Statistical Analysis and Modeling of the Internet Traffic

Toni Janevski[1], Dusko Temkov[2], Aleksandar Tudjarov[3]

Abstract – **In this paper we show empirical data from Internet traffic measurements. Collected measurements are analyzed for different protocols, such as TCP and UDP. We perform statistical analysis through the correlation coefficients, covariance, and self-similarity degree i.e. Hurst parameter. Our experimental studies captured traffic with Hurst parameter around 0.7-0.75, which is near half way between values of 0.5 (it is not a self-similar) and 1 (strong self-similar properties). We use Maximum Likelihood approach to fit the obtained time series to existing distributions, such as Pareto and exponential distribution, where the first one is a self-similar process and the second is not. The analysis pointed out that Internet traffic with such values for the Hurst parameter could be modeled with similar accuracy using either distribution, Pareto and exponential.**

Keywords – **Internet, traffic, analysis, Pareto distribution, exponential distribution**

I. Introduction

Internet (IP) traffic is shown to be self-similar and bursty by nature [1,2]. Aggregate IP traffic is consisted of different traffic types, which are based on different protocols or applications. One may perform classification of the Internet traffic by analyzing different traffic types [1].

IP packets have varying length and they are generated with varying data rates, which is dependent upon the transport and application protocols, as well as link capacity. Experimental studies have shown that Internet traffic can have different characteristics than traditional voice traffic [3]. Further, some authors have shown that IP traffic properties are dependent upon the buffer size, because small buffers cannot capture self-similar behavior [4].

In this paper we present statistical analysis of measured traces for main traffic types in today's IP networks: TCP traffic, UDP traffic. Our main goal is targeted to statistical analysis and modeling of the IP traffic intensity. We capture traffic traces from a live network and then examine the self-similarity of IP traffic with so-called Hurst parameter and autocorrelation function. Also, we obtain probability distribution function (PDF) of the captured traffic for each of the Internet traffic types. Furthermore, we compare Pareto and exponential distributions for modeling the Internet traffic by fitting traffic types to each of them.

[1]Toni Janevski is with the Faculty of Electrical Engineering, University "Sv. Kiril i Metodij", Karpos 2 bb, 1000 Skopje, Macedonia, E-mail: tonij@cerera.etf.ukim.edu.mk.

[2]Dusko Temkov is with Macedonian Telecommunications, MT-net department, Orce Nikolov bb, 1000 Skopje, Macedonia, E-mail: duletem@mt.net.mk.

[3]Aleksandar Tudzarov is with Komercijalna Banka, KBnet department, 1000 Skopje, Macedonia, E-mail: tucko@kbnet.com.mk.

The paper is organized as follows. Section 2 gives the background. We present IP traffic measurements and modeling in Section 3. Finally, Section 4 concludes the paper.

II. Background

In this section we provide mathematical basis for the Internet traffic statistical analysis and modeling.

IP packets arrivals at a given network node, are described mathematically as point processes, consisting of arrivals at instants in time $T_0, T_1, ..., T_n$. A mathematically equivalent description is the interarrival time process $\{A_n\}_{n=0}^{\infty}$, where the continuous function $A_n = T_n - T_{n-1}$ is the time separating the n-th arrival from the previous one. If all An are identically and independently distributed, one gets a renewal process.

We will compare the measured traffic traces with two type of distributions: exponential distribution, which is widely used in teletraffic modeling of voice traffic; and Pareto distribution, which is almost standardized for modeling the self-similar packet traffic. In the following, we describe both distributions, as well as a notion of self-similarity.

Self-similar properties of a random process are slow-decay variance and long-tailed autocorrelation. By a definition, long-tailed distribution is a distribution which complementary cumulative distribution function has the following asymptotic behavior (regardless of its shape for small values of the random variable):

$$P(T \geq t) \sim t^{-\alpha} \quad \text{for} \quad t \to \infty, \text{ and } 0 < \alpha < 2. \qquad (1)$$

The Pareto CDF is a power curve. It is given by:

$$F(t) = \begin{cases} 1 - \left(\frac{t}{k}\right)^{-\alpha-1} & \text{if } t > k \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

The probability for $t < k$ is zero; k is the minimum value which can occur in a sample set. The probability density function (PDF) is given by:

$$f(t) = \begin{cases} \alpha k \left(\frac{t}{k}\right)^{-\alpha-1} & \text{for } t > k \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

A stationary time series $X = (Xt, t = 1, 2, 3, ...)$ is statistically exact second-order self-similar if it has the same autocorrelation function $r(k) = E[(X_t - \mu)(X_{t+k} - \mu)]$ as the series X^m for all m, where $X^m = (X_k^{(m)} : k = 1, 2, 3, ...)$ is the m-aggregated series obtained by summing the original series X over non-overlapping blocks of size m:

$$x_k^{(m)} = \frac{1}{m}(X_{km-m+1} + X_{km-m+2} + \cdots + X_{km}). \qquad (4)$$

A stationary time series X is statistically asymptotically second order self-similar if autocorrelation $r^{(m)}(k)$ of X^m, for large k, agrees asymptotically with the autocorrelation $r(k)$ of X. Self-similar traffic patterns can be detected by visual observation of traffic plots on different time scales. It looks similar over many time scales whereas short-range dependent time series look like noise after aggregating them. The degree of self-similarity can be defined using the so-called Hurst parameter H, which expresses the speed of decay of the autocorrelation function. The range of values is $0.5 < H < 1$. For $H \to 0.5$ the time series is short range dependent, while for $H \to 1$ the process becomes more and more self-similar. Since slow decaying variance and long range dependence (i.e. slow decaying autocorrelation functions) are both related to self-similarity, it is possible to determine the degree of self-similarity using either of these properties. In this work, we obtain the Hurst parameter of traffic traces by using the so-called variance-time plot, which relies on the slow decaying variance of every self-similar process. For a self-similar time series, the variance of an aggregated process decreases linearly (for large m) in log-log plots over m. The slope of the curve β can be estimated using a linear regression. Then, the Hurst parameter is determined by the following equation:

$$H = 1 - \frac{\beta}{2}. \tag{5}$$

On the other side, exponential distribution is defined with a single parameter λ, and its CDF is given by:

$$F(t) = 1 - e^{-\lambda t}. \tag{6}$$

III. IP Traffic Measurements and Modeling

For the purpose of the analysis we created traces with application CommView, scanning the Internet traffic on user Network interface toward Internet Service Provider (ISP) on client side. Actually the traces are made on Ethernet interface on local PC attached on network which has direct connection to Internet by using a leased line.

CommView text file is parsed and stored in Microsoft access base and after that it is converted in an SQL base for analysis purposes. All mathematical statistics and processing are made in Matlab. Figs. 1 to 3 illustrate the measured traffic intensity for aggregated traffic, TCP and UDP traffic. It can be noticed that TCP traffic has highest volume. This was expected due to the popularity of TCP-based applications today such as WWW. On the other side, UDP traffic intensity is respectively smaller then TCP, because in most cases it is due to DNS traffic. Each session is identified by unique source/destination address and source/destination port.

In practice we characterize statistical processes by their first two moments. Hence, we target our statistical analysis of IP traffic traces to autocorrelation function and variance. We obtain correlation coefficients from the traffic trace in the following manner: for a given measurement with N samples, $y_1, y_2, ..., y_N$, at time moments $x_1, x_2, ..., x_N$, the lag k correlation coefficient is defined as:

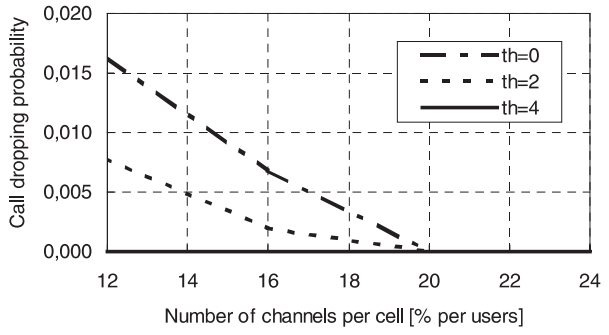

Fig. 1. Aggregated traffic trace

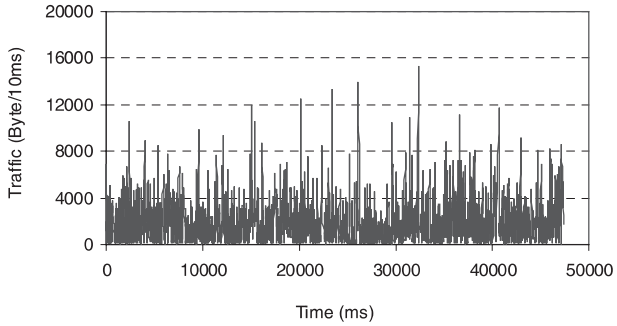

Fig. 2. TCP traffic trace

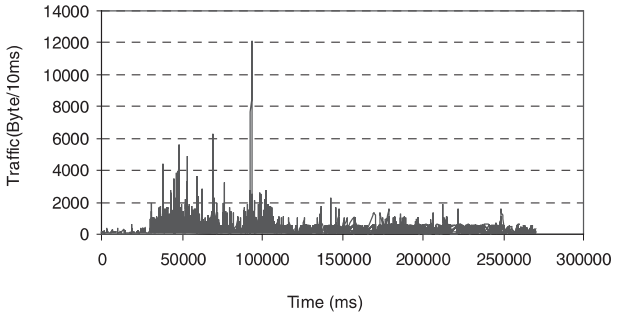

Fig. 3. UDP traffic trace

$$r_k = \frac{\sum_{i=1}^{N-k}(yi - \bar{y})(y_{i+k} - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{7}$$

Autocorrelation functions of the captured traffic types are shown on Figs. 4 to 6. One can observe that TCP and aggregated traffic traces have autocorrelation function which oscillates around zero, while UDP traffic experiences long-tailed autocorrelation. It leads to stronger self-similar properties of UDP traffic compared to the TCP traffic from the measurements.

We obtain the Hurst parameter from traffic traces by us-

Table 1. Hurst parameter and slope β

Traffic	β	H
Aggregate	0.632055	0.6840
TCP	0.533801	0.7331
UDP	0.465238	0.7674

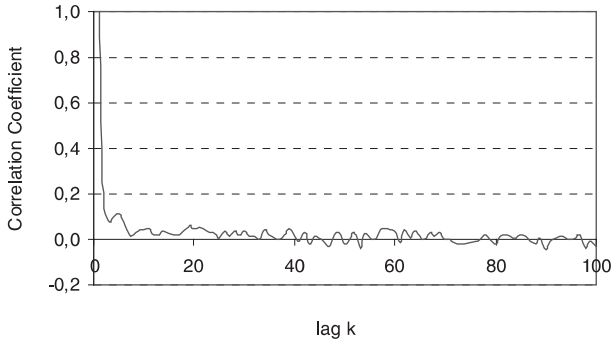

Fig. 4. Correlation coefficients for aggregated traffic

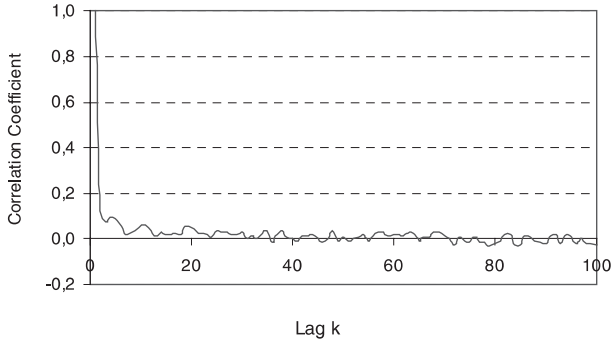

Fig. 5. Correlation coefficients for TCP traffic

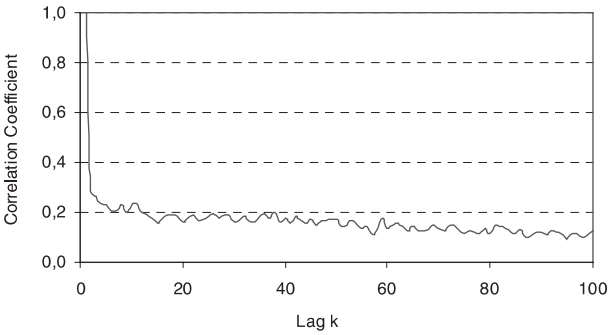

Fig. 6. Correlation coefficients for UDP traffic

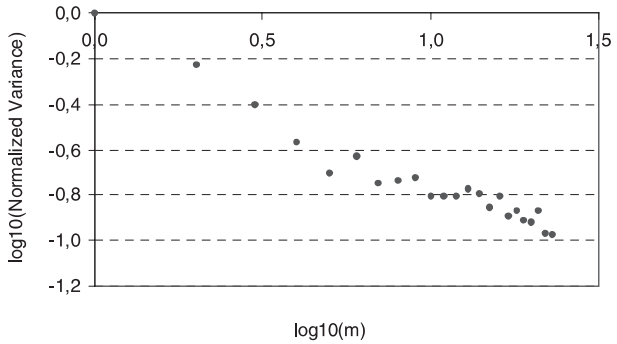

Fig. 7. Variance-Time Plot for aggregate traffic

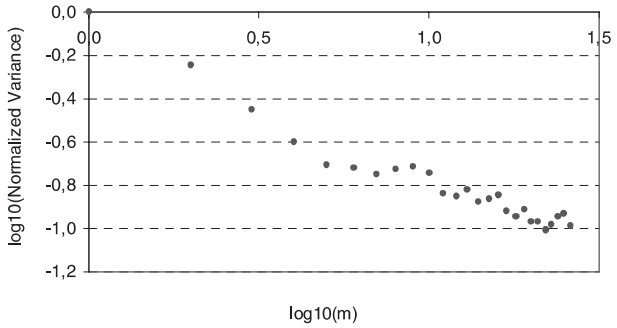

Fig. 8. Variance Time Plot for TCP traffic

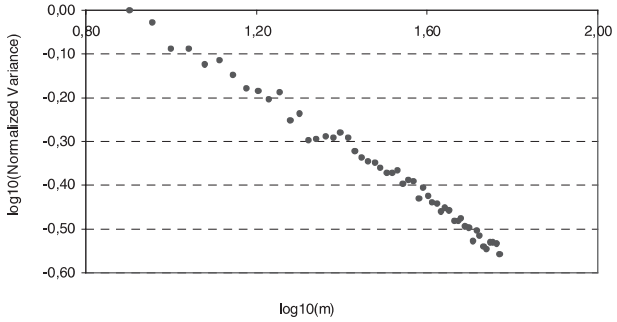

Fig. 9. Variance Time Plot for UDP traffic

ing variance time plot method. For a self-similar process, the variance of an aggregated process decreases linearly (for large m) in log-log plots over m. The slope β can be estimated using linear regression, leading to the Hurst parameter defined as above. Hurst parameter can be calculated from the slope β using Eq. (5).

Figs. 7 to 9 show variance-time plots for traffic traces, and the obtained parameters, β and H, are given in Table 1.

Further, our aim is to fit the first two moments of the measured Internet traffic traces with the two distributions. In other words, we need to obtain the optimal parameters of the Pareto distribution as well as exponential distribution that best model the measured traffic data.

We start with fitting the measured traces to Pareto distribution. With analysis of traffic traces (shown in Figs. 1-3) we obtain the value for α parameter of the Pareto distribution, using its relation with the Hurst parameter, Eq. (5). It results in $\alpha = 1.16$. Then, for $k = 60$ to 5000 we perform minimization of mean square error between the measured samples and Pareto distribution.

Further, we continue with fitting the measurements data to the exponential distribution. In this case, our aim is to find the optimal parameter λ of the exponential distribution that provides the best match between the exponential PDF and the measured traffic. For this purpose we search for $\lambda_{\min}$ that minimizes mean square error between the measured traffic PDF and the exponential PDF. These results are shown in Fig. 10-12 for aggregated, TCP and UDP traffic, respectively.

The comparison of fitting the measured data to exponential and Pareto distributions is given in Table 2. One can notice that mean square error is slightly slower for the exponential distribution than for Pareto. However, such result may be expected due to lower values of the Hurst parameter for

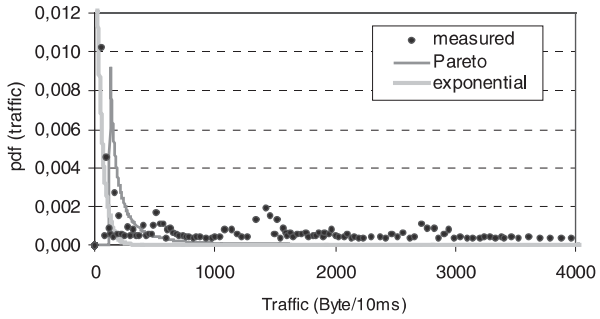

Fig. 10. Measured, Pareto and exponential PDF for aggregated traffic

Table 2. Mmean square error for EXP and PARETO

Traffic	$\lambda_{\min}$	MSE for EXP	$k_{\min}H$	MSE for Pareto
Aggregate	0.021	1.038E-05	1500	1.0998E-05
TCP	0.016	1.585E-05	1510	1.6753E-05
UDP	0.016	1.429E-04	550	1.4729E-04

the measured traffic ($H \approx 0.70$-0.75). So, when $H \rightarrow 1$ we should use Pareto for modeling bursty data traffic, while when $H \rightarrow 0.5$ then exponential distribution should be the best model. However, if we are in the middle of the range for the Hurst parameter, then according to our results presented in this paper we may equally choose either, exponential or Pareto distribution, for modeling the Internet traffic.

IV. Conclusions and Future Work

We performed statistical analysis of the captured Internet traffic from a real network. We analyzed the traffic per protocol, i.e. TCP and UDP, as well as aggregated traffic.

Then, we used obtained statistics to fit the measured traffic data to the exponential and to the Pareto distribution. The results showed that for traffic with Hurst parameter in the range 0.7-0.75 we can use each of the distributions for modeling the Internet traffic with equal accuracy.

Possible future extension to this work is to use a linear combination of the two distributions for modeling the Internet traffic.

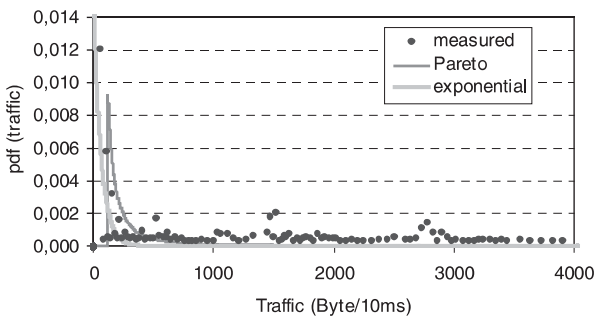

Fig. 11. Measured, Pareto and exponential PDF for TCP traffic

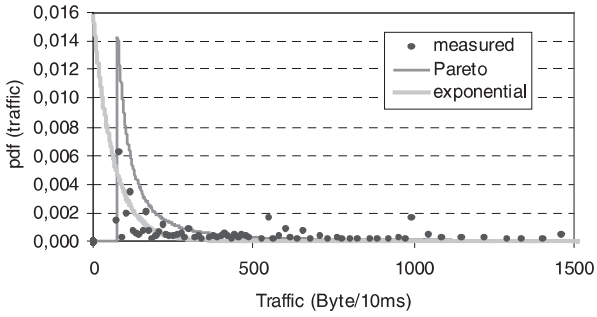

Fig. 12. Measured, Pareto and exponential PDF for UDP traffic

References

[1] Toni Janevski, *Traffic Analysis and Design of Wireless IP Networks*, Artech House Inc., 2003.

[2] Vern Paxson and Sally Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling", *IEEE/ACM Transactions on Networking*, June 1995, pp.226-244.

[3] Walter Willinger and Vern Paxson, "Where Mathematics meets the Internet", *Notes of the American Mathematical Society*, Vol.45, No.8, August 1998, pp.961-970.

[4] F. Huebner, D. Liu and J.M. Fernandez, "Queuing Performance Comparison of Traffic Models for Internet Traffic", *GLOBECOM'98*, Sydney, Australia, November 8-12, 1998, pp.1931-1936.