Rostyslav STRUBYTSKYI[*]

MODELLING AND FORECASTING OF CLOUD DATA WAREHOUSING LOAD

Abstract

Cloud data storages in their internal structure are not using their full potential functionality because of the complexity of behavior of network traffic, which affects the quality of service. The paper describes various models of network traffic and analyzes the most promising models for cloud data storages that take into account the phenomenon of self-similarity. The result of research found the frequency of cloud data warehouse traffic and that the intensity of storage load mainly depends on the incoming and outgoing traffic. Sufficiently high value of Hurst parameter indicates the potential for modelling and prediction of congestion cloud data storage in the long run.

1. INTRODUCTION

During the processing and storage of information we can observe a need to exchange data between the players of the process. At the end of 70s the rapid development of computer networks and corresponding network equipment began. Local and global networks continue to evolve, new data transfer protocols appear, the hardware capabilities of network equipment expand the number of connected users, and the total volume of traffic increase.

Existing modern cloud data storages by its internal structure do not use its potential capabilities functionality to the fullest. One of the reasons is complexity in behavior of network traffic, both within in and outside the cloud that affects the quality of service.

In terms of current trends in telecommunications and not only cloud data warehousing the topical task is to build a converged multi-service network. Such a network ought to provide an unlimited range of services to provide flexibility for the management and creation of new services. The latter requires the implementation of universal transport network with distributed switching,

[*] Lviv Polytechnic National University, Ukraine, Ternopil, Monastyrskogo str., 42/4, +380961875205, e-mail: strubytsky@gmail.com

where the interaction between devices and applications is done by creating of virtual connections. Intricacies of the management of which significantly affect the features of the stochastic dynamics of packet switching processes influence on its management to a great lengths.

On the other hand, intensive development of the industry entails a number of problems. One of them is the increasing number of consumers of information services together with the increasing demands for network and server equipment needed to maintain the proper level of service quality.

2. RELEVANCE OF THE TOPIC

Development of network equipment and transport protocols ought to be based on appropriate mathematical models and parameters of traffic simulation tools network processes. The nature of the network traffic is determined by several factors starting with the behavior of users or application software and finishing with the transmission protocols and equipment used. It goes without saying that macro-parameters of network traffic on a relatively large time intervals are defined by the man. However, the nature of the traffic at the intervals of microseconds is mainly determined by the order of transport protocols, network equipment and server software. Thus, the study of the basic characteristics of cloud server, such as the allocation of memory, CPU usage and process status of the operating system set against the intense network traffic is a topical task.

One of the most relevant problems of the study of cloud data warehouses temporal probability characteristics is consideration of the features of the network traffic. The aim of the study is to examine different models of network traffic and analyze the most promising models for cloud data warehouses, which takes into account the properties of self-similar traffic as a time series.

On the basis of real cloud data warehouse the dynamic characteristics of incoming and outgoing traffic as well as the distribution of hardware capacity of cloud server worked out. For all processes the self-similarity was defined. It confirms the applicability of fractal models to work with cloud data warehouse, particularly to predict the behavior of the servers of cloud data storage.

3. PURPOSE AND PROBLEM STATEMENT

The classical theory of remote traffic approaches are based on the assumption that the input streams are stationary Poisson flow type, that are superposition of a large number of independent stationary ordinary flow evenly without aftereffect of low intensity. For telephone networks with channel switching this assumption is valid. However, studies show that modern telecommunication

network traffic cloud data storage packet has a special structure that can not be used in the design of conventional methods that are based on Markov models and formulas of Erlang. This is a manifestation of the effect of self-similarity remote traffic is always present in the implementation of a certain amount of emissions sufficiently strong against the background of relatively low average. This phenomenon significantly affects the properties (increasing losses, delay and jitter) of the self-similar traffic passing through the network.

Until recently, the theoretical basis for the design of information distribution provided teletraffic theory which is a branch of queuing theory.

This theory sufficiently describes the processes occurring in such systems distribution information like telephone networks built according to the principle of switching channels. The most common call flow model (of data) in teletraffic theory is the simplest torrent (ordinary stationary stream without aftereffect), also known as a stationary Poisson flow.

The current state of rapid development of high technology has led to the emergence and spread of ubiquitous networks, packet data, which gradually began to force the system switching network, but still they were designed on the basis of the general provisions of the theory of teletraffic.

Thus “the teletraffic problem of self-similarity” formed, to which in recent years more than a thousand works were devoted, and that still has not lost its relevance. Despite the considerable popularity of the subject and a long period of active learning, we have to admit that there are still a lot of questions and unsolved problems.

The main ones are:

- virtually no theoretical framework that would come to replace the classical queuing theory in the design of modern information distribution systems with self-similar traffic, there is no single universally accepted model of self-similar traffic;
- there are no credible and recognized methods of calculation of the burst for a given flow, which corresponds to the ratio of peak intensity to the admission process service requests to its median;
- parameters and indicators of quality of information distribution subject to the influence of the effect of self-similarity;
- no algorithms and mechanisms providing quality of service in terms of self-similar traffic.

The aim is to determine the characteristics of fractal processes of different data streams in cloud data warehouses to make relevant decisions about how to manage them.

To achieve this goal the following problems have to be solved:

- to examine traffic cloud data repositories;
- to analyze combined data streams;
- to conclude that the dependence of the total flux of fractal self-similar properties of individual flows which it contains.

4. NETWORK TRAFFIC IN THE CLOUD MODELS REVIEW

Stochastic traffic models, which were widely used in the past [7], based on Markov-Term represented processes that have short-term dependence. Such models were described by Poisson distribution with variable-length messages according to law of exponent, and were based on queuing theory. These models were formed during the early networks of ARPANET. Simulation results of traffic based on theory of queuing were corresponding to time distribution of calls in telephone networks.

However, with time we can observe the works in which the tendencies of data gateway exchange and the total volume of traffic were indicated. Similarly, the nature of the traffic began to affect new data transfer protocols on the network, particularly in the internal environment of cloud storage data. Based on similar researches the concept of "chain of packets" was developed in 1986. In the model it is considered that the network packets are transmitted together but at the same time each package of the Poisson model is worked separately [10].

In recent years a model within which the volume of traffic has a long-term relationship has became popular. It was also found that traffic is self-similar and is characterized with heavy tails distribution [9].

Within the classical traffic model it is considered that the data sources work shifts [2]. That means that periods of high activity are changed by long delays. Thus, it was defined that an incoming message length and its time obey the exponential distribution, and the process of incoming messages of data sources is Poisson process. All the processes are stationary and independent.

Poisson model does not take into consideration that real network traffic has periods of strong bursts of activity. For classical model autocorrelation function tends to zero for large samples, the same time as the presence of bursts of activity in the real test traffic leads to positive autocorrelation.

Traffic patterns as a message thread was formulated and became popular in the 80s [5]. Within the model it is considered that the traffic packets are transmitted together and can be processed as a unit. Networking at each point in the network may decide to further processing chain for the first message. Such an algorithm would prevent the network from unnecessary operations analysis staff. However, it should be noted that this is the model of source messages.

The model can be applied only to messages that have the same destination. It is evident that the implementation of transport protocols and network equipment model for a message thread and classic models will be radically different.

In many modern works it is observed [6, 9] that combining traffic from multiple sources of variables leads to the fact that traffic becomes auto-correlated of long-term dependence [8]. This leads to the fact that the stability of correlation structures doesn't vanish not even for large values of lag. In other words, a set of a plurality of data sources that exhibit infinite variance syndrome, as a result provides integrated self-similar network traffic, which is close to the fractal Brownian motion. Moreover the studies of various traffic sources show that strongly variable behaviour is a property that is inherent to client/server architecture.

The problems of self-similarity of network traffic were exploited by many scientists. In particular the paper [14] studied the properties of actual traffic in networks with packet switching. Using of the R/S analysis shows self-similar nature of network traffic in information networks. Based on this approach, a model of traffic generator that implements the multifractal behaviour of data streams in real-world information systems which allows simulating traffic with given performance self-similarity has been developed. The work [16] shows that the standard 802.16b network traffic self-similar properties are shown at data link as well as transport layers. The values of the main exponents of fractal network traffic were obtained and the methods of aggregating the output statistics were suggested [15].

The work [6] is devoted to experimental removal of network traffic of a major Internet - providers, as well as the results of the analysis of structural features of the given traffic are provided. The authors have shown that self-similar properties manifest themselves as data link and transport layers. In paper [1], U.S. researchers studied the processes of long-term dependence. To generate such processes the authors propose the use of the fractal model integrated moving average.

Self-similarity is a property of the object the parts of which are similar to the whole object as unit. Many objects in nature have the following properties, e.g. coast, clouds, the circulatory system of a person or animal.

Informally self-similar (fractal) process can be defined as a stochastic process whose statistical characteristics exhibit scaling property. Self-similar process does not change significantly when considering species at different scales on the time scale. In contrast to processes that do not have fractal properties, there is no quick “smoothing” process at a scale averaging process saves time, that is the process is penchant for bursts.

Let $\{X_k; k = 0,1,2,...\}$ is a stationary random process. Taking into consideration the assumption of stationary and the existence and finiteness of the first two points, let us introduce the notation:

When averaging timeline let us understand the transition to process $\{X^{(m)}\}$, as such that

$$X_k^{(m)} = \frac{1}{m} \sum_{i=km-m+1}^{km} X_i \qquad (1)$$

where: $m = E[X_t]$ – the average value or expectation,

$\sigma^2 = E[X_t - m^2]$ – dispersion,

$R(k) = E[(X_{t+k} - m)(X_t - m)]_\infty$ – correlation function,

$r(k) = R(k)/R(0) = R(k)/\sigma^2$ – correlation coefficient.

When modelling network traffic value X_k is interpreted as the number of packets (less frequently as the total volume of data in bytes) that entered the channel or network for k-th time interval. The output process is thus already averaged. In some cases where there is a need to avoid this initial averaging either a point process or flow of events are considered, that is a sequence of points in receipt of individual packets in the network.

There is no single causal factor that causes self-similarity. Different correlations that exist in the self-similar network traffic and affect the different time scales can occur due to various reasons, manifesting themselves in characteristics to specific time scales.

The reasons for the long-term dependence in network traffic can be the following factors:

- user behaviour and application software;
- generation, structure and retrieval of data;
- combining of traffic;
- network administration tools;
- optimization mechanisms based on feedback;
- complexity of network structure, increasing the number of subscribers.

The process X is called self-similar with parameter $H = 1 - (\beta/2)$ if its autocorrelation coefficient is

$$r(k) = \frac{1}{2}[(k+1)^{2-\beta} - 2k^{2-\beta} + (k-1)^{2-\beta}] = g(k), k \in N \qquad (2)$$

where the function is:

$$g(k) = \frac{1}{2}\delta^2 k^{2-\beta} \qquad (3)$$

expressed by means of the central difference operator of 2nd order $\delta^2(f(x))$, which affects on the function $f(x) = x^{2-\beta}$ such that $\delta(f(x)) = = f(x+1/2) - f(x-1/2)$.

The essence of self-similarity is evident in the fact that for a process that satisfies the first condition, the equality $rm(k) = r(k)$ is valid, which means that this process does not change the coefficient of autocorrelation after averaging over blocks of any length m. Therefore, for the process of self-similar statistical characteristics of the second-order normalized aggregate process $X(m)$ is not different from the characteristics of the output process X with a significant range of change m.

The parameter H is an indicator of degree of self-similarity process and indicates that he has such properties as persistence/anti persistence and long term memory [4]. The parameter can accept values from 0 to 1. For white noise (Markov process) Hurst parameter is equal to 0.5, which means a complete lack of long-term or short-term dependence and the process is completely random, respectively, the simplest (Poisson) flow is called "the stream of pure randomness of the first kind".

If $H \in [0.5,1]$ the process is persistent, has a long-term dependency [3]: if for some time in the past, there has been an increase in process parameters, then in the future there will be the average growth. In other words, the probability that at step $k+1$ the process deviates from the average in the same direction, as at \$k\$ step is as large as the parameter H is close to 1.

If $H \in [0,0.5]$ process inherent anti persistence, it has short-term dependence [4]: high values of the process are followed by low and vice versa. That is, the probability that at step $k+1$ the process deviates from the average in the opposite direction (relative deviation k step) is as large as the parameter H is close to 0.

Long-term dependence is causing sharply pronounced fluctuation process, but gives the opportunity to discuss some predictability within narrow limits of time. From the point of view of the theory of queues, an important consequence of correlation flow is unacceptability of parameter estimation queues that are based on forecasts of identical and independent distribution of intervals in the input stream.

5. PRACTICAL RESEARCH OF NETWORK TRAFFIC OF DATA STORAGE

In order to confirm the existence of self-similarity properties of the different data streams of multiservice network, it is necessary to measure some of the characteristics of different types of network traffic. This requires statistics on traffic flows and data, and a study of the combined flow and variables cloud data storage should be conducted.

For studying purposes the cloud data storage was used. The physical server is divided into multiple virtual areas using the Solaris operating system, each of which is used to perform a number of tasks. Most of the traffic is transmitted by HTTP/HTTPS, FTP/FTPS and SFTP protocols.

For remote monitoring of data warehouse parameters in real-time application Zabbix is used [13]. Zabbix is a client-server application used to collect, store and process information about the network status, network load, and the state of the operating system of the data warehouse server in real-time.

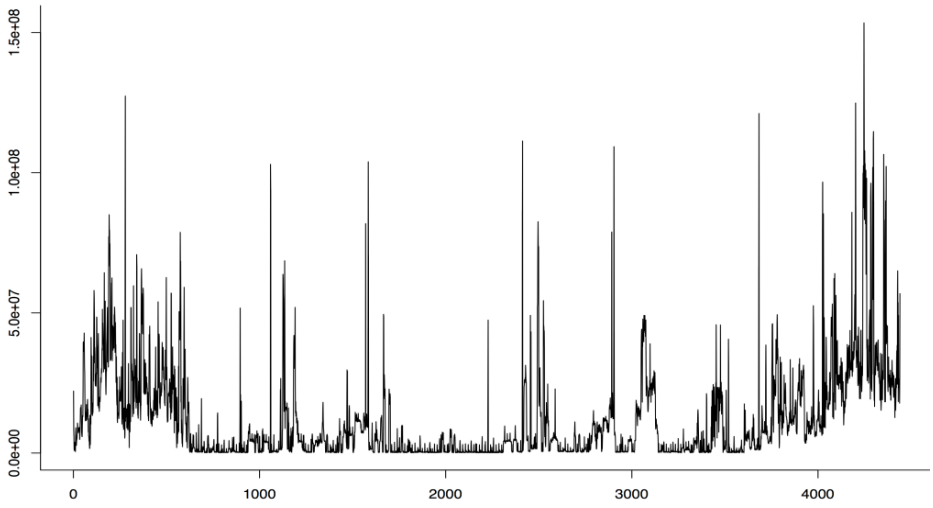
For further processing the following parameters of a data warehouse were used:

- incoming / outgoing traffic;
- number of running processes;
- load and idle processors;
- the average load on the processor;
- the amount of cache.

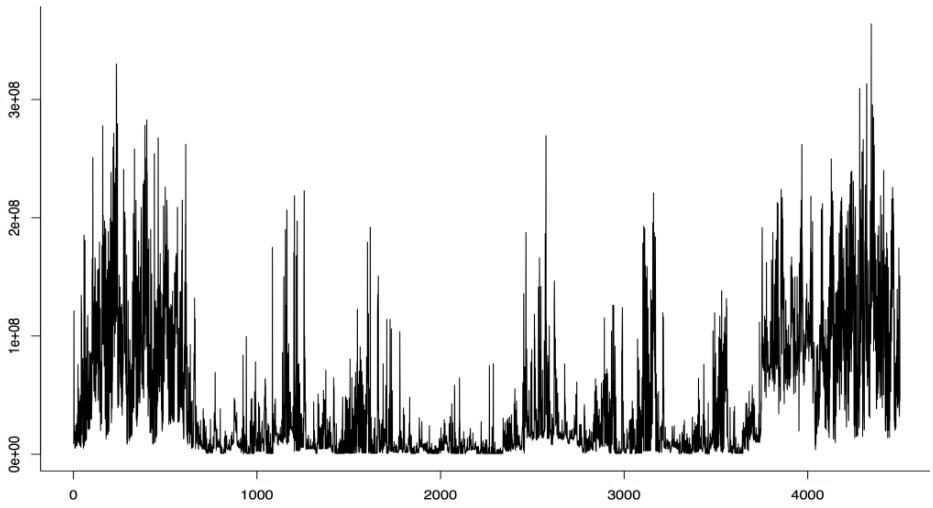
The received data is consolidated during the week, so we can assume that they represent the real picture of cloud data storage usage. The time dependence of the volume of traffic is shown in Figure 1 for input (a) and output (b).

Similarly to the data graph of the incoming and outgoing traffic (Figure 1), the other dependencies of other dynamics repeat. This is due to the fact that the main function of a cloud data warehouse is input, storage and output of users' data. Therefore the main load on the servers of data warehousing represents incoming and outgoing traffic. Thus the analysis and modeling of cloud storage can be limited to traffic patterns only or congestion model of processor.

Graphical representation of the autocorrelation coefficient allows visual verifying that the analyzed traffic has a long-term dependency.



a)



b)

Fig. 1. The time dependence of the incoming (a) and outgoing (b) traffic [source: own study]

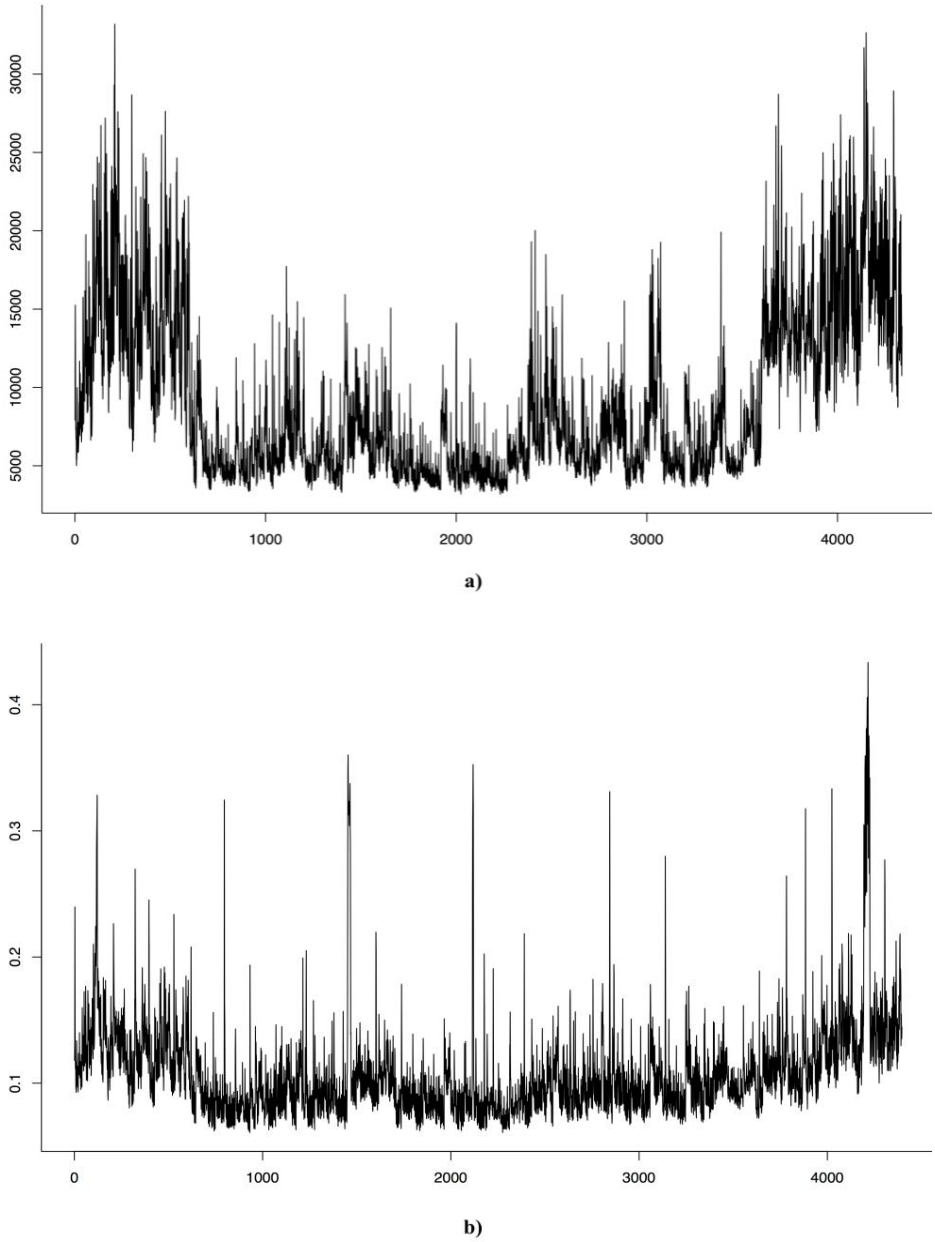
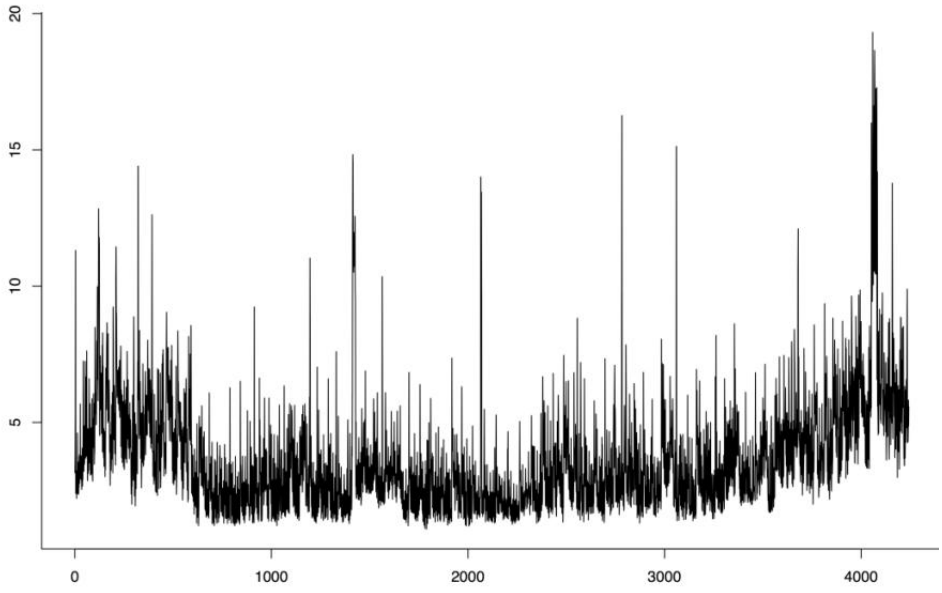
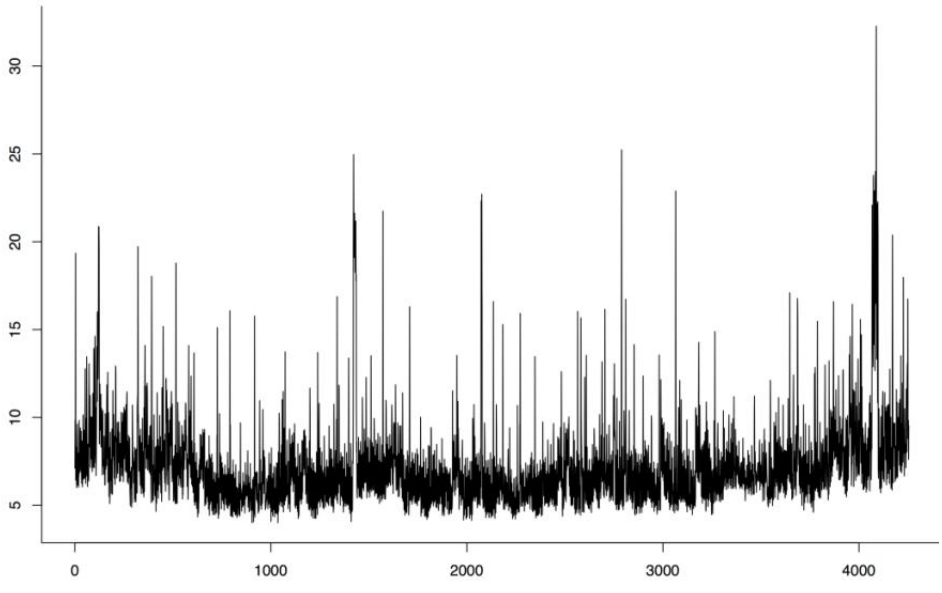


Fig. 2. Time dependence of switching (a) and load (b) of processors [source: own study]



a)



b)

Fig. 3. Time dependence of system CPU (a) to user (b) usage [source: own study]

In Figure 4 the graph of correlation coefficient for the process that matches incoming traffic on a logarithmic scale is given. It is evident that the points on the picture as a whole are grouped around a straight line whose slope can be determined by linear regression.

If the process is self-similar, then according to (2) slope coefficient $\beta = 2(H-1)$. If the resulting value $\beta = -0.2125$ Hurst parameter was found to be 0.8937.

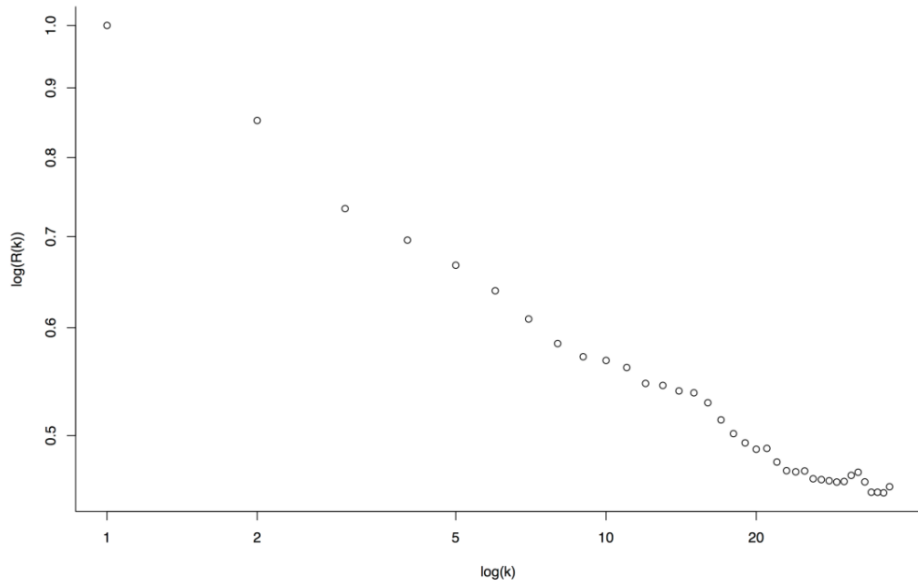


Fig. 4. The logarithmic dependence of the autocorrelation coefficient of incoming traffic [source: own study]

For outgoing traffic the graph of correlation coefficient – Figure 5, as well as for incoming traffic points are generally grouped around a straight line whose slope can be determined by linear regression. For a self-similar process slope coefficient $\beta = -0.1986$ Hurst parameter was at 0.9007.

Traditionally, self-similarity in any stochastic process is revealed through the definition of Hurst parameter H. The fact that $0.5 < H < 1$, thus Hurst parameter value is other than 0.5 is considered a sufficient basis for the recognition of self-similar process. It should be noted that the value of H, which is approaching to 1 could mean that the process is deterministic, that is not accidental: for some strictly deterministic processes structure is strictly repeated on any scale, leading to Hurst parameter of 1.

Observing the time dependence of traffic the presence of a periodic component in it that also leads to a large value of the Hurst was noted. The proximity of Hurst parameter to 1 allows performing more accurate predictions. In order to investigate the traffic details one needs to consider the temporal dependence separately within the same day.

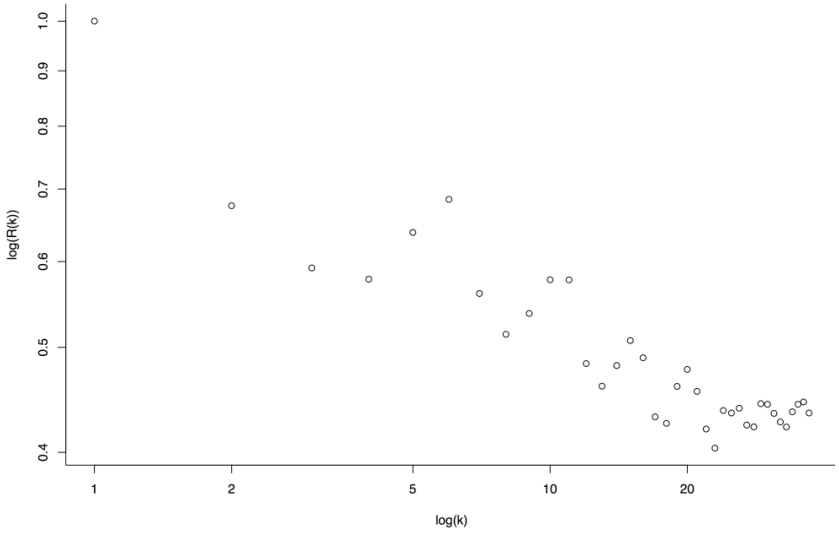


Fig. 5. The logarithmic dependence of the autocorrelation coefficient of outgoing traffic [source: own study]

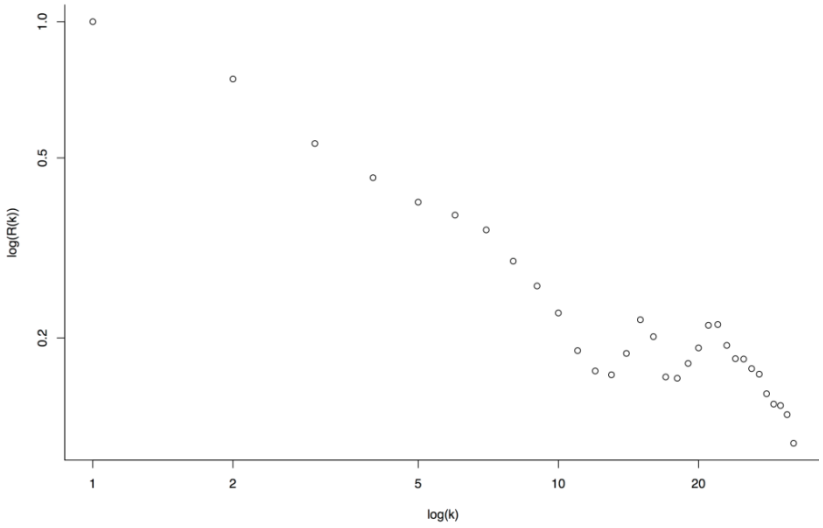


Fig. 6. The logarithmic dependence of autocorrelation coefficient for one-day traffic [source: own study]

Excluding the daily periodic component of traffic analysis and separating it within one day (Figure 6) shows a bit smaller Hurst parameter value, it ranges from 0.70 – 0.87.

5. CONCLUSIONS

As a result of this study the frequency of traffic cloud data warehouse, which has daily character was found. Intensity of storage load mainly depends on the incoming and outgoing traffic. Sufficiently high value of Hurst parameter indicates potential possibility of modelling and forecasting workload cloud data storage in the long term.

REFERENCES

1. HARMANTZIA F. C., HATZINAKOS D.: *Heavy Network Traffic Modeling and Simulation using Stable FARIMA Processes*. IEEE Trans. Signal Proc. Lett., 2000, pp. 48-50.
2. HEYMAN D. P., SOBEL M. J.: *Stochastic Models in Operations Research: Stochastic optimization*. Dover Publications, 2003.
3. HURST H.: *Transaction of the American society of civil*. Long term storage capacity of reservoirs, New York, 1951, pp. 770-799.
4. HURST H. E., BLACK R. P., SIMAIKA Y. M.: *Long-term storage: an experimental study*. Constable, 1965.
5. JAIN R., ROUTHIER S.: *Packet Trains--Measurements and a New Model for Computer Network Traffic*. IEEE J. Sel. A. Commun., vol. 4, 2006, pp. 986-995.
6. LELAND W. E. et al.: *On the self-similar nature of Ethernet traffic (extended version)*. Volume. IEEE Press, Piscataway, NJ, USA, 1994, pp. 1-15.
7. BASKETT F. et al.: *Open, Closed, and Mixed Networks of Queues with Different Classes of Customers*. ACM, vol. 22, New York, NY, USA, 1975, P. 248-260.
8. KHAYARI R. et al.: *The Pseudo-self-similar Traffic Model: Application and Validation*. Elsevier Science Publishers B. V., vol. 56, Amsterdam, The Netherlands, 2004, pp. 3-22.
9. WILLINGER W. et al.: *Self-similarity Through High-variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level*. IEEE Press, vol. 5, Piscataway, USA, 1997, pp. 71-86.
10. SONG CH.: *Packet Train Model: Optimizing Network Data Transfer Performance*. The University of Wisconsin – Madison, 1989.
11. STALLINGS W.: *High-speed networks and internets: performance and quality of service*. Pearson Education, 2002.
12. TANENBAUM A. S., WETHERALL D. J.: *Computer Networks*. Prentice Hall, 2011.
13. VACCHE A.D., LEE S.K.: *Mastering Zabbix*. http://books.google.co.uk/books?id=d1ZwAgAAQBAJ, 2013.
14. БЕССАРАБ В.І., ІГНАТЕНКО Е.Г., ЧЕРІВНСЬКИЙ В.В.: *Генератор самоподібного трафіку для моделей інформаційних мереж*. ДонНТУ, vol. 15, no. 130, Донецьк, 2008, pp. 23-29.
15. БЄЛЬКОВ Д.В.: *Дослідження мережевого трафіку*. ДонНТУ, vol. 10, no. 153, Донецьк, 2009, pp. 212-215.
16. ПЛАТОВ В.В., ПЕТРОВ В.В.: *Дослідження самоподібної структури телетрафіку бездротової*. vol. 3, 2004, pp. 38-49.