On Multimedia Networks: Self-Similar Traffic and Network Performance

Zafer Sahinoglu and Sirin Tekinay, New Jersey Institute of Technology

ABSTRACT The main objective in telecommunications network engineering is to have as many happy users as possible. In other words, the network engineer has to resolve the trade-off between capacity and QoS requirements. Accurate modeling of the offered traffic load is the first step in optimizing resource allocation algorithms such that provision of services complies with the QoS constraints while maintaining maximum capacity. In recent years, as broadband multimedia services became popular, they necessitated new traffic models with self-similar characteristics. In this article we present a survey of the self-similarity phenomenon observed in multimedia traffic and its implications on network performance. Our current research aims to fill the gap between this new traffic model and network engineering. An immediate consequence of this study is the demonstration of the limitations or validity of conventional resource allocation methods in the presence of self-similar traffic.

The future will bring a wide variety of multimedia applications each with different traffic characteristics at optimized performance, to be carried by both wireless and wireline networks. In wireless mobile networks the offered traffic varies both temporally and spatially, with the spatial variation significantly higher than in wired networks. Models of the traffic offered to the network or a component of the network will be critical to providing high quality of service (QoS). Traffic models are used as the input to analytical or simulation studies of resource allocation strategies.

We may view traffic at the application or packet level, where an application-level view may simply describe the offered traffic as "a videoconference between three parties," while the packet-level view is given by a stochastic model that mimics the arrival process of packets associated with this application reasonably well. Clearly, in order to quantify traffic, packet-level representation of applications will be used. An important feature of multimedia traffic at the packet level that has a significant impact on performance is traffic correlation. The complexity of traffic in a multimedia network is a natural consequence of integrating, over a single communication channel, a diverse range of traffic sources such as video, voice, and data that significantly differ in their traffic patterns as well as their performance requirements. Specifically, "bursty" traffic patterns generated by data sources and variable bit rate (VBR) real-time applications such as compressed video and audio tend to exhibit certain degrees of correlation between arrivals, and show long-range dependence in time (*self-similar traffic*). The questions that arise here are how prevalent such traffic patterns are and under what conditions performance analysis is critically dependent on taking self-similarity into account. There are different studies pointing out either the importance of self-similarity to network performance [1–4] or the irrelevance of the need for capturing self-similarity in traffic modeling [5]. To clarify this dilemma, a through understanding of *QoS* and *resource allocation* in a network environment is necessary. Optimal resource allocation is determining optimal buffer sizes, assignment of bandwidth, and other resources in order to get the desired QoS expressed in terms of parameters such as queuing delay, retransmission time, packet loss probability, and bit error rate.

Quality of Service

The International Organization for Standards (ISO) defines QoS as a concept for specifying how good the offered networking services are [6]. A layered model of the multimedia communication system (MCS) with respect to QoS is presented in Fig. 1. Generally, QoS parameters are performance measures such as bit error rate, frame error rate, cell loss probability, delay, and delay variation or guarantee, which is the maximum difference between end-to-end delays experienced by any two packets.

The user and application requirements for the MCS are mapped into a communication system that tries to satisfy the requirements of the services, which are parameterized. Parameterization of the services is defined in ISO standards through the notion of QoS.

In this article we concentrate on QoS in the network layer since the characteristics of network traffic and its effects on network performance are to be discussed. The set of chosen parameters for a particular service determines what will be measured as the QoS.

Network QoS parameters describe requirements for network services. They may be specified in terms of:

- Network load, characterized by average/minimal interarrival time on the network connection, packet cell size and service time in the node for the connection's packet/cell [7].
- Network performance, describing the requirements that the network services must guarantee. The performance might be expressed through a source-to-destination delay bound for the connection's packet loss rate [7].

Resource Allocation

Services for multimedia networked applications need resources to perform their functions. Of special interest are resources that are shared among application, system, and network. There are several constraints that must be satisfied during multimedia transmission:

- Time constraints, which include delays, computing time, and signaling delay
- Space constraints such as system buffers
- Frequency constraints, which are network bandwidth and system bandwidth for data transmission

The best usage (or, in other words, utilization) of resources in a network environment is only possible by first characterizing the traffic, then determining the parameters such as buffer size and bandwidth to maximize the performance.

0163-6804/99/$10.00 © 1999 IEEE

WHAT IS SELF-SIMILARITY?

A self-similar phenomenon displays structural similarities across a wide range of timescales. Traffic that is bursty on many or all timescales can be described statistically using the notion of self-similarity. Self-similarity is the property associated with "fractals," which are objects whose appearances are unchanged regardless of the scale at which they are viewed [8]. In the case of stochastic objects like time series, self-similarity is used in the distributed sense: when viewed at varying timescales, the object's relational structure remains unchanged. As a result, such a time series exhibits bursts at a wide range of timescales.

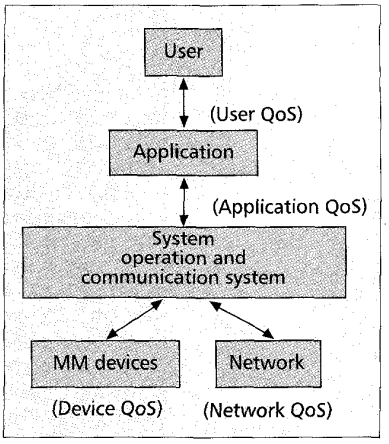


Figure 1. *A QoS layered model for the MCS [6].*

SELF-SIMILARITY IN NETWORK TRAFFIC

In 1993, a seminal event in the field of network performance modeling occurred with the publication of the paper titled "On the Self-Similar Nature of Ethernet Traffic [9]. " Ethernet is a broadcast multi-access system for local area networking with distributed control. The authors reported the results of a massive study of Ethernet traffic and demonstrated that it had a self-similar (i.e., fractal) characteristic. This meant the traffic had similar statistical properties at a range of timescales: milliseconds, seconds, minutes, hours, even days and weeks. Another consequence is that the merging of traffic streams, as in a statistical multiplexer or an asynchronous transfer mode (ATM) switch, does not result in smoothing of traffic. Again, bursty data streams that are multiplexed tend to produce a bursty aggregate stream. This first paper sparked a surge of research around the globe. The results show the self-similarity in ATM traffic, compressed digital video streams, and Web traffic between browsers and servers. Although a number of researchers had observed over the years that network traffic didn't always obey Poisson assumptions used in queuing analysis, this paper for the first time provided an explanation and a systematic approach to modeling realistic data traffic patterns. Following the announcement of the fractal nature of data traffic, network theorists split into two camps; one advocated that the entire network theory has to be rewritten, and the other disagreed.

Traditionally, networks have been described by generalized Markovian processes that are statistical models and rely on postulates framed by the Russian mathematician A. A. Markov. Markovian models of networks have limited memory of the past. They reflect short-range dependence. In a Markovian model, smoothing of bursty data is possible. Averaging of bursty traffic over a long period of time gives rise to a smooth data stream. A network based on fractal nature will have very different parameters and congestion control techniques. Let us view in detail what kind of statistical properties self-similar patterns present.

PROPERTIES OF SELF-SIMILARITY

X is defined to be a wide sense stationary random process with mean μ, variance σ, and autocorrelation function ρ. In particular, $\rho(\tau)$ is of the form $\rho(\tau) \rightarrow \tau^{\beta}$, as $\tau \rightarrow \infty$ where $L(\tau)$ is slowly varying at infinity [4], that is, $\mathrm{Lim}_{\tau \rightarrow \infty} L(\tau x)/L(\tau) = 1$ for all $x > 0$. Let $X^{(m)}$ denote the new process obtained by averaging the original series X in nonoverlapping subblocks of size m. That is, $X^{(m)}(t) = (1/m)(X_{tm-m+1} + X_{tm-m+2} + \ldots + X_{tm})$.

For each m, $X^{(m)}$ defines a wide-sense stationary random process. Process X is said to be second-order self-similar with self-similarity parameter H if the aggregated processes have the same autocorrelation structure as X [4]. That is, $H = 1 - \beta/2$ and $\rho^{(m)}(\tau) = \rho(\tau)$ for all $m = 1, 2, \ldots$.

In other words, X is exactly second-order self-similar if the aggregated processes are indistinguishable from X with respect to their first- and second-order properties.

The most striking feature of self-similarity is that the correlation structures of the aggregated process do not degenerate as $m \rightarrow \infty$. This is in contrast to traditional models, all of which have the property that the correlation structure of their aggregated processes degenerates as $m \rightarrow \infty$; that is, $\rho^{(m)}(\tau) \rightarrow 0$, as $m \rightarrow \infty$ for $\tau = 1, 2, 3, \ldots$.

LONG-RANGE DEPENDENCE AND HEAVY-TAILED DISTRIBUTIONS

Long-range dependent processes are characterized by an auto-correlation function which decays hyperbolically. This implies that the auto-correlation function is nonsummable, unlike more conventional short-range dependent processes, which have auto-correlation functions that decay exponentially [4].

Also, a distribution is heavy-tailed if $P[X > x] \sim x^{-a}, x \rightarrow \infty$, where $0 < a < 2$ [3]. That is, the asymptotic shape of the distribution follows a power law. A random variable that follows a heavy-tailed distribution can take on extremely large values with nonnegligible probability. Heavy-tailed distributions can be used to characterize probability densities that describe traffic processes such as packet interarrival times and burst length [10].

THE HURST PARAMETER: THE MEASURE OF SELF-SIMILARITY

The Hurst parameter H is a measure of the level of self-similarity of a time series. H takes values from 0.5 to 1. In order to determine if a given series exhibits self-similarity, a method is needed to estimate H for a given series. Currently, there are three approaches to doing that:

- Analysis of the variances of the aggregated processes $X^{(m)}$
- Analysis of the rescaled range (R/S) statistic for different block sizes
- A Whittle estimator

The first method, the variance time plot, relies on the slowly decaying variance of a self-similar series. The variance of $X^{(m)}$ is plotted against m on a log-log plot. Then a straight line with a slope ($-\beta$) greater than -1 is indicative of self-similarity. and the parameter H is given as above [8].

The second method, the R/S plot, uses the fact that for self-similar data, the rescaled range or R/S statistic grows according to a power law with exponent H as a function of the number of points included, n. Thus, the plot of R/S against n on a log-log plot has a slope which is an estimate of H.

While the preceding two graphical methods are useful to estimate H, they may be biased for large H. The third method, a Whittle estimator, does provide a confidence interval. This technique uses the property that any long-range dependent process approaches fractional Gaussian noise(FGN) when aggregated to a certain level, and so should be coupled with a test of the marginal distribution of the aggregated observations to ensure that it has converged to the normal distribution [8]. As m increases, short-range dependences are averaged out of the data set. If the value of H remains relatively constant, it is almost certain that this H value measures a true level of self-similarity of the data set.

The Hurst Effect — Self-similar processes provide an elegant explanation of an empirical law known as Hurst's law or the Hurst effect. For a given set of observations $X_1, X_2, \ldots, X_n$ with sample mean $\overline{X}(n)$ and sample variance $S^2(n)$, the rescaled adjusted range or R/S statistic is given by

$$R(n)/S(n) = (1/S(n)) \,(\max(0, W_1, W_2, \ldots W_n) - \min(0, W_1, W_2, \ldots W_n))$$

with $W_k = X_1 + X_2 + \ldots X_k - k\overline{X}, k = 1,2, \ldots, n$.
Hurst found that many naturally occurring time series are well represented by the relation $E[R(n)/S(n)] \sim cn^H$ as $n \to \infty$ with Hurst parameter H normally around 0.73, and c a finite positive constant independent of n. However, if the observations come from a short-range dependent process, it has been shown that $E[R(n)/S(n)] \sim dn^{0.5}$ as $n \to \infty$ with d a finite positive constant independent of n [4]. This discrepancy is referred to as the Hurst effect or Hurst phenomenon.

Slowly Decaying Variances

The most important feature of self-similar processes is that the variance of the arithmetic mean, μ, decreases more slowly than the reciprocal of the sample size n. That is equal to saying $\text{var}(X^{(m)}) \sim am^{-\beta}$ as $m \to \infty$, where a is a finite constant independent of m. On the other hand, for short-range dependent processes $\text{var}(X(m)) \sim bm^{-1}$, $m \to \infty$, where b is a finite positive constant independent of m.

What Causes Self-Similarity?

Since self-similarity is believed to have a significant impact on network performance, understanding the causes of self-similarity in traffic is important.

Research done by M. E. Crovella *et al.* [8] has revealed that the traffic generated by World Wide Web transfers shows self-similar characteristics. Comparing the distributions of ON and OFF times, they found that the ON time distribution was heavier-tailed than the OFF time distribution. The distribution of file sizes in the Web might be the primary determiner of Web traffic self-similarity. In fact, the work presented by K. Park *et al.* [1] has shown that the transfer of files whose sizes are drawn from a heavy-tailed distribution is sufficient to generate self-similarity in network traffic. The ON and OFF periods do not need to have the same distribution. These results suggest that the self-similarity of Web traffic is not a machine-induced artifact; in particular, changes in protocol processing and document display are not likely to remove the self-similarity of Web traffic [8].

In a realistic client/server network environment, the degree to which file sizes are heavy-tailed can directly determine the degree of traffic self-similarity at the link level [1, 3]. This causal relation is proven to be robust with respect to changes in network resources (bottleneck bandwidth and buffer capacity), network topology, the influence of cross-traffic, and the distribution of interarrival times. Specifically, measuring self-similarity via the Hurst parameter H and the file size distribution by its power law exponent α, it has been shown that there is a linear relationship between H and α over a wide range of network conditions.

Network Performance

Well-defined metrics of delay, packet loss, flow capacity, and availability are fundamental to measurement and comparison of path and network performance. In general, users are most interested in metrics that provide an indication of the likelihood that their packets will get to the destination in a timely manner. Therefore, estimates of past and expected performance for traffic across specific Internet paths, not simply measures of current performance, are important. Users are also increasingly concerned about path availability information, particularly as it affects the quality of multimedia applications requiring higher bandwidth and lower latency, such as Internet phone and videoconferencing. Availability of such data could help in scheduling online events such as Internet-based distance education seminars, and also influence user willingness to purchase higher service quality and associated service guarantees.

Given the ubiquity of scale-invariant burstiness observed across diverse networking contexts, finding effective traffic control algorithms capable of detecting and managing self-similar traffic has become an important problem.

The control of self-similar traffic involves modulating the traffic flow in such a way that the resulting performance is optimized. Scale-invariant burstiness (i.e., self-similarity) introduces new complexities into optimization of network performance and makes the task of providing QoS together with achieving high utilization difficult.

The Effects of Self-Similarity on Network Performance

Many analytical studies have shown that self-similar network traffic can have a detrimental impact on network performance, including amplified queuing delay and packet loss rate [1, 2, 4]. On the other hand, Heyman *et al.* [5] found that long-range dependence was unimportant for buffer occupancy when there was strong short-range dependence and the Hurst parameter was not very large ($H < 0.7$). However, they did not touch the case where there was strong long-range dependence with a larger Hurst parameter.

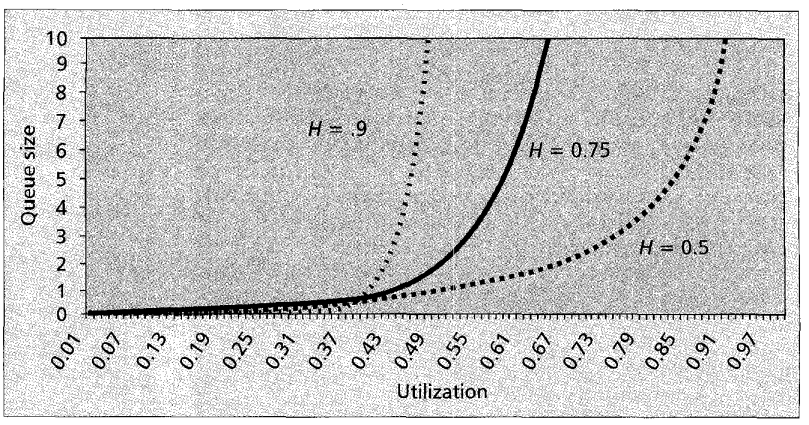
One practical effect of self-similarity is that the buffers needed at switches and multiplexers must be bigger than those predicted by traditional queuing analysis and simulations. These larger buffers create greater delays in individual streams than were originally anticipated [3, 4]. The delay-bandwidth product problem arising out of high-bandwidth networks and QoS issues stemming from support of real-time multimedia communication have added further complexities to the problem of optimizing performance.

How much self-similarity affects network performance is modulated by the protocols acting at the transport/network layer. An exponential trade-off relationship was observed between queuing delay and packet loss rate [2].

It is certain that a linear increase in buffer sizes will produce nearly exponential decreases in packet loss, and that an increase in buffer size will result in a proportional increase in the effective use of transmission capacity. With self-similar traffic, these assumptions do not hold. The decrease in packet loss with buffer size is far less than expected, and as can be seen from Fig. 2, the buffer requirements begin to explode at lower levels of utilization for higher degrees of long-range dependence (higher values of H).

Heyman *et al.* showed that for sources with large Hurst parameters, Markov chain models estimated the buffer occupancy well when the buffer sizes were not too large (no larger than 10 ms for a single source) [5], but these models might not estimate the cell loss rate and mean buffer size accurately for larger buffers. Also, another study has shown that queuing delay exhibited a superlinear dependence on self-similarity when buffer capacity was large [2]. The queue length distribution decayed more slowly for long-range dependent sources than short-range dependent sources.

Moreover, scale-invariant burstiness implies the existence of concentrated periods of high activity at a wide range of timescales, which adversely affects congestion control and is an important correlation structure which may be exploitable for congestion control purposes [4]. Network performance as captured by throughput, packet loss rate, and packet retransmission rate degrades gradually with increasing heavy-tailedness. The degree to which heavy-tailedness affects self-similarity is determined by how well congestion control is able to shape its source traffic into an on-average constant output stream while conserving flow [2].



■ **Figure 2.** *Queue size–utilization trade-off as self-similarity changes defined by* H, *the Hurst parameter [10].*

A dynamic congestion control strategy is difficult to implement. Such a strategy is based on measurement of recent traffic and can fail utterly to adapt to rapidly changing conditions. Also, congestion prevention by appropriate sizing of switches and multiplexers is difficult because data network traffic does not exhibit a predictable level of busy traffic periods; patterns can change over a period of days, weeks, or months, and congestion can occur unexpectedly with dramatic intensity. On the other hand, predictive congestion control was studied for improving network performance by Tuan *et al.* [3]. In their algorithm, information about the future is utilized to make traffic control decisions. They called this *Selective Aggressiveness Control* (SAC), and it is aimed to be robust, efficient, and portable so that it can easily be incorporated into existing congestion control schemes.

SAC tries to aggressively soak up bandwidth if it predicts the future network state to be "idle," adjusting the level of aggressiveness as a function of the predicted idleness and its confidence. They showed that the performance gain due to SAC is higher the more self-similar the network traffic is [3]. Although in real life the perfect prediction of future traffic congestion is not possible, SAC achieves the highest throughput with perfect future information among other congestion control algorithms such as generic feedback congestion control.

As the traffic self-similarity (described by the α parameter of Pareto file size distribution, $\alpha = 3 - 2H$) and network resources (buffer capacity, bottleneck bandwidth) vary, a gradual change in the packet loss rate is observed: as α approaches 1, along with a decrease in buffer capacity, packet loss rate increases. This relation is shown in Fig. 3 for different link buffer (LB) sizes in the range of 2–46 kbytes.

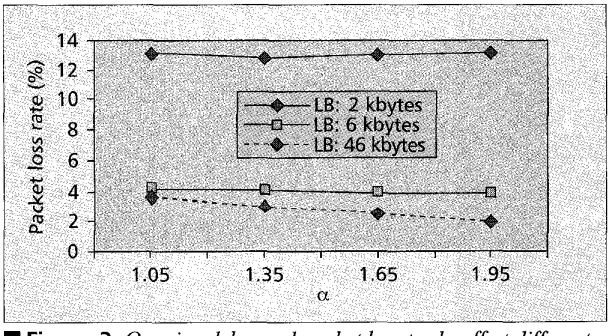
Packet loss and retransmission rate decline smoothly as self-similarity is increased under reliable flow-controlled packet transport [1]. The only performance indicator exhibiting a more sensitive dependence on self-similarity is mean queue length, and this concurs with the observation that queue length distribution under self-similar traffic decays more slowly than with Poisson sources. Increasing network resources such as link bandwidth and buffer space results in a superlinear improvement in performance. However, large buffer sizes are accompanied by long queuing delays. In the context of facilitating multimedia traffic such as video and voice in a best-effort manner while satisfying their diverse QoS requirements, low packet loss, on average, can only be achieved at a significant increase in queuing delay and vice versa.

Increasing link bandwidth, given a large buffer capacity, has the effect of decreasing queuing delay much more drastically under highly self-similar traffic conditions than when traffic is less self-similar (Fig. 4). Therefore, high-bandwidth communication links (for multimedia network applications) should be employed to alleviate the exponential trade-off between queuing delay and packet loss (throughput) for supporting QoS-sensitive traffic.

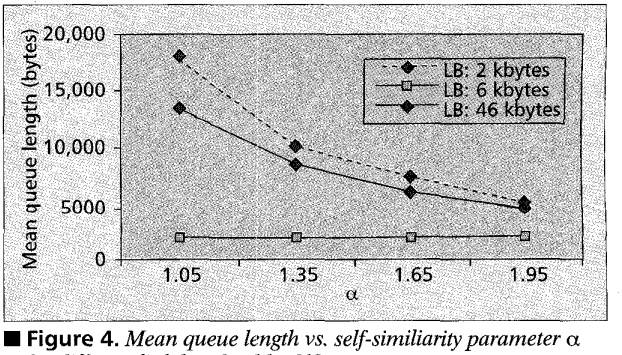
One important discovery is that the higher the load on the Ethernet, the higher the degree of self-similarity [10]. When the network load is in the range of 30–70 percent, waveforms of the traffic display self-similarity for which H was approximately 1. Furthermore, a load between 80 and 99 percent produces waves with a strong periodic component, and calculation of H becomes unreliable.

CURRENT RESEARCH

Studies performed in the last couple of years have presented convincing evidence that multimedia network traffic exhibits self-similar traits. Therefore, the ongoing research



■ **Figure 3.** *Queuing delay and packet loss trade-off at different ranges of self-similar traffic [1].*



■ **Figure 4.** *Mean queue length vs. self-similiarity parameter* α *for different link bandwidths [1].*

on characterization of self-similar processes and their application to teletraffic modeling is increasingly important. Also, impacts of second-order self-similar processes on ATM networking is a subject matter of recent research work. The results will have a significant impact on the correct dimensioning of ATM networks, in particular ATM multiplexers and switches.

Due to the inherent bursty nature of multimedia traffic, packet loss and network delay are common problems experienced by multimedia applications such as video on demand. Hence, optimal allocation of buffers in a network in order to smooth the bursty traffic caused by multimedia data is also subject to further study.

Our ongoing research aims to detect self-similarity in real time, and come up with a measure of self-similarity such that this measure can be input for the optimization of resource allocation algorithms. Our aim is to demonstrate the limitations or validity of conventional resource allocation methods in the presence of self-similar traffic. In general, self-similar traffic, as established in this article, exhibits a higher level of complexity than conventional traffic models, which has caused existing network engineering tools and methods to be inadequate for such traffic. The new approach will simplify self-similarity by reducing its modeling to a single measure, and generate new network engineering tools and methods that will adaptively operate on this measure to provide optimal performance and capacity.

References

[1] K. Park, G. Kim, and M. Crovella, "On the Relation Between File Sizes, Transport Protocols, and Self-Similar Network Traffic," *Proc. IEEE Int'l. Conf. Network Protocols*, Oct. 1996, pp. 171–80.

[2] K. Park, G. Kim, and M. Crovella, "On the Effect of Traffic Self-Similarity on Network Performance," *Proc. SPIE Int'l. Conf. Perf. and Control of Network Sys.*, 1997, pp. 296–310.

[3] T. Tuan and K. Park, "Congestion Control for Self-Similar Network Traffic," Dept. of Comp. Sci., Purdue Univ., CSD-TR 98-014, May 1998, to be published.

[4] P. R. Morin, "The Impact of Self-Similarity on Network Performance Analysis," Ph.D. dissertation, Carleton Univ., Dec. 1995.

[5] D. P. Heyman and T. V. Lakshman, "What are the Implications of Long-Range Dependence for VBR-Video Traffic Engineering?" *IEEE Trans. Networking*, vol. 4, no. 3, June 1996, pp. 301–17.

[6] R. Steinmetz and K. Nahrstedt, *Multimedia: Computing Communications & Applications*, Prentice Hall., 1995, pp. 420–45.

[7] D. Ferrari and D. C. Verma, "A Scheme for Real Time Channel Establishment in Wide-Area Networks," *IEEE JSAC*, vol. 8, no. 3, Apr. 1990, pp. 368–79.

[8] M. E. Crovella, "Self-Similarity in WWW Traffic: Evidence and Possible Causes," *IEEE Trans. Networking*, vol. 5, no. 6, Dec. 1997, pp. 835–45.

[9] W. E Leland *et al.*, "On The Self-Similar Nature of Ethernet Traffic," *IEEE Trans. Networking*, vol. 2, no. 1, Feb. 1994, pp. 1–15.

[10] W. Stallings, *High Speed Networks; TCP/IP ATM Design Principles*, Prentice Hall, 1998, pp. 181–207.

Biographies

ZAFER SAHINOGLU (zxsl602@red.njit.edu) graduated with a B.S. degree in electrical and electronics engineering from Gazi University, Ankara, Turkey in 1994. He focused on bioelectronics signal processing applications and received his M.S. degree at the New Jersey Institute of Technology, Newark, in 1997. Afterwards, he joined the New Jersey Center for Multimedia Research at NJIT to pursue his Ph.D. in telecommunications and networking engineering. His current research interests include analysis of network traffic, self-similarity, and adaptive resource allocation schemes.

SIRIN TEKINAY [M] (stekinay@megahertz.njit.edu) holds a Ph.D. (1994) degree with concentration in telecommunications from the School of Information Technology and Engineering, George Mason University. She served as a visiting scientist at CONTEL from 1991 until 1993. In 1994 she joined NORTEL as a senior member of scientific staff where she led several projects, including the capacity and performance evaluation of GSM systems, wireless network planning for CDMA PCS systems, and external research projects with universities. In 1996 she joined Bell Laboratories, Lucent Technologies, where she was appointed technical prime on wireless radiolocation. During this appointment, she has served on CDMA Development Group task forces, TIA 45.5 standards groups, and contributed to the CTIA. In September 1997, she joined the Department of Electrical and Computer Engineering at the New Jersey Institute of Technology and New Jersey Center for Multimedia Research. She is director of the recently founded New Jersey Center for Wireless Telecommunications. Her research interests include teletraffic modeling and management, resource allocation, mobility management for wireless and wireline networks, computer communications networks, wireless geolocation systems, propagation environment characterization, and wireless and wireline multimedia networking.