Journal of Computing and Information Technology - CIT 16, 2008, 1, 15–21
doi:10.2498/cit.1000878

Analysis of Ethernet Traffic Statistical Properties

Michal Kociský, Jozef Lasz and Ivan Kotuliak

Slovak University of Technology, Bratislava, Slovakia

Traffic profile, mainly its self-similarity properties, can have crucial impact on the network performance. In this regard, we evaluate traffic profile of Ethernet traffic. We have performed a measurement of the traffic on Ethernet network. Captured data has been analyzed from the protocol point of view, with the stress on the self-similarity, LRD and SRD properties. To evaluate these characteristics, properties of wavelet transform (DWT) are deployed and, based on α parameter, scaling property of traffic is estimated. We show that self-similarity is present in analyzed data and that it depends on analyzed time scale and on analyzed protocol.

Keywords: traffic profile, self-similarity, wavelet transform, hurst parameter, SRD, LRD

1. Introduction

Traffic properties are in the center of interest for few decades. They can significantly influence performances of the switches [2,3] and can have disastrous impact mainly on loss and delay. The phenomenon of self-similarity is well known and studied in different areas of research, including network performance.

Statistical properties of traffic are changing with new applications and new Internet deployment. In this article, we present detailed study of statistical properties of Ethernet traffic. Analyzed Ethernet traffic was captured on university network and under two aggregation levels. This study is based on two main branches: the first one focuses on a classical statistical properties like variance and protocol volume rates. The second part is devoted to self-similarity evaluation deploying properties of wavelet transform.

Our previous works [7,9] focused on the protocol representation in the measured data. In this article, we extend this view to the scaling properties of the traffic. In [2], similar study was presented using R/S approach. We prefer Abry-Veitch estimator [4,5,6] based on properties of wavelet transform. Comparing to these works, we extend the study of scaling properties to upper layers too and we also compare scaling coefficients under different levels of aggregation.

The rest of the text is organized as follows: Section 2 deals with self-similarity (SS), Long-Range dependence (LRD) and Short Range Dependence (SRD). Definitions and theoretical bases are also given here. Section 3 describes network model used for Ethernet traffic measurement. In Section 4, we analyze captured data and provide results for parameters α and H for whole traffic as well as for different protocols. Section 5 provides concluding remarks.

2. Self-similarity and Long-range Dependence

It is now widely accepted that network traffic exhibits LRD [1]. Even though practical implications of this discovery are not completely understood yet, it is known that traditional Poisson models of network traces cannot capture the behavior of LRD traffic. Often, LRD data is also SS [2,3], indicating that it possesses similar statistical properties on multiple time scales.

Self-similarity and long-range dependence are closely related phenomena. Self-similarity is the property that, as you scale a process (zoom in on the details), you see the same structure repeated, and the statistical properties (for instance mean, variance, or marginal distribution)

are the same (under a transformation). LRD relates to correlations in the data which, though it is decreasing over wider ranges, never becomes insignificant.

Consider a discrete time stochastic process or time series $X_t \in Z$ where X_t is interpreted as the traffic volume measured in packets or bytes at time instance t. We denote cumulative process (total traffic volume from time 0 up to time t) by Y_t and we reserve X_t to be increment process corresponding to Y_t , that is,

$$X_t = Y_t - Y_{t-1}.$$

Following is a definition of self-similarity for continuous time process in the sense of finite dimensional distributions. Consider the cumulative process Y_t, albeit in continuous time $t \in R$. Y_t is *self-similar* with *self-similarity parameter* H (*Hurst parameter*), $0 < H < 1$, if for all $a > 0$ and $t \geq 0$

$$Y_t =_d a^{-H} Y_{at} \tag{1}$$

Thus Y_t and its time-scaled version Y_{at} after normalizing by a^{-H} must follow the same distribution.

Stationary process X_t is said to be *long-range dependent* if its autocorrelation function $\rho(k)$ has the asymptotic form:

$$\rho(k) \sim c_p k^{-\beta} \; as \; k \to \infty \tag{2}$$

where $0 < \beta < 1$ and real number $c_p > 0$. Afterwards,

$$H = 1 - \beta/2 \tag{3}$$

Autocorrelation function of the process which exhibits LRD sums to infinity $\Sigma_k \rho(k) = \infty$.

This non-summability of the correlation captures the intuition behind LRD, namely that while high-lag correlations are all individually small, their cumulative effect is important and gives rise to features which are drastically different from those of the more conventional, i.e., SRD.

The latter are characterized by an exponential decay of the autocorrelations, i.e.,

$$\rho(k) \sim \delta^k, \; as \; k \to \infty \tag{4}$$

$0 < \delta < 1$, resulting in a summable autocorrelation function $0 < \Sigma_k \rho(k) < \infty$.

When working in the frequency domain, LRD manifests itself in a spectral density that obeys a power-law behavior near the origin. Thus, from the spectral analysis point of view, LRD implies that $f(0) = \Sigma_k \rho(k) = \infty$, that is, it requires a spectral density which tends to $+\infty$ as the frequency λ approaches 0 ("$1/f$-noise"). On the other hand, SRD is characterized by a spectral density function $f(\lambda)$ which is positive and finite for $\lambda = 0$.

LRD is also commonly defined as the power-law divergence at the origin of its spectrum:

$$f_x(\nu) \sim c_f |\nu|^{-\alpha}, |\nu| \to 0. \tag{5}$$

for some positive constant c_f and some real $\alpha \in (0, 1)$.

The power-law decay is such that the sum of all correlations (out from any lag) is always appreciable, even if, individually, the correlations are small. The past, therefore, exerts a long-range influence on the future.

The main parameter of LRD is the dimensionless scaling exponent α. It describes the qualitative nature of scaling – how behavior on different scales is related. The value of α helps us determine which kind of scaling is present:

$\alpha = 0$	suggest short-range dependent	
$\alpha = (0, 1)$	long-range dependent	(6)
$\alpha > 1$	self-similar	

In the case of LRD, α is a directly relevant parameter, however this is sometimes given as Hurst parameter H. However, H is not the parameter of strict self-similarity, it is merely a convention to rewrite alpha in this way for LRD processes and it is defined (for LRD traffic) as:

$$H = (\alpha + 1)/2. \tag{7}$$

In addition, the Hurst parameter h for SS is defined as:

$$h = (\alpha - 1)/2. \tag{8}$$

Related parameters c_p and c_f are quantitative parameters giving a measure of magnitude of LRD-induced effects. The parameters may be estimated jointly using the Abry-Veitch wavelet-based estimator [4,5,6], or, separately, by a number of other techniques [3].

Heavy-tailedness [12] of certain network related variables e.g., file sizes and connection durations can be shown to underlie the root cause

of long-range dependence and self-similarity in network traffic.

A random variable Z has a *heavy-tailed distribution* if

$$Pr\{Z > x\} cx^{-\varphi}, \qquad x \to \infty \qquad (9)$$

where $0 < \varphi < 2$ is called the *tail index* or *shape parameter* and c is a positive constant.

That is, the tail of the distribution, asymptotically, decays hyperbolically. This is in contrast with *light-tailed distributions* e.g., exponential and Gaussian which possess an exponentially decreasing tail. A distinguishing mark of heavy-tailed distributions is that they have infinite variance for $0 < \varphi < 2$, and if $0 < \varphi \leq 1$, they have also unbounded mean. In the networking context, we will be primarily interested in the case $1 < \varphi < 2$. A frequently used heavy-tailed distribution is the *Pareto distribution* whose distribution function is given by

$$Pr\{Z \leq x\} = 1 - \left(\frac{b}{x}\right)^{\varphi}, \qquad b \leq x, \quad (10)$$

where $0 < \varphi < 2$ is the shape parameter and b is called the *location parameter*. The mean is given by $\varphi k/(\varphi - 1)$. We say that there are distributions e.g., Weibull and log-normal that have *subexponentially* decreasing tails, but possess finite variance.

3. Considered Model of Measurement

Traffic measurement has been performed on university Ethernet network using transparent bridge. Bridging was implemented using Linux kernel 2.4 and using bridging extensions in kernel. Two 100 Mbps network adapters, which served as bridge ports, were used during measurements. This transparent bridge was integrated into existing network to be able to monitor all passing traffic. Traffic was measured and later analyzed by implemented network analyzer based on pcap [8] library. To compare different traffic aggregations and their impact on the traffic properties, we have performed two measurements. The first one on 100 Mbps department to university backbone connection (aggregation of about 100 computers) and the second one on a small 100 Mbps LAN segment (consisting of 10 computers).

4. Captured Data Analysis

This Section provides the analysis of measured traffic and its graphical presentation. Figure 1 shows the output of backbone data capturing including basic characteristics. This output is generated directly by our analyzer and displayed statistics are measurement length, packet volumes, protocol overhead (headers volume to frames volume ratio) and protocol representation. The protocol representation shows the distribution of measured frames over various layers in their amount (absolute value and percentage) and volume (in bytes).

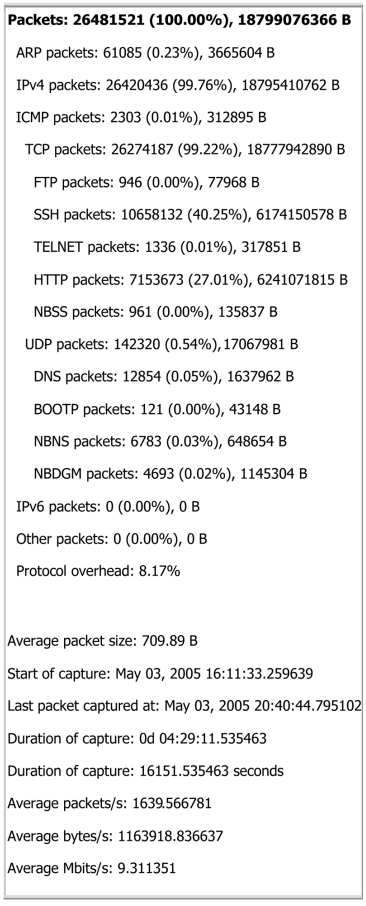

Figure 1. General statistics of captured frames.

We have evaluated traffic in data link, network, transport and application layer of the TCP/IP model. Nevertheless, presented are only graphs for application layer packets for better illustration (the ARP, ICMP and UDP traffic is minor in comparison with IP and TCP, respectively).

In Figure 2, the comparison of applications over TCP in the backbone is given (volume of transferred packets in Bytes). The values are sums of captured packets (and their lengths) during 30 s periods. We have observed that majority of data was transferred using HTTP and SSH protocols. FTP, TELNET and NetBIOS protocols remain only poorly represented in provided measures. The reason of this fact is that nearly all communication, including data transfer, is made using secured connection (SSH) and by WWW data (HTTP). Representation in packet volume is similar and its detailed description can be found in [7].

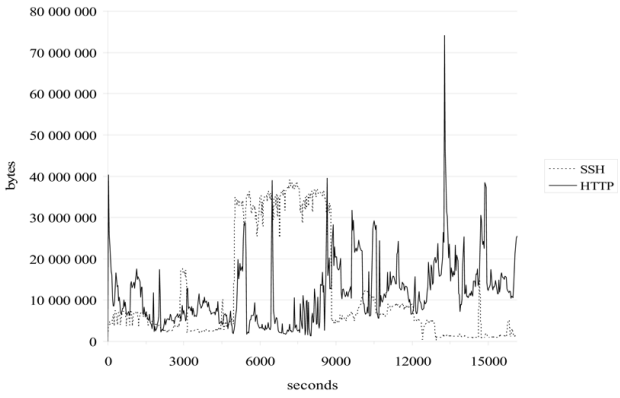

Figure 2. Volume of captured application layer packets over TCP in time scale (statistics record period 30 s).

In Figure 3, the time scales of applications over UDP in the backbone are given (volume of transferred packets). In comparison with the previous graph for TCP applications (Figure 2), we can see that TCP is more dominant and more important protocol. UDP is used mainly for RT traffic transporting and for service packets like DNS or BOOTP (DHCP).

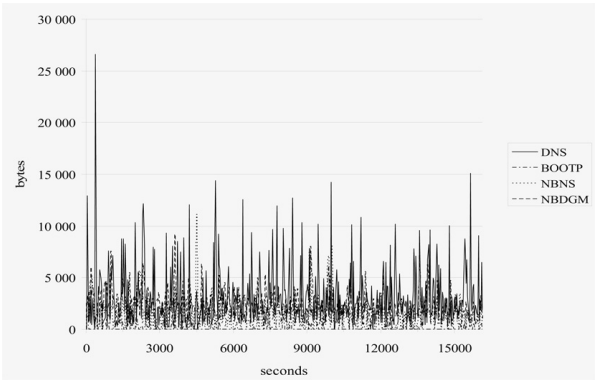

Figure 3. Volume of captured application layer packets over UDP in time scale (statistics record period 30 s).

In Figure 4 and Figure 5, the distribution of frame length is given in volume (Figure 4) and in packet amount (Figure 5). The most present are 60 B long frames (35%) and frames longer than 1400 B (39.5%). It is obvious that frames longer than 1400 B transport majority of data (82.5%). Other lengths are not so important. The important lengths are given by defined minimum Ethernet II frame length (60 B) and maximum frame length (1514 B). Service and control packets (ARP, ICMP, routing info, TCP ACK) usually use the minimum length. The maximum length frames serve for data transfer.

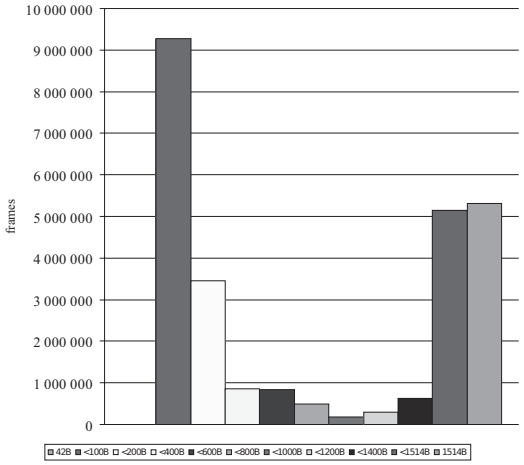

Figure 4. Frame size distribution (amount).

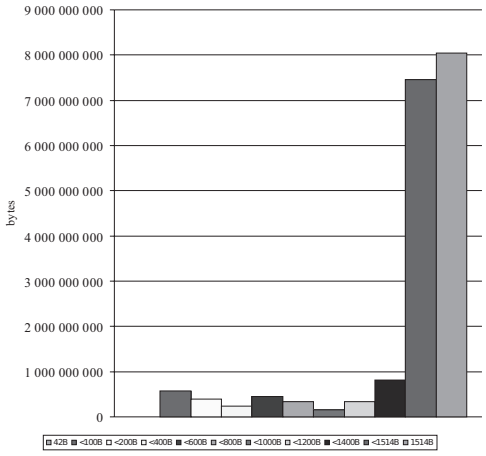

Figure 5. Frame size distribution (volume).

In Figure 6, the comparison of applications over TCP in the LAN segment is given (volume of transferred packets). The main difference between local and backbone traffic is presence of

NetBIOS protocol (in this case NetBIOS Session Service). When we look at the local segment results, HTTP [10] and SSH [11] protocols are in volume preceded by NetBIOS packets (48.2% of entire local traffic are NBSS packets). On the other hand, nearly none of NetBIOS packets is routed to the backbone of the network.

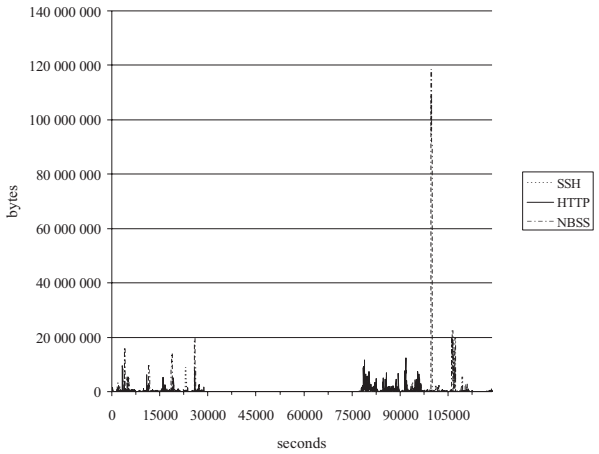

Figure 6. Volume of captured application layer packets over TCP in time scale (statistics record period 30 s) – LAN segment.

NetBIOS packets in Windows networks serve for transporting, logging information, shared disk file transfer etc. Note that volume of the traffic differs significantly comparing backbone connection (Figure 2) and LAN measurement (Figure 6).

4.1. Self-similarity Evaluation

In this subsection, we proceed the analysis of measured Ethernet traces with evaluation of their scaling properties (see Section 2). For that purpose, we use Abry-Veitch wavelet-based estimator [4,6], which generates the Logscale Diagram (LD). LD allows us the estimation of key parameter, the scaling exponent α (Eq. 2,3). We estimate α on time series, which we get from traces as a sum of packets sizes on 2 and 20 seconds intervals. We apply such analysis on the whole traffic and on TCP and UDP protocols separately.

In Figure 7 and Figure 8 there are LD's for the whole traffic, where the volume of transferred bytes are divided into 2 seconds (Figure 7) and 20 seconds intervals (Figure 8). In LD, the x-axis is represented by octaves j, where j corresponds to dilatation operator of discrete wavelet transform. Due to the dyadic nature of wavelet basis, there are half as many points (number of wavelet coefficients) at scale j+1 as at the finer scale j. The solid line in LD's represents estimation of variance of wavelet details, which is on each octave completed with confidence intervals (vertical bars). The slope of the straight dashed line over selected range of scales, using weighted regression, yields an accurate estimate for the LRD exponent α. Number of vanishing moments N is needed to ensure that the wavelet

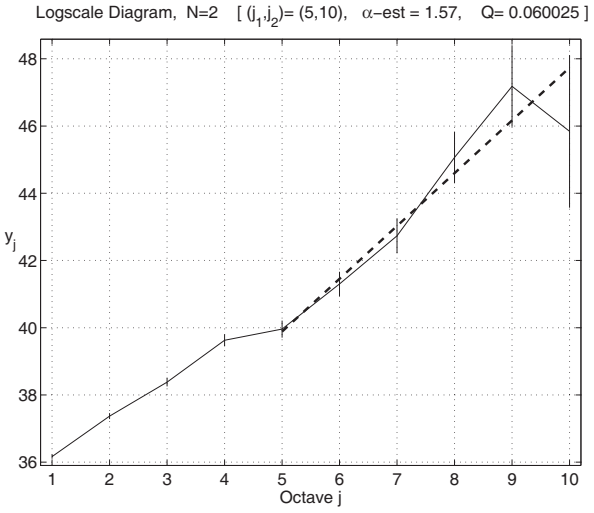

Figure 7. LD for the whole traffic bytes measured on backbone of LAN network divided into 2 s.

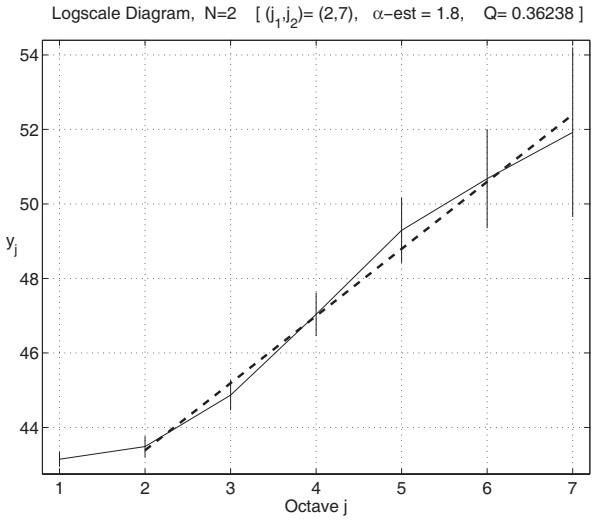

Figure 8. LD for the whole traffic bytes measured on backbone of LAN network divided into 20 s intervals.

details are well defined and Q is a goodness of fit statistic, which helps with the choice of scaling range.

In Table 1 and Table 2, we present estimated parameters α, H (for LRD traffic) and h (for SS traffic). Values are given for different intervals (2 s and 20 s) and for whole traffic, TCP traffic and UDP traffic. Table 1 deals with LAN segment, which represents low aggregation traffic (about 10 of computers) and Table 2 with backbone representing high aggregation (100 of computers).

	α	H for LRD	h for SS
TCP bytes 2s intervals	2.2	–	0.6
TCP bytes 40ms intervals	2.24	–	0.62
UDP bytes 2s intervals	0.246	0.623	–
UDP bytes 40ms intervals	0.097	0.549	–
whole traffic 2s intervals	2.15	–	0.575
whole traffic 40ms intervals	2.13	–	0.565

Table 1. Estimations of coefficients α, H and h on data measured on segment of LAN network.

	α	H for LRD	h for SS
TCP bytes 2s intervals	1.57	–	0.285
TCP bytes 20ms intervals	1.8	–	0.4
UDP bytes 2s intervals	0.75	0.875	–
UDP bytes 20ms intervals	0.73	0.865	–
whole traffic 2s intervals	1.57	–	0.285
whole traffic 20ms intervals	1.8	–	0.4

Table 2. Estimations of coefficients α, H and h on data measured on backbone network.

Values of parameter α for protocol TCP and for the whole traffic are very close, while TCP traffic represents 99% of the whole traffic. Evaluated data are of SS nature (h=0.6 for segment and 0.29 for backbone). In the case of protocol UDP, the value of parameter α for both time units is lower than 1 and, therefore, it identifies LRD character of data. The parameter H yields to 0.62 for LAN segment and 0.88 for backbone.

Comparing results of local segment and of backbone, we can see the similar presence of scaling process in both traces. While the whole traffic parameters and TCP traffic were Self Similar, the UDP traffic was LRD. Another important point of this evaluation is that the SS phenomenon seems to be more important in local segment than in backbone. This property should be confirmed by another, more extensive study.

5. Conclusions

One of the most important phenomena discussed in the last years in networking research is the self-similarity (SS). SS phenomenon is present in all traffic with different power. This statement served as a frame of the research provided in this paper.

Considering SS, long-range dependence (LRD) and short-range dependence (SRD), we have measured Ethernet traffic on backbone connection and in LAN and have evaluated its statistical properties. Firstly, we evaluated traffic in terms of protocols representation in volume and in number of packets, taking into account statistical properties of measured traffic. This study focused on data link, network, transport and application layer of the TCP/IP model, nevertheless presented are only graphs for application layer. Secondly, we focused on characterization of the traffic, using wavelet transform to study its scalability (SS, LRD or SRD). This evaluation was done on different protocols and in different study environments (LAN and backbone). We have shown that scalability phenomenon depends on deployed protocol (TCP yields h=0.6, while the whole traffic h=0.575) and on the environment and aggregation level (where segment yields h=0.575 and backbone h=0.215 only), which is also one of the major contributions of this article.

It should be emphasized that this work subscribes to the wider research focusing on performance of the networks. In this scope, the next step of this work is to use wavelet transform for traffic with defined statistical properties generation (synthesis). Such approach can improve analytical and simulation-based evaluations of various network models.

6. Acknowledgment

Research described in the paper was financially supported by the Slovak Grant Agency under grant No. 1/1048/04 and 4/4084/07.

References

[1] D. R. COX, Long-range Dependence: A Review. In *Statistics: An Apprasial*, H. A. David and H. T. David (Eds.), The Iowa University Press, Ames, Iowa, (1984), 55–74.

[2] W. E. LELAND, M. S. TAQQU, W. WILLINGER, D. WILSON, On the Self-similar Nature of Ethernet. ACM SIGCOMM *Computer Communication Review*, **23**(4) (1993), ISSN: 0146-4833.

[3] P. R. MORIN, J. NEILSON, The Impact of Self-similarity on Network Performance Analysis. Thesis, Carleton University, Computer Science, 1995.

[4] P. ABRY, D. VEITCH, Wavelet Based Joint Estimator of the Parameters of Long-range Dependences. In *IEEE Transactions on Information Theory*, **3**(45), (1998).

[5] P. ABRY, P. FLANDRIN, M. S. TAQQU, D. VEITCH, Wavelets for the Analysis, Estimation, and Synthesis of Scaling Data, in *Self-similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds. New York: Wiley, 2000.

[6] P. ABRY, D. VEITCH, Wavelet analysis of long-range-dependent Traffic. *IEEE Trans. on Information Theory*, **44**(1) (1998), 2–15. `http://citeseer.nj.nec.com/abry98wavelet.html`

[7] R. BENKOVIČ, J. LASZ, I. KOTULIAK, Measurement of Traffic Profile in Ethernet Network. In *Proceedings of RTT 2004*, (2004), Cesky Raj (CR).

[8] *Pcap library*, 2005. `http://www.tcpdump.org`

[9] I. KOTULIAK, J. LASZ, Measurement of traffic profile in ethernet network. In *Proceedings of EC-SIP-M 2005*, (2005), Smolenice (SR).

[10] R. FIELDING ET AL. Hypertext Transfer Protocol – HTTP/1.1. *RFC 2616*, June (1999). `http://www.ietf.org/rfc/rfc2616.txt`

[11] D. J. BARRET, R. E. SILVERMAN, R. G. BYRNES, *SSH: The Secure Shell (The Definitive Guide)*. O'Reilly, 2005, ISBN 0-596-00895-3.

[12] K. PARK, W. WILLINGER, Self-similar network traffic: An Overview in *Self-similar Network Traffic and Performance Evaluation*. Wiley-Interscience, New York, 2000.

Received: June, 2006
Revised: April, 2007
Accepted: May, 2007

Contact addresses:
Michal Kociský
Department of Telecomunications
Slovak University of Technology
Ilkovicova 3
Bratislava, Slovakia

Jozef Lasz
Department of Telecomunications
Slovak University of Technology
Ilkovicova 3
Bratislava, Slovakia

Ivan Kotuliak
Department of Telecomunications
Slovak University of Technology
Ilkovicova 3
Bratislava, Slovakia
e-mail: `Ivan.Kotuliak@stuba.sk`

MICHAL KOCISKÝ was born in 1983 in the Slovak Republic. From 2001, he is a graduate student at STU, in Telecommunication specialization. At the same time, he cooperates on several research projects in networking area at the Dept. of Telecommunications.

JOZEF LASZ was born in 1981 in Bratislava, Slovak Republic. He received his BSc. and MSc. degrees in telecommunication from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (MSc. in 2006). His scientific interest is network performance analysis.

IVAN KOTULIAK was born in 1974 in the Slovak Republic. He completed his BSc. and MSc. studies of information technology at Slovak University of Technology (STU). During his research at the National Institute of Telecommunications in Evry, France, he prepared his PhD, which he received from both Versailles University and STU in 2003. Since 2003, he has been a researcher at STU and his scientific interest are networks and their performance analysis, using modeling and simulation tools. He is author of more than 20 scientific papers.