statistics and sample autocorrelations of data sets. We found that the upstream traffic packet size distribution is significantly different than the other three data sets. The fraction of data packets *i.e*. packets with full data payloads are much smaller in upstream traffic. Packet-Inter arrival times are well fitted with a Pareto distribution. We examine the packet count process (packets/sec) and byte count process (byte/sec) of the underlying packet trace. We found that example auto-correlations decay slowly in their data sets. Parameters that are greater than 0.5 indicate that underlying data sets have long-range dependent property. We used FGN and FARIMA to model the long-range dependent property of the underlying traffic. We found that among the models tested, FARIMA more accurately synthesizes the long memory characteristics of the underlying traffic. The results of our work can be used in many areas. They include synthetic traffic generation, capacity planning in various network related hardware, and *etc*.

ACKNOWLEDGEMENT

Authors would like to thank Seongjin Lee for his helpful comments and discussions on this work.

REFERENCES

1. A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing user behavior and network performance in a public wireless LAN," in *Proceedings of ACM SIGMETRICS Performance Evaluation Review*, 2002, pp. 195-205.
2. J. Beran, "Statistics for long-memory processes," *Jerey Glosup Technometrics*, Vol. 39, 1997, pp. 105-106.
3. C. Oliveira, J. B. Kim, and T. Suda, "Long-range dependence in IEEE 802.11b wireless LAN traffic: An empirical study," in *Proceedings of IEEE 18th Annual Workshop on Computer Communications*, 2003, pp. 17-23.
4. J. J. Collins and C. J. de Luca, "Upright, correlated random walks: A statistical-biomechanics approach to the human postural control system," *Chaos*, Vol. 5, 1994, pp. 57-63.
5. D. Cox, *Computer Experiments with Fractional Gaussian Noises*, A Review, in Sta-

tistics and Appraisal Iowa State Statistical Library, 1984.
6. D. M. Etter, *Engineering Problem Solving with C*, Prentice Hall, New Jersey, 2004.
7. R. F. Felter, "Processes stochastiques fractals avec applications these de doctorat," Ph.D. Thesis, Department of Information and Communication, University of Paris, 1998.
8. U. Frisch, *Turbulence: The Legacy of A.N. Kolmogorov*, Cambridge University Press, Cambridge, UK, 1995.
9. M. Grossglauser and J. C. Bolot, "On the relevance of long- range dependence in network traffic," *IEEE/ACM Transactions on Networking*, Vol. 7, 1999, pp. 629-640.
10. J. R. M. Hosking "Fractional differencing," *Biometrika*, Vol. 68, 1981, pp. 165-176.
11. J. R. M. Hosking, "Modeling persistence in hydrological time series using fractional differencing," *Water Resources Research*, Vol. 20, 1984, pp. 1898-1908.
12. J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, Vol. 5, 1995, pp. 1566-1579.
13. W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, Vol. 2, 1994, pp. 1-15.
14. B. A. Mah, "An empirical model of http network traffic," in *Proceedings of the INFOCOM, 16th Annual Joint Conference of the IEEE Computer and Communications Societies, Driving the Information Revolution*, 1997, pp. 592-600.
15. M. McNett and G. M. Voelker, "Access and mobility of wireless PDA users," Technical Report No. CS'04, University of California, San Diego, USA, 2004.
16. W. Willinger, M. S. Taqqu, and V. Teverovsky, "Estimators for long-range dependence: An empirical study, fractals," *Fractal*, Vol. 3, 1995, pp. 785-798.
17. A. Tudjarov, D. Temkov, T. Janevski, and O. Firfov, "Empirical modeling of internet traffic at middle-level burstiness," in *Proceedings of the 12th IEEE Mediterranean*, Vol. 2, 2004, pp. 535-538.
18. O. Shiluhin, S. Smolskiy, and A. Ocin, *Fractial Process in the Telecommunication*, Radio Tehnika, Russian, 2003.
19. O. Shiluhin, S. Smolskiy, and A. Ocin, *Modeling for Information System*, Radio Technika, Moscow, Russian, 2005.
20. K. Park, G. Kim, and M. Crovella, "On the relationship between file sizes, transport protocols, and self-similar network traffic," Technical Report No. 1996-016, Boston University, 1996.
21. V. Paxson, "Fast, approximate synthesis of fractional Gaussian noise for generating self-similar network traffic," *ACM SIGCOMM Computer Communication Review*, Vol. 27, 1997, pp. 5-18.
22. J. M. Pitts and J. A. Schormans, *Introduction IP and ATM Design and Performance With Applications Analysis Software*, British Library Cataloguing in Publication Data, 2000.
23. V. Ribeiro, R. Riedi, M. S. Crouse, and R. G. Baraniuk, "Multiscale queueing analysis," *IEEE/ACM Transactions on Networking*, Vol. 14, 2000, pp. 1005-1018.
24. J. S. R. Lee, H. W. Park, and H. D. J. Jeong, "Simulation run-lengths of self-similar queueing processes," in *Proceedings of the 12th IASTED International Conference*

on Applied Simulation and Modelling, 2003, pp. 526-531.
25. W. Stallings, *High-Speed Networks and Internets: Performance and Quality of Service*, 2nd ed., Prentice Hall, New Jersey, 2002.
26. S. Teymori and W. Zhuang, "Queue analysis and multiplexing of heavy-tailed traffic in wireless packet data networks," *Mobile Networks and Applications*, Vol. 12, 2007, pp. 31-41.
27. B. Tsybakov and N. D. Georganas, "On self-similar traffic in ATM queues: Definitions, overflow probability bound, and cell delay distribution," *IEEE/ACM Transactions on Networking*, Vol. 5, 1997, pp. 397-409.
28. P. M. Krishna, V. M. Gadre, and U. B. Desai, *Multifractal Based Network Traffic Modeling*, Kluwer Academic Publishers, Boston, 2003.

DETAILS OF LONG RANGE DEPENDENCE

The proof is as follows: Let us define a new process $Y(t) = X(t) - X(t-1)$ which is the increment process corresponding to $X(t)$ (sampled at every integer instant). The Long range dependence characteristics of the increment process can be obtained from the analysis of the covariance of the process $Y(t)$.

$$\begin{aligned} E[Y(t+k)Y(t)] &= E[(X(t+k)-X(t+k-1))(X(t)-X(t-1))] \\ &= E[X(t+k)X(t)]-E[X(t+k)X(t-1)] \\ &\quad -E[X(t+k-1)X(t)]+E[X(t+k-1)X(t-1)] \end{aligned} \tag{1}$$

The above expression can be simplified to obtain

$$\gamma_Y(k) = E[Y(t+k)Y(t)] = \frac{\rho^2}{2}\{(k-1)^{2H} - 2k^{2H} + (k+1)^{2H}\}. \tag{2}$$

The autocorrelation of the increment process $\rho(k)$ is obtained as

$$\rho_Y(k) = \frac{\gamma_Y(k)}{\rho^2} = \frac{1}{2}\{(k-1)^{2H} - 2k^{2H} + (k+1)^{2H}\}. \tag{3}$$

The asymptotic behavior of $\rho(k)$ can be obtained by using Taylor series expansion. Eq. (3) can be modified as

$$\rho(k) = \frac{k^{2H}}{2}[(1+\frac{1}{k})^{2H} - 2 + (1-\frac{1}{k})^{2H}] = \frac{k^{2H}}{2}\, g(k^{-1}), \tag{4}$$

where $g(x) = [(1+x)^{2H} - 2 + (1-x)^{2H}]$.

Taking the first and second derivatives of $g(x)$, we get

$$g'(x) = 2H[(1+x)^{2H-1} - (1-x)^{2H-1}],$$
$$g''(x) = 2H(2H-1)[(1+x)^{2H-2} + (1-x)^{2H-2}].$$

The Taylor series expansion of $g(x)$ can be written as

$$g(x - x_0) = g(x_0) + xg'(x_0) + \frac{x^2}{2} g''(x_0) + \ \ldots,$$

so that

$$g(x) = g(0) + xg'(0) + \frac{x^2}{2} g''(0) + \ \ldots.$$

Substituting the expressions for $g'(x)$ and $g''(x)$, the first non zero term in the expansion of $g(x)$ is seen as

$$g(x) = \frac{x^2}{2} 2H(2H - 1). \tag{5}$$

The expression for $\rho(k)$ will now become

$$\rho(k) = \frac{k^{2H}}{2} g(k^{-1}) = k^{2H} H(2H - 1)k^{-2},$$
$$\lim_{k\to\infty} \frac{\rho(k)}{H(2H - 1)k^{2H-2}} \to 1. \tag{6}$$

Dashdorj Yamkhin is currently Ph.D. candidate at Division of Electrical and Computer Engineering of Hanyang University, Seoul in Korea. He received the B.S. degree in 1991 from Novosibirsk State Technical University in Russia and M.S. degree in 1996 from Mongolian University of Science and Technology in Mongolia. His current research interests are network communication, network traffic modeling and analysis, network system design and embedded system.

Youjip Won is currently Associate Professor at Division of Electrical and Computer Engineering, Hanyang University, Seoul Korea. He is leading Distributed Multimedia Computing Lab. He did his B.S. and M.S. in Department of Computer Science, Seoul National University, Seoul, Korea in 1990 and 1992, respectively. He received his Ph.D. in Computer Science from University of Minnesota in 1997. Before joining Hanyang University in 1999, he worked at Intel Corp. as Server Performance Analyst. His research interests include internet traffic measurement, modeling and analysis, mining for actionable knowledge, multimedia system and networking, performance modeling and analysis, intelligent storage subsystem, and software support for low power system.