Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

Annales UMCS Informatica AI XIII, 2 (2013) 69–81
DOI: 10.2478/v10065-012-0040-0

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI
http://www.annales.umcs.lublin.pl/

Computer network traffic analysis with the use of statistical self-similarity factor

Paweł Dymora[1*], Mirosław Mazurek[1†], Dominik Strzałka[1‡]

[1] *Rzeszów University of Technology, Faculty of Electrical and Computer Engineering, Department of Distributed Systems, ul. Wincentego Pola 2, 35-959 Rzeszów, Poland*

Abstract – The optimal computer network performance models require accurate traffic models, which can capture the statistical characteristic of actual traffic. If the traffic models do not represent traffic accurately, one may overestimate or underestimate the network performance. The paper presents confirmation of the self-similar nature of the selected protocols in the computer network communication layer. It shows that the good measure of self-similarity is a Hurst factor.

1 Introduction

Statistical analysis of network traffic measurements shows a clear presence of the fractal or self-similar properties in computer network [**1**]. This means that similar statistical patterns may occur at different time scales which can vary by many time orders. The statistical characteristics of computer network traffic have been of greate interest to scientists for many years, not least to obtain a better understanding of the factors that affect the performance and scalability of large systems such as the Internet. Network traffic is inherently fractal or long-range dependent (LRD). That fact leads to question the extent to which the results of these studies are applicable in practice. Is it possible to diagnose network traffic and provide congestion risk? At the time being, there is mounting evidence that LRD is of fundamental importance for a number of engineering problems, such as traffic measurements [**2, 3, 4**] and queuing behaviour. The similar processes have been observed and analyzed in a number of other areas like,

*pawel.dymora@prz.edu.pl
†mirekmaz@prz.edu.pl
‡strzalka@prz.edu.pl

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

for instance hydrology, economics, biophysics. A self-similar phenomenon represents a process displaying structural similarities across a wide range of scales of a specific dimension. Recent measurements of network traffic have shown that traffic exhibits variability in a wide range of scales. The reference structure repeats itself over a wide range of scales of diverse dimensions (geometrical, statistical or temporal), and the statistics of the process do not change with time. In reality, simple systems do not exist. In the case of real, complex systems, in contradiction to simple systems, one can indicate the following features of processes: thermodynamic non-equilibrium, heterogeneous topologies, small-worlds phenomenon, long-range dependencies, bursty and self-similar traffic, scale-free (power law) distributions, packet switching, structure hierarchy, percolation, clustering, self-organization, parameters degradation and collapse [**5, 6, 7**].

The statistical characteristic of teleinformatic systems changed a lot when the human behaviour was replaced by the hierarchical, complex system, so-called computer. On the other hand, the voice traffic was quite static and low variable (short-range dependent) but now the data traffic is much more variable with both the extremely short and extremely long calls (self-similar and long-range dependent) [**8, 9**]. Thus, it can be noticed that both the original "pure form" human behavior and the original "pure" traffic nature (a simple stream) are lost when higher (nested) layers of stack are successively added and the simple computer system becomes a complex system that has a fractal nature. It is well known that the nesting concerns all areas of computer engineering (networks, computer hardware, operating systems, programming language and queuing system as well) and inevitably leads to the long-term dependent processes from the short-range dependent processes. It is particularly intensified in complex large-scale systems, i.e. distributed systems, computer networks. The positive features of the system namely: heterogeneity, openness, security, scalability, failure handling, concurrency and transparency are understood by negative complex system features such as degradation and collapse [**10, 5, 11, 12**].

2 Self-similarity statistical factor

Self-similarity and fractals are the notions pioneered by Benoit B. Mandelbrot. Self-similarity can be associated with "fractals", which are objects with unchanged appearances over different scales. In the case of statistical fractals, this is the probability density that repeats on every scale. On the other hand a dynamical fractal is generated by a low-dimensional dynamical system with chaotic solutions. The research related to traffic self-similarity can be classified into four categories: measurement-based traffic modelling, physical modelling, queuing analysis and traffic control as wll as resource provisioning [**1, 13**]. In order to review the LRD processes several definitions are introduced.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

A self-similar time series has the property that when aggregated (leading to a shorter time series in which each point is the sum of multiple original points), the new series have the same autocorrelation function as the original.

That is, given a stationary time series $X = (X_t;\ t = 0, 1, 2, \ldots)$, we define the m-aggregated series $x^{(m)} = (x_k^{(m)} :\ k = 1, 2, 3, \ldots)$ by summing the original series X over the nonoverlapping blocks of size m. Then if X is self-similar, it has the same autocorrelation function $r(k) = R[(x_t - \mu)(x_{t+k} - \mu)]$ as the series $X^{(m)}$ for all m. This means that the series is self-similar: the distribution of the aggregated series is the same (except changes in scale) as that of the original **[10, 3]**.

A process with long-range dependence has an autocorrelation function $r(k) \sim k^{-\beta}$ as $k \to \infty$ where $0 < \beta < 1$. Thus the autocorrelation function of such a process decays hyperbolically (as compared to the exponential decay exhibited by the traditional traffic models). Hyperbolic decay is much slower than the exponential decay, and since $\beta < 1$, the sum of tile autocorrelation values of such a series approaches infinity. This has a number of implications. First, the variance of n samples from such a series does not decrease as a function of n (as predicted by basic statistics for uncorrelated datasets) but rather by the value $n^{-\beta}$. Second, the power spectrum of such a series is hyperbolic, rising to infinity at frequency zero-reflecting the "infinite" influence of long-range dependence in the data.

The main advantage of using models of self-similar patterns of the time series is that the degree of self-similarity of the series is expressed by only one parameter. The parameter expresses the speed of decay series autocorrelation function. For historical reasons, the parameter used is the *Hurst* parameter $H = 1 - \beta/2$. For self-similar series, $1/2 < H < 1$, as $H \to 1$ the degree of self-similarity increases. Thus, the main criterion for a series of self-similarity reduces the question of whether H is significantly different from $1/2$.

There are many ways to determine the variance. We can use the variance-time plot, basing on the slowly decaying variance of a self-similar series. The variance of $X^{(m)}$ is plotted against m on a log-log plot; a straight line with the slope (β) greater than -1 is indicative of self-similarity, and the parameter H is given by $H = 1 - \beta/2$. We can use the R/S method. The R/S plot, uses the fact that for a self-similar dataset, the rescaled range or R/S statistic grows according to a power law with the exponent H as a function of the number of points included (n). Thus the plot of R/S against n on a log-log scale has a slope which is an estimation of H. The last approach, the periodogram method, uses the slope of the power spectrum of the series as frequency approaches zero. On a log-log plot the periodogram slope is a straight line with the slope $\beta - 1 = 1 - 2H$ close to the origin.

These methods are not resistant to faulty assumptions (such as non-stationarity in the dataset) and they do not provide confidence intervals. The fourth method, called the Whittle estimator does provide a confidence interval, but has the drawback that the form of the underlying stochastic process must be supplied.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

3 Network architecture and subject of analysis

In our study, we use data collected in a private computer network (small company). The collected data were the result of the normal operation of programming between the hours of 10 am to 10 am the following day. The company has an eight-hour working time in two hourly intervals from 7:00 to 15:00 and from 8:00 to 16:00. It is possible that employees stay after regular working hours. At night, all computers should be turned off, but this is not strictly obeyed. The network uses 19 computers and network devices. The analyzed network topology is shown in Fig. 1.

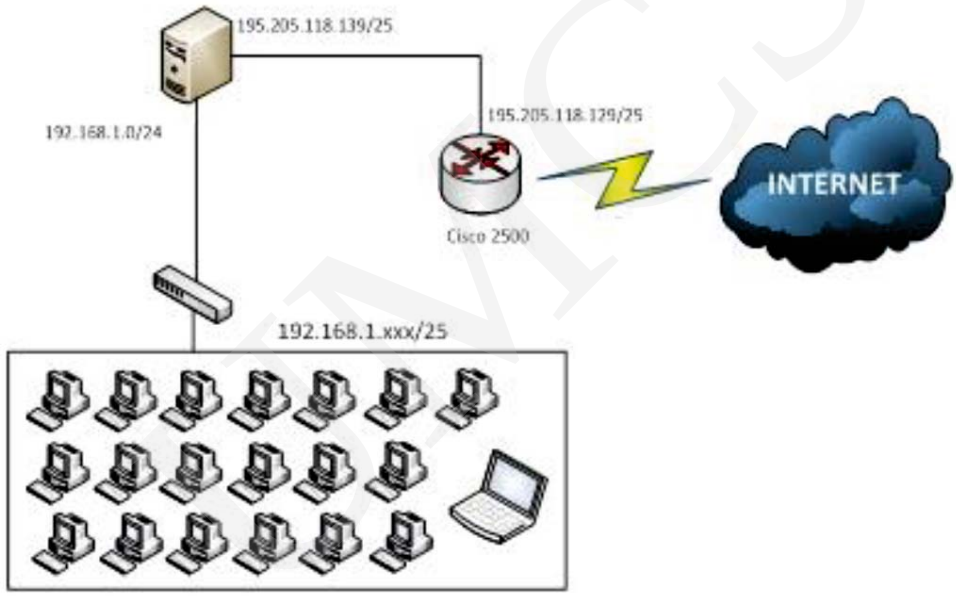


Fig. 1. Analyzed network architecture model.

To collect the data we used one of the sniffer programs to capture packets - Wireshark. This program can save the movement from the level of the data link layer [**5**]. The captured traffic samples contain information such as the location of the file, its size, format, type of encapsulation and packet size limit, time of the first packet that is the start time of the test procedure and its completion, the total length of the work. In addition, it provides the information about the number and type of packages. For 24 hours the analysed network recorded 7 818 848 packets, the average speed was 90.447 packet per second. The examplary statistics is shown in Fig. 2.

The study aimed at observing network traffic and determining whether there are long-term dependencies in all network working time and above-hour intervals. In order to carry out work of all captured packets we isolated those that had the greatest impact on the network. They were divided in the terms of services and protocols into five main groups:

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

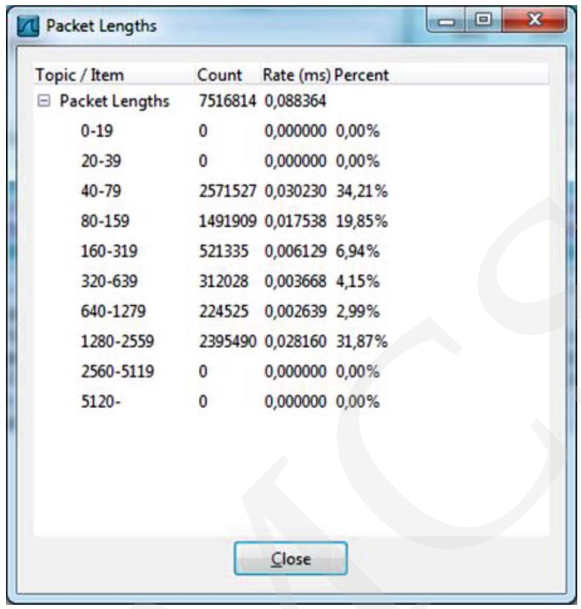


Fig. 2. Examplary statistics of packets lengths in the collected data.

- Web – *HTTP*;
- E-mail – *POP3,IMAP, SMTP*;
- SSL;
- Unknown – *IPSec, DG Gryphon*;
- IPv6.

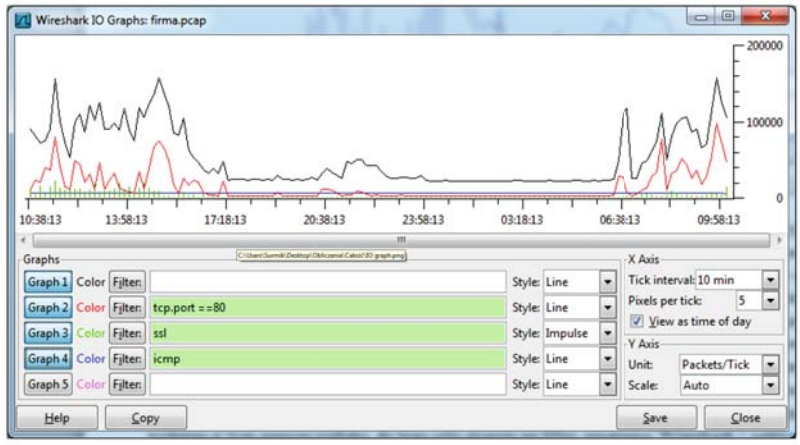


Fig. 3. 24 hours period of analyzed network traffic.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

For each group number of packets, the total length of the packet and the average packet length in hourly intervals were calculated. There were also the largest and the smallest size packages. The next step was to calculate the Hurst factor by earlier estimation of β using the Benoit and Power Spectrum method.

We studied using the HTTP protocol. The HTTP is the most commonly used protocol that supports the Internet. The data are used only with TCP and the default configuration uses port 80. Each object (e.g. website, video, audio) downloaded from a Web server sends through a single session.

An essential element of communication and business users is the e-mail. Messages that are sent can use traffic spam. We send text messages using standard protocols (POP3 and SMTP) which do not significantly affect network traffic. However, if the message contains a large attachment, sending and retrieving it may have a significant impact on the operation of the entire network. We analyzed the total traffic generated by the company employees within 24 hours (Fig. 4 b). With the analysed samples of traffic packets we estimate the exchange of electronic mail 2.52% of the total. In total, protocols responsible for the correspondence were sent up to 878.44 MB. The largest increase occurred in the traffic business hours.

The third analyzed service is a Secure Sockets Layer (SSL). SSL technology was originally developed by Netscape Communications to ensure the safety and privacy on the Internet compiled session. It enters data stream encryption. In 1999 the Transport Layer Security (TLS) standard was published, which provides security at the transport layer and solves some of the SSL problems. It is used to encapsulate higher-level applications traffic such as HTTP, Lightweight Directory Access Protocol (LDAP), FTP, SMTP, POP3 and IMAP. It provides authentication and integrity through certificates and digital signatures.

Programs for the analysis of traffic networks are not always able to recognize all the protocols that exist in the captured files. Groups of such unrecognized packets are described as unknown (Fig. 4). Typically, the data are sent by the programs that have their copyrights protocols recognized as unknown. But they can also be sent by recognized data protocols, but the port numbers must be changed. In the case of enterprise network tests, the packets marked as unknown are the largest share of traffic which is 51.2% and 17.88 GB of data sent within 24 hours.

Sample graph power-spectrum density for the same hours is shown in Fig. 5.

4 Results of investigations

Obviously, the choice of method is connected with theoretical considerations presented in Section 2. Usually it is hard to calculate a real value of spectral density slope because of the usage of the last mean square method that is not necessarily good for log-log plots, but it is commonly accepted that such an approach can be taken for rough estimation of H parameter.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

Fig. 4. Traffic network in 24 hours period.

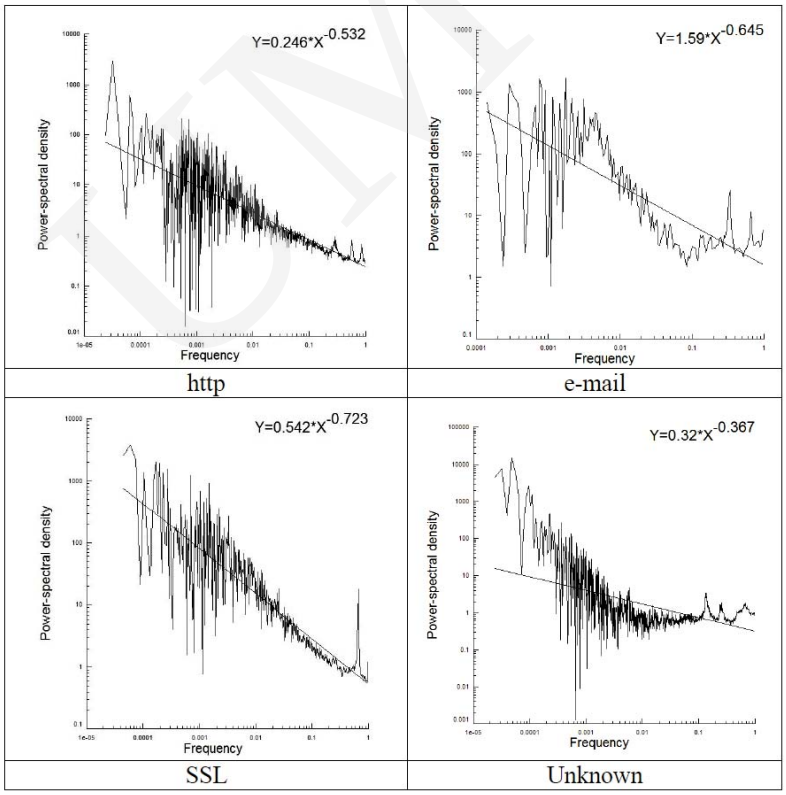
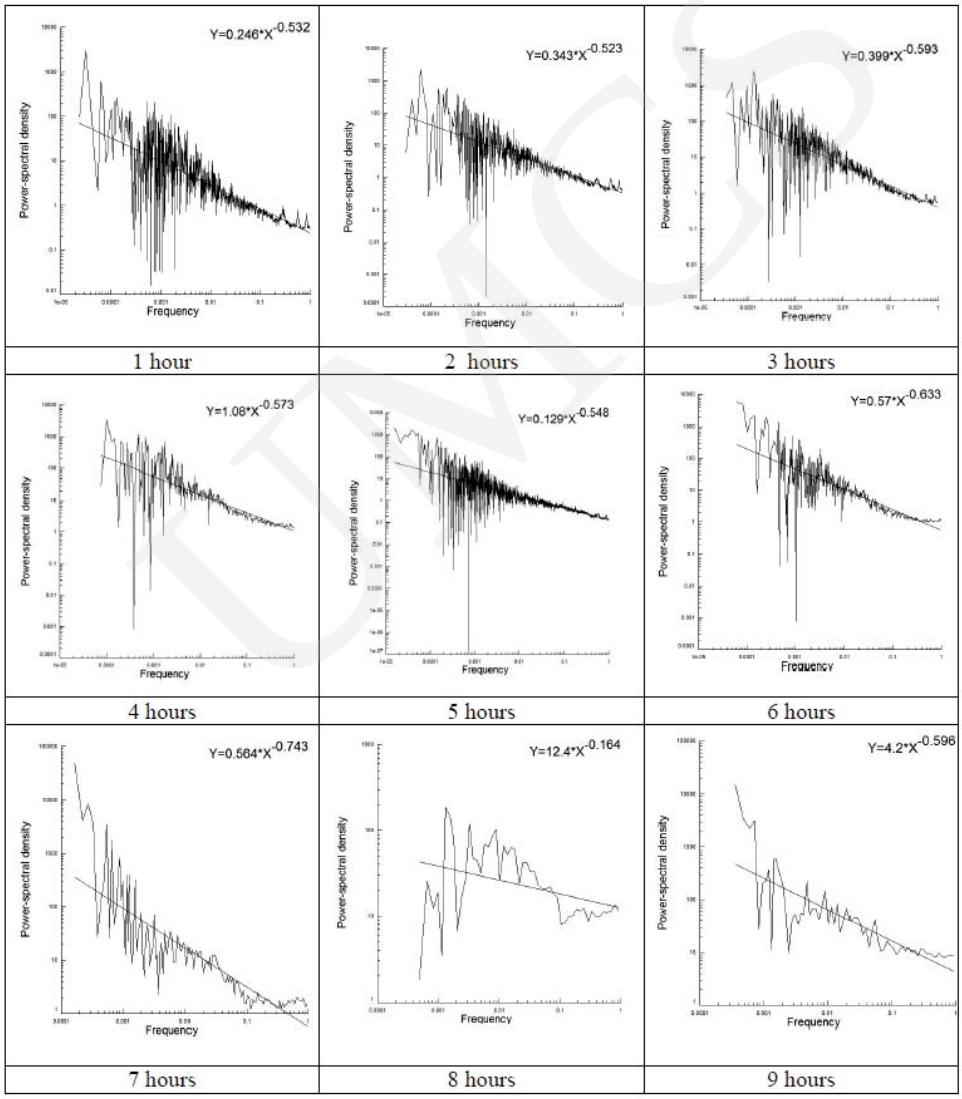


Fig. 5. Power-spectrum density in the same hours for different services.

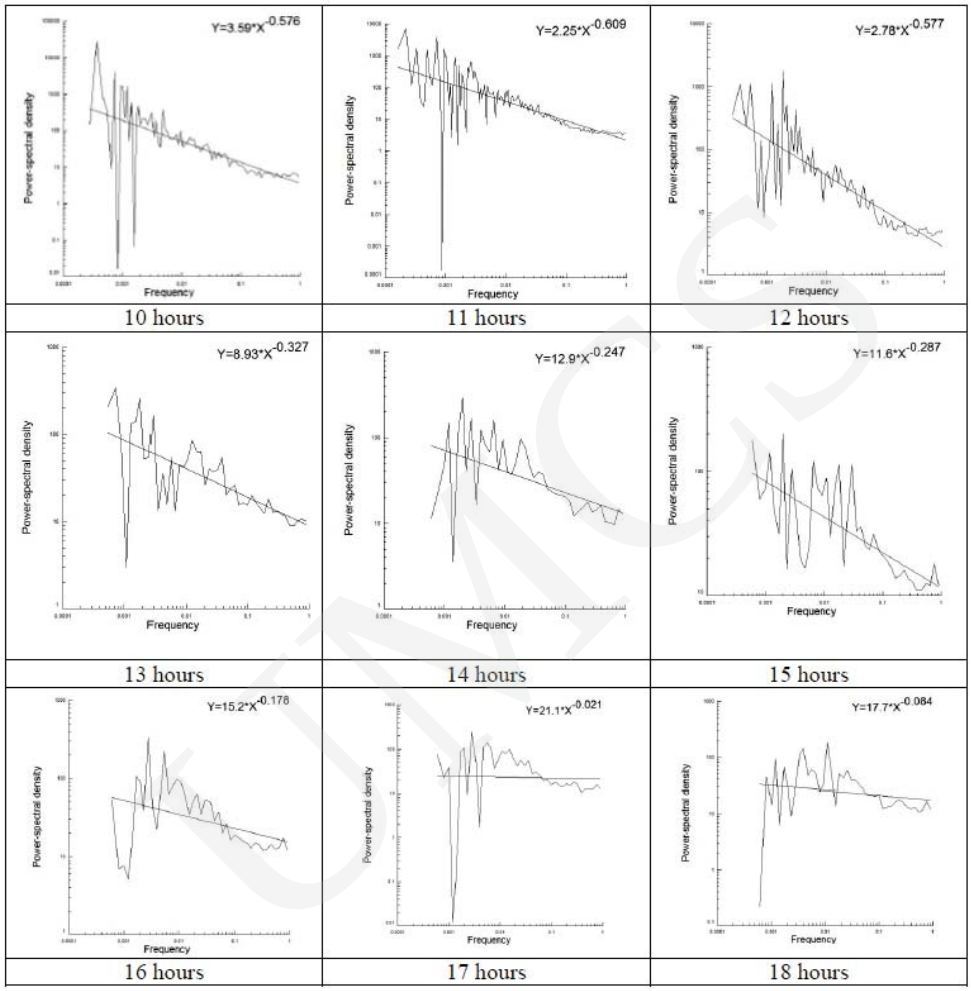
Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

In order to calculate the possible existence of long-range dependencies, the data are divided into one-hour intervals. We decided to divide our collection into 24 subcollections (they represent traffic during 1 hour) and calculate for them the slope of spectral density. This was shown in Fig. 6. Generally, it is considered that this method calculates the degree of long-range dependencies, regardless of whether a process belongs to the Gaussian or power-law probability distributions domains of attraction.



The obtained results of this experiment show that in contradiction to so far existing belief the traffic can have a self-similar nature. As it can be seen (Table 1) such a

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16



property is well visible. Power spectral density can be used not only for proposal of analytical model but also to confirm the validity of the proposed model based on experiment.

The analysis of selected protocols shows that the network traffic is self-similar. The degree of self-similarity Hurst exponent is specified by the range 0.5 to 1. The shorter the average length of packet including the Hurst exponent tends to 0.5 (white noise). The average value of the Hurst exponent for the e-mail traffic is 0.799, with a maximum value of 0.976 and 0.513 minimum. It can be seen that reducing the flow in the network (for example, overnight) causes the large fluctuations of the Hurst exponent, which tends to a value of 0.5. For SSL the average Hurst exponent is 0.721. In the entire traffic range of the test, it oscillates in the range of 0.54 to 0.98. The analysis by the Hurst exponent for the flow of packets marked as unknown shows that this traffic

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

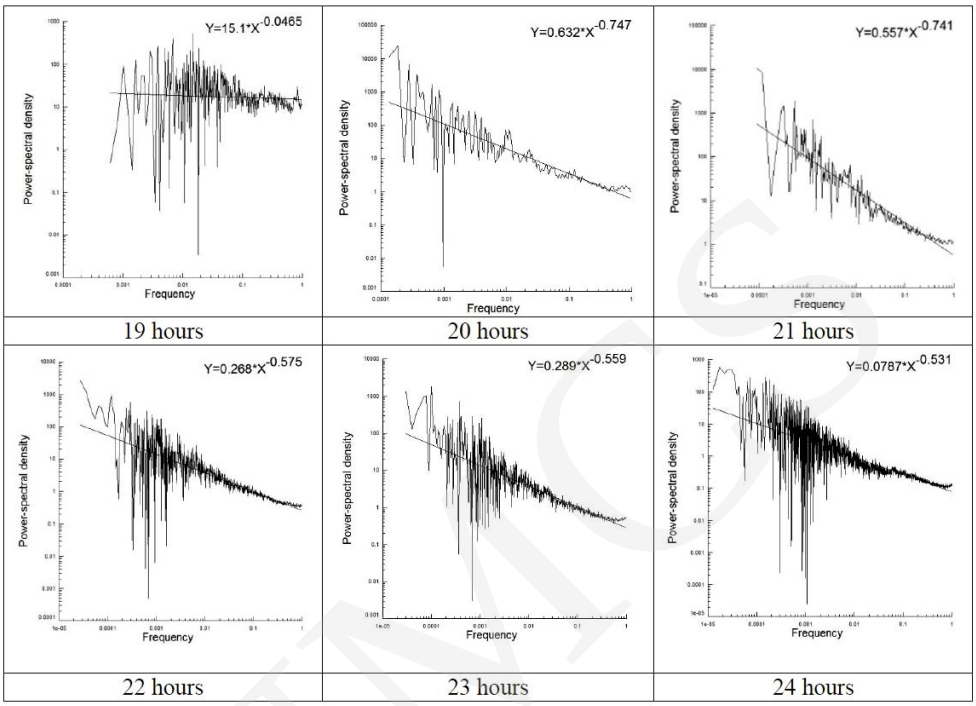


Fig. 6. Power-spectral density for 24 hour–intervals.

does not show a large variation. Even in the event of fluctuations in the number of packages, the results are very similar in each hour. Its nature is self-similar, and the value audited Hurst exponent is in the range of 0.5 to 0.7. The studies show that a new version of the Internet protocol IPv6 is very good in the LAN environment, even in the case of large fluctuations in the number of packages, the Hurst exponent does not change much. The analysis for the IPv6 traffic presents the calculated value of the Hurst exponent at the level of 0.665 and shows a self-similar nature to the degree of self-similarity in the range of 0.5 to 1.

An interesting result is provided by the dependence of the average packet length Hurst factor. The analysis of the traffic for specific protocols shows that the value of the Hurst exponent takes the values between 0.5 and 1.0, which confirms the self-similar nature of the analyzed protocols. The degree of self-similarity is dependent on the volume of traffic and the type of service. In the case of minor traffic approaching the value of the exponent equal to 0.5 (white noise), it is characterized by the complete randomness and a lack of correlation between packages. The changes in the Hurst coefficient in the range of 24 hours are shown in Fig. 7.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

Table 1. Summary of the Hurst factor values for different services.

Hour	HTTP		e-mail		SSL		Unknown		IPv6	
	β	H	β	H	β	H	β	H	β	H
1	-0,532	0,766	-0,645	0,8225	-0,723	0,8615	-0,367	0,6835	-0,671	0,8355
2	-0,523	0,7615	-0,765	0,8825	-0,572	0,786	-0,247	0,6235	-0,597	0,7985
3	-0,593	0,7965	-0,717	0,8585	-0,613	0,8065	-0,2	0,6	-0,152	0,576
4	-0,573	0,7865	-0,774	0,887	-0,635	0,8175	-0,233	0,6165	-0,341	0,6705
5	-0,548	0,774	-0,74	0,87	-0,66	0,83	-0,0777	0,53885	-0,205	0,6025
6	-0,633	0,8165	-0,663	0,8315	-0,608	0,804	-0,182	0,591	-0,395	0,6975
7	-0,743	0,8715	-0,565	0,7825	-0,835	0,9175	-0,0906	0,5453	-0,507	0,7535
8	-0,164	0,582	-0,357	0,6785	-0,556	0,778	-0,146	0,573	-0,192	0,596
9	-0,596	0,798	-0,952	0,976	-0,0924	0,5462	-0,0984	0,5492	-0,133	0,5665
10	-0,576	0,788	-0,249	0,6245	-0,22	0,61	-0,0974	0,5487	-0,206	0,603
11	-0,609	0,8045	-0,642	0,821	-0,563	0,7815	-0,269	0,6345	-0,442	0,721
12	-0,577	0,7885	-0,026	0,513	-0,637	0,8185	-0,221	0,6105	-0,176	0,588
13	-0,327	0,6635	-0,193	0,5965	-0,101	0,5505	-0,192	0,596	-0,152	0,576
14	-0,247	0,6235	-0,876	0,938	-0,0789	0,53945	-0,115	0,5575	-0,213	0,6065
15	-0,287	0,6435	-0,273	0,6365	-1,6	0,3	-0,115	0,5575	-0,307	0,6535
16	-0,178	0,589	-0,888	0,944	-0,503	0,7515	-0,033	0,5165	-0,195	0,5975
17	-0,021	0,5105	-0,896	0,948	-0,146	0,573	-0,136	0,568	-0,206	0,603
18	-0,084	0,542	-0,412	0,706	-1,34	0,17	-0,024	0,512	-0,231	0,6155
19	-0,046	0,523	-0,127	0,5635	-0,661	0,8305	-0,172	0,586	-0,265	0,6325
20	-0,747	0,8735	-0,79	0,895	-0,963	0,9815	-0,0897	0,54485	-0,247	0,6235
21	-0,741	0,8705	-0,71	0,855	-0,629	0,8145	-0,169	0,5845	-0,695	0,8475
22	-0,575	0,7875	-0,687	0,8435	-0,654	0,827	-0,181	0,5905	-0,451	0,7255
23	-0,559	0,7795	-0,713	0,8565	-0,592	0,796	-0,101	0,5505	-0,393	0,6965
24	-0,531	0,7655	-0,705	0,8525	-0,628	0,814	-0,118	0,559	-0,524	0,762

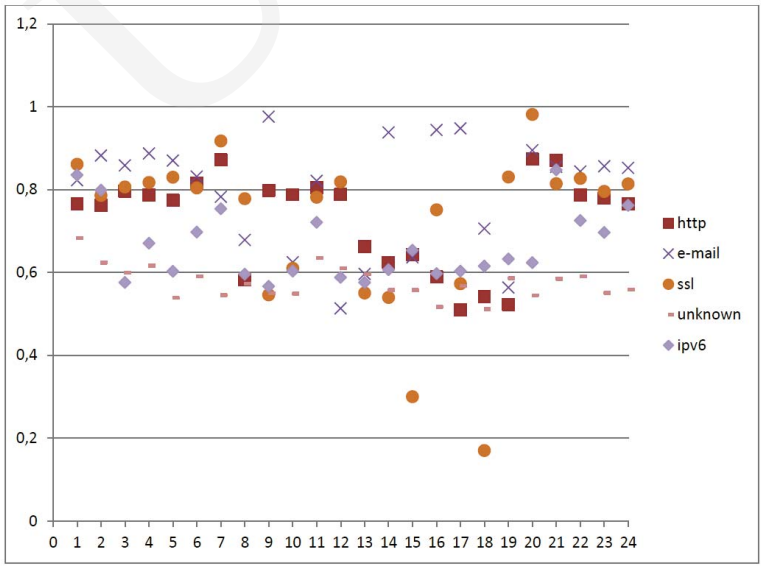


Fig. 7. The Hurst coefficient in 24 hour–period for selected services.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

5 Conclusions

In this paper we have reported the results from the analysis of the computer network traffic using the statistical self-similarity factor. The results confirmed that the analyzed traffic has a self-similar nature to the degree of self-similarity in the range of 0.5 to 1. The measurement and analysis have shown that the self-similar nature of computer network traffic expressed by the fractional Brownian motion or the fractional Gaussian noise and the holistic approach to queuing analysis made it possible to determine the power spectral density which can be an internal level measure of high variable traffic in the whole system. We can observe that burstiness is present across many time scales. The parameter H is larger when network utilization is higher. The network performance is dominated by the self-similarity property in the network traffic. Some of the most significant physical phenomena may give a significant raise to LRD of user behaviour, data generation, organization and retrieval, traffic aggregation, network controls etc. The results of analytical considerations and experiment show that the self-similarity factor can be successfully used in the computer network traffic analysis.

The research was carried out on the equipment purchased in the project No POPW.01.03.00-18-012/09 from the Structural Funds, The Development of Eastern Poland Operational Program co-financed by the European Union, the European Regional Development Fund.

References

[1] Park K., Willinger W., Self-similar Traffic and Performance Evaluation, John Wiley & Sons, Inc. (2000).

[2] Crovella M. E., Bestavros A., Explaining World Wide Web Traffic Self-Similarity, Technical Report TR-95-015 (1995).

[3] Grabowski F., Strzałka D. Dynamic behavior of simple insertion sort algorithm, Fundamenta Informaticae 72 (2006): 1653.

[4] Field A. J., Harder U., Harrison P. G., Measurement and modeling of self-similar traffic in computer network, IEE Proc. Commun. 151(4) (2004).

[5] Grabowski F., Strzałka D., Simple, complicated and complex systems – the brief introduction. in: 2008 Conference On Human System Interactions 1, 2 (2008): 576.

[6] Dymora P., Mazurek M., Strzałka D., Statistical mechanics of memory pages reads during man–computer system interaction, Metody Informatyki Stosowanej 1 (26) (2011): 15.

[7] Strzałka D., Szurlej P., Power-law distributions in hard drive behavior, Journal of Software Engineering and Applications 04(12) (2011): 710.

[8] Strzałka D., Non-extensive statistical mechanics – a possible basis for modeling processes in computer memory system, Acta Physica Polonica A 117(4) (2010): 652.

[9] Strzałka D., Grabowski F., Non-Extensive Thermodynamics of Algorithmic Processing - the Case of Insertion Sort Algorithm, in Thermodynamics, ed. Tadashi Mizutani, InTech (2011): 121.

[10] Strzałka B., Mazurek M., Strzałka D., Queue Performance in Presence of Long-Range Dependencies – an Empirical Study, International Journal of Information Science 2(4) (2012): 47.

[11] Dymora P., Mazurek M., Strzałka D., Long-range dependencies in memory pages reads during man-compute system interaction, Annales UMCS Informatica XII (2) (2012): 49.

Pobrane z czasopisma Annales AI- Informatica **http://ai.annales.umcs.pl**
Data: 23/03/2020 12:41:16

[12] Dymora P., Mazurek M., Strzałka D., Piękoś M., Influence of batch structure on cluster computing performance - complex systems approach, Annales UMCS Informatica XII (1) (2012): 57.
[13] Eberbach E., Wegner P. Beyond Turing machines, Bulletin of the European Association for Theoretical Computer Science 81 (2003): 279.

UMCS