On network traffic statistical analysis

Liudas KAKLAUSKAS, Leonidas SAKALAUSKAS (MII)
e-mail: liukak@fm.su.lt

Abstract. The present article deals with statistical university network traffic, by applying the methods of self-similarity and chaos analysis. The object of measurement is Šiauliai University LitNet network node maintaining institutions of education of the northern Lithuania region. Time series of network traffic characteristics are formed by registering amount of information packets in a node at different regimes of network traffic and different values of discretion of registered information are present. Measurement results are processed by calculating Hurst index and estimating reliability of analysis results by applying the statistical method. Investigation of the network traffic allowed us drawing conclusions that time series bear features of self-similarity when aggregated time series bear features of slowly decreasing dependence.

Keywords: self-similarity, computer network, fractality.

1. Introduction

Empirical research of computer network packet traffic shows that it is attributed with self-similarity [1, 2, 6, 8]. After estimating the latter feature, it is possible to adequately prognosticate the change of traffic and to apply the prognosis results in increase of network throughput and improvement of its QoS quality of service, while regulating packet latency, fluctuation restriction and packet loss transportation on data and physical OSI layers [3, 10]. Quality of Service (QoS) refers to the capability of a network to provide better service to selected network traffic over various technologies. These technologies allow you to measure bandwidth, detect changing network conditions (such as congestion or availability of bandwidth), and prioritize or throttle traffic.

The self-similarity phenomenon is explained by network usage order attributed with burstiness. In fact, data is inherently "bursty" as it occurs in short bursts of communications followed by long periods of silence. Indeed, one can characterize data communication users who wish network resources to send their data as follows: users don't warn you exactly when they will demand access; one cannot predict how much they will demand, most of the time users do not need access to network; when users ask for it, they want immediate access [9]. Such situation is frequently faced in distance learning networks when students receive tasks and send theirs answers almost at the same time.

In contemporary university studies, computer networks are widely applied; they often undergo non-prognosticated overload. For effective network control, it is necessary to perform monitoring of network nodes in order to prognosticate network node load and overload. On the base of A. Erramilli, O. Narayan and W. Willinger, in 1989, by empirical research of Ethernet local area network of 10 Mbps which was carried out at Bellcore laboratory, it was estimated that Ethernet traffic characterisations bear fractal characteristics and are attributed with self-similarity with long-range dependence [1]. I. Kaj [5] in the monographs suggests the methods of statistical analysis of characteristics of modern communication traffic, by applying possibilities of contemporary mathematical modelling. J. Beran, analyse network traffic as a fractal process attributed with a second order statistical self-similarity which is characterised by a fractal measure [6]. Methods of non-linear (chaos) theory are applied for modelling and description of network processes, while estimating the heavy-tails which characterise large burstiness of network traffics.

The aim of this research is to analyse measurement results of Šiauliai University LitNet network node traffic and to estimate its self-similarity. It should be noted that analogous analysis of region's various educational network data has not been carried out yet in Lithuania in order to find out about the self-similarity. Programme and device tools for monitoring the network were used in analysis of network traffic; these tools registered data packets at the indicated interval. Data was registered while applying different levels of time discretisation, at different levels of network load present, while forming aggregated time series. Measurement results were processed by estimating the fractal measure and calculating Hurst coefficient and statistically estimating reliability of analysis results [6].

2. Composition of Empirical Data

For measurement of network traffics, Šiauliai University LitNet network node with the highest intensity of traffic load was chosen. In this node, received inter-city channel traffic of 1 Gbps is distributed to the university and educational institutions of Šiauliai region. Only data packets arriving at the node M, while disregarding sent packets, were analysed. Obtained information was collected in external data base Porstgree SQL (DB). Initial measurement was carried out with exactness of one microsecond. Record on the data base was formed right after receiving TCP or other protocol's data frame. Service information was not withdrawn while saving framework data: title, feature of framework beginning, addresses of a sender and receiver, etc. The biggest length of fixed transport frameworks was up to 1518 bytes [10].

ulogd software for Linux operational system distributed under GPL licence was used for measurement [12]. Data was being fixed in incoming data traffic drives of the router. Every pre-routed packet is registered by *ulogd* daemon in PostgreeSQL database (see Fig. 1). Data from January 4, 2008 13:30:35 to April 16, 2008, 12:00:00 was chosen from the data base for analysis. Within this period, more than three billion records were accumulated in the data base; they corresponded to 8936965 seconds or 103 days 10 hours 29 minutes and 25 seconds. Data for analysis was selected according to days of the week and part of the day, while estimating intensity of data traffic, i.e., those time series were selected when data traffic was the least (Sundays), medium (Saturdays) or the highest (weekdays). While investigating the change of load during the day, hours when data traffic is the highest, medium and the least were indicated for every day of the week. The method of k-means of cluster analysis was applied in analysis of intensity of data traffic in selected time intervals, while using Statistical Package for Social Sciences (SPSS). The method of non-hierarchic cluster analysis was applied, when the amount of cluster figures ($k = 24$) is known, and distances

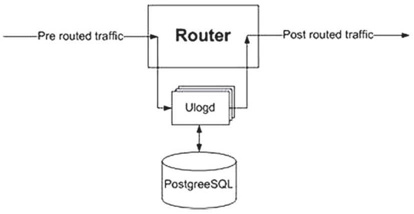


Fig. 1. The scheme of network traffic registration.

between clusters and objects are calculated by using Euclidean square range metrics. Series of one hour measurement consists of up to one-and-a-half million records. 309 hours of when data traffic is the highest, medium and the least were selected for further analysis.

In order to analyse such time series, it must be aggregated, i.e., to calculate data traffics in equal time intervals. For aggregation of data, two methods were chosen: 1 – the method of smoothing of moving surfaces was applied when an average traffic for a data series are calculated in a chosen time interval Δt: $x_k^{\Delta} = \frac{\sum_{\tau_i \in [t_k, tk-1]} x_i}{\Delta t}$, here $t_k = k\Delta t + \tau_1$. Obtained time series characterise average changes of data traffic in time moments Δt; 2 – transferred data traffic amount within the time interval Δt: $x_k^{\Sigma} = \sum_{\tau_i \in [t_k, tk-1]} x_i$, here $t_k = k\Delta t + \tau_1$, τ_i is data traffic measure interval, where $\tau_i - \tau_{i-1} \neq \tau_{i+1} - \tau_i$.

While forming the series for the research, the time intervals $\Delta t \in [100\,ms, 500\,ms, 1\,s]$ were chosen. Out of 309 selected measurement sequences, 6 queue groups were formed, totally 1854 series. Aggregated time series, while estimating network load, are marked as follows: $x_t^{\Sigma_{\min}}$, $x_t^{\Delta_{\min}}$ – the lowest load, $x_t^{\Sigma_{\Delta}}$, $x_t^{\Delta_{\Delta}}$ – average load and $x_t^{\Sigma_{\max}}$, $x_t^{\Delta_{\max}}$ – the highest load.

For estimation of time series, the programme Fractan 4.4 was applied [10]. In the aggregated series, the programme calculates the following: Hurst coefficient and fractal measure, presents graphic image of numerical values and draws obtained attractors.

3. Network self-similarity estimated by using Hurst statistics

As time series formed of lengths of data frameworks transferred via the computer network do not satisfy the normal distribution, this section investigates their Hurst statistics. Hurst coefficient characterises whether the series analysed is random, whether it has a short-range or long-range, also called Markov, dependence. If Hurst coefficient $H = 0.5$, it means that sequence members are random and its every subsequent member does not depend on previous series members; in an opposite case, we can state that previous events recorded in time series have constant influence on further processes and this influence is the stronger the closer the event is to the past. Such series are invariant from the viewpoint of time. Influence of the current process on future events is calculated by estimating its correlation [6, 3]: $C = 2^{2H-1} - 1$, where C – correlation measure, o H – Hurst coefficient. While evaluating self-similarity of a time series, the value of Hurst coefficient, i.e., interval where it occurs, is very important.

If $0 \leqslant H < 0.5$, then the process characterised by the time series is anti-dispersive, i.e., we can state that if increase is observed in one period, in other period decrease will definitely follow, and the probability is the higher the closer H is to 0. In this case, correlation is negative and draws closer to 0.5. Such series usually bear a feature of high changeability and are formed of frequent increases and decreases.

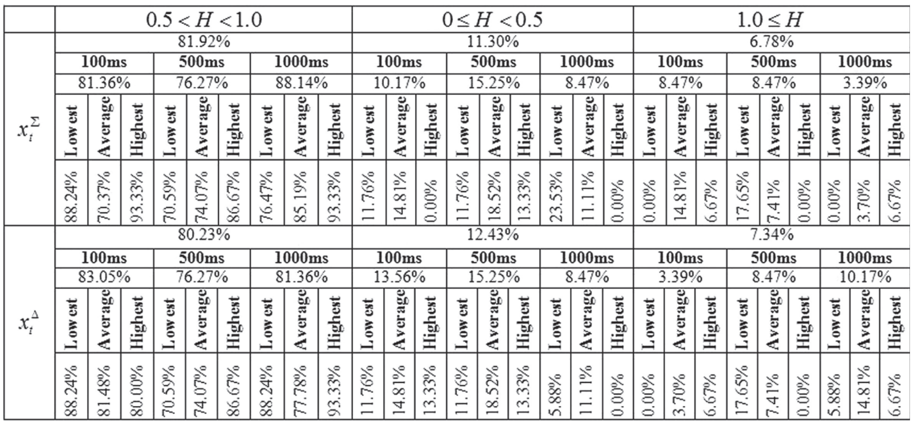
If $0.5 < H < 1.0$, then it is a persistent process with long-term memory, i.e., in the past, the process had a feature for increase, and it will retain this in future with the higher probability the closer H is to 1, and correlation will draw closer to 1. Usually, such series are called trend-resistant, while H draws closer to 0.5, the amount of trends (noises) increases in the series.

For formed and aggregated time series x_t^{Θ} in the network node, Hurst coefficient is calculated according to the formula $H = \log(R/S)/\log(n/2)$, where H – Hurst coefficient, R/S – r/s statistics acquired according to the formula:

$$R/S = \frac{R(n)}{S(n)} = \frac{Max\left(\sum_{i=1}^{\tau}(x_i^{\Theta} - \overline{x_t^{\Theta}})\right) - Min\left(\sum_{i=1}^{\tau}(x_i^{\Theta} - \overline{x_t^{\Theta}})\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{\Theta} - \overline{x_t^{\Theta}})^2}},$$

here $1 \leqslant \tau \leqslant n$, where n – amount of sequence members, $\overline{x_t^{\Theta}}$ – average value of the series x_t^{Θ}, and $\sum_{i=1}^{\tau}(x_i^{\Theta} - \overline{x_t^{\Theta}})$ – the formed cumulative series describing sum of changes throughout time τ. According to Hurst [12], we can state that the expression is suitable for majority of natural phenomena: $M\left(\frac{R(n)}{S(n)}\right) \approx cn^H, n \to \infty$, where c – constant independent value [6]. Hurst coefficient is closely related with the fractal measure D which characterises local features of computer network data traffic, and Hurst coefficient describes characteristics of the whole process – memory of the process. In self-similar processes, local features are reflected in global ones and vice versa; because the time series measure $N = 1$, therefore the connection can be estimated by using the formula: $D = 2 - H$, where D – fractal measure, the so called attractors' dimension, H – Hurst coefficient. For estimation of attractors' dimension, we calculate

Table 1. Dispersal of Hurst coefficient values

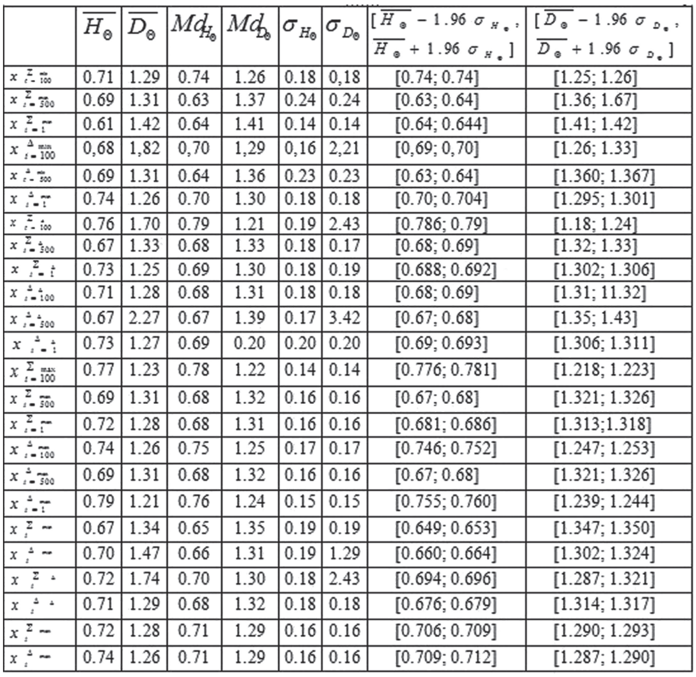


	$0.5 < H < 1.0$									$0 \leq H < 0.5$									$1.0 \leq H$								
x_t^{Σ}	81.92%									11.30%									6.78%								
	100ms			**500ms**			**1000ms**			**100ms**			**500ms**			**1000ms**			**100ms**			**500ms**			**1000ms**		
	81.36%			76.27%			88.14%			10.17%			15.25%			8.47%			8.47%			8.47%			3.39%		
	Lowest	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**
	88.24%	70.37%	93.33%	70.59%	74.07%	86.67%	76.47%	85.19%	93.33%	11.76%	14.81%	0.00%	11.76%	18.52%	13.33%	23.53%	11.11%	0.00%	0.00%	14.81%	6.67%	17.65%	7.41%	0.00%	0.00%	3.70%	6.67%
x_t^{Δ}	80.23%									12.43%									7.34%								
	100ms			**500ms**			**1000ms**			**100ms**			**500ms**			**1000ms**			**100ms**			**500ms**			**1000ms**		
	83.05%			76.27%			81.36%			13.56%			15.25%			8.47%			3.39%			8.47%			10.17%		
	Lowest	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**	**Lowest**	**Average**	**Highest**
	88.24%	81.48%	80.00%	70.59%	74.07%	86.67%	88.24%	77.78%	93.33%	11.76%	14.81%	13.33%	11.76%	18.52%	13.33%	5.88%	11.11%	0.00%	0.00%	3.70%	6.67%	17.65%	7.41%	0.00%	5.88%	14.81%	6.67%

Hausford measure which is obtained by analysing the strange Lorenz's attractor. For estimation of the system, we calculate Hausford measure D, the so called fractal measure [2]: $D = \lim_{\varepsilon\to 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon}$, here N is minimum amount of n-time blocks with facet length which cover points of a set, when facet length draws close to zero. We analyse the system, when $1 < D < 2$, then the formula is as follows: $D = \frac{\ln N}{\ln(1/(2*r)}$, here N – amount of elements used fro measuring fractal's undulation, $N \to 2$ when a fractal is in plane, r – radius of a circle used in 2-dimensional space. In computer networks, fractal measure characterises dynamics of formed data traffic time series changes, when one variable is used [13]. Hurst coefficient distribution with percentage estimation is displayed in Table 1. It suggests that more than 80% of time series $0.5 < H < 1.0$ and they preserve this feature when $\Delta t \in [100ms, 500ms, 1000ms]$ and different network loads $x_t^{\Theta_{\min}}, x_t^{\Theta_\delta}, x_t^{\Theta_{\max}}$.

The calculated numerical expressions are estimated analytically as well, i.e., for every series group, according to the time interval, the average, median, standard deviation and dependent interval with reliability of 95% were calculated. Estimation results are presented in Table 2. Hurst coefficient changes from 0.61 to 0.79, thus, the process of data transferred via computer network which is described by aggregated series is a persistent process with long-term memory. Fractal measure D changes from 1.23 to 1.82, and dependent intervals are very narrow; this proves that calculated values are reliable, and fractional expression of the fractal measure suggests that the series bear

Table 2. Estimations of Herst coefficient and fractal measure reliability



	$\overline{H_\Theta}$	$\overline{D_\Theta}$	Md_{H_Θ}	Md_{D_Θ}	σ_{H_Θ}	σ_{D_Θ}	$[\overline{H_\Theta} - 1.96\,\sigma_{H_\Theta};$ $\overline{H_\Theta} + 1.96\,\sigma_{H_\Theta}]$	$[\overline{D_\Theta} - 1.96\,\sigma_{D_\Theta};$ $\overline{D_\Theta} + 1.96\,\sigma_{D_\Theta}]$
$x_{t=100}^{\Sigma\,\min}$	0.71	1.29	0.74	1.26	0.18	0,18	[0.74; 0.74]	[1.25; 1.26]
$x_{t=500}^{\Sigma\,\min}$	0.69	1.31	0.63	1.37	0.24	0.24	[0.63; 0.64]	[1.36; 1.67]
$x_{t=1}^{\Sigma\,\min}$	0.61	1.42	0.64	1.41	0.14	0.14	[0.64; 0.644]	[1.41; 1.42]
$x_{t=100}^{\Delta\,\min}$	0,68	1,82	0,70	1,29	0,16	2,21	[0,69; 0,70]	[1.26; 1.33]
$x_{t=500}^{\Delta\,\min}$	0.69	1.31	0.64	1.36	0.23	0.23	[0.63; 0.64]	[1.360; 1.367]
$x_{t=1}^{\Delta\,\min}$	0.74	1.26	0.70	1.30	0.18	0.18	[0.70; 0.704]	[1.295; 1.301]
$x_{t=100}^{\Sigma\,\delta}$	0.76	1.70	0.79	1.21	0.19	2.43	[0.786; 0.79]	[1.18; 1.24]
$x_{t=500}^{\Sigma\,\delta}$	0.67	1.33	0.68	1.33	0.18	0.17	[0.68; 0.69]	[1.32; 1.33]
$x_{t=1}^{\Sigma\,\delta}$	0.73	1.25	0.69	1.30	0.18	0.19	[0.688; 0.692]	[1.302; 1.306]
$x_{t=100}^{\Delta\,\delta}$	0.71	1.28	0.68	1.31	0.18	0.18	[0.68; 0.69]	[1.31; 11.32]
$x_{t=500}^{\Delta\,\delta}$	0.67	2.27	0.67	1.39	0.17	3.42	[0.67; 0.68]	[1.35; 1.43]
$x_{t=1}^{\Delta\,\delta}$	0.73	1.27	0.69	0.20	0.20	0.20	[0.69; 0.693]	[1.306; 1.311]
$x_{t=100}^{\Sigma\,\max}$	0.77	1.23	0.78	1.22	0.14	0.14	[0.776; 0.781]	[1.218; 1.223]
$x_{t=500}^{\Sigma\,\max}$	0.69	1.31	0.68	1.32	0.16	0.16	[0.67; 0.68]	[1.321; 1.326]
$x_{t=1}^{\Sigma\,\max}$	0.72	1.28	0.68	1.31	0.16	0.16	[0.681; 0.686]	[1.313;1.318]
$x_{t=100}^{\Delta\,\max}$	0.74	1.26	0.75	1.25	0.17	0.17	[0.746; 0.752]	[1.247; 1.253]
$x_{t=500}^{\Delta\,\max}$	0.69	1.31	0.68	1.32	0.16	0.16	[0.67; 0.68]	[1.321; 1.326]
$x_{t=1}^{\Delta\,\max}$	0.79	1.21	0.76	1.24	0.15	0.15	[0.755; 0.760]	[1.239; 1.244]
$x_t^{\Sigma\,\min}$	0.67	1.34	0.65	1.35	0.19	0.19	[0.649; 0.653]	[1.347; 1.350]
$x_t^{\Delta\,\min}$	0.70	1.47	0.66	1.31	0.19	1.29	[0.660; 0.664]	[1.302; 1.324]
$x_t^{\Sigma\,\delta}$	0.72	1.74	0.70	1.30	0.18	2.43	[0.694; 0.696]	[1.287; 1.321]
$x_t^{\Delta\,\delta}$	0.71	1.29	0.68	1.32	0.18	0.18	[0.676; 0.679]	[1.314; 1.317]
$x_t^{\Sigma\,\max}$	0.72	1.28	0.71	1.29	0.16	0.16	[0.706; 0.709]	[1.290; 1.293]
$x_t^{\Delta\,\max}$	0.74	1.26	0.71	1.29	0.16	0.16	[0.709; 0.712]	[1.287; 1.290]

features of fractals. After generalising research data of this section, we can state that the investigated time series characterise persistent processes with long-term memory.

4. Conclusions

1. Estimations of Hurst coefficient reliability proved that the aggregated series describe a persistent process with long-term memory. It was proved by analysis of Hurst coefficient charts.
2. The data traffic of Šiauliai University LitNet network node is attributed with self-similarity with long-term memory.

References

1. A. Erramilli, O. Narayan, W. Willinger, *Experimental Queuing Analysis with LongnRange Dependent Packet Traffic* [interactive], accessed 2008-01-25: `<http://www-net.cs.umass.edu/cs691s/narayan96.ps>`
2. В.Ф. Разумов, *Соотношение определенности и случайности в физицеских законах* [interactive], accessed 2008-01-25: `<http:// www.icp.ac.ru /structure/departments /pho/labs/supramol/reports /Razumov_Lecture_ MFTI_2002.pdf>`
3. E. Leland, S. Taqqu, W. Willinger, D. V. Wilson, On the self-similar nature of Ethernet Traffic, *IEEE/ACM Transactions on Networking*, **2**(1) (1994).
4. G. He, Y. Gao, J.C. Hou, K. Park, A case for exploiting self-similarity of network traffic in TCP congestion control, *Computer Networks*: *The International Journal of Computer and Telecommunications Networking*, **45** (2004).
5. I. Kaj, *Stochastic Modeling in Broadband Communications Systems*, SIAM, Philadelphia, USA (2002).
6. J. Beran, *Statistics for Long-Memory Processes*, Capman & Hall/CRC, USA (1998).
7. K. Park, W. Willinger, *Self-Similar Network Traffic and Performance Evaluation*, John Wiley & Sons, USA (2000).
8. L. Kaklauskas, *Kompiuterių tinklai*, 1, 2 d., Šiaulių universiteto leidykla, Šiauliai (2003–2005).
9. L. Kleinrock, Creating a mathematical theory of computer networks, *Operations Research*, **50**(1), 125–131 (2002).
10. *Self-Similarity h.u*, accessed 2008-01-25: `<http://www.teletraffic.ru/index.php>`
11. *Userspace Logging Daemon*, accessed 2008-01-25: `<http://www.netfilter.org>`
12. H.E. Hurst, The longnterm storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers*, **116**, 770–799 (1951).
13. J. Theiler, Estimating fractal dimension, *Journal of the Optical Society of America*, **A7**, 1055–1073 (1990).

REZIUMĖ

L. Kaklauskas, L. Sakalauskas. Kompiuterių tinklo apkrovos statistinės analizės klausimu

Šiame straipsnyje nagrinėjama statistinė universitetinio tinklo apkrova, pasinaudojant savastingumo analizės metodu. Matavimo objektu pasirinktas Šiaulių universiteto LitNet tinklo mazgas, aptarnaujantis Šiaurės Lietuvos švietimo įstaigas. Tinklo apkrovos charakteristikų laiko eilutės formuojamos registruojant priimamos informacijos paketų skaičių mazge, esant skirtingoms tinklo apkrovos režimams bei įvairiomis registruojamos informacijos diskretizavimo reikšmėms. Matavimo rezultatai apdorojami, skaičiuojant Hersto indeksą bei įvertinant analizės rezultatų patikimumą statistiniu būdu. Tyrimai leido padaryti išvadas, kad nagrinėjamos laiko eilutės pasižymi savastingumu, kai agreguotos laiko eilutės pasižymi lėtai mažėjančia priklausomybe.

Raktiniai žodžiai: savastingumas, kompiuterių tinklas, fraktališkumas.