Helix Vol. 9 (2): 4858- 4863

DOI 10.29042/2019-4858-4863

Systematic Performance Analysis of Bit-Torrent Traffic

[1]Rupesh C. Jaiswal, [*2]Mousami V. Munot, [3]G. S. Mundada, [4]M. P. Turuk, [5]S. D. Lokhande

[1-4] Dept. of Electronics and Telecommunication, Pune Institute of Computer Technology, Pune

[5] Principal, Sinhgad College of Engineering, Pune

*Email: rcjaiswal@pict.edu, *mvmunot@pict.edu, gsmundada@pict.edu, mpturuk@pict.edu, sdlokhande.scoe@sinhgad.edu*

Received: 18th March 2019, Accepted: 10th April 2019, Published: 30th April 2019

Abstract

The eruption of the World Wide Web as a medium for information dissemination, movies, music and entertainment, has made it extremely important to understand the characteristics of its traffic. Bit-Torrent (BT), communication protocol for peer-to-peer (P2P) file sharing, continues to still remain a dominant source of upstream traffic worldwide. This wide spread popularity of BT has attracted exponentially increasing attention of the researchers in the networking domain. Existing studies on BT systems have reported analytical performance modeling, which includes Markov chain model, Deterministic model, Queuing network model and Fluid flow model. This research study presents exhaustive survey of BT protocol and various mathematical modeling reported for its performance evaluation under transient and steady condition. Precise and apt comparison of the performance parameters in terms of complexity, precision and flexibility is initially summarized in this research. This study further incorporates rigorous experimentation on a huge dataset (BT- data packet traffic) explicitly created for this research to evaluate and analyze other significant parameters like heavy tailedness, self-similarity, autocorrelation, power spectral density and burstiness (packet inter-arrival time and packet size), which conform the self-similar behavior of BT traffic . Analysis of such parameters is crucial and inevitable during the simulation of the data traces, designing accurate models and is also required by the network provider to design and manage internet traffic efficiently, in order to avoid unnecessary deterioration of quality of service caused by heavy-tailedness. However, such rigorous and wide-ranging analysis of BT traffic, has received meager attention in the reported literature, and is the key contribution of this research study.

Keywords

Bit-Torrent Traffic, Peer to Peer, Heavy Tailedness, Modeling.

Introduction

In today's era, the use of digital images is increasing [1]. P2P communication, which uses simple strategy of collecting the data from all sources and sharing it to respective peers, has therefore become more popular over conventional client server approach [2,3]. This new protocol brought millions of music admirer together to exchange their own music assets [4,5]. In 2001 Bram Cohen has made the revolution in P2P communication by writing Bit-Torrent (BT) protocol [6,7]. BT protocol has almost replaced all others protocols in P2P network. Today, BT can be credited for a quarter of all upstream Internet traffic in North America, more than any other traffic source[8]. The scenario remains same across the globe. Nowadays most of the social networking sites also use BT network to share hundreds of Megabytes data to all worldwide nodes in a very systematic way[9-11]. BT is a simple protocol that consists of *Leeches* (peers who download the file but do not share the whole file with the others), *Seed or seeder* (peers with complete copy of the file), *Swarm* (network of peers and seeds -peers are downloading the file and simultaneously uploading), *.torrent file* (a metadata file that contains the address of server), *Tracker* (a server that handles the BT data sharing process) and *Seeding process* (uploading the file after completion of downloading)[12,13]. As depicted in fig. 1, two peers, peer1 and peer2 are downloading the same file *xyz.torrent*. Peer 2 was already in the network and peer1 has downloaded the *xyz.torrent* file from web browser by using BT client. Once it has .torrent file, it sends request to the tracker for the peer list. Tracker responds it back with the peer list. When peer gets the list, it will form a peer set containing all its neighboring peers from the list. Peer 1starts downloading from this peer set. As peer 2 was already in the network, it starts uploading to peer 1 along with seed. This seed uploads to peer 2 as well. [14]

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

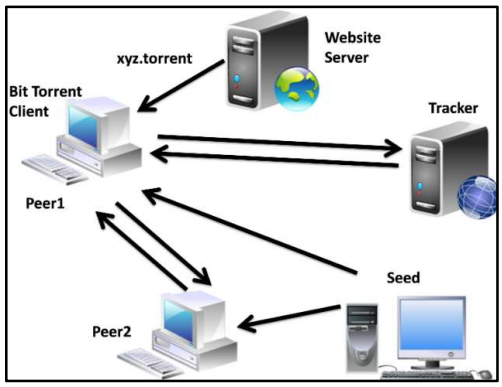

Fig. 1: Bit Torrent Network Architecture

For characterizing BT packet traffic, this research has focused on heavy tailed analysis, self-similarity analysis, long range dependence analysis and checked for burstiness. Heavy tailedness is a significant property of BT traffic which follows power law [14] .The distribution of file sizes follows power law and therefore very large file transfer can be expected with non-negligible probability. BT traffic shows self-similar behavior because of long range dependency. To study self-similarity, Hurst parameter is calculated. Long range dependency is explicated using autocorrelation analysis and power spectrum analysis. Visual inspection of burstiness and calculation of Index of Dispersion for Counts (IDC) and Peak to Mean Ratio (PMR) further emphasizes self-similar behavior of BT traffic. Next section details the dataset explicitly created for this research followed by comprehensive comparison of various analytical modeling methods for BT traffic in section 3. Section 4 highlights the major contribution of this research study and details the results of the exhaustive experimentation for the wide-ranging and full-fledged analysis of the BT traffic, followed by the conclusion section.

Dataset and Methodology

In this research work, 3.30 GHz Intel i5-4590 CPU workstation with 8GB of RAM and Ubuntu 18.04 (LTS) version operating system is used. A packet capturing tool, Wire shark, [15] is installed on the said client system to capture real time torrent traffic along with Bit Torrent client. Wireshark is specifically configured in non-promiscuous mode with "no broadcast and no multicast", filter option selected. This is to avoid the interference of local broadcast traffic as well as any updates related to background system software or application software. The data used was collected in January, 2019 with the help of machine with given specification. Multiple downloads including open source OS installable files, free Multimedia contents with authenticated user login is used in campus network at different periods of time in a day. It includes independent seeding process, independent leeching process and simultaneously both operations. The traces comprise of 29 files (29 Traces) with *pcap* file format is developed during this research study. Responses of few are depicted in table 1 for indicative purpose.

Dataset	Volume of Traffic
Trace-1	366.22 MB
Trace-2	227.29 MB
Trace-3	394.81 MB

Dataset	Volume of Traffic
Trace-4	244.32 MB
Trace-5	375.67 MB
Trace-6	786.23 MB

Dataset	Volume of Traffic
Trace-7	127.7 MB
Trace-8	329.7 MB
Trace-9	197.79 MB

Table 1: Bit Torrent Traces/Datasets Used for Analysis.

Modeling Techniques for BT Traffic

The Markov chain model, Deterministic model, Queuing network model and the Fluid flow model are the popular modeling techniques reported in the literature for BT network [16,17]. Performance of BT network considering these analytical modeling methods, as reported in the literature is summarized in this research study. Review of all the modeling techniques along with their estimated parameters is presented in Table 2. This table gives idea about the modeling technique and the parameters they can estimate. Various parameters such as regime in which the model is applied, complexity, precision and flexibility are used to compare these methods. When there are plenty of requests coming for a newly introduced file, deterministic model can be used for analyzing the BT network in the initial transient regime. In deterministic model, most of the parameters are generally assumed, so deterministic model is considered to be an idealistic one. This model does not consider the parameters like different upload and download speed of peers, heterogeneous environment, arbitrary arrival of peers and the influence of free riders. The performance metric parameter estimated by this model is file download delay [18,19].

Once the performance of the system becomes stable, service capacity goes in to steady state. To evaluate such systems in steady state, Markov chain modeling is used. In Markov chain modeling different states is involved and the transition from one state to another state is considered. To improve the accuracy, number of states could be extended. Different studies have extended this modeling technique and created their own models to estimate different performance metrics [20-22]. Generally download throughput per peer, service capacity, delay for downloading, accessibility of file and the complete lifespan for a torrent etc. parameters can be obtained using Markov chain modeling. Fluid flow model is used

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

to study time fluctuating system performance. Fluid model can also be used to model the non Markovian queues[23-26]. Generally it consists of two differential equations indicating arrival and departure of seeds at different time instants. These differential equations can be extended, if we consider multi-class peers but it increases the complexity of solving these equations. Queuing modeling is also used to model steady state performance of the system. It can model many realistic computer networks system [27]. Advantage of queuing modeling over Markov chain and fluid flow modeling is that it models the system in more detailed way with appropriate assumptions of arrival and service process.
Table 2 presents a comprehensive comparison of all the parameters, Functional Regime, Complication, Precision and Flexibility as reported in the literature. The major highlight of this study incorporates rigorous experimentation on a huge dataset to evaluate and analyze other significant parameters like heavy tailedness, self-similarity, autocorrelation, power spectral density and burstiness (packet inter-arrival time and packet size) as detailed in the next section.

Mathematical Modeling Technique	Parameters Estimated	Functional Regime	Complication	Precision	Flexibility
Markov chain model	• Service capacity[3][9][23] • Download throughput per peer[3] • File download delay [3] • File accessibility[7] • Total lifespan of torrent[25]	Steady state	Low	Depends on conditions	Possible
Deterministic model	File download delay [3][9][23]	Transient	Medium	Not good	Restricted
Queuing network model	File download delay [16]	Steady state	High	Depends on conditions	superior
Fluid flow model	• Service capacity [14] • Download throughput per peer [17] • File download delay [14] • File accessibility [14][17] • Total lifespan of torrent [17]	Time fluctuating	Low	Good	Good

Table 2: Comparison of All Four Modeling Techniques

Result and Discussion

Our earlier research has reported the experimental analysis of bit torrent traffic highlighting the significance of heavy tailed probability distributions [24]. Other various important parameters like inter-arrival times and lengths of packets are analyzed and used to plot the CDF. This research study aims to evaluate the BitTorrent traffic in terms of heavy-tailedness, self-similarity, autocorrelation, power spectral density and burstiness of parameters like inter-arrival time and packet size. Literature reports meager attention to the analysis of BT traffic from the perspective of these parameters, which are of vital importance for simulation and modeling. The experimental results and the obtained analysis are detailed as follows:

Heavy Tail Distribution [HTD]

For a Distribution to be Heavy Tailed it must follow power law. Mathematically it can be represented as

$$P[X > x] \sim x^{-\alpha} \quad \text{as } x \to \infty, \qquad (1)$$

Where $0 < \alpha < 2$ expressed by Crovella [26].

The most important property of Heavy tailed Distributions (HTD) is that its rate of decay is relatively slower when compared to exponential distributions (Poisson traffic). Moreover, If complementary CDF (CCDF) function is defined as $\overline{F}(x) = 1 - F(x)$, then the CCDF is slower by some power of x given as $\overline{F}(x) \to cx^n e^{-\lambda}$. In the experimentation of the dataset generated in this research, CCDF has been utilized to analyze the behavior of BT traffic and validate its distribution. As illustrated in fig. 2, when Poisson distribution is plotted along with an HTD on log-log CD plots, then HTD spreads above Poisson distribution. It can be observed that the graphs of BT data traffic for inter arrival time lies above Poisson distribution. This response, thus clearly shows the property of heavy tailedness.

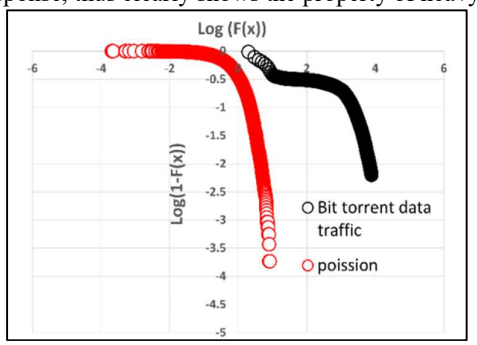

Fig. 2: Log –Log CD Plots for BT Traffic and Poisson's Distribution.

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

Self-Similarity

Self-similarity is a significant phenomenon which is considered by the researchers while modeling real internet traffic. It is a measure of resemblance (exact or approximate) exhibited by an entity to a part of itself or whole, irrespective of scaling. This section presents the results of the self-similarity tests on the captured traces. Firstly, we discuss the properties of self-similar traffics, which have been concluded at several studies that are established practically to measure and analysis the statistical characteristics of self-similar traffic of a packet based networks [6, 7, 20, 21]. For Bit torrent traffic time series, it exhibits self-similar behavior. Understanding of such behavior analysis is crucial in designing of computer networks for resource sharing, queue management and routing management.

For a continuous time series self-similarity can be defined as

Let G = { $G(j), j \geq 1$} be a stationary sequence.

$$G^{(n)}(k) = \frac{1}{n} \sum_{j=(k-1)n+1}^{kn} G(j), \quad k = 1,2,\ldots, \tag{2}$$

then $G^{(n)}(k)$ will the aggregated sequence. It has aggregation level of m which is obtained by averaging over non-overlapping blocks of size n. So we can state self-similar process, for all integers n as

$$G \overset{\text{def}}{=} n^{1-H} G^{(n)} \tag{3}$$

The testing of self-similarity is based on estimating the Hurst exponent value. Hurst parameter can quantify self-similar behavior of BT traffic. Hurst parameter value lies between 0 and 1 for a self-similar process. For range of H values between 0 and 0.5, the process is short range dependent. Values of H between 0.5 and 1 indicate long range dependent. In our research work, we calculated H values using various methods like R/S method, Absolute moment method, Time variance method, Difference variance method and boxed periodogram method [27]. Fractals are specifically interesting class of self-similar objects. Fractal Dimension [FD] shows a measure of complexity of a self-similar figure. There are several methods for calculating FD like pair counting, box counting, tug of war and correlation integrals. The methods have already been implemented in our previous research work [28].We also carried out fractal dimension (FD) analysis [29] and demonstrated in this paper using box counting method. The Hurst values are calculated for both IAT and packet length parameters. The average values for data traffic from the self-similarity tests on the captured traces are shown in Table 2. The value of the Hurst exponent from various tests is close to 1, which implies that the tested trace is self-similar. This research experimentally validates that the BT traffic is self-similar by running the same tests on all the traces in the dataset created for the study.

Data Set	R/S Method	Abs. Moment Method	Time Variance Method	Difference Variance Method	Boxed Periodogram	Box Counting
IAT	0.7081	0.6885	0.65293	0.61585	0.5913	1.2424
Packet Length	0.7012	0.6532	0.65829	0.56824	0.5369	1.6416

Table 3: Average Hurst Parameter Values for IAT and Packet Length of BT Traffic

Long Range Dependence [LRD]

Long Range Dependence [LRD] estimates reliance between the present values and the old estimations of any random process. The autocorrelation function has to fall hyperbolically for a process to be LRD. Conversely, the autocorrelation function of a short range dependence process decreases exponentially. LRD can be expressed by a time series. For such a series autocorrelation function can be written as

$$\boldsymbol{s(m) \sim m^{-\beta} \text{ as } m \to \infty} \tag{4}$$

the value of β is between 0 and 1.The relation between Hurst parameter and β is as follows

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

Thus for LRD time series, Hurst parameter [30] is given as, $1/2 < H < 1$, as $H \to 1$. The degree of long-range dependence increases. In this research, autocorrelation analysis has been carried out and obtained plots are illustrated in Fig. 3(a) and Fig. 3(b). It can be seen that lower degree of autocorrelation is present as the lag increases. This results to lower Hurst parameter values.

The power spectrum of a self-similar process adheres to power law and it is centered at the origin. It is important to note that the existence of self-similarity will affect the power spectrum at the band of the low frequencies i.e. as $\omega \to 0$. This indicates that the power spectrum of self-similar process follows a power law distribution as $\omega \to 0$.Self-similar traffic can be characterized by power spectral density (PSD). For LRD time series PSD follows a power law near origin.

$$Q_x(v) \approx |v| \text{ as } v \to \infty, 0 < \gamma < 1, \tag{6}$$

Where v is frequency, $Q_x(v)$ is the spectrum density and

$$\gamma = H - 1.$$

PSD graph are plotted for both IAT and Packet length, they demonstrated analogous characteristics; representing a little 1/f type power spectrum behavior as shown in Fig. 3(c) and Fig.3 (d). Also Gaussian type power spectra are observed for both of them. Fig. 3(c) and Fig.3 (d) illustrates only one plot of PSD for a data trace, but similar plots are obtained for other data traces.

Burstiness

The variation in network traffic is called as burstiness of traffic. Burstiness can be studied using Peak to mean ratio (PMR) and Index of Dispersion for Counts (IDC) as defined below:

$$\boldsymbol{PMR = \frac{Max(K_t)}{Mean(K_t)}} \tag{7}$$

$$I_t = \frac{Variance(K_t)}{E(K_t)} \quad (8)$$

Where, K_t indicates the number of arrivals in an interval of time t and $E(K_t)$ is mean number of arrivals in t. Both IDC and PMR shows extent of burstiness present in the given traffic. IDC value for Poisson traffic is 1. A process when defined as self-similar is expected to demonstrate same characteristics at any time scale. BT traffic being self-similar shows same burstiness at varied time scales. In this research, IDC values are calculated for both IAT and packet length at varied time scales and indicated in table 3. It is evident that the values are quite stable and constant over larger time scales. Thus, even if we magnify or scale our data, we still get burstiness and that too of same level. This analysis is clearly demonstrated in time series plot of BT traffic (10 sec and 100 sec) in fig. 3 (e) and fig. 3(f) respectively.

Time Duration* (seconds)	1	10	50	100
IAT	2337	2693	2619	2355
Packet Length	1521	1465	1387	1254

Table 4: IDC Values for IAT and Packet Length of BT Traffic

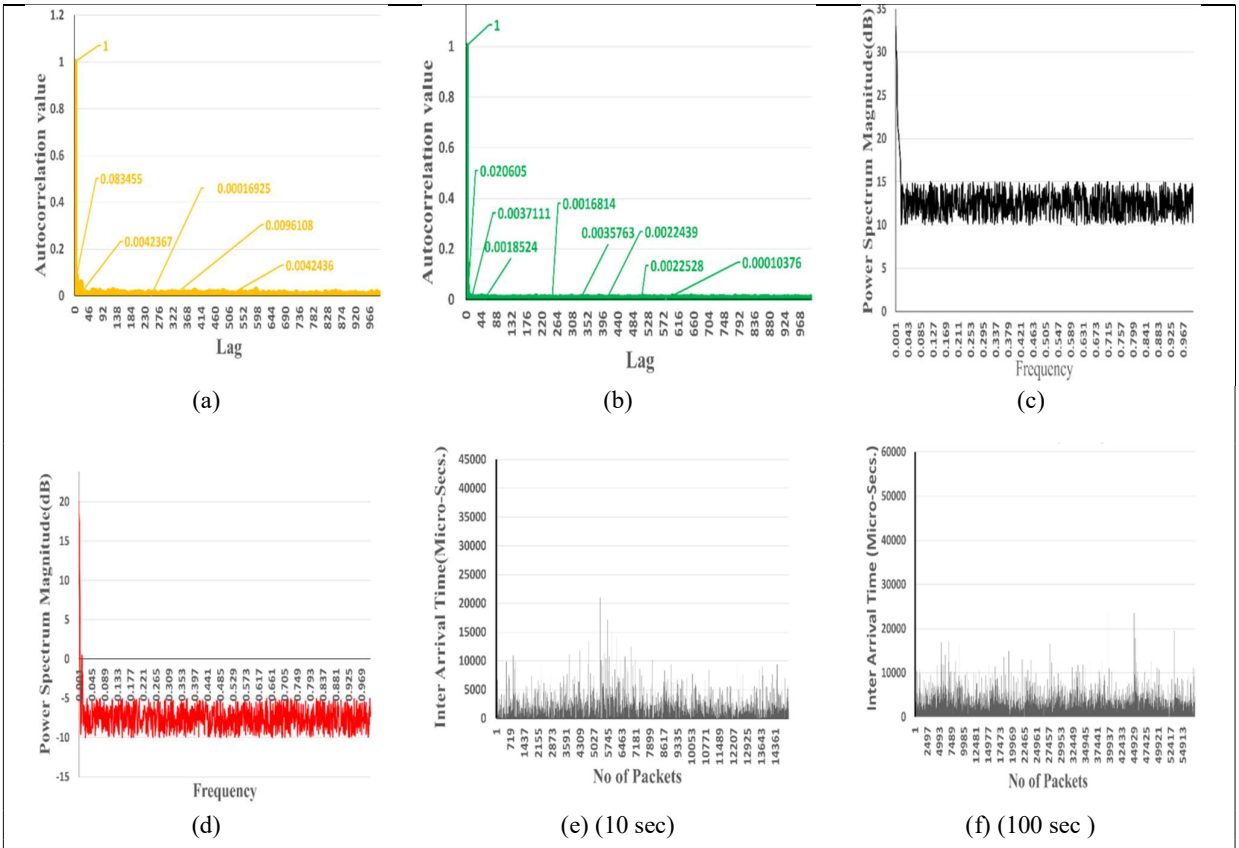

Fig 3. (a) Autocorrelation Function Plot for IAT (b) Autocorrelation Function Plot for BT Traffic (Packet Length) (c) PSD Plot for IAT (d) PSD for BT Traffic (IAT) (e) Time Series Plot for BT Traffic – 10s (f) Time Series Plot for BT Traffic – 100s

Conclusion

A substantial survey for evaluating the performance using various analytical modeling techniques for BT P2P file sharing network is presented in this research (Markov chain model, Deterministic model, Queuing network model and Fluid flow model). To evaluate the performance of BT traffic, this research focused on file accessibility parameters, delay for downloading the file, downloading throughput for each peer and service capacity. From the analysis it is observed that deterministic model are more idealized so it is less accurate compared to all other models. Also it is observed, that analysis of Queuing network model is more complex compare to other models. In addition, these analytical methods are compared on the basis of functional regime, complication, precision and flexibility. Further our experimentation with BT traffic demonstrated that it is heavy tailed. Thus, the data packet communication that takes place among BT system shows heavy tailedness. The presence of heavy tailedness is one of the reasons for the presence of self-similarity in BT traffic. The autocorrelation analysis and PSD graphs evidenced the long range dependent behavior. All the above outcomes pointed out that the BT traffic is bursty. An investigation of such parameters is significant and inescapable for the network-administrator / engineer to structure and oversees web traffic proficiently. Such

 Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495

thorough and wide-going examination of BT traffic, which has received meager attention in the literature is the key commitment of this research study.

References

[1] B. N. Mahajan, A. R. Mahajan, A. Thomas, "Image Compression Analysis in Torrent Based Wireless Sensor Network", Helix Journal, Vol. 8(5): 3774- 3780 (2018)

[2] Liaoa W.C., Papadopoulosa F., Psounisb K. : Performance analysis of BitTorrent-like systems with heterogeneous users. Elsevier Journal on Performance Evaluation 64, pp. 876-891 (2010).

[3] Yang X. and Veciana G. : Service Capacity of Peer to Peer Networks. In : IEEE INFOCOM, Vol.4, pp. 2242-2252 (2004).

[4] Rollins K.S, Khambatti M.: From the Editors: Peer-to-Peer Community:Looking beyond the Legacy of Napster and Gnutella. In: IEEE Computer Society, Vol.7, No. 3 (2006).

[5] Application usage and threat report. Technical report. eleventh edition (2014).

[6] Bharambe A. R, Herley C., Padmanabhan V. N. : Analyzing and Improving a BitTorrent Network' Performance Mechanisms. In: IEEE INFOCOM (2006).

[7] Tian Y, Di W., and Kam W. : Modelling, Analysis and Improvement for Bit-Torrent Like File Sharing Networks. In : IEEE INFOCOM (2006).

[8] Cohen B. : Incentives build robustness in BitTorrent. In *:* First Workshop on Economics of Peer-to-peer Systems, Berkeley, CA, USA (2003).

[9] Yang X.. and Veciana G : Fairness, incentives and performance in peer-to-peer networks. In: Forty-first Annual Allerton Conference on Communication*,* Control and Computing, USA (2003).

[10] Cooper R. B *:* Introduction to Queueing Theory: 2nd edition , pp. 176-300, North Holland , New York, Oxford.

[11] Yao Z., Leonard D. , Wang X., and Loguinov D: Modeling Heterogeneous User Churn and Local Resilience of Unstructured P2P Networks. In: IEEE ICNP, pp.32-41 (2006)

[12] Leon-Garcia A.*:* Probability, Statistics and Random Processes for Electrical Engineering*,* 3rd edition, Prentice Hall, pp. 647-692 (2008).

[13] Susitaival R. Aalto S.: Analyzing the file availability and download time in a P2P file sharing system. In: IEEE International Conference on Next Generation Internet Networks, pp.88-95 (2007).

[14] Srikant R. and Qiu D.: Modeling and performance analysis of BitTorrent peer-to-peer networks. In: ACM SIGCOMM, pp.367-378, New York, USA (2004).

[15] Wireshark, Available: http:// www.wireshark.org

[16] Sikdar B, Ramachandran K.K.: A queuing model for evaluating the transfer latency of P2P systems. Transactions on Parallel and Distributed Systems, Vol. 21, No. 3, pp. 367-378 (2010).

[17] Lei Guo, Zhen Xiao, Enhua Tan: A Performance Study of BitTorrent-like Peer-to-Peer Systems. IEEE journal on selected areas in communications*,* Vol. 25, No. 1 (2007).

[18] Arnaud Legout G. UrvoyKeller and Michiardi P: Rarest First and Choke Algorithms Are Enough. IMC'06 (2006).

[19] Zihui Ge, Daniel R. Figueiredo, Jaiswal S., Jim KuIJFrose, Towsley D. : Modeling Peer-Peer File Sharing Systems. IEEE INFOCOM, pp. 2188-2198 (2003).

[20] Weber R. and Mundinger J., Efficient file dissemination using peer-to-peer technology. Univ. of Cambridge, Cambridge, U.K., Technical. Report (2004).

[21] Yao Yue and Chuang Lin, Zhangxi Tan : Analyzing the Performance and Fairness of BitTorrent-like Networks Using a General Fluid Mode. In: IEEE GLOBECOM (2006).

[22] Bob Kinicki: Performance Evaluation of Computer Networks*. WPI.*

[23] Veciana G, Yang X: Performance of peer-to-peer networks: Service capacity and role of resource sharing policies. Performance Evaluation in P2P Computing Systems, Vol. 63, No. 3, pp.175-194 (2006).

[24] Aishwarya Gaikwad, Rupesh Jaiswal, Experimental Analysis of Bit-torrent Traffic based on Heavy-Tailed Probability Distributions, International Journal of Computer Applications, 155 (2), 1-5 (2016)

[25] Virtamo J., Susitaival R. and Aalto X.: Analyzing the dynamics and resource usage of P2P files haring systems by a spatio-temporal model. In *:* International Workshop on P2P for High Performance Computational Sciences (P2P-HPCS'06) in conjunction with ICCS (2006).

[26] Mark E. Crovella and Azer Bestavros: Self-Similarity in world wide web traffic: evidence and possible causes. In: IEEE/ACM Transactions on Networking, Vol 5, No. 6 (1997).

[27] Thomas Karagiannis, MichalisFaloutsos, SELFIS: A Tool for Self-Similarity and Long-Range Dependence Analysis, 1st Workshop on Fractals and Self-Similarity in Data Mining: Issues and Approaches (in KDD) Edmonton, Canada, July 23, 2002.

[28] Jaiswal R., Lokhande S: Measurement, Modelling and Analysis of HTTP Web Traffic, ICCC-2014, Elsevier.

[29] Cervantes-De la F Torre, Gonz´alez-Trejo J I Real-Ram´ırez , C A and Hoyos-Reyes L F: Fractal dimension algorithms and their application to time series associated with natural phenomena. J. Phys.: Conf. Ser. 475 012002 doi:10.1088/1742-6596/475/1/012002.

  Helix E-ISSN: 2319-5592; P-ISSN: 2277-3495