ROMAI J., **3**, 2(2007), 161–175

ASPECTS OF ETHERNET TRAFFIC ANALYSIS

Monica Iacob, Corina Săraru

University of Piteşti

monicaiacob11@yahoo.com,corina_sararu@yahoo.com

Abstract As the need of communication increases, traffic analysis takes a more and more important role in the concern of specialists. Traffic behavior is usually described using traffic monitoring.

This paper has two main parts. The first part describes the theoretical foundations of network traffic modelling and reviews different kinds of mathematical approaches which are useful for modelling and simulating the behavior of network traffic. Models that try to improve the Poisson based models are presented here. The self-similar nature of the traffic data is also emphasized through the statistical models presented.

The second part contains an analysis made on captured ethernet traffic from several points of view. During the traffic monitoring there are some particular aspects that are important for this study: the type of each packet, the number of packets that are captured at fixed intervals of time, the dimension of each group of packets in that interval. The data is then analyzed in order to obtain several features and possible anomalies in the network. Statistical tools are used to analyze the behavior of the traffic.

1. INTRODUCTION

Packets are blocks of data sent through a computer network. Besides the specific message, which may be represented by data, each packet can contain, controls for the management of the connection, or a demand addressed to the service and other additional information like the sender, the receiver and information about the control errors that may occur. The data packets may have a fixed or variable length and can be rearranged, if necessary, at the destination. The format of a packet depends on the protocol that creates it.

Protocol is a term used in networking and communication and it defines the procedures executed for transmitting and receiving data. It is the standard that controls and makes possible the connection, the communication and the data transfer between two computers. Protocols can be implemented at hardware or software level, and also as a combination between the both [8].

The mostly used protocols are:

1 transmission Control Protocol (TCP). It is the protocol from the TCP/IP model that manages the fragmentation of the data messages in packets which will be sent via the IP (Internet Protocol), and also the reassembly and verification of the correctness of the messages received. TCP adds a heading to the information that is to be sent. The source and destination ports allow data to be sent backwards to the right process that is being executed on each computer;

2 another important protocol in the TCP/IP model is the Internet Protocol (IP) which is data oriented. Even if it does not ensure the security of the data and it does not check if it reached destination, the protocol makes sure that the heading of the message does not contain any errors;

3 user Datagram Protocol (UDP) is a protocol used in the TCP/IP protocol suite. The UDP converts the messages generated by an application in packets which are then being sent via IP, but it is not very reliable because it does not ensure the communication path between the source and the destination and it does not verify that the message was transmitted without errors. Multicast applications like Mbone, that provide audio and video streams, use UDP as a transmitting mechanism because the retransmitting services offered by TCP are not necessary;

4 address Resolution Protocol (ARP) is used to determine the physical address of a node from a network connected to the internet, when only the IP address (or the logical address) is known. When a node needs to send a packet, it is first checked that the information from the physical

address is already sent. If this is the case, then that address is used and the traffic in the network is therefore reduced;

5 Real-Time Transport Protocol (RTP) is a standard transport protocol on internet, used for data transfer in real time. RTP uses unicast services as well as multicast ones. It was created to be used in virtual conferences and it uses UDP as a transmitting mechanism;

6 Real-Time Control Protocol (RTCP) is designed for scalable transport that uses RTP for monitoring transmissions in real time between more computers. RTCP sends at regular time intervals information control packets and is used to determine how efficient the data is sent and received;

7 Hypertext Transfer Protocol (HTTP) handles the transportation of web requests from the client to the web servers and the receiving of the web pages requested by the client browsers;

8 the transfer of the files over a network that uses TCP/IP, such as Internet is done via File Transport Protocol (FTP). It handles the possibility of copying files in remote system. FTP uses a client/server model in which a program is being executed on the receiving computer and accesses an FTP server which runs on the host.

The modelling of the network traffic involves the selection of the characteristics to be modelled while taking into account the relations between them.

Moreover, in order to provide a good simulation for the real situation, it is necessary that the model built be close to the reality.

The construction of a model involves different levels of characterization. The data stream is usually described by a sequence of observations which can be structured in the form

$$\ldots, X(t_n), X(t_{n+1}), X(t_{n+2}), \ldots$$

at time moments

$$\ldots, t_n, t_{n+1}, t_{n+2}, \ldots$$

These observations may describe the traffic from various points of view. Generally, $X(t_i)$ is modelled by a family of random variables with a known distribution. Considering the time at which the measurements are being made, the development of the model, can be made under various criteria

1 the nature of the time series:

   (a) if the sequence t_n is considered to be finite or numerable, then the described process, S, will be called a discrete time process ($S = \{X_n\}_{n=0}^{\infty}$);

   (b) if the sequence t_n is considered to be infinite, then the described process, S, will be called a continuous time process ($S = \{X_t\}_{t=0}^{\infty}$);

2 the way in which the arrival of the packets to a destination are considered (or of the entities with which one works):

   (a) if the process is a punctual one (i.e. at a time moment T a finite number of packets reach destination), then the model will be build taking into account a sequence of constants, $T_0 = 0, T_1, \ldots, T_n, \ldots$;

   (b) if one considers that the process can be written like this: $\{N(t)\}_{t=0}^{\infty}$ and that the arrivals are contained in time intervals like: $(0, t]$, then the stochastic processes $N(t)$ that compose the initial process may be expressed in this way: $N(t) = \max\{n : T_n \leq t\}$;

   (c) if, instead of working with time intervals, one will use the time between two arrivals, then the process can be written like this:$\{A_n\}_{n=1}^{\infty}$, where, for the sequence $T_0 = 0, T_1, \ldots, T_n, \ldots$, the relation $A_n = T_n - T_{n-1}$ is true.

For a network traffic many types of models have been proposed. Some of them, though, are not complex enough to describe in a convenient manner the existent processes.

For example, in the twentieth century, the Poisson distribution was used to model telephony traffic. Initially, it was thought that it would be appropriate also for network traffic.

The modelling involved the following aspects:

- arrivals of data at certain time moments are described. For an interval A_n one builds the relation

$$P\{A_n \leq \tau\} = 1 - e^{-\lambda\tau},$$

  where the average arrival rate (the average number of arrivals in a time unit) is λ;

- for an interval A_n of length τ, the arrivals number is described by

$$P\{N(\tau) = n\} = \frac{(\lambda\tau)^n e^{-\lambda\tau}}{n!}.$$

Working with a Poisson model assumes the identification of the average arrival rate, λ.

The matter that determined the orientation towards other types of models was the fact that, in a network, the time intervals between arrivals are represented by correlated random variables and the use of a Poisson model would not be useful in this case. The Markov modelling assumes that the dependencies among the variables $X(t)$ are described by Markov processes. Due to the fact that, generally, the processes depend on discrete states, this type of modelling will resume to using Markov chains. Markov Arrival Processes (MAP) generalize the Poisson processes by having a non exponential distribution at each state. A classification of special cases of MAPs may be this one:

1 phase type renewal processes; the time until the next arrival of an entity is given by the necessary time of a Markov chain to reach absorption;

2 MMPP (Markov modulated Poisson processes); the different states of the process are represented by Poisson processes. Each state will determine the average arrival state, λ_k;

3 Batch Markov arrival processes; they generalize Markov arrival processes by having an arrival time greater than one.

In 1993, though, the self-similarity nature of a network traffic was emphasized [1].

Definition 5. *The series* $X = \{X_n\}_{n=0}^{\infty}$ *can be m-agregated by means of* $X^{(m)} = \{X_k^{(m)}\}_{k=0}^{\infty}$ *by summing over a partition of elements of dimension m.*

Definition 6. X *is called H-self-similar if for* $\forall m > 0, X^{(m)}$ *has the same distribution like* X*, rescaled at dimension m.*

Thus,

$$X_n = m^{-H} \sum_{i=(n-1)m+1}^{nm} X_i = m^{-H} X^{(m)}, \forall m \in N,$$

where H is the Hurst parameter. This parameter has values ranging between 0 and 1. For characterizing self-similarity by means of it, values between 0.5 and 1 are enough. If $H = 0.5$, then there is no self-similarity, while $H = 1$ indicates perfect self-similarity. In practice, the Hurst parameter may be estimated by different methods. Another instrument used for self-similarity studies is the self-correlation function.

The cross-correlation is a measure of two signals. The characteristics of one of them (unknown signal) are being estimated by comparing it with the other (known) signal.

Self-correlation is a cross-correlation of one entity with itself.

Some of the existing models are:

1 Fractional Brownian Motion (FBm). This is a Gaussian process of mean zero, denoted by $B_H(t)$, that can be described by the following properties:

(a) i $E[B_H(t)] = 0$;
ii $B_H(0) = 0$;
iii $B_H(t+\delta) - B_H(t)$ is distributed $N(0, \sigma\,|\delta|^H)$;
iv $B_H(t)$ has independent increments;
v $E[B_H(t)B_H(s)] = \sigma^2/2(|t|^{2H} + |s|^{2H} - |t-s|^{2H})$.

Remarks:

(a) in addition, the Hurst parameter can offer information about the correlation of the increments of the FBm process.
If $H = \frac{1}{2}$ then the process is actually a Brownian motion.
If $H > \frac{1}{2}$ then the increments of the process are positively correlated.
If $H < \frac{1}{2}$ then the increments are negatively correlated;

(b) i $B_H(t)$ is a self-similar process if $B_H(at) = |a|^H B_H(t)$.
ii $B_H(t)$ is a self-similar process because from the point of view of the distribution, the above relation holds [9].
iii A stochastic process $\{X(t)|t \in T\}, X(t) \in R$, with T linear ordered, has independent increments if $\forall a, b, c, d \in T$, such that $a < b < c < d \Rightarrow X(b) - X(a)$ and $X(c) - X(d)$ are independent variables;

2 Fractional Gaussian Noise (FGn). The increments of a FBm are known as Fractional Gaussian Noise. This type of model, in which a stationary process, called $G_H(t)$, is considered, is described by the following properties:

(a) $G_H(t) = \frac{1}{\delta}\left(B_H(t+\delta) - B_H(t)\right)$;

(b) $G_H(t)$ is distributed $N(0, \sigma\,|\delta|^{H-1})$;

(c) $E[G_H(t+\tau)G_H(t)] = \sigma^2 H(2H-1)\,|\tau|^{2H-2}$, for $\tau >> \delta$.

For the same process in discrete time, one may use the following self-correlation function

$$\rho_X(k) = \frac{1}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right), k \geq 1.$$

Other interesting types of models that emphasize the self-similarity of the traffic are: ARFIMA (Fractional AutoRegressive Integrated Moving Average), wavelet models, Poisson-Zeta processes, $M/G/\infty$ models, self-similar Markov models, aggregation models.

2. MODELLING PROCESS

The data was captured, using Iptraf under Linux operating system, during one week and some important aspects for the traffic modelling were monitored: the packet type (it is relevant to know to which protocol does a certain packet belongs), the number of packets that are captured in one period of time, the dimension of this number of packets and the time period mentioned before. The packets captured belong to the following protocols: IP, TCP and ICMP.

For making a statistical analyze we used SPSS, a computer statistics program. The two plots from the figs. 1, 2 provide a visual representation of the model-estimated mean of the number of packets, and to the dimension of each group of packets belonging to each protocol for each period of time respectively.

In order to model the data it is very important to find the regression function. This function describes the relationship between the dependent variable Y and the explanatory variable(s) X.

In this case the model will define the relationship between the number of packets and the period of time. Only the IP protocol data will be used. Using

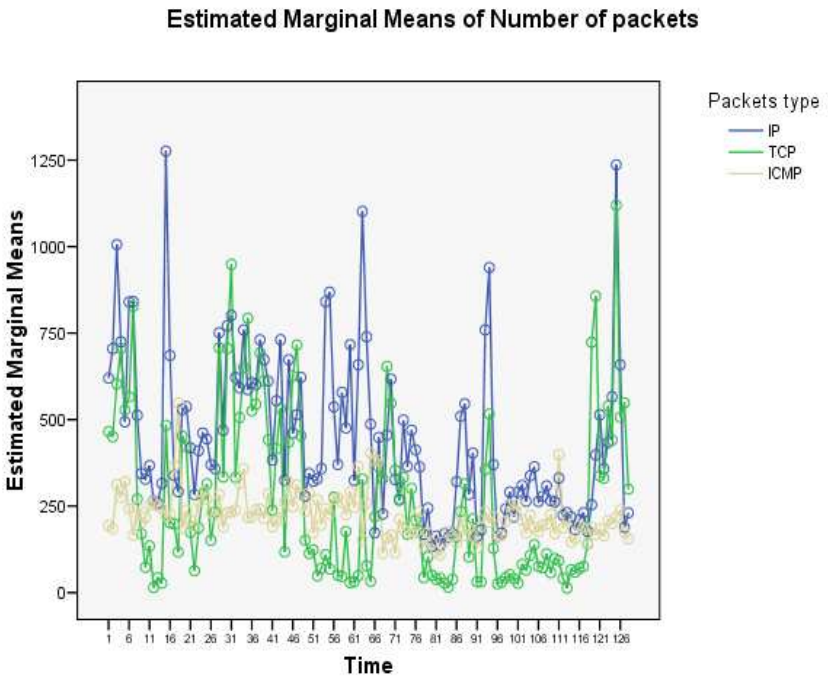

Fig. 1. Estimated marginal means of number of packets.

this type of data, a choice of getting a model is to consider the time to be the explanatory variable and the number of packets that are captured in one period of time for the dependent variable.

In order to get the right function of the model, we will use the histogram of the data.

From fig. 3 it can be seen that the regression function is not linear. The frequency is very high at the beginning, but then it decreases.

In order to see exactly the data behavior a plot of the number of packets on the unit of time will be used. In the fig. 1 it can be seen that as the time is passing the number of packets captured is decreasing, but after a short

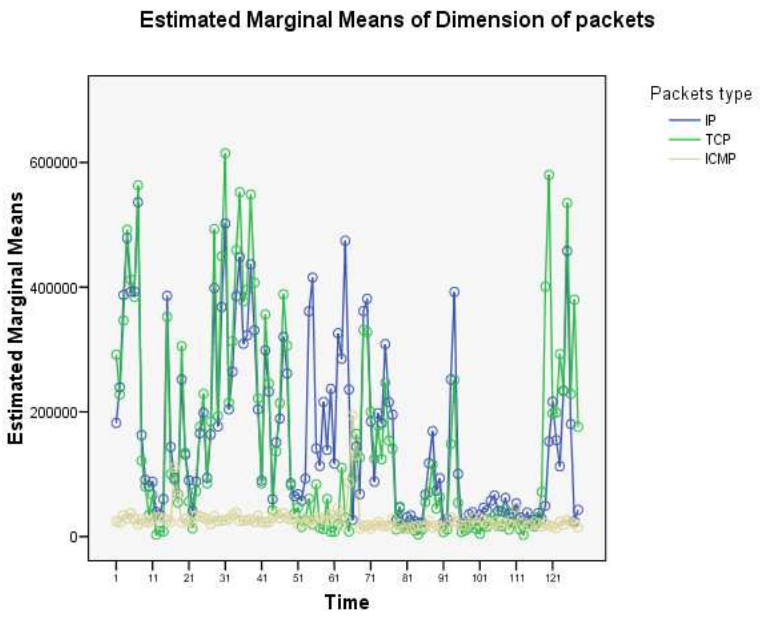

Fig. 2. Estimated marginal means of dimension of packets.

stabilization it increases again. An appropriate model for this kind of pattern is the asymptotical regression model

$$y = b_1 + b_2 \cdot e^{b_3 \cdot x},$$

where $b_1 > 0, b_2 > 0$, $b_3 < 0$.

The Nonlinear Regression Procedure requires some starting values for all parameters in the regression function:

- b_1 represents the lower asymptote for the number of packets. The lowest value of it is 0, so that is a reasonable starting value;
- b_2 is the difference between the value of y when x=400 and the lower asymptote. A reasonable starting value is the maximum value of y minus b_1. For this a good starting value is 2800;

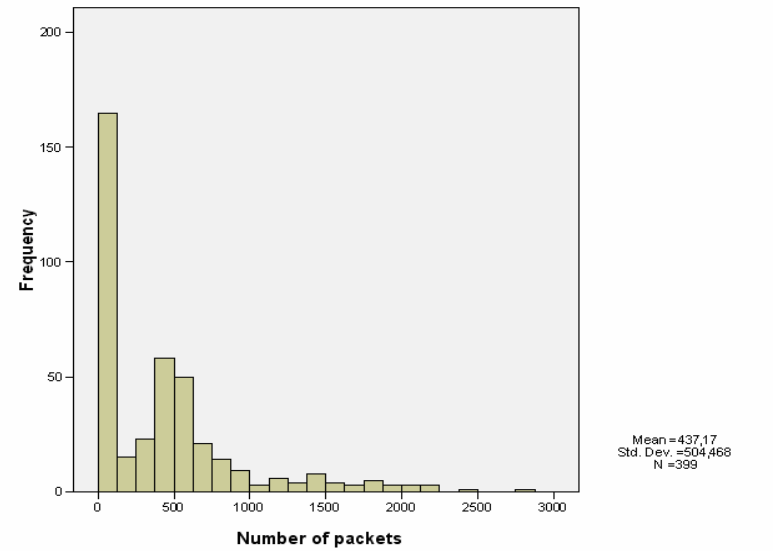

Fig. 3. The histogram of number of packets.

- initially b_3 can be roughly estimated by the slope between two "well separated" points on the plot. Looking at the chart there are points about x=30, y=2800, and about x=380, y=0. The slope between these points is $\frac{2800-0}{30-380} = -8$, thus a rough initial estimate for b_3 is -8.

From fig. 4 we can get the estimation of the parameters from the regression function:

- b_1 represents the minimum possible number of packets, even if infinite times were available. Its small standard error with respect to the value of the estimate suggests that the value is confident in the estimate;

- b_2 is the difference between the maximum and the minimum possible number of packets. Its standard error is large and confidence interval is wide compared to the value of the estimate, so there is some uncertainty here;

- b_3 controls the rate at which the maximum is reached, the so-called "rate constant". Like b_1 its small standard error with respect to the value of the estimate suggests that the value is confident in the estimate.

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1	,000	,665	-,050	,050
b2	1295,644	1,350	1147,507	1443,780
b3	-,007	,001	-,009	-,004

Fig. 4.

In fig. 5 the Uncorrected Total represents the entire variability in the dependent variable, while the Corrected Total is adjusted to only reflect variability about "average" number of packets.

ANOVA

Source	Sum of Squares	df	Mean Squares
Regression	1,2E+008	3	40715929
Residual	52149641	396	131691,0
Uncorrected Total	1,7E+008	399	
Corrected Total	98914057	398	

Dependent variable: Number of packets

Fig. 5. ANOVA table.

It is important to know how good the model is, so the plot of the residuals is useful for this. The scatterplot of the residuals is showed in fig. 6. It is

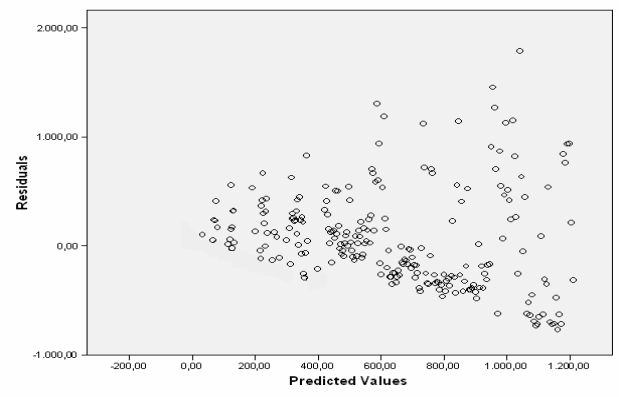

Fig. 6. Scatterplot of the residuals.

very easy to note that the residuals do not exhibit a pattern, so the asymptotic model is acceptable in the sense the residuals are independent of the fit values.

To emphasize this, the histogram of the residuals shows a normal probability distribution in the fig. 7.

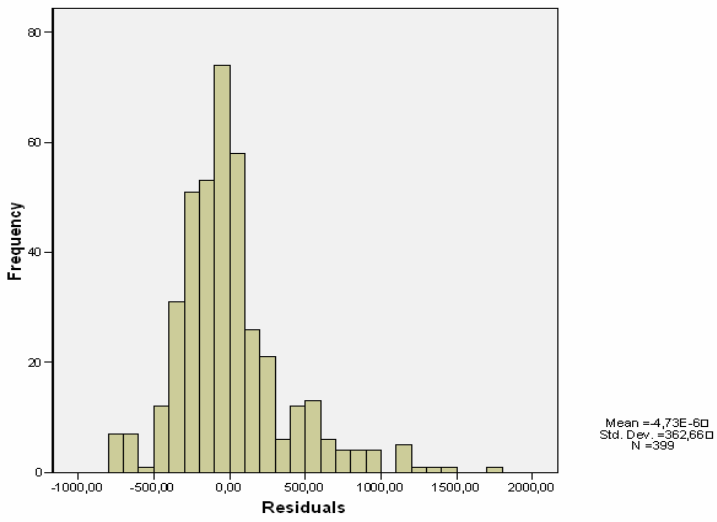

Fig. 7. Histogram of the residuals.

3. CONCLUSIONS

As the network traffic is quite a complex system to be modelled, various attempts have been made in order to describe its behavior.

Nevertheless, several properties and characteristics rose from the existent studies, such as self-similarity.

Our analysis attempts to provide a description of an ethernet traffic observed along one week, using Iptraf as monitoring tool. The data was used to derive a statistical description based on an asymptotic model.

The residuals have a normal probability distributions. This means that the model is suitable for the traffic data captured.

References

[1] Hlavacs, H., Kotsis, G., Steinkellner, C., *Traffic source modeling*, Technical Report No. TR-**99101**, 1999.

[2] Field, T., Harder, U., Harrison, P., *Analysis of network traffic in switched ethernet systems*, Performance Evaluation, **58** (2004), 243-260.

[3] Paxson, V., *Fast approximation of self-similar network traffic*, Computer Communications Review, **27** (1997), 5-18.

[4] Paxson, V., *Wide-area traffic: the failure of Poisson modeling*, IEEE/ACM Transactions on Networking, **3** (1995), 226-244.

[5] Poortinga, R., Van de Meent, R., Pras, A., *Analyzing campus traffic using the meter-MIB*, Proceedings of the Passive and Active Measurement workshop (2002), 192-201.

[6] Pras, A., Van der Meent, R.,Mandjes, M., *Gaussian traffic everywhere?*, Proceedings of the 2006 IEEE International Conference on Communications, (2006), 573-578.

[7] Leland, W.,Taqqu, M.,Willinger, W.,Wilson, D., *On the self-similar nature of ethernet traffic*(extended version), IEEE/ACM Transactions on Networking, **2** (1994), 1-15.

[8] http://en.wikipedia.org/wiki/Protocol_(computing)

[9] http://en.wikipedia.org/wiki/Fractional_Brownian_motion

[10] http://economics.about.com/od/economicsglossary/g/regressionf.htm

[11] http://www.itl.nist.gov/div898/handbook/pmd/pmd.htm