International Journal of Novel Research in Engineering and Applied
Sciences (IJNREAS) 1(1) February 2014
©Avi-D Publishers 2014
www.avidpublishers.ca

The Desirability of Pareto Distribution for Modeling Modern Internet Traffic Characteristics

Alakiri Harrison. O,
Department Computer Technology,
Yaba College of Technology,
Lagos, Nigeria.
halakiri@yahoo.com

Oladeji Florence. A,
Department Computer Science,
University of Lagos, Nigeria.
foladeji@unilag.edu.ng

Benjamin Benjamin. C,
Department Computer Science,
University of Jos, Nigeria.
benyx6@gmail.com

Okolie Cletus. C,
Department Computer Science,
University of Lagos, Nigeria.
okoliec@gmail.com

Okikiola Folasade.M
Department Computer Technology,
Yaba College of Technology,
Lagos, Nigeria.
sade.mercy@yahoo.com

ABSTRACT

The Internet is rapidly growing in number of users, traffic levels and topological complexities. A vital component to understand the requirements, its capabilities and extension of the network is its traffic analysis. In literature, conventional traffic model does not capture today's burstiness in Internet traffic with the assumption that inter-arrival time is independent and identically distributed. The Internet forwarding scheme has given provision for correlated inter-arrival times which are rampant when streams of packets arrive to a destination at the same time. Modern researchers are contemplating of modulating Poisson process with another distribution in form of Compound Poisson Process, Markov Modulated Poisson Process or use of Pareto model to tract actual traffic characteristics. Pareto is assumed in literature to have facility to cater for burstiness. This paper uses goodness of fit test to detect the exponential nature of the inter arrival times of Pareto generated network packets. A network simulator was used to generate statistics of arrival times of traffics were traced, inter-arrival times calculated, ranked and subjected to goodness of fit test. The analysed result fell into acceptable region which implies that Pareto can stand in place of conventional exponential distribution in modeling inter-arrival times of a bursty network.

Keywords: *Internet traffic, Poisson process, traffic burstiness, Exponential distribution and Goodness of Fit test*

1.0 Introduction

The Internet is rapidly growing in number of users, traffic levels and topological complexities since it has become a universal communication network for all kinds of information ranging from simple transfer of binary data to the transmission of voice, video, or interactive information in real time. Internet traffic is the flow of data coming from different users or applications and is routed towards the same or different destinations across the network. Due to the distributed nature of the Internet, there is no single point of measurement of its total traffic (Williamson, 2001). Even though, the Internet traffic data from public peering points can give an indication of Internet volume and growth, the figures exclude traffics that remain within a single service provider's network as well as traffic that crosses private peering points. Therefore, traffic analysis is a vital component to understand the requirements and capabilities of a network. Nevertheless, there is no single traffic model that can efficiently capture the traffic characteristics of all types of networks, under every possible circumstance (Adas, 1997). For example, the conventional traffic model (Poisson) does not capture today’s burstiness in Internet traffic (Paxson& Sally, 1995). Modern researchers are contemplating of using Pareto distribution in place of Poisson or combine other models with Poisson such as the Compound Poisson Process, Markov Modulated Poisson (Du, 1995) so as to study the inter-arrival times when packets arrive in bulk. Pareto is defined to assume correlated inter-arrival times when synonymous streams of packets are forwarded to a destination at the same time (Yang & Petropulu, 2001). This paper uses GOF (a choice of Chi-square) test to detect the exponential nature of Pareto generated inter-arrival time data. A simulation was carried out using network simulator to collect Pareto-controlled traffics and statistics of arrival times of traffics were traced. Inter-interval arrival times were calculated, ranked and subjected to goodness of fit test.

The simulation and analysis in this report has shown that Pareto described as a double exponential distribution can be used in place of Poisson when it comes to burstiness in items arrivals .Thus this studies involving burstiness in items arrival and corrected inter arrival time can be validated using Pareto distribution.

The paper is presented as follows: section one is the introduction, section two is related works on traffic models and in section three the experimentation is documented. Sections four and five present the discussion of result and conclusion respectively.

2.0Related Works

Traffic modeling is used in finding stochastic processes to represent the behavior of any network traffic, telephone or computer network. A study at Copenhagen Telephone Company in the twentieth century famously characterized telephone traffic at the call level by certain probability distributions for arrivals of new calls and their holding times (Erlang et al, 1948). The author applied the traffic models to estimate the telephone switch capacity needed to achieve a given call blocking probability. Teletraffic theory for packet networks has seen considerable progress in recent decades (Frost & Melamed, 1994),(Park & Willinger 2000), (Willinger & Paxson,1998). The first traffic model, based on Poisson processes, was born in the context of telephony, where call arrivals could be considered independent and identically distributed and "holding times" followed an exponential distribution (Jain & Routhier , 1986). Although initially successful and analytically simple, the Poisson model has proven not suitable to describe data traffic in modern Local Area Networks (LANs) and Wide Area Networks (WANs), where batch arrivals, event correlations and traffic burstiness are important factors. The stochastic process shows some limitations when applied to tract today's Internet traffic for example. The model is characterized by assuming that the packet arrivals A_n and inter-arrival time A_T have the following characteristics (Cao et al., 2002):

1. they are independent,
2. they are exponentially distributed with rate parameter λ,
3. $P\{A_T \leq t\} = 1\text{-}e^{-\lambda t}$.

Alternatively, this implies that the traffic is described through a counting process and the probability of having n arrivals in time t is described in the equation:

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

where *N(t)* is the number of arrivals at time *t*. Internet traffic can be modeled as a sequence of arrivals of discrete entities, such as packets or cells. Mathematically, this leads to the usage of two equivalent representations: counting processes and inter-arrival time processes. A counting process $\{N(t)\}_{t=0\ldots\infty}$ is a continuous-time, integer-valued stochastic process, where N(t) expresses the number of arrivals in the time interval (0,t). An inter-arrival time process is a non-negative random sequence $\{A_T\}$, where $A_T=T_n\text{-}T_{n\text{-}}$

$_1$indicates the length of the interval separating arrivals n-1 and n.

Traffic modeling comprises of three steps:

(i) selection of one or more models that may provide a good description of the traffic type

(ii) estimation of parameters for the selected models

(iii)statistical testing for election of one of the considered models and analysis of its suitability to describe the traffic type under consideration(Anderson &Nielsen, 1998). Parameter estimation is based on a set of statistics (e.g. mean, variance, density function or auto covariance function, multifractal characteristics) that are measured or calculated from observed data. In literature there are two major parameters generated by network traffic models: packet length distributions and packet inter arrival time distributions. Other parameters, such as routes, distribution of destinations, etc., are of less importance. Simulations that use traces generated by network traffic models usually examine a single node in the network, such as a router or switch; factors that depend on specific network topologies or routing information are specific to those topologies and simulations. Internet traffic is appealing as a problem hindering the administration of congestion and flow control quality of service and switch performance in modern network. In literature, Pareto is a double exponential, power law probability distribution that coincides with social, scientific, geophysical, actuarial, and many other types of observable phenomena. If $f_{(X)}$ is a random variable with a Pareto distribution, then the probability that X is greater than some number x, i.e. the survival function(also called tail function), is given by

$$\mathbf{f\ (x) =} \frac{a * b^{a}}{x^{(a+1)}} \textbf{ for } \boldsymbol{x \geq b}$$

and

$$\boldsymbol{E(X) =} \frac{b * a}{(a - 1)} \boldsymbol{\text{ for } a > 1,}$$

where ***b*** is the (necessarily positive) minimum possible value of ***b***, and ***a*** is a positive parameter. The distribution is characterized by a scale parameter ***b*** and a shape parameter ***a***, which is known as the tail index for Pareto Type1.

(Froot et al., 2008) applied Pareto distribution to severity distribution in a context of catastrophe reinsurance and can be used to model the income of an individual. It was noted that Pareto is good for modeling situations where items concerned are highly skewed e.g. the current Internet traffics are more of few chunks of real time applications packets and few chunks of large files. The exponential distribution is a family of continuous probability distributions. It describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. It is the continuous analogue of the geometric distribution. Its probability distribution function is given as:$f(t) = e^{-\lambda t}$, for t is > 0 and λ> 0. Its cumulative probability distribution function is$F(\mathrm{t}) = 1 - \mathrm{e}^{-\lambda \mathrm{x}}$.

The goodness of fit test of a statistical model describes how well a model fits a set of observation. Measuring goodness of fit typically summarizes the discrepancy between observed values and the values expected under the model in question. Goodness-of-Fit test (GOF) uses a randomly selected sample of observations to determine if the true distribution of the random variable is as hypothesised. Pearson in (Pearson, 1905) was an early user of this method. The author proposed the use of squared differences to overcome the cancelling of the positive and negative differences between the observed and expected frequencies. To give a meaningful interpretation to each squared difference, he used the expected frequency as a scaling factor. Partial contributions from each category were then added to provide a global test statistic known as the Chi-Square or Pearson's Chi-Square test statistic as defined as:

$$Tx^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei}$$

where: k is the number of categories, O_i and E_i are the observed and expected frequencies for category i respectively, $(1 \leq \ i \leq k)$.

3.0 Simulation Experiments

This paper attempts to study if traffics that are generated by Pareto process have exponential characteristics. The parameter used is the inter-arrival time which is modeled using a negative exponential continuous distribution. For the experimentation, ns 2 simulator was used. Traffics were generated using the Pareto traffic generating process embodied in the OTcl class Application/Traffic/Pareto. The network topological setting for this research work is given in the Figure 1below:

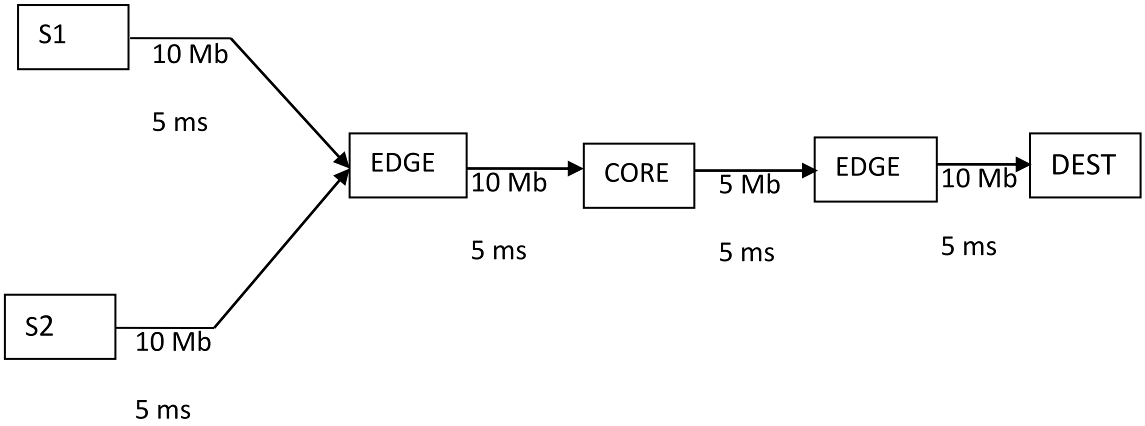

Fig. 1 Simulation topology diagram

In the topology, there are two sources S1 and S2, each generates packets randomly using Pareto model and sends to edge router EDGE1at 10MB per 5ms. The edge EDGE1 sends metered traffics to a core router CORE which forwards to egress router EDGE2 and finally to a destination DEST at the rate of 10Mb per 5ms.Drop tail buffer mechanism was employed at the edge routers and priority scheduling scheme was employ at the CORE for forwarding the packets. Traffic details were traced and analyzed in order to obtain the inter-arrival times that were used for the GOF test.

4.0 Discussion of Findings

During the simulation, events on each packet were traced as given in Table 1 below:

Table 1: **Sample of traced data**

Event	Time	from node	To node	Traffic generator	Pck size	Flags	fid	Scr Addr	dest addr	Seq Num	pkt id
+	0.12225	3	4	Pareto	1000	----------	0	0	5	0	0
-	0.12225	3	4	Pareto	1000	-----------	0	0	5	0	0
+	0.12425	3	4	Pareto	1000	-----------	0	0	5	1	1
-	0.12425	3	4	Pareto	1000	-----------	0	0	5	1	1
+	0.12625	3	4	Pareto	1000	----------	0	0	5	2	2
-	0.12625	3	4	Pareto	1000	----------	0	0	5	2	2
+	0.12825	3	4	Pareto	1000	---------	0	0	5	3	3
-	0.12825	3	4	Pareto	1000	-----------	0	0	5	3	3

R	0.12885	3	4	Pareto	1000	------------	0	0	5	0	0

From Table 1, for instance, Packet 2 arrived at time 0.12625ms and packet 3 arrived at 0.12825, the inter arrival time between packet 2 and 3 is 0.00200ms. This calculation was repeated for all the arrived packets and a sample was shown in Table 2.

Table 2: Sample of calculated Inter-arrival time

S/N	1	2	3	4	5	6	7	8	9	10
Interval Arrival Time	0.0012	0.0008	0.0012	0.0008	0.0012	0.0008	0.0012	0.0008	0.0012	0.0008

These inter-arrival times are subjected to goodness of fit test in order to study if the data fit into negative exponential function. Chi square was used for the test as shown with sample data in Table 3below:

Table 3 Packets subjected to GOF test

S/N	Fi Observed requency	Xi Data obtained	fx_i	Average Inter-arrival time	Probability Distribution function $F(x)$ Expected Frequency	$\sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei}$
1	90	0.00085	0.0765	0.00133	0.491704	122.9243
2	0	0.00105	0		0.070966	17.7393
3	90	0.00125	0.1125		0.061058	15.26191
4	0	0.00145	0		0.052533	13.13035
5	0	0.00165	0		0.045199	11.29633
6	0	0.00185	0		0.038888	9.71831
7	70	0.00205	0.1435		0.033459	8.360553
Total	250	0.01015	0.3325		0.793807	198.431053

Using the Chi-square goodness of fit test at 5% level of significance and k-r-1 degrees of freedom gives 11.07 while the calculated estimates in Table 3 gives 198.43. This shows that the data falls within the acceptable region. It is then concluded that Pareto which the literature defined as a double exponential distribution can be used instead of exponential distribution in modeling bursty arrivals in order to predicts the future characteristics of a QoS network.

5.0Conclusion

This report made use of ns-2.31 simulator to mimic a real life networking scenario to study an end to end traffic management and congestion control that will help in improving traffic on the Internet. The study attempted to study whether Pareto generated data has properties of exponential distribution. In ns 2, packets were generated and inter-arrival times between packets were calculated and subjected to Chi-square goodness of fit approach. It was discovered that Pareto can be used to model current Internet bursty traffic having tested with exponential probability distribution, if falls within the acceptable region.

In conclusion, it was realized that the Pareto generated data using ns 2 under the goodness of fit test, using exponential probability distribution function and the chi square value validates that it falls within the acceptance region. Therefore Pareto which allows traffic burst can be used in place of exponential distribution to model current internet traffic.

6.0 DIRECTION FOR FUTURE WORK

For future research, to model traffic control on the Internet they can compare the performance various internet traffic generators and use statistical methods and tool to actually proof and validate the results of experiments.

REFERENCES

1. Anderson, A., &Nielsen, B. (1998). A Markovian Approach for Modeling Packet Traffic with Long-Range Dependence. IEEE Journal on Selected Areas in Communications, vol. 16, no. 5, pp. 719-732.
2. Erlang, A.K., Brockmeyer, E., H.L. Halstrom, H.L., & Arns, J.(1948). The Copenhagen Telephone Company. Proceedings of the C.C.I.F. ("Le comité consultatif international des communications téléphoniques à grande distance"), Montreux.
3. Adas, A. (1997). Traffic Models in Broadband Networks", IEEE Communications Magazine.
4. Du, Q. (1995).Monotonicity result for a single-server queue subject to a Markov-modulated Poisson process. J. Appl. Probability 32, 1103–1111.

5. Cao, J., Cleveland, W., Lin, D., & Sun, D. (2002). Internet traffic tends toward Poisson and independent as the load increases in Nonlinear Estimation and Classification. New York, NY: Springer Verlag.
6. Park, K., & Willinger, W. (2000). Self-similar network traffic. An Overview. In Self-similar network traffic and performance evaluation. Chichester, England: J.Wiley & Sons.
7. Jain, R., & Routhier, S. (1986). Packet Trains--Measurements and a New Model for Computer Network Traffic. IEEE Journal on Selected Areas in Communications, Vol. 4, Issue 6, pp. 986-995.
8. Frost, V., & Melamed, B. (1994). Traffic Modeling for Telecommunication Networks. IEEE Communications Magazine, 32(3), pp. 70-80.
9. Paxson, V., & Sally, F. (1995). Wide-area Traffic: The Failure of Poisson Modeling. IEEE/ACM Transactions on Networking, pp.226-244.
10. Willinger, W., & Paxson, V. (1998). Where Mathematics Meets the Internet. Notices of the AMS, Vol 45, No.8, P961-970.
11. Williamson, C. (2001). Internet Traffic Measurement. *IEEE Internet Computing* 5 (6): pp 70–74.
12. Yang, X., & Petropulu, A.P. (2001). The Extended Alternating Fractal Renewal Process for Modeling Traffic in High-Speed Communication Networks," IEEE Trans. Sig. Proc., vol. 49, no. 7.York, NY.
13. Anderson, A.,&Nielsen. (1998). A Markovian Approach for Modeling Packet Traffic with Long-Range Dependence. IEEE Journal on Selected Areas in Communications, vol. 16, no. 5, pp. 719-732.
14. Froot, K. A., O'Connell., & Paul G.J. 2008. On the pricing of intermediated risks: Theory and application to catastrophe reinsurance. Journal of Banking.
15. Pearson, K. (1905). On the General Theory of Skew Correlation and Non-linear Regression. London: Dulau & Co.