Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 2, 2025 (New Series), pp 181–189
http://www.ssca.org.in/journal

Time Dependent Performance Analysis of Asynchronous Internet Switch - Multi Server Queueing System and Markovian QBD Process

Malla Reddy Perati, Shivaji Arepelly and Abhilash Vollala

Department of Mathematics, Kakatiya University, Warangal- 506009

Received: 14 February 2024; Revised: 25 November 2024; Accepted: 29 December 2024

Abstract

In this paper network nodes namely routers and switches with self similar input traffic are modelled as multi-server finite buffer queueing system. Time dependent Markovian Arrival Process (MAP) is employed for input process, as it emulates the nature of traffic. The network traffic is asynchronous, and of variable packet lengths, service times(packet lengths) are assumed to follow Phase type (PH) distribution, The Phase type distribution is more general, in the sense, it includes all kinds of exponential distributions namely Erlang, Hyper exponential and Coxian distribution. The transition rate matrix of buffer occupancy of the system is obtained using Quasi Birth and Death (QBD) process. Here, time dependent behaviour of queueing system is evaluated through the performance metrics, namely, packet loss, and mean waiting time. This kind of investigation is useful in dimensioning the network nodes to guarantee Quality of Service (QoS).

Key words: Internet switch; Self-similarity; Multi-server queue; Markov modulated Poisson process; Loss probability.

AMS Subject Classifications: 60G18, 60K25, 68M20

1. Introduction

In this era of technology, performance of communication networks play a significant role in all aspects of day to day life. Typically, network systems are characterized by queueing systems using Markovian approach. Due to inflated demand, network traffic is bursty, which causes congestion, thereby loss of information. For the said reasons, network traffic modeling and performance analysis have become crucial issues to guarantee the QoS. The networks namely LAN and WAN traffic display fractal-like characteristics, which can significantly impact network performance. The aggregation of this traffic reveals a self-similar pattern, and the degree of self-similarity measured in terms of the Hurst parameter, which is used for measuring the burstiness of traffic (Paxson and Floyd (1995); Crovella and Bestavros (1996)). Self-similar traffic exhibits the same statistical behaviour over a wide range of time scales (Leland *et al.* (1994)), hence, it is necessary to consider the said nature in modeling.

Corresponding Author: Shivaji Arepelly
Email: sivajiarepalli@gmail.com

To model self-similar behaviour efficiently, various traffic models have been proposed earlier, like chaotic maps, Fractional Brownian Motion (FBM), and FARIMA, *etc.* These models characterize the traffic patterns in a parsimonious manner. Later on, several Markovian models were proposed to emulate the self-similar nature of traffic based on second order characteristics over certain ranges of time scales. However, they are not realistic because, the traffic parameters in the models were assumed to be homogeneous with respect to time and the performance metrics are computed at the steady state (Shao *et al.* (2005); Chang *et al.* (2006); Chen *et al.* (2007)). In this regard, a new procedure was proposed recently, based on the second order characteristics for fitting the time dependent Markovian process, and presented that it emulates self-similar traffic (Abhilash and Perati (2022)). On the other hand, Wavelength Division Multiplexing (WDM) is a promising technology to guarantee the Quality of Service (QoS) and provides a highly efficient way to increase network capacity by transmitting multiple signals over a single fiber. In WDM systems, converting asynchronous data (such as IP traffic) into synchronous traffic is often done through statistical multiplexing, it is used to dynamically allocate bandwidth based on traffic demand, and is suitable for circuit-switched networks (Chen *et al.* (2007); Perati *et al.* (2007); Venkataramani *et al.* (1997)). In asynchronous networks, packets typically have variable lengths, and routers must be capable of handling variable length packets. Using an MMPP/D/1/C queueing system (with deterministic service time) to model router performance is not suitable for variable-length packets. A better approach is the MMPP/M/1/C system, where service time is exponentially distributed. However, it is often more appropriate to model the service time (packet length) using a more general distribution, such as a Phase type distribution. This approach is particularly relevant in Broadband Integrated Services Digital Networks (B-ISDN), where communication services are designed to offer differentiated services (Diff-Serv) and ensure Quality of Service (QoS). The asynchronous router consists of N input fiber lines, N output fiber lines, and each fiber line has K (say) wavelength channels, and a wavelength converter pool of size c (say), $(0 \leq c \leq K)$ dedicated to each output fiber line. In the papers Krieger and Wagner (1992); Artalejo *et al.* (2001); Kim *et al.* (2016) network node is modelled as multi-server queueing system, and considered underlying Markov chain with its quasi birth-death structure. However, performance evaluation in the above papers was made in steady state which is again not realistic. For the above reasons, in the current work, asynchronous switch with self-similar input traffic is modeled into MMPP(t)/PH/c/K queueing system and transient performance analysis is made by using QBD process. The performance measures, namely, mean waiting time, packet loss are presented numerically.
The paper is organized in the following way. The fundamental definitions of self-similar and Long Range Dependent(LRD) processes are placed in section 2. The queueing model description is presented in section 3. In section 4, performance analysis is given, and numerical results are presented in section 5. Finally, conclusions are given in section 6.

2. Self-similar and long range dependent (LRD) process

Self-similarity is a key characteristic in fractals. It is a property, wherein a certain feature of the object is preserved with respect to scaling in space and time. It is statistically defined as follows. Let X be a second order process with variance σ^2, and the time axis is splitted into disjoint sub intervals of unit length. Let $X = \{X_t : t = 1, 2, ...\}$ be the points (packet arrivals) in the t^{th} interval. Let $X^{(m)} = \{X_t^{(m)}\}$ be a new sequence obtained by averaging the original sequence over non-overlapping blocks of size m. *i.e.*,

$$X_t^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \quad t = 1, 2, 3, ... \tag{1}$$

The obtained sequence is a second order process, and is called exactly second order self similar with Hurst Parameter $H = 1 - \beta/2$, if

$$Var(X^{(m)}) = \sigma^2 m^{-\beta}, \quad \forall\ m \geq 1 \tag{2}$$

If the parameter H lies in $(0, 0.5)$, then the process X is called Short Range Dependant (SRD). The process X is called Long Range Dependant (LRD), if $0.5 < H < 1$. Self-similar traffic displays burstiness across a wide range of time scales, which distinguishes it from traditional SRD traffic models. However, it is possible to replicate the self-similar behaviour over a specified time-scale range by using Markov Arrival Processes (MAPs) to model, and by matching the second-order self-similarity characteristics of the process accurately (Shao *et al.* (2005); Abhilash and Perati (2022)).

3. Queueing model

Network switch under consideration is the WDM asynchronous $N \times N$ router, where each output port with self-similar input traffic is modelled into MMPP(t)/PH/c/K multi-server queueing system. Self-similar input traffic here is modeled as MMPP with time dependent sinusoidal arrival rates *i.e.*, $a + b_i sint$, where a, b_i are constants and $0 < b_i < 1$. The arrival process here is assumed to be superposition of d two state IPPs, and a classical Poisson process. The i^{th} IPP is characterized by following matrices:

$$Q_i = \begin{bmatrix} -d_{1i} & d_{1i} \\ d_{2i} & -d_{2i} \end{bmatrix}, \ \Lambda_i(t) = \begin{bmatrix} \lambda_{1i}(t) & 0 \\ 0 & 0 \end{bmatrix}, (i = 1, 2, \cdots, d). \tag{3}$$

The superposition of IPPs is given by

$$\begin{aligned} Q &= Q_1 \oplus Q_2 \oplus \cdots \oplus Q_d, \\ \Lambda(t) &= \Lambda_1(t) \oplus \Lambda_2(t) \oplus \cdots \oplus \Lambda_d(t) \oplus \lambda_p(t) \end{aligned} \tag{4}$$

where $\oplus$ indicates kronecker's sum, and $\lambda_p(t)$ denotes arrival rate of classical Poisson process at time t. The mean arrival rate $\lambda(t)$ is $\frac{1}{t}$ $(\pi \int_0^t \Lambda(t)dte)$, where e is the vector of ones with appropriate dimension, and π is steady state vector of Q. Here, each server follows continuous-time Phase type service distribution characterized by (β, S) with dimension k, where β is vector of order $1 \times k$, $S^0 = -Se$. The mean service time is $\mu = -\beta S^{-1}e$. Service discipline is assumed to be FIFO, and the service time (here packet lengths) are independent, identically distributed random variables governed by Phase type distribution of order k. System capacity is finite, and is taken to be K. The resultant MMPP(t)/PH/c/K queueing system can be described by using stochastic process, $Z(t) = (R(t), H(t), Y(t)), t \geq 0$. Here, $R(t)$ denotes the number of packets in the system at time t, and $H(t)$ denotes the number of customers just served in different phases of the service process at time t. According to assumptions, the process $Z(t)$ is a CTMC on the finite state space $S = \{z = (r, h_1, h_2, \cdots, h_k, y)\}, 0 \leq r \leq K, 0 \leq h_j \leq c, 1 \leq j \leq k, 1 \leq y \leq m$. In the case of quasi birth and death process with lexicographical ordering of state space S, transition rate matrix is given by

$$A(t) = \begin{bmatrix} A_{00}(t) & A_{01}(t) & 0 & \cdots & \cdots & \cdots & 0 \\ A_{10} & A_{11}(t) & A_{12}(t) & \ddots & \cdots & \cdots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \ddots & A_{cc-1} & A_{cc}(t) & A_{cc+1}(t) & \ddots & \vdots \\ \vdots & \cdots & \ddots & A_{c+1c} & A_{c+1c+1}(t) & \ddots & 0 \\ \ddots & \cdots & \cdots & \ddots & \ddots & \ddots & A_{K-1K}(t) \\ \vdots & \cdots & \cdots & \cdots & 0 & A_{KK-1}(t) & A_{KK}(t) \end{bmatrix}$$

where each block matrix $A_{ij}(t)$ is an $n_i \times n_j$ matrix. The block matrix $A_{ii+1}(t)$ corresponds to arrival of IP packet, Q_{ii-1} associate with departure of IP packet and, $A_{ii}(t)$ are associate with internal phase changes with no arrival and service completion. The block matrices are as follows:

$$\begin{aligned} & A_{00}(t) = D_0(t) \\ & A_{ii+1}(t) = D_1(t) \otimes \beta \otimes I, && i = 0, 1, \cdots, c-1 \\ & A_{ii-1} = I \otimes S^{\oplus i}, && i = 0, 1, \cdots, c-1 \\ & A_{ii}(t) = D_0(t) \otimes S^{\oplus i}, && i = 0, 1, \cdots, c-1 \\ & A_{ii}(t) = D_0(t) \otimes S^{\oplus c}, && i = c, c+1, \cdots, K-1 \\ & A_{KK}(t) = D_0(t) \otimes S^{\oplus K} + D_1(t) \otimes I \\ & A_{ii+1}(t) = D_1(t) \otimes I, && i = c, c+1, \cdots, K \\ & A_{ii-1} = I \otimes (S^0\beta)^{\oplus c}, && i = c, c+1, \cdots, K, \end{aligned}$$

where $D_0(t) = Q - \Lambda(t)$, $D_1(t) = \Lambda(t)$ and I is the identity matrix of appropriate order.

4. Performance analysis

Let $\pi(t) = (\pi_0(t), \pi_1(t), \pi_2(t), \cdots, \pi_K(t))$ be transient state probability vector of $A(t)$, that is, $\pi(t)$ satisfies Stewart (1994)

$$\frac{d}{dt}\pi(t) = A(t)\pi(t), \tag{5}$$

$$\implies \pi(t) = \pi(0) + \int_0^t A(s)\pi(s)ds \tag{6}$$

By applying Picard's iterative method, one can get k^{th} iteration, and is given by

$$\pi_k(t) = \pi(0)\left(I + \sum_{r=1}^{k} \int_0^t \int_0^{s_r} ... \int_0^{s_2} A(s_r)...A(s_1)ds_r...ds_1\right) \tag{7}$$

The solution of Eq. (7) is given by $\pi(t) = \lim_{k\to\infty} \pi_k(t)$, that is,

$$\pi(t) = \pi(0)\left(I + \sum_{r=1}^{\infty} \int_0^t \int_0^{s_r} ... \int_0^{s_2} A(s_r)...A(s_1)ds_{r+1}...ds_1\right) \tag{8}$$

If each element of arrival rate matrix $\Lambda(t)$ is Riemann integrable, then each element of $A(t)$ is Riemann integrable. By using Theorem 2.4.3 in Slavík (2007), one can get

$$\pi(t) = \pi(0) \prod_{k=0}^{n} \left(I + A(t_k)h\right), \tag{9}$$

where $h = t_k - t_{k-1}, n$ is the number of partitions of the interval $(0, t]$, and $\pi(0)$ is state probabilty vector at time $t = 0$. The performance measures, namely, Mean Waiting Time (MWT) and Packet loss Probability are derived from Zhao *et al.* (2015) as follows:

By Little's Law, The Mean Waiting Time (MWT) is given by

$$MWT = \frac{E[N_L(t)]}{\lambda(t)} = \frac{\sum_{i=c+1}^{K} (i-c)\pi_i(t)}{\lambda(t)}.$$

where $E[N_L(t)]$ represents the average queue length of the system. In a finite buffer queueing system, there is a chance of loss of packets due to buffer is being full. The packet loss in Δ time is given by

$$P_{loss}(\Delta) = \frac{1}{\lambda(t+\Delta)} \left(\pi_K(t) \left[\int_t^{t+\Delta} (D_1(x) \otimes I)dx\right] e\right).$$

5. Numerical results

In this study, a real-time Internet switch is considered for the analysis. Wireshark, an open-source software tool, is commonly used for network protocol analysis. It enables the capture and real-time inspection of data packets transmitted across the network. The data collected (which is presented in Appendix) using Wireshark serves as the basis for numerical analysis, from which key network traffic parameters such as the arrival rate of the whole process (λ_w), the Hurst parameter(H), and the variance (σ^2) are calculated using the NumXL tool in MS-Excel. The NumXL tool in MS-Excel is used to calculate the traffic parameters. It is a statistical and time series analysis add-in that offers advanced functions for computing metrics, making it easy to analyze network traffic directly within Excel. Based on the tool, the values are obtained as follows $\lambda_w = 27.964$, $\sigma^2 = 138.39$, $H = 0.72$ in the interval of $[10, 10^3]$. Since $H > 0.5$, the input traffic exhibits long-range dependence (self-similar nature). Based on the obtained network parameters, the transient MMPP fitting has been done using the procedure given in the paper Abhilash and Perati (2022), *i.e.*, by matching variance of MMPP and Self-similar traffic (real time). The number of superposed MMPPs (IPPs) are taken to be two, *i.e.*, $d = 2$. The sinusoidal arrival rates are considered for modelling the asynchronous input traffic, and is given in the form of $a + b \sin t$, where b varies in between $(0, 1)$. The arrival rates referenced are drawn from the work of Eick *et al.* (1993) and used to analyze the behaviour of $Mt/G/\infty$ queue. Based on the fitting procedure for the given input parameters the following values are obtained, which are used in the current numerical analysis. The transition rates of packet arrivals are $d_{11} = 0.0401126$, $d_{21} = 0.0087$, and arrival rates are $\lambda_1(t) = 1 + 0.3 \sin t$, $\lambda_2(t) = 1 + 0.7 \sin t$. Here, it is assumed that service time follows phase distributions with two phases, namely, Erlang (E_2), Coxian (C_2), and Hyper Exponential (H_2) with varying rates. The numerical results were obtained using MATLAB programming and are presented in the form of figures. In which, the comparison of performance measures, namely, Mean Waiting time and Packet Loss Probability is presented with respect to traffic intensity pertaining to E_2, C_2, and H_2.

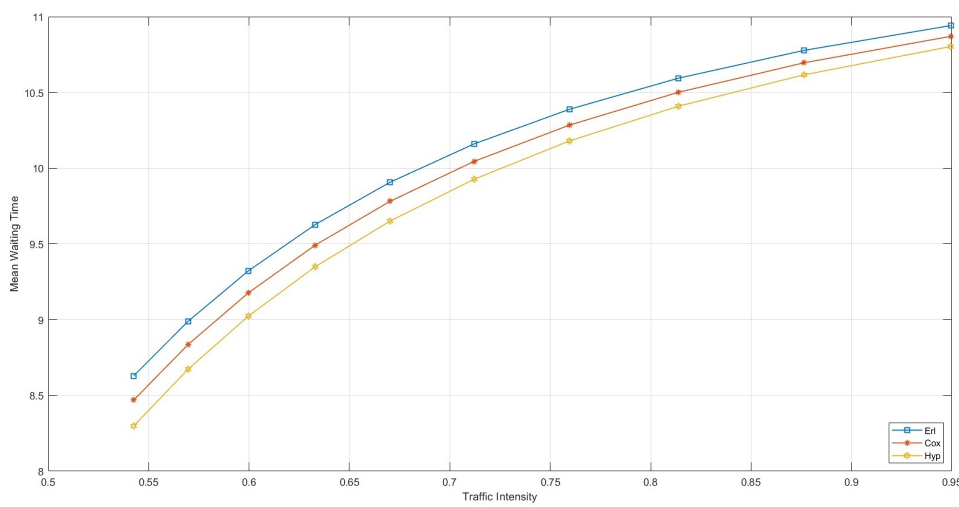

Figure 1: Mean waiting time versus traffic intensity for *t=5, K=20, c=3*

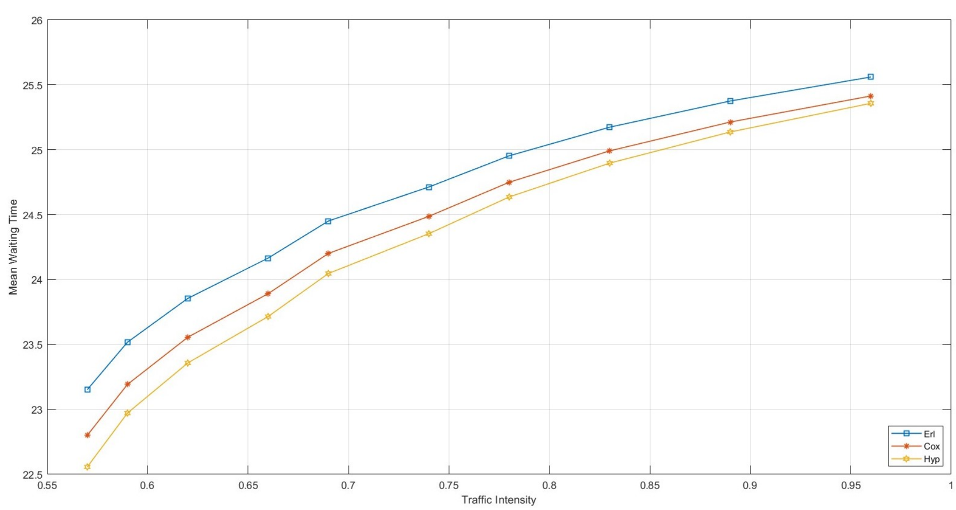

Figure 2: Mean waiting time versus traffic intensity for *t=10, K=20, c=3*

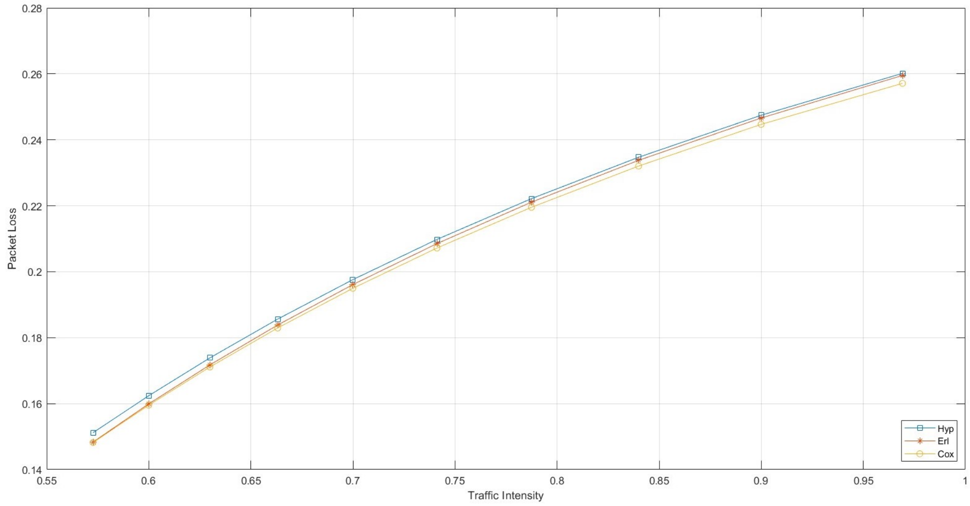

Figure 3: Packet loss probability versus traffic intensity for $\Delta = 1$, *t=10,K=20, c=3*

The figures show that as traffic intensity increases, both the average waiting time and the Packet loss probability increases. This suggests that, under heavier traffic conditions, the system faces longer delays and a greater chance of buffer overflow. The results presented in the figures reveals that the average waiting time for the $E2$ service type is consistently higher than $C2$ and $H2$, which means that $E2$ tends to experience more delays under similar traffic conditions. Also, packet loss probability, varies across different service types, *i.e.*, $H2$ has the lowest loss probability, followed by $E2$, and $C2$ experiences the highest packet loss. This indicates that, as traffic volume increases, $H2$ handles the load with the least packet loss, while $C2$ is more prone to packet loss.The insights gained from these results are crucial for understanding how each service type copes with varying traffic intensities. These type of analysis is useful for designing and optimizing networks to reduce delays and packet loss, particularly when traffic levels are high.

In this paper, the work is distinguished by its integration of time-dependent analysis, asynchronous behaviour, and Markovian models in a multi-server queueing system. The Internet switches with self-similar input traffic are modelled as multi-server Phase type queueing system, and the queueing behaviour is studied using time-dependent input parameters which obtained for a real time network traffic. Transient performance measures are obtained at different time instants using numerical and analytical techniques which are complimentary to steady state analysis. This approach offers a more accurate method for modeling and forecasting the performance of Internet switches, which is vital for modern networks dealing with fluctuating and unpredictable traffic. By using advanced queueing models like the QBD process, the analysis captures the complex dynamics between servers and queues, providing valuable inputs to improve network performance and reliability. This sort of analysis is appropriate in dimensioning the asynchronous internet switch.

Acknowledgements

Second author wishes to acknowledge the Ministry of Social Justice and Empowerment, Government of India (MOSJE, GoI) for their funding under the scheme NFSC (National Fellowship for Scheduled Caste) (UGC-Ref. No.:4545/(CSIR-UGCNET JUNE2019).

Conflict of interest

Second author received financial support from the Ministry of Social Justice and Empowerment, Government of India (MOSJE, GoI) under National Fellowship for Scheduled Caste (NFSC) scheme for this study. The authors declare that this funding did not influence the study design, data analysis, or manuscript preparation, and that no other conflicts of interest exist.

References

Abhilash, V. and Perati, M. R. (2022). Fitting model for self-similar traffic-time dependent markovian process and second order statistics. *Statistics and Applications*, **20**, 297–309.

Artalejo, J. R., Gomez Corral, A., and Neuts, M. F. (2001). Analysis of multiserver queues with constant retrial rate. *European Journal of Operational Research*, **135**, 569–581.

Chang, C. H., Shao, S. K., Perati, M., and Wu, J. (2006). Performance study of various packet scheduling algorithms for variable-packet-length feedback type wdm optical packet switches. In *2006 Workshop on High Performance Switching and Routing*.

Chen, C. Y., Chang, C. H., Perati, M. R., Shao, S. K., and Wu, J. (2007). Performance analysis of WDM optical packet switches employing wavelength conversion under markovian modeled self-similar traffic input. In *2007 Workshop on High Performance Switching and Routing*, pages 1–6.

Crovella, M. and Bestavros, A. (1996). Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, **5**, 835–846.

Eick, S. G., Massey, A. W., and Whitt, W. (1993). Mt/G/infty queues with sinusoidal arrival rates. *Management Science*, **39**, 241–252.

Kim, C., Dudin, A., Dudin, S., and Dudina, O. (2016). Analysis of two-server queueing model with phase-type service time distribution and common phases of service. In *Distributed Computer and Communication Networks*, pages 19–29, Cham. Springer International Publishing.

Krieger, U. R. and Wagner, D. (1992). Analysis and application of an MMPP/PH/n/m multi-server model. In *Performance of Distributed Systems and Integrated Communication Networks*, pages 323–342. North-Holland, Amsterdam.

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

Paxson, V. and Floyd, S. (1995). Wide area traffic: the failure of poisson modeling. *IEEE/ACM Transactions on Networking*, **3**, 226–244.

Perati, M. R., Shao, S. K., Chang, C. H., and Wu, J. S. (2007). An efficient approximate markovian model for optical packet switches employing partial buffer sharing mechanism under self-similar traffic input. In *2007 IEEE Workshop on High Performance Switching and Routing, doi:10.1109/HPSR.2007.4281232*.

Shao, S. K., Perati, M. R., Tsai, M. G., Tsao, H. W., and Wu, J. (2005). Generalized variance-based markovian fitting for self-similar traffic modelling. *IEICE Transactions on Communications*, **E88B**.

Slavík, A. (2007). *Product Integration, Its History and Applications.* Matfyz press.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, New Jersey.

Venkataramani, B., Bose, S. K., Srivathsan, K., and R, K. (1997). Queuing analysis of a non-pre-emptive MMPP/D/1 priority system. *Computer Communications*, **20**, 999–1018.

Zhao, N., Lian, Z., and Wu, K. (2015). Analysis of a MAP/PH/1 queue with discretionary priority based on service stages. *Asia-Pacific Journal of Operational Research*, **32**, 1550042.

ANNEXURE

The data presented below is used for performing numerical analysis and for calculating key traffic parameters like arrival rate of whole processes, Variance, and Hurst Parameter.

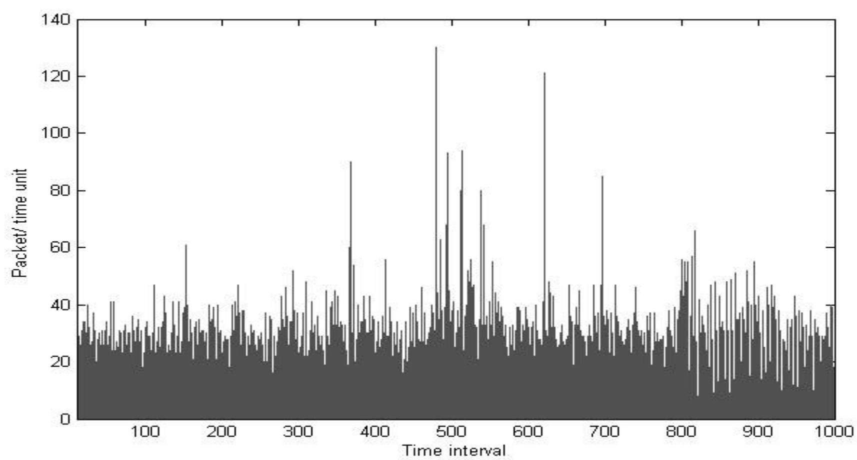

Figure 4: Packets per unit time of real-time data