STATISTICAL MODELS FOR LONG TERM NETWORK TRAFFIC IN ENTERPRISE NETWORKS

A.W.C.K. Atugoda

(148451H)

Degree of Master of Science

Department of Electronic and Telecommunication Engineering

University of Moratuwa
Sri Lanka

April 2019

STATISTICAL MODELS FOR LONG TERM NETWORK TRAFFIC IN ENTERPRISE NETWORKS

Atugoda Walawwe Chathurangi Kumari Atugoda

(148451H)

Dissertation submitted in partial fulfillment of the requirements of the degree Master of Science in Electronics and Automation Engineering

Department of Electronic and Telecommunication Engineering

University of Moratuwa

Sri Lanka

April 2019

Declaration

"I declare that this my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other university or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also I hereby grant to University of Moratuwa the non-exclusive right to produce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Name of student: A.W.C.K.Atugoda

Signature: Date:

The above candidate has carried out research for the Masters dissertation under my supervision.

Name of the supervisor: Dr.Upeka Premarathne

Signature: Date:

Abstract

With the rapid development of the internet it has converted the world into a global village and now a day we cannot even think of a micro second down time. For an instance, user demand has caused the internet to successfully combined with other networks. This expanded development has caused for huge internet traffic loads and network congestion.

For solving this key issue of the networks it is important to predict the traffic peaks in the network. These traffic peak is caused by a large amount of data being requested like in a download. If these traffic peaks are predictable then non critical traffic from another network can be scheduled to avoid peak to reduce the congestion and maximize utilization.

This dissertation introduces a method to solve that key issue. Curve fitting technique in Matlab and distributing fittings are used to build statistical models of predicting traffic. Once that identifying some drawbacks through curve fitting methodology it has been rejected and statistical models for long term network traffic in Enterprise network is used as the proposed technique. Pareto distribution, Beta-Prime distribution and Exponential distribution are derived as the statistical models to predict the traffic peak in Enterprise network. The analysis is conducted by looking at the predictability of a peak in terms of level crossing of a given level.

According to the available literature there was no such technique for predicting traffic peaks. As per the results curve fitting methodology error is significantly high. Beta-Prime and Exponential distribution are not good statistical models of predicting traffics due to huge error occurred when compared to the actual behavior of the network. But Pareto distribution is the best model of prediction on traffics in the network as it has vey less error when compared to the actual behavior of the network.

According to the results Pareto distribution is the best statistical model of predicting traffic peak. Once predicting the traffic peak can be scheduled the large data from other network for maximum utilization and to avoid the traffic congestion.

Key Words: traffic peak, level crossing, statistical model

Acknowledgement

I would like to acknowledge and express my heartfelt gratitude to my supervisor Dr.Upeka Premarathne, Senior Lecturer, Department of Electrionic and Telecommunication Engineering, University of Moratuwa for being with me in each and every step which took this research into a success while giving vital encouragement.

Further, I am heartly grateful to Dr. Roshan Regal, Director, Lanka Education and Research Network(LEARN), Dinesh Gunawardena, Secretory,LEARN and Senevi Herath, Network Engineer,LEARN for providing the data.

I am also heartily thankful to the coordinators of the MSc in Electronics and Automation Engineering pragramme and, all the staff members of Department of Electronics and Telecommunication Engineering, University of Moratuwa.

I am grateful to my parents and husband for encouraging me toward this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PPS	Packet Per Second
PDF	Probability Density Function
JPF	Joint Probability Density Formula
LLB	Local Link Bandwidth
UD	Uplink Data
DD	Downlink Data
SD	Single machine Data
ID	Institutional Data

CHAPTER 01

INTRODUCTION

1.1 Background and motivation

Internet traffic has been constantly increasing with the complete developments in communication networks and applications. This expanded development of communication methods has not only increased the demand for internet access, but also brought heavier network traffic loads. As revealed in [1], most IP traffic will be doubled with more integrating devices to the network in the next few years.

The greatly increased user demands have caused the internet to successfully develop in the automation network and enterprise network [2]. Historically both of these two networks were isolated and due to internet usage they are combined for some particular extent. This caused for huge internet traffic loads and network congestion. On the other hand, the main reason is packet losses specially in UDP type protocols [3] used by industrial automation. Therefore, there is a need to consider probable network management solutions for the future convenience. According to [42], new protocol also defined to overcome network congestion. So different techniques also has been derived as a solution for network management in order to improve the quality of the network [24],[26]. This will be enhanced the accuracy and efficiency of the network [27]-[29].

The question of how to avoid packet losses and maximum utilization of the bandwidth. An efficient method to address network traffic issue is to monitor the network performance based upon long term network traffic analysis for non-identical data collection from different fields. This would facilitate to identify the network traffic patterns. Then could be proposed effective and fair solution for avoiding traffics in the network.

The traffic characteristic rely on when and where on the internet the traffic is investigated. The traffic behavior differs from in the different network to network and the traffic characteristics changes with new applications, new types of networks and with changing user behavior.

Figure 1.1 shows examples of internet traffic behavior. Both graphs are completely different on traffic distributions. So that two graphs are traffic distributions on

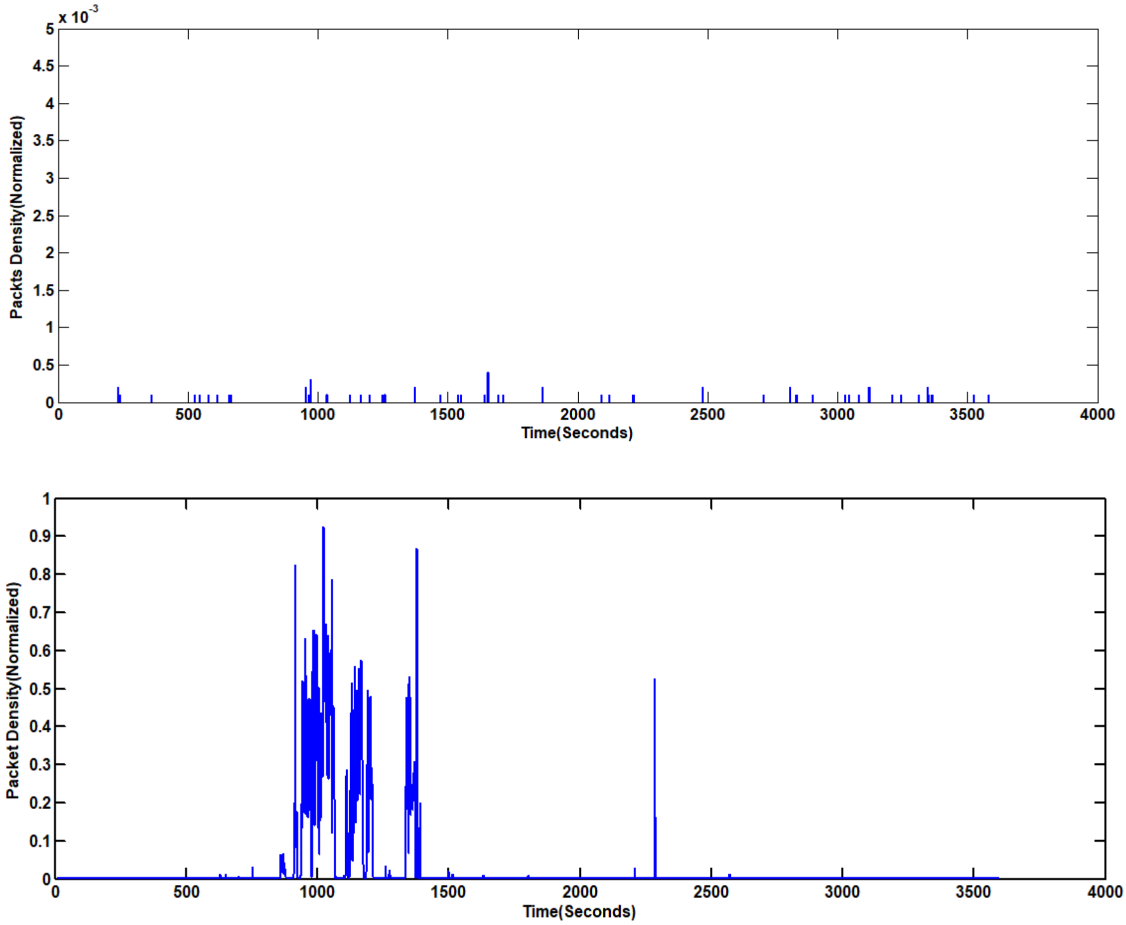


Figure 1.1: Example of Internet traffic behavior. Top: Packets per second during one hour for uploading(normalized). Bottom: packets per second during one hour for downloading(normalized).

uploading and downloading. The top one which indicates the uploading average traffic peak is about 0.0004 packet per second (PPS). The bottom graph for downloading average traffic peak is about 0.92 PPS and graph's peaks indicate some burstiness in the traffic distributions.

These spikes in the graph for downloading probably indicate periods when individual large files are downloaded or when a server on the network is accessed by a user who downloads multiple files. So peaks of the traffic may just reflect of the occupancy of the bandwidth [40]. So network planning by evaluating traffic [36], [37] is very essential to provide the best service without any barriers.

As per the figures traffic peak is low for uploading and while the downloading peaks remains quite high. It just reflects downloads have bursty properties and bandwidth greedy behavior, but don’t last long. In this situation has to be focused on internet traffic management. As revealed in [4] internet traffic management is avoiding network congestion and making good use of available network resources. By controlling the network congestion, it will cause to increase the performance of the network [43]. When a packet is lost this is recalled as network congestion and the transmission rate is decreased. So many techniques have been used to control the network congestion [44].

1.2 Problem statement

With the rapid development of network technology, and network information flow increasing network traffic control has become increasingly onerous [5]. The internet usage is more Automation Network in industrial automation and Enterprise Network are more or less combined [2]. This complex integration caused to develop network congestion and packet losses mainly occurs in industrial automation network due to UDP/IP.

So main role such a situation is network traffic control and management. Earlier if we can accurately predict the trend of traffic and then can be scheduled resources to reduce or avoid the occurrence of congestion, improve the utilization of network resources. However, establish the corresponding prediction model is the key to network traffic prediction. This mechanism directly can be applied for industrial automation network and enterprise network.

1.3 Aim and Objectives

Aim:

Investigation of the predictability of network traffic peaks in Enterprise Network to schedule non critical traffic from an Industrial Automation Network avoid peaks to reduce the congestion and maximize utilization.

Objectives;

i. Predict the traffic peaks in terms of level crossings of a given level in Enterprise Network.
ii. Analyze the collected data sets from Enterprise Network using statistical models.
iii. Predict the best statistical model to improve the analysis in future.

1.4 Contribution

As mentioned in the objectives, the analysis is carried out for Enterprise Network to identify the traffic behavior in different fields. The prediction of network traffic is examined in two ways. One is identified the traffic peak in terms of level crossings in a given level. Second is, analyzing using three statistical models as Probability Density Functions (PDF). Ultimately, these two methods are compared to predict the best statistical model.

1.5 Report Outline

The thesis consists of five chapters. The first chapter is the introduction which briefly introduces the basic concepts of traffic monitoring and analysis, the concept of traffic distributions in the network and the question to be addressed. Specifically, the need of the predictability of traffic peak is introduced in this chapter. Chapter two presents related work and background information relevant to this thesis including previous works in the area, related technologies, prediction of network traffic, traffic distributions and self-similarity. The methodology that used to analyze the data is described in chapter three. This chapter also introduces the different theorems which related on analysis, software to extract the data and processing. In the fourth chapter, the analysis that was performed is presented and the obtained result is interpreted in detail. The thesis project's results are given as conclusions in the fifth chapter, along with a discussion of possible future work.

CHAPTER 02

LITRATURE SURVEY

Chapter 1 is described that with the internet usage is more everyone has connected as globally. This has become a major issue of occurring network congestion as well as the packet loss of the data transmission. As a result of more internet usage traffic level is also become very high. So there are few of characteristics of the traffic in terms of packet sizes, flow duration and percentage composition by protocol and applications [30]. Also actual network is more and more complex characterized by long range dependence and self-similarity and so on [10]. To identify these generated traffic in the network mathematical tool also has been derived to make the smooth system [23].

Thus several techniques have been investigated by researches as the prediction of network traffic [45]-[50]. Using these different methods of predicting network traffic are facilitated to identify the behavior of network traffic and for more accurate planning in future demands. So this chapter will be discussed some techniques/methods regarding on network traffic prediction are done by researches.

2.1 Related Works

2.1.1 Study based on packet length and inter arrival time

When predict the network traffic statistical evaluation and analyzing are popular. Two parameters are used as the statistical characteristics of the network [6], [7],[38]. They are;

- Packet length (size)
- Inter arrival time

According to [6], a new statistical model has been used for packet length. The main contribution of this investigation was to propose a new Probability Density Function (PDF) for packet length which can be used to identify and classify internet traffic. Eworton has used the usual configuration for a Local Area Network (LAN) with internet as shown in Figure 2.1. Packet length was measured between the network server and the Internet connection as illustrated in Figure 2.1. He has mentioned that the traffic produced by the user presents Uniform distribution, Normal distribution and Beta distribution (Figure 2.1a,Figure 2.1b,Figure 2.1c).

Further when the data traffic flows through the aggregation point (router, gateway or server) it suffers a non-linear transformation (Figure 2.1d, Figure 2.1e) and finally produced the bimodal distribution (Figure 2.1f). So with bimodal distribution, he has proposed to model the packet length probability density function after the aggregation point. The proposed probability density function is compared with measurements presented in several articles [11]- [13]. According to those articles he used Tafvelin measurements, Rastin measurements, CAIDA measurements and sprint measurements and their Sum of Squares due to error (SSE), Root Mean Square Error (RMSE), R-square(RS) and Adjusted R-Square (ARS) were used to compare with proposed cumulative Distribution Function Model Beta Distribution, Exponential Distribution, Log-Normal Distribution, Pareto Distribution and Weibull Distribution. Further he has observed numeric and graphically beta distribution represents very good results by keeping SSE and RMSE values very close to zero. Ultimately he has concluded that beta distribution has a good fitting and that it performs better than the Exponential, Log normal, Pareto or Weibull distributions, for bimodal traffic. This result is important because one of these distributions may be used in network traffic simulators.

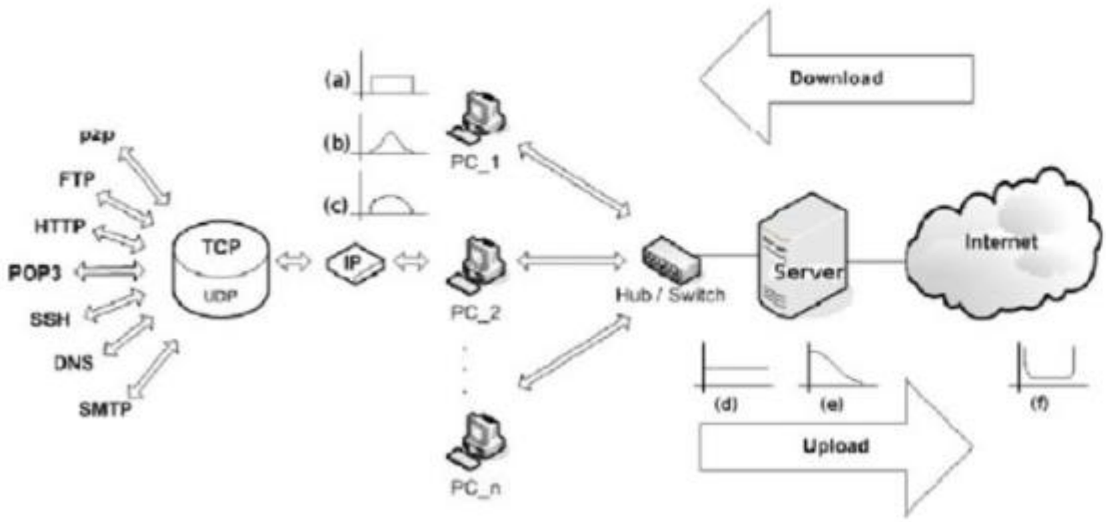


Figure 2.1: Packet aggregation by the network server

Source: Adapted from [6].

Inter-arrival time analysis and model

According to [7], packet inter arrival time follows Power Law and that can be modeled by heavy tail distribution (Pareto Distribution) and has also been presented hybrid mathematical model of packet size.

Sajjad has mentioned that for a finite observation of a point process can be easily generated a histogram that approximates the probability density function of inter – arrival time. He used the Origin Lab(OriginPro 7.5E) for further analysis and graphs. Then he plots the inter arrival time histograms (IIH) on log-log scale with bin size as shown in Figure 2.2.

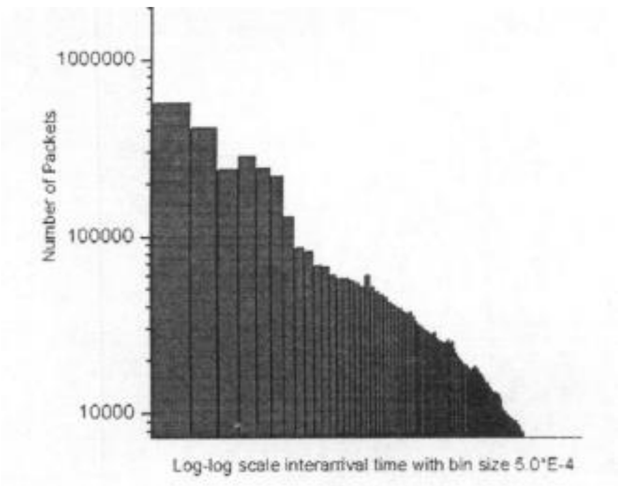


Figure 2.2: Packet inter arrival time histogram

Source: Adapted from [7]

Next computed the probabilities and plot the probability density function for packet inter-arrival time. For most results of this type plots he has found that a large part of the resulting graph can be fitted to straight line as shown in Figure 2.3.

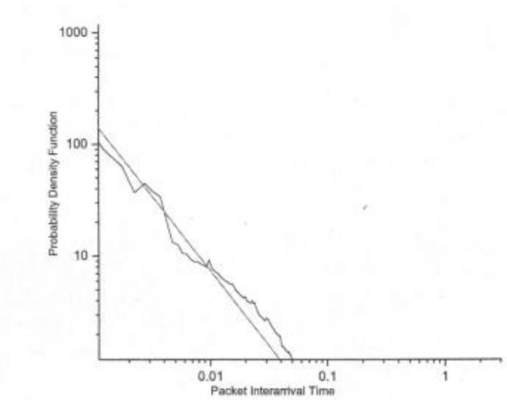


Figure 2.3: Log-log scale Iner-Arrival Time Plot for combined Data Set

Source: Adapted from [7]

This implies that there is a power law behavior of the probability density function

$y = p(x) = ax^b$. This means; $\log(y) = \log(a) + blog(x)$.

This refers to the heavy tail Pareto distribution by looking at the values of a and b. Further he has mentioned all the data set that he aggregated yield almost similar Probability Density Function.

Packet size analysis and model

Further Musthaq has done packet size analysis as well. Here he has used the Cumulative Distribution Function to analyze the data. Also he has used the segmentation algorithm in the case of extracting the probability density function from the cumulative distribution function.

2.1.2 Study based on self-similarity

The oldest models were based on simple probability distributions with the assumptions. For instant, Poisson traffic distributions frequently used as the traffic models [8],[31]. Most particularly, the earliest models like Poisson model ignored bursts completely [32]. They were derived by focusing on simplicity of analysis. These type models generally operated under the assumption that aggregating traffic from sources tended to smooth out bursts. According to [43], new algorithm is also defined to detect the burst. Apart from that when the network more and more complex it exhibits self-similarity behavior of the network. Self-similarity refers to distributions that shows the same characteristics at all scales [15]. For example, a self-similar network trace would look the same aggregated in 10ms bins as aggregated in 10second bins [14]. The following analysis on self-similar model which is widely used in network [10].

a) Fractional Brownian Motion(FBM) and Fractal Gaussian Noise

b) ON/OFF model of Heavy Tails Distribution

c) Fractional Autoregressive Integrated Moving Average (FARIMA)

According to Xiaoguang research he has mentioned about these three models like this.

- **FBM and FGN model**

  When estimating parameters such as Hurst parameter which describes the self-similar characteristic these two models can be used relatively simply. But these models are strictly self-similar and couldn't analyze the traffic of short term correlation structure very well.

- **ON/OFF model of Heavy Tails Distribution**

  He has investigated that when the file size is consistent with heavy tail distribution, the corresponding file transportation leads to the self-similarity of link layer. This kind of model helps to learn the nature of the self-similarity deeply. But it has its own disadvantages.

i) Each source must be independent

ii) Identically distributed

So the problem is most of the network distribution cannot be built on this premise.

- **FARIMA**

Here FARIMA model is used to fit to the actual traffic trace and then calculate the required parameter for prediction of the traffic. FARIMA model can describe the long-range dependence characteristics of network traffic effectively. Meanwhile it can represent traffic with short range dependence structure very well. However, the main issue of the model is, computation quantity of this algorithm is too large.

CHAPTER 03

METHODOLOGY

This chapter explains about the data analysis part of the research. The key part of this research is predicting traffic peaks in the Enterprise Network. Therefore, for the analysis several data have been collected from different networks as Institutional data and Individual data. In this research three statistical models are introduced for predicting the traffic peaks. Therefore, as the statistical models Pareto Distribution, Beta-Prime Distribution and Exponential Distribution are used to represent the distribution of traffic.

3.1 Data Collection

Collected data sets for analysis summarized as shown in the table 3.1

Table 3.1: Summary of data collection

Source	No.Samples	Duration of traffic slot	No.of Packets
i) Single Machine	100	1hr	6.2M pkts
ii) Downlink	100	1hr	2.1M pkts
iii)Uplink	100	1hr	1.5M pkts
iv)Institutes	73	1second	85.5G pkts
	73	6second	125.7G pkts
	73	24second	167.1G pkts
	73	288second	244.4G pkts

3.1.1 Institutional Data collection

Institutional Data is collected through backbone link of the Institutional Network and which is captured using RRD tool. There seventy-three institutes are considered to collect the required data for analysis. These data have been collected in different time scales as within 1 second, 6 second, 24second and 288 second as shown in the table 3.1.

3.1.2 What is RRD tool.

RRD tool is a data base which is correlated with time series data like network bandwidth utilization. The main function of this data base rather than storing data is creation of the graph according to the stored data.

3.1.3 Using RRD tool for capturing the data

Basically there are three main steps to set RRD tool for graphing according to the data sets.

1. Initialize the data base
2. Collect the data sets over time
3. Create the graph

3.1.4 Capture the graph through RRD tool

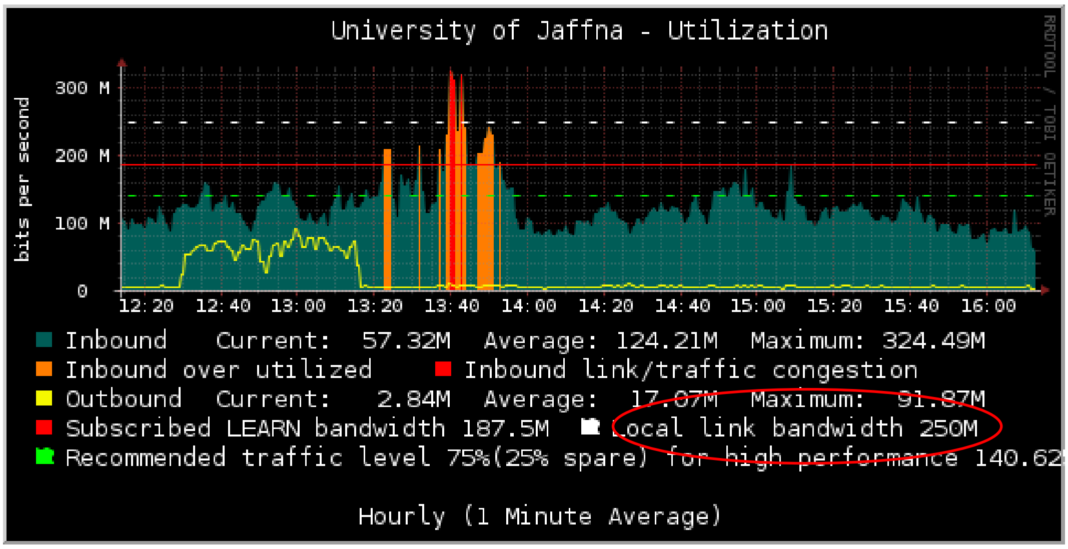


Figure 3.1: Captured graph for bandwidth utilization

The graph reveals that bandwidth utilization by the appropriate Institute. There Local Link Bandwidth is 250M. The graph clearly indicates that the time when it takes peak spikes. These spikes in the graph it probably indicates when the large files are downloaded or multiple files are requested.

The Local Link Bandwidth (LLB) for seventy-three institutes is included in Table 3.2.

Table 3.2: LLB for Institutes

Inst: No	LLB	Inst: No	LLB	Inst: No	LLB	Inst: No	LLB	Inst: No	LLB
1	10M	16	100M	31	10M	46	20M	61	5M
2	50M	17	10M	32	1000M	47	700M	62	5M
3	80M	18	20M	33	10M	48	5M	63	5M
4	400M	19	150M	34	20M	49	5M	64	20M
5	50M	20	150M	35	50M	50	5M	65	10M
6	50M	21	800M	36	100M	51	5M	66	100M
7	150M	22	20M	37	100M	52	5M	67	20M
8	80M	23	1000M	38	200M	53	5M	68	150M
9	50M	24	10M	39	150M	54	5M	69	150M
10	10M	25	20M	40	150M	55	5M	70	20M
11	20M	26	20M	41	100M	56	5M	71	150M
12	50M	27	80M	42	250M	57	5M	72	150M
13	80M	28	20M	43	300M	58	5M	73	150M
14	100M	29	150M	44	80M	59	5M		
15	250M	30	50M	45	150M	60	5M		

3.15 Individual Data Collection

Individual data is collected from a Wireshark installed on a single computer. This single machine data consists with one-hour traffic slots. Wireshark application is used to filter that downlink and uplink data.

3.16 Wireshark application

Wireshark tool is a packet analyzer. It is used for capturing packets and filtering them. In addition to that can be used to analyze the traffic flow on the network. So Wireshark tool is used in this research to filter out downlink and uplink data.

3.17 Using Wireshark tool for capturing data

Following figure that represents how is the Wireshark is showing details are appearing on Wireshark according to the time.

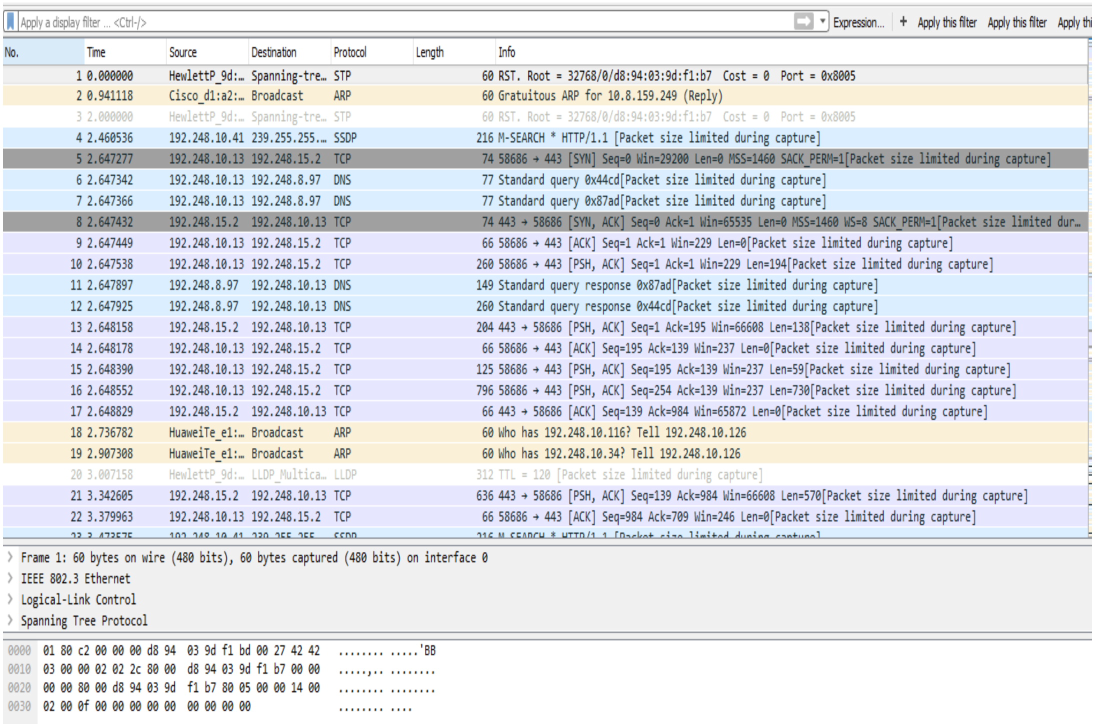


Figure 3.2: Wireshark is capturing the data

3.2 Background Analysis

3.2.1 Rearranging data in feasible way for analysis

It is very important preprocess data what collected before analyzing large volume of traffic flow [25],[41]. Single Machine Data set is main tool which is used for extracting the data for uplink and downlink. Single Machine Data is consisting as the PCAP files. So used Wireshark network protocol analyzer to extract the data [16]. This data file contains one-hour traffic slot.

No.	Time
1	0.000000
2	0.381457
3	0.468891
4	0.477152
5	0.477286
6	0.551520
7	0.581626
8	0.581650
9	0.645445
10	0.645947
11	0.659141
12	0.703819
13	0.776988
14	0.784717
15	0.825279
16	0.859892
17	0.919841
18	0.954510
19	1.000538
20	1.047549
21	1.234987
22	1.235112
23	1.235273
24	1.287996

Figure 3.3: Extracted data from PCACP files Wireshark Network Protocol Analyze

Single machine data use to extract the data for uplink and downlink. Using by the Wireshark network analyzer filter tool, extracted data for uplink and downlink as shown in Figure 3.2 and Figure 3.3.

- For up-link data, filter is applied as " tcp and ip.src==192.248.10.13
- For down-link data, filter is applied as " tcp and ip.dst==192.248.10.13

No.	Time
27	1.359868
28	1.360166
29	1.360192
32	1.362257
172	7.349448
177	7.349639
178	7.349725
183	7.350376
184	7.350568
185	7.350693
189	8.114950
260	13.077264
261	13.077307
264	13.077599
673	54.480743
675	54.522244
676	54.522399
679	54.564752
680	54.564967
681	54.565140
686	54.651000
708	59.148209
710	59.196296
772	63.091759

4 packts within 1 second

Figure 3.4: Extracted data for uplink

No.	Time
30	1.362170
31	1.362231
33	1.362264
176	7.349620
182	7.350356
186	7.350955
188	8.076257
262	13.077506
263	13.077565
674	54.522201
677	54.564210
678	54.564719
684	54.606747
685	54.612079
709	59.196232
773	63.133450
776	63.175620
777	63.176841
782	63.221013
783	63.237644
786	63.365802
789	63.571450
790	63.571474
792	63.571486

3 packts within 7 th second

Figure 3.5: Extracted data for downlink

Single machine data is used to extract downlink and uplink data for 1hr time slots. Data consists with its serial Number and time. So, the time defines no of data packets flow through the link within that time duration. These extracted results used to count no of data packets flow in each seconds. Using by the MATLAB code counted no of packets for each seconds. The Matlab code regarding on counting no. of data packets is attached in Appendix A.

time(s)	No.of pkts
0	18
1	24
2	36
3	37
4	16
5	19
6	16
7	21
8	6
9	10
10	13
11	9
12	30
13	25
14	24
15	12
16	16
17	22
18	11
19	22
20	6
21	16

Figure 3.6: Counting No. of pkts in each seconds.

As per the Figure 3.4 it shows distribution of the data packets within that time series. Also can be determined how is that data packets are distributed with packet inter-arrival time. So for that purpose compute the difference of each consecutive packets. These two distributions are illustrated in Figure 3.5 and Figure 3.6.

Let's take that packet arrival event happening at times t_i,

Then packet inter-arrival time Δt_i;

$$\Delta t_i = t_{i+1} - t_i \qquad (1)$$

According to equation 1 difference of each consecutive packets are computed as shown in Figure 3.7. Also these same steps were carried throughout the other two data sets (Downlink and Uplink Data).

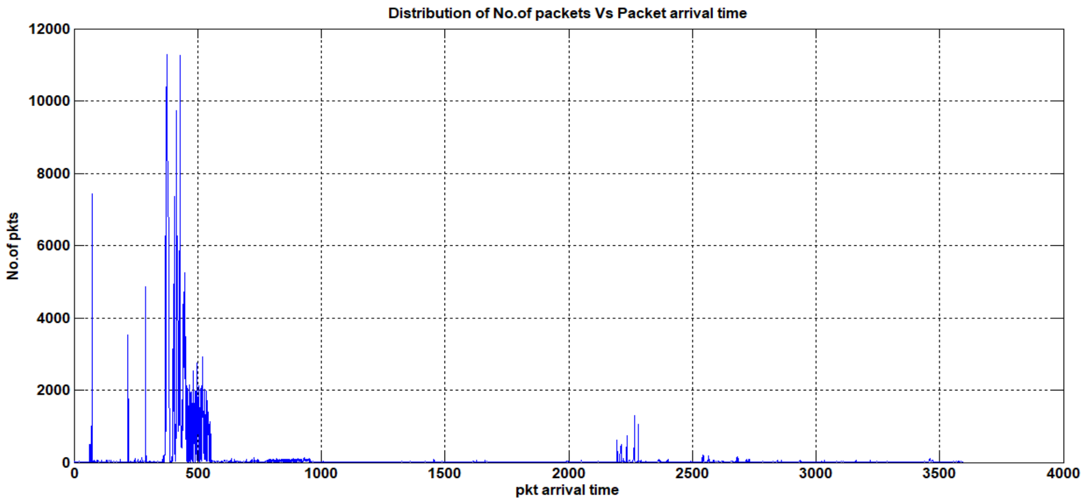


Figure 3.7: Distribution of packets with its arrival time

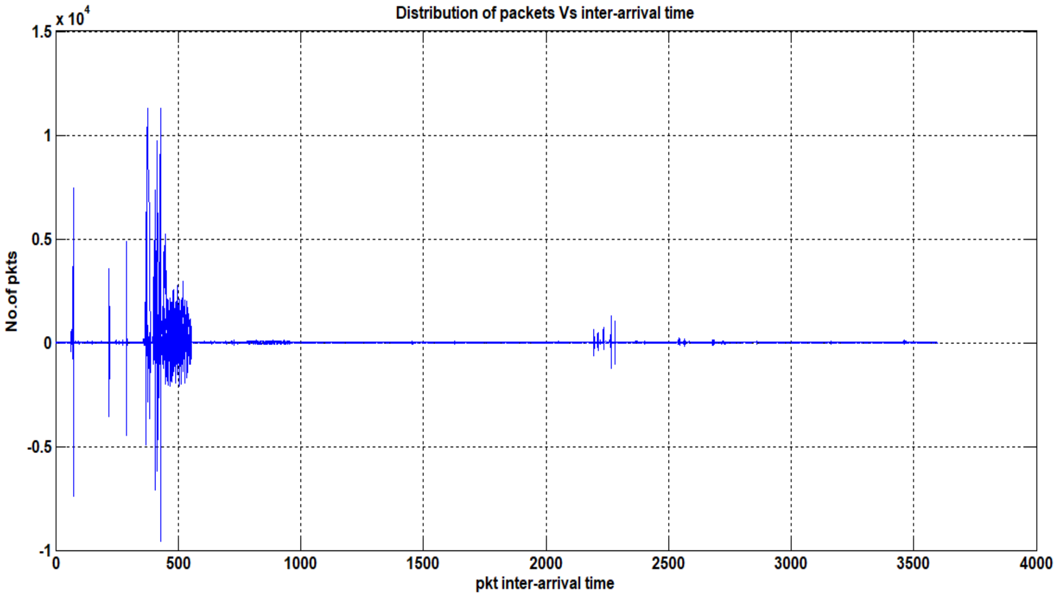


Figure 3.8: Distribution of packets with its packet inter-arrival time

time(s)	No.of pkts	diff:
0	18	6
1	24	12
2	36	1
3	37	-21
4	16	3
5	19	-3
6	16	5
7	21	-15
8	6	4
9	10	3
10	13	-4
11	9	21
12	30	-5
13	25	-1
14	24	-12
15	12	4
16	16	6
17	22	-11
18	11	11
19	22	-16
20	6	10
21	16	18

Figure 3.9: Computing the difference of each consecutive pkts.

3.2.2 Developing variables

According to the data which is mentioned as in Figure 3.7 can be defined as two data series with the time. One data series is data packet distribution with the packet arrival time and other data set is defined as data packet distribution with packet inter arrival time. Initially before modeling the prediction model it is essential to look at the correlation of this two data series. So It is very essential to use hypothetical test for categorical variables to exam whether two variables are independent or not. In this case Chi-Square test for independence is used as the hypothetical test.

3.2.3 chi-square test for independence

The Chi-Square Test for independence of two variables [17], [23] is a test which uses a cross classification table to examine the nature of the relationship between these variables. The test for independence examines whether the observed pattern between the variables in the table is strong enough to show that the two variables are dependent on each other or not.

The test for independence of No. of packets(X) and difference of consecutive packets(Y) begins by assuming that there is no relationship between the two variables. The alternative hypothesis states that there is some relationship between two variables. If the two variables in the cross classification are X and Y the hypotheses are;

H_0: no relationship between X and Y

Ha: some relationship between X and Y

In terms of independence and dependence this hypothesis could be stated

H_0: X and Y are independent

Ha: X and Y are dependent

Chi-Square Equation

$$E_{i,j} = \frac{\sum_{k-1}^{c} O_{i.j} \ \sum_{k-1}^{r} O_{k,j}}{N}$$

Where,

$E_{i.j}$ = expected value

$\sum_{k-1}^{c} O_{i.j}$=sum of i^{th} coulumn

$\sum_{k-1}^{r} O_{k,j}$=sum of k^{th} row

N=total number

$$x^2 = \sum_{i-1}^{r} \sum_{j-1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

$x^2 = \text{chisquare} - \text{test of indpendence}$

$O_{i,j}$ = Observed value of two nominal variables

$E_{i,j}$=Expected value of two nominal variables

Degree of freedom is calculated by using following formula,

DF=(r-1) (c-1)

Where DF=degree of freedom

r=number of rows c=number of columns

reject the null hypothesis = calculated chi-square value > tabulated chi-square value.

3.2.4 Applying chi-square test for data set

State the hypothesis:

H_0: No.of pkts and difference of pkts are independent.

H_a: No of pkts and difference of pkts are not independent.

Analyze Sample data:

Table 3.3: Computing required values for Chi-Square Test

Variable1	Variable 2	Total(row)
18	6	24
24	12	36
36	1	37
37	21	58
16	3	19
19	3	22

16	5	21
21	15	36
6	4	10
10	3	13
13	4	17
9	21	30
30	5	35
25	1	26
24	12	36
12	4	16
16	6	22
22	11	33
11	11	22
22	16	38
6	10	16
16	18	34
Column total =409	Column total = 192	Column total = 601

Calculating the expected values

$$E_{i,j} = \frac{\sum_{k-1}^{c} O_{i.j} \ \sum_{k-1}^{r} O_{k,j}}{N}$$

$E_{1,1}$ = (24*409) /601

= 16.332

Calculating the observed chi-square values

$$x^2 = \sum_{i-1}^{r} \sum_{j-1}^{c} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

$X_1^2 = (18\text{-}16.332)^2/16.332$

$= 0.17035$

Calculated Chi-square value = $x^2 = x_1^2 + x_2^2 + x_3^2 + \ldots\ldots = 10.08566$

df= (22-1) (2-1) =21

let consider 95% level of confidence.

critical chi-square value for df=21 and at 0.05 =32.671

Calculated chi-square value < table chi-square value

So it rejects the alternative hypothesis.

Then can be expected the null hypothesis.it means that two variables are independent.

3.2.5 Estimating of Probability Density Functions

As mentioned in the beginning of the chapter 3 the main purpose of the research is predicting the traffic peak in the network. When make the prediction has to be worked with few of formulas as Joint Probability Formula and Rice's Formula. These formulas are integrated with PDF. In order to develop appropriate PDF needs to obtain the histograms for two independent variables. Histograms for two variables are shown in Figure 3.8 and Figure 3.9.

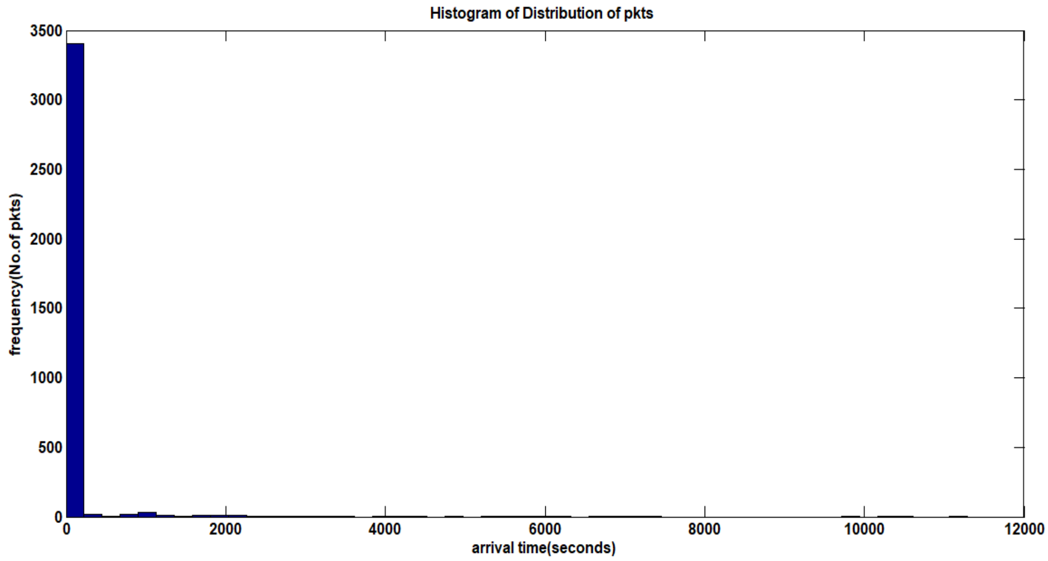


Figure 3.10: Histogram for packet distribution

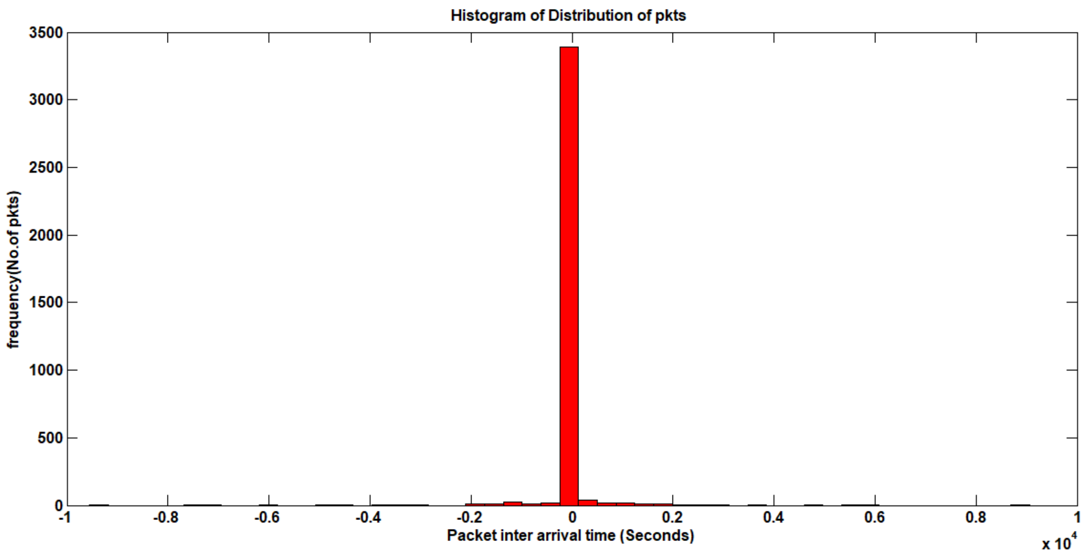
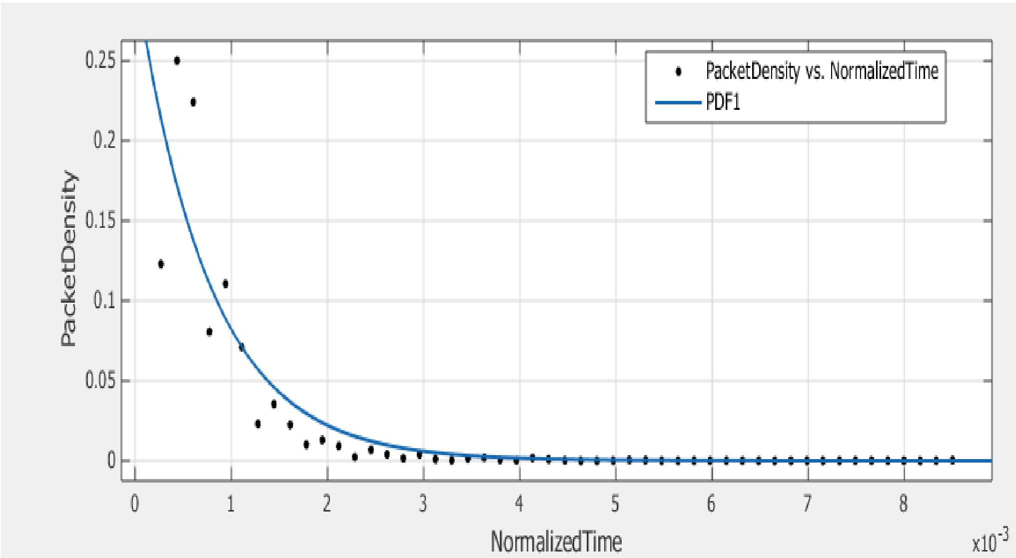


Figure 3.11: Histogram for packet inter arrival time

After getting that two histograms for two independent variables it’s essential to find the appropriate PDF for each histograms. Method A and Method B indicates how PDF are statistically modeled for histograms.

Method A: Using Matlab Curve Fitting Tool

In order to obtain the PDF of the histogram which is shown as Figure 3.8 used Matlab curve fitting tool to estimate the function. According to the pattern of distributed points most appropriate function is determined as the exponential curve. Figure 3.10 indicates how that curve is fitted with distributed points.

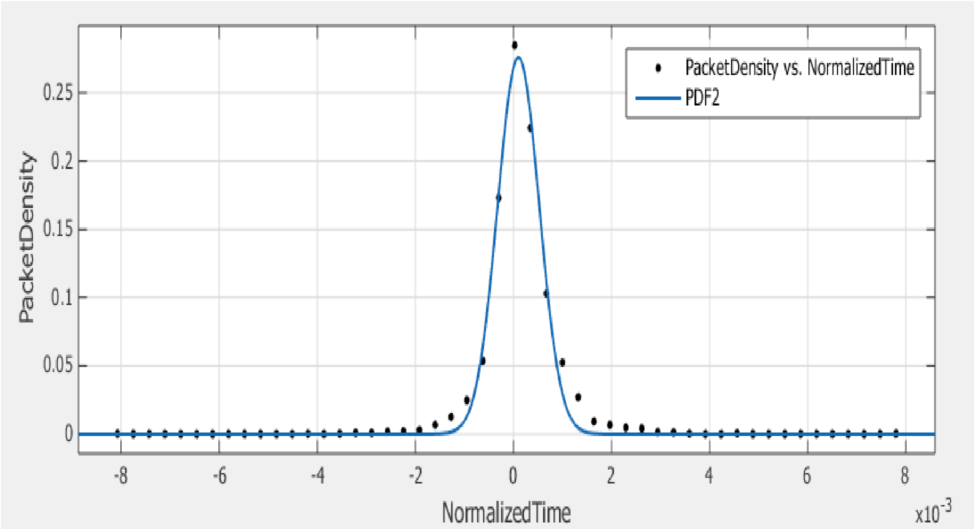


General model Exp1:
  f(x) = a*exp(b*x)
    where x is normalized by mean 0.004384 and std 0.002449
Coefficients (with 95% confidence bounds):
    a = 0.0009772 (-0.0003036, 0.002258)
    b = -3.207 (-4.045, -2.369)

Goodness of fit:
  SSE: 0.02568
  R-square: 0.809
  Adjusted R-square: 0.805
  RMSE: 0.02313

Figure 3.12: Exponential function as PDF

As shown in Figure 3.9 most appropriate function is estimated as the Gaussian function as illustrated in Figure 3.11.




```
General model Gauss1:
  f(x) = a1*exp(-((x-b1)/c1)^2)
Coefficients (with 95% confidence bounds):
    a1 =     0.2765  (0.266, 0.2869)
    b1 =  0.0001006  (8.213e-05, 0.0001192)
    c1 =     0.0006  (0.0005738, 0.0006262)

Goodness of fit:
  SSE: 0.001963
  R-square: 0.9877
  Adjusted R-square: 0.9872
  RMSE: 0.006462
```

Figure 3.13: Gaussian function as PDF.

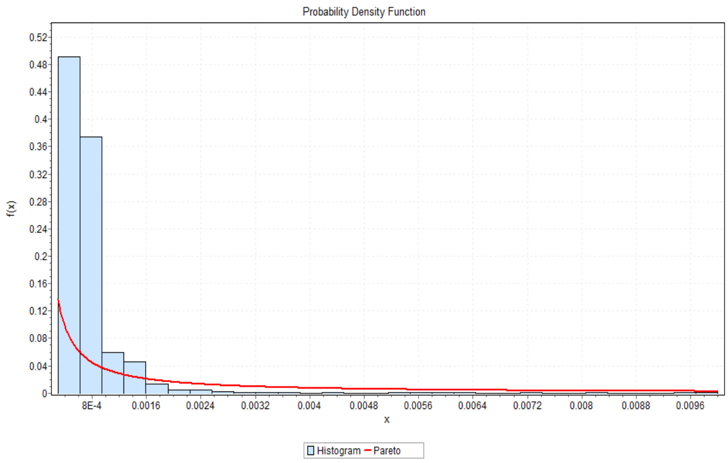
Obtained results from curve fitting method are tabulated as shown in the Table 3.2.

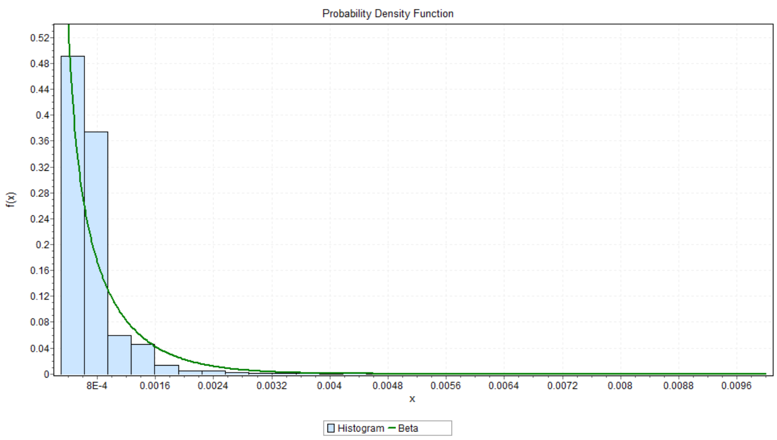
Table 3.4: Functions with appropriate parameters

Curve fitting function	**Parameter.1**	**Parameter.2**	**Parameter.3**	**Function**
Exponential	a=0.0009772	b=-3.207	-	$f_1(x) = a.e^{b.x}$
Gaussian	a1=0.2765	b1=0.0001006	c1=0.0006	$f_2(x) = a1.e^{-(x-b1)^2/c1^2}$

Method B: Distributing fitting using Easy Fit software

In order to get PDF for two independent variables, used an another method. There are some advantages of using method B using by Easy fit software [19] as it has various PDFs for distributing fittings and also easily can be plot the appropriate distributing fittings with its parameters. Four types of probability distribution models are selected to detect the best distribution model. So specially determined that pareto distribution, beta prime distribution, exponential distribution and normal distribution for two independent variables. Obtained distributions models are shown below with their parameters.





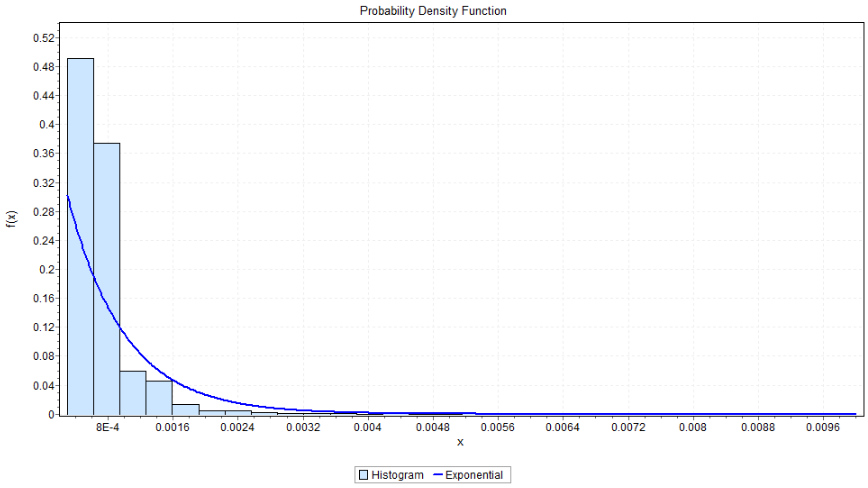


Figure 3.14: Probability Distribution models for packet distribution. Top: Pareto Distribution. middle: Beta-Prime Distribution. bottom: Exponential Distribution.

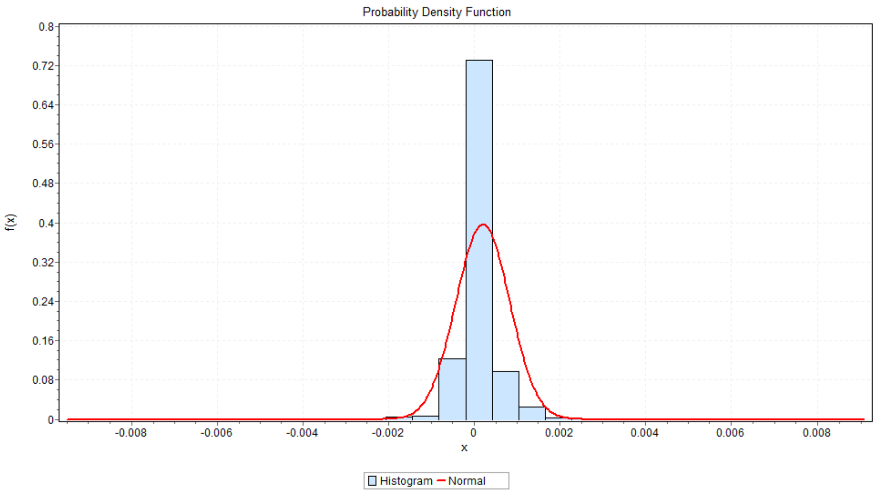


Figure 3.15: Probability Distribution Model for distribution of packets within inter arrival time. Normal Distribution

In order to obtain the best PDFs 30 number of bins is selected from data.

3.2.5.1 verify the graphs obtain through the easy fit software using MATLAB distributing fitting tool

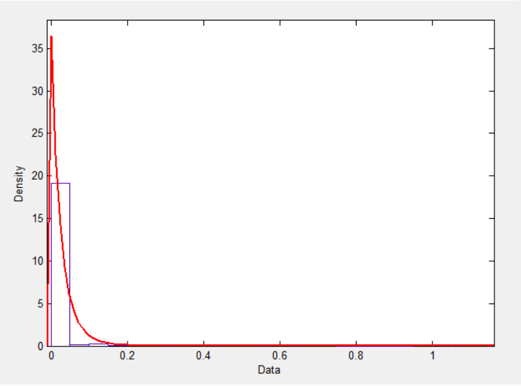


Figure 3.16 : Pareto distribution

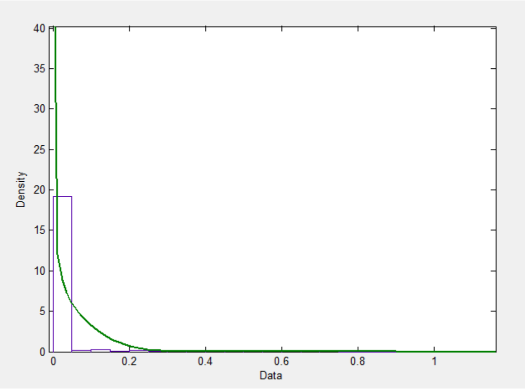


Figure 3.17 : Beta distribution

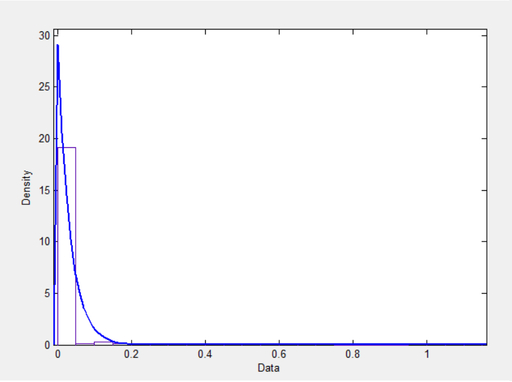


Figure 3.18 : Exponential distribution

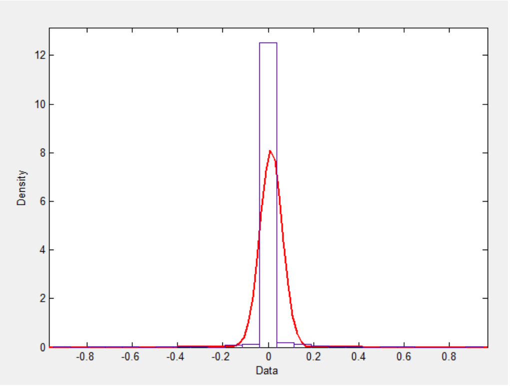


Figure3.19: Normal distribution

MATLAB results also provided parameters for distributions. Following factors indicate the evaluation of data for generate proper graph of PDFs.

No.of bins of data = 30

Bin size = 0.2

Table 3.5: PDF with its parameters

Distribution Model	Parameter1	Parameter2	Parameter3	PDF
Pareto $P_1(x)$	α(shape) = 0.12546 $\alpha > 0$	β(scale)=0.000298 $\beta > 0$	-	$P_1(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}$
Beta-Prime $P_2(x)$	α_1(shape) =0.57249 $\alpha_1 > 0$	α_2(shape)=0.13395 $\alpha_2 > 0$	B (beta function)	$P_2(x) = \frac{x^{\alpha_1-1}(1+x)^{-\alpha_1-\alpha_2}}{B(\alpha_1,\alpha_2)}$
Exponential $P_3(x)$	λ $(scale)$=1434.6 $\lambda > 0$	-	-	$P_3(x) = \lambda e^{-\lambda x}$
Normal $P_4(y)$	μ(location) = 0.00019381	σ^2(scale) =0.00062371	-	$P_4(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$

3.2.6 Joint probability formula and its relevancy in this research

According to [18] if two continuous random variables are independent

$$f_{X,Y(x,y)} = f_X(x).f_Y(y) \qquad (2)$$

According to the chi-square test it proved both variables are independent. Joint probability formula represents the product of PDFs of two variables. Joint Probability Density Function is computed using MATLAB code. The code is attached under Appendix A.

For an instance, Joint Probability Density Formula (JPF) for Pareto Distribution and Normal Distribution,

$$P_{1,4}(x,y) = P_1(x).P_4(y) \qquad (3)$$

$$P_1(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}} \;\; and \;\; P_4(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$$

$$P_{1,4}(x,y) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2 / 2\sigma^2} \qquad (4)$$

$$= \frac{\alpha\beta^{\alpha} x^{-(\alpha+1)}}{\sqrt{2\pi\sigma^2}} \cdot e^{-(y-\mu)^2 / 2\sigma^2} \qquad (5)$$

According to equation 2 can be derived JPF using other PDFs. After getting that JPF the graph is shown in Figure 3.14.

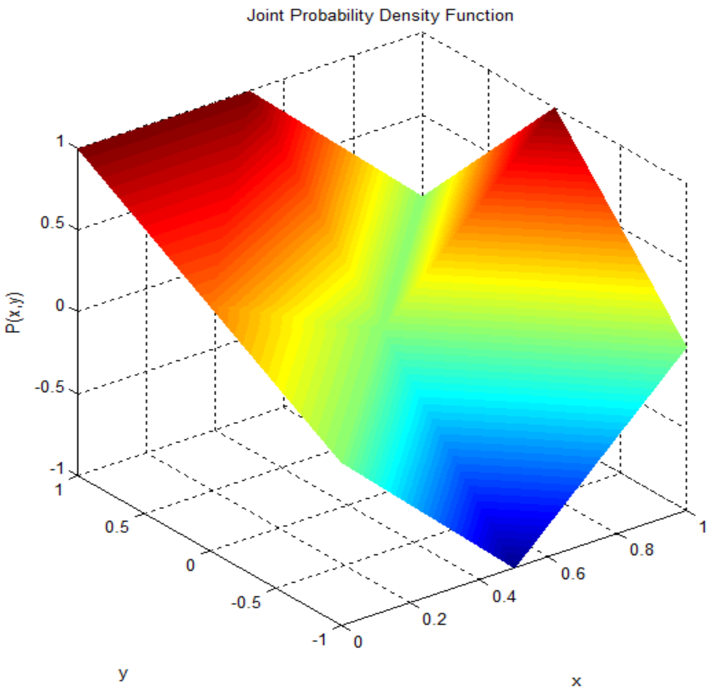


Figure 3.20: JPF for Pareto and Normal Distribution

3.2.7 Rice's Formula

Rice's formula counts the average number of times stochastic stationary process X(t) per unit time ($t \in [0,1]$) crosses a fixed level u [20],[34] . Then Rice's formula states that,

$$v_{(u)} = \int_{-\infty}^{\infty} |\dot{x}|\, P_{x,\dot{x}}\, (u,\dot{x}) d\dot{x} \qquad (6)$$

Here,

$v_{(u)}$ = average number of times stochastic process$x(t)$ crosses the fixed level u.

$\dot{x}$ = first derivative of x with respect to time.

$p(x, \dot{x})$ = joint probability density of $x(t)$ and $x(\dot{t})$ at time t.

3.2.8 Predictability of level crossings.

As mentioned in chapter 1 predicting traffic peak is identified as the one of main objective as it integrates with future developments of the network. In this point predicting traffic peak is recognized in terms of the level crossings. These levels are considered as 0.7,0.8,0.9 and 0.95 times of the peak level. That is emphasized number of times these levels are crossed by stochastic process (traffic distribution). That predictability can be simplified using by the Rice's formula as indicated under equation 6.

According to equation 6;

Let's take that fixed level as u. Then;

$$v_{(u)} = \int_{-\infty}^{\infty} |\dot{x}|\, P_{x,\dot{x}}\, (u, \dot{x}) d\dot{x}$$

$$x(t) = P_1(x) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}}$$

$$\dot{x}(t) \quad = P_4(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2 / 2\sigma^2}$$

$$v_{(u)} = \int_{-\infty}^{\infty} |y|\, P_1(u). P_4(y) dy$$

$$= \int_{-\infty}^{\infty} |y| \frac{\alpha\beta^{\alpha}}{u^{\alpha+1}} . \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2 / 2\sigma^2} dy$$

$$v_{(u)} \quad = \int_{-\infty}^{\infty} |y| \frac{\alpha\beta^{\alpha} u^{-(\alpha+1)}}{\sqrt{2\pi\sigma^2}} . e^{-(y-\mu)^2 / 2\sigma^2} dy \qquad (7)$$

As solve in equation 7 predictability of level crossing can be determined for other JPFs as well. Computing of level crossing has been done using Matlab and that code is attached in Appendix A.

3.2.9 Actual level crossing rate

Statistical models of distributing (Pareto, Beta-Prime, Exponential and Normal distributions) are used for predicting the traffic peak. So it is essential to determine the most appropriate distributing fitting among the above statistical models for making the future plans on network with that statistical model. So the best statistical model can be specified by comparing each statistical model output with the actual output. "Output" is referred to as level crossing rate of the traffic distribution.

Figure 3.15 reveals of estimating the number of actual level crossings for stochastic process.

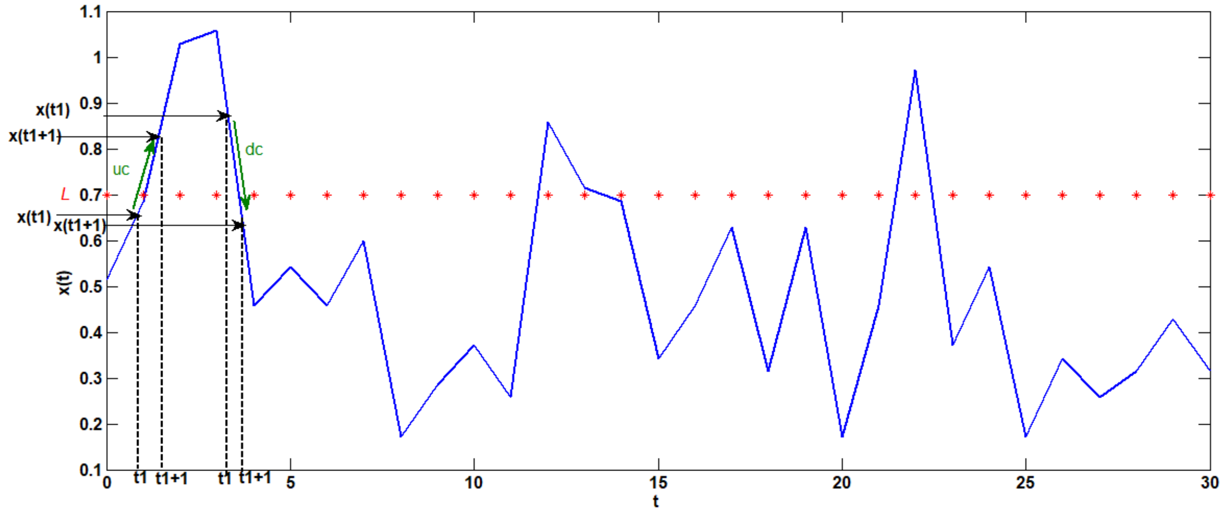


Figure 3.21: Counting actual number of level crossings.

Two constraints are derived for counting that number of up crossings and down crossings.

Constraint 1: $if\ x(t_1) < L\ \ and\ x(t_1\ +1) \geq L \dots\dots\dots\dots ndc = ndc + 1$

or

Constraint 2: $if\ x(t_1) > L\ \ and\ x(t_1\ +1) \leq L\ \ \dots\dots\dots nuc = nuc + 1$

According to these constraints the Matlab code is attached in Appendix A.

When specifying the best statistical model for predicting traffic peak it's essential to determine the precision of the obtained results through the prediction. The precision of the predicted results can be compared with the actual results what has been obtained for traffic distribution. This precision of the results is determined using error formula as mentioned equation 8 [21].

$$\%error = \left|\frac{experimental\ value - theoritical\ value}{theoritical\ value}\right| \times 100 \qquad (8)$$

According to equation 8, experimental value is the predicted results that is obtained through equation 7. And also theoretical value is the results that is obtained through constraint1 and constraints 2. The calculated error is included in chapter 4 results and analysis.

3.2.10 Institutional data analysis

As mentioned in beginning of the chapter the background analysis is done for single machine data, up-link data and down-link data. This analysis is facilitated to decide the most appropriate method from method A and method B. Method B is the best method for analysis due to huge error occurred in method A. Average error for method A is 96.87%. Then method A is rejected and choose method B for Institutional data analysis.

Further analysis is verified to find the best distributing fitting model for traffic distribution. Method B is applied for single machine data, uplink data, down-link data and institutional data. So the results of these sources will be identified the best distributing fitting model.

CHAPTER 04

RESULTS

As mentioned in chapter 3 data analysis has been done for collected data sets such as single machine data, up-link data, down-link data and Institutional data. The analysis is carried through Matlab and Easy fit software. Statistical models are developed using by those soft wares for predicting the traffic. The results are described in this chapter.

4.1 Background analysis

Manily background analysis is used for analysis in single machine data and data extracted from it such as uplink- data and down-link data. This analysis is done using two ways as mentioned in chapter 3 method A and method B.

4.1.1 Method A – using matlab curve fitting tool

Matlab curve fitting tool is used to obatain the parameters for PDFs. Those results were tabulated in Table 4.1.

Table 4.1: Single Machine Data-predicted level crossing rate

Level	SD1	SD2	SD3	SD4	SD5	SD6	SD7
0.7	1.53702	0.92923	2.7264E-06	0.07272	0.124069	3.40442E-54	1.2309E-54
0.8	1.28393	0.62357	1.5718E-07	0.05226	0.014124	1.74523E-50	8.97974E-62
0.9	1.08521	0.42687	1.0140E-08	0.04841	0.014617	1.78991E-68	1.5305E-68
0.95	0.00159	0.35563	2.6844E-09	0.34982	0.008852	2.3120E-72	8.47094E-72

4.1.2 Actual crossing rate

Traffic distribution characteristic is used to obtain the actual crossing rate for same single machine data SD1, SD2, SD3, SD4, SD5, SD6 and SD7.

Table 4.2: Actual crossing rate for single machine data

Level	SD1	SD2	SD3	SD 4	SD5	SD6	SD7
0.7	46	16	0	1	3	0	0
0.8	48	16	0	1	2	0	0
0.9	41	15	0	1	2	0	0
0.95	41	15	0	0	2	0	0

4.1.3 comparison of results method A

According to equation 8 error percentage is calculated as described in Table 4.3. This comparison is done for the single machine data analysis for using by the curve fitting tool.

Table 4.3: Single machine data - method A results comparison

Sample No:	Level	Actual Level Crossing rate	Predicted Level Crossing rate	Error
SD1	0.7	46	1.53702	96.65865
	0.8	48	1.28393	97.32514
	0.9	41	1.08521	97.35314
	0.95	41	0.00159	99.99611
SD2	0.7	16	0.92923	94.19234
	0.8	16	0.62357	96.1027
	0.9	15	0.42687	97.15421

	0.95	15	0.35563	97.62911
SD3	0.7	0	2.7264E-06	NaN
	0.8	0	1.5718E-07	NaN
	0.9	0	1.0140E-08	NaN
	0.95	0	2.6844E-09	NaN
SD4	0.7	1	0.07272	92.728
	0.8	1	0.05226	94.774
	0.9	1	0.04841	95.159
	0.95	0	0.34982	NaN
SD5	0.7	3	0.12407	95.86437
	0.8	2	0.01412	99.2938
	0.9	2	0.01462	99.26915
	0.95	2	0.00885	99.5574
SD6	0.7	0	3.40442E-54	NaN
	0.8	0	1.74523E-50	NaN
	0.9	0	1.78991E-68	NaN
	0.95	0	2.3120E-72	NaN
SD7	0.7	0	1.2309E-54	NaN
	0.8	0	8.97974E-62	NaN
	0.9	0	1.5305E-68	NaN
	0.95	0	8.47094E-72	NaN

Average error for single machine data using by method A is 96.87%. The error is usually very huge. Then method B is used for analysis of single machine data and extracted data from single machine data as up-link data and down-link data.

4.1.4 Method B: using Easy Fit software

Method B is analyzing data using Easy Fit software. Distributing fittings are considered as Pareto, Beta-Prime, Exponential and Normal distributing.

4.1.5 Predicted Level Crossing Rate

Table 4.4: Predicted results of Pareto distributing

Predicted Level Crossing Rate							
Pareto Distribution							
Level	SD1	SD2	SD3	SD4	SD5	SD6	SD7
0.7	54.776	29.27	3.1358	0.86621	1.6296	0.44281	0.113127
0.8	52.473	28.741	2.6624	0.74107	1.5654	0.37621	0.096681
0.9	50.74	24.61	2.3045	0.64579	1.3675	0.32584	0.084171
0.95	50.028	22.918	1.1567	0.60625	1.0867	0.30503	0.078983

Table 4.5: Predicted results of Beta-Prime distributing

Predicted Level Crossing Rate							
Beta-prime Distribution							
Level	SD1	SD2	SD3	SD4	SD5	SD6	SD7
0.7	1.28266	3.045814	0.039483	0.76874	0.184434	0.084824	0.070252
0.8	1.075738	2.461568	0.032487	0.431696	0.158647	0.067994	0.055843
0.9	0.91851	2.029247	0.027245	0.252178	0.138758	0.055629	0.045325
0.95	0.853482	1.85398	0.0251	0.195253	0.130437	0.050642	0.041103

Table 4.6: Predicted results for Exponential distributing

Predicted Level Crossing Rate							
Exponential Distribution							
Level	SD1	SD2	SD3	SD4	SD5	SD6	SD7
0.7	3.25E-03	1.222691952	4.00E-03	3.56E-06	0.001977263	2.99E-06	3.69E-06
0.8	1.72E-04	0.425689507	6.05E-04	7.07E-07	0.000281201	2.11E-07	3.76E-07
0.9	9.07E-05	0.148207041	9.16E-05	1.40E-08	3.99916E-05	1.49E-08	3.87E-08
0.95	2.08E-06	0.08744939	1.13E-06	1.97E-10	1.50815E-05	1.25E-09	3.87E-09

Determined distributing fittings such as Pareto, Beta Prime, Exponential and Normal distributing are applied for uplink data which is extracted from single machine data.

Table 4.7: Predicted level crossing rate for uplink data (UD)

Predicted Level Crossing Rate						
	Pareto Distribution		Beta-Prime Distribution		Exponential distribution	
Level	UD1	UD2	UD2	UD3	UD4	UD5
0.7	1.049221	3.6893	1.107467734	0.18382	0.301796227	2.78E-03
0.8	1.042276	3.117	0.698500095	0.15829	0.10497813	1.24E-04
0.9	1.036968	0.6864	0.395947726	0.13858	0.036516056	5.57E-05
0.95	1.03476	0.5091	0.273362905	0.13033	0.021536562	1.18E-06

Table 4.8: Pareto distribution for downlink data (DD).

Predicted Level Crossing Rate				
Pareto Distribution				
Level	DD1	DD2	DD3	DD4
0.7	22.3315	7.650685098	0.42721	0.543250244
0.8	18.8128	6.460795599	0.364	0.46423347
0.9	15.4329	5.56583051	0.31605	0.404132513
0.95	10.2761	5.197617532	0.29621	0.379213443

Table 4.9: Beta-Prime distribution for downlink data (DD)

Predicted Level Crossing Rate				
Beta-Prime Distribution				
Level	DD1	DD2	DD3	DD4
0.7	10.74315028	2.315284467	0.170501828	0.456986633
0.8	9.610127427	1.870925215	0.146841776	0.39591464
0.9	6.51045355	1.54152209	0.128576042	0.348616426
0.95	4.469668785	1.407854026	0.120929385	0.328767772

Table 4.10: Exponential distribution for downlink data (DD)

Predicted Level Crossing Rate				
Exponential Distribution				
Level	DD1	DD2	DD3	DD4
0.7	1.63E-04	0.249391635	4.59031E-06	1.85E-03
0.8	6.72E-05	0.346215139	3.15104E-07	8.38E-04
0.9	2.77E-06	0.141118557	2.16305E-08	3.79E-05
0.95	5.61E-07	0.090095514	5.66725E-09	8.06E-06

4.1.6 Actual Level Crossing rate

Actual crossing rate is discovered for single machine data, uplink data and downlink data.

Table 4.11: Actual level crossing rate for single machine data

Actual Level Crossing Rate							
Level	SD1	SD2	SD3	SD4	SD5	SD6	SD7
0.7	46	16	5	1	3	0.3	0.4
0.8	48	16	3	1	2	0.3	0.2
0.9	41	15	2	1	2	0.2	0.2
0.95	41	15	1	0	2	0.2	0.2

Table 4.12: Actual level crossing rate for uplink data and down link data

Level	UD1	UD2	DD1	DD2	DD3	DD4
0.7	4	3	34	13	1	1
0.8	3	3	30	15	1	1
0.9	1	1	18.5	15	1	1
0.95	0	1	14	15	1	1

4.1.7 Error Percentage

Table 4.13: Error percentage of Pareto distribution for single machine data

Error Percentage %							
Pareto Distribution							
Level	SD1	SD 2	SD 3	SD 4	SD 5	SD 6	SD 7
0.7	19.07826	82.9375	37.284	13.379	45.68	47.60333	71.71821
0.8	9.31875	79.63125	11.25333	25.893	21.73	25.40333	51.65956
0.9	23.7561	64.06667	15.225	35.421	31.625	62.92	57.91468
0.95	22.01951	52.78667	15.67	NaN	45.665	52.515	60.50832

Table 4.14: Error percentage of Beta-Prime distribution for single machine data

Error Percentage %							
Beta-Prime Distribution							
Level	SD 1	SD 2	SD 3	SD 4	SD 5	SD 6	SD 7
0.7	97.21161	80.96366	99.21034	23.126	93.8522	71.72522	82.43696
0.8	97.75888	84.6152	98.91709	56.83036	92.06764	77.33543	72.07843
0.9	97.75973	86.47169	98.63773	74.7822	93.0621	72.18553	77.33759
0.95	97.91834	87.64013	97.49001	NaN	93.47814	74.67889	79.4484

Table 4.15: Error percentage of Exponential distribution for single machine data

Error Percentage %							
Exponential Distribution							
Level	SD 1	SD 2	SD 3	SD 4	SD 5	SD 6	SD 7
0.7	99.99293	92.35818	99.92008	99.99964	99.93409	99.999	99.99908
0.8	99.99964	97.33944	99.97983	99.99993	99.98594	99.99993	99.99981

0.9	99.99978	99.01195	99.99542	100	99.998	99.99999	99.99998
0.95	99.99999	99.417	99.99989	NaN	99.99925	100	100

Table 4.16: Error percentage of distributions for Uplink data

Error Percentage %						
	Pareto Distribution		Beta-Prime Distribution		Exponential Distribution	
Level	UD1	UD 2	UD 1	UD 2	UD 1	UD 2
0.7	73.769475	22.976667	72.31331	93.87244	92.4551	99.9073
0.8	65.257467	3.9000000	76.71666	94.72352	96.5007	99.9959
0.9	3.6968	31.360000	60.40523	86.14154	96.3484	99.9944
0.95	NaN	49.090000	NaN	86.96661	NaN	99.9999

Table 4.17: Error percentage of Pareto distribution for Downlink data

Error Percentage %				
Pareto Distribution				
Level	DD1	DD 2	DD 3	DD 4
0.7	34.31888235	41.14857617	68.67495	57.279
0.8	37.29066667	56.92802934	11.07256	63.6
0.9	16.57891892	62.89446326	32.21127	68.395
0.95	26.59928571	65.34921645	50.01039	70.379

Table 4.18: Error percentage of Beta-Prime distribution for Downlink data

Error Percentage %				
Beta-Prime Distribution				
Level	DD1	DD2	DD3	DD4
0.7	68.40249918	82.19011948	82.9498172	95.81836987
0.8	67.96624191	87.52716523	85.31582239	96.54544197
0.9	64.80835919	89.72318607	87.14239577	97.09209601
0.95	68.07379439	90.6143065	87.9070615	97.31640067

Table 4.19: Error percentage of Exponential distribution for Downlink data

Error Percentage %				
Exponential Distribution				
Level	Data1	Data2	Data3	Data4
0.7	99.99951988	98.08160281	99.99954097	99.81481983
0.8	99.99977597	97.69189908	99.99996849	99.91621494
0.9	99.99998504	99.05920962	99.99999784	99.99620913
0.95	99.99999599	99.39936324	99.99999943	99.99919365

4.1.8 Average Error Percentage

Table 4.14 to Table 4.19 describes the error percentage of distributions when compared with its actual rate. These error shows how much it deviates from actual value. This obtained results will help to identify the most appropriate statistical model for analyzing traffic distributions. So Table 4.20 illustrates average error percentage of above results to make better decision.

Table 4.20 shows the average error percentage for above data samples.

Sample Type	Average Error Percentage %		
	Pareto Disribution	Beta-Prime Distribution	Exponential Distribution
Single Machine Data	40.14963881	83.66738889	99.55291741
Uplink Data	37.20312371	80.11888042	86.28818602
Downlink Data	47.67063805	84.33706733	99.62233099

According to the error percentage can be determined the Pareto Distributing has the minimum error. So Pareto distributing is the best distributing fitting for all data. This back ground analysis is facilitated to determine the proper distributing fittings among the other distributing fittings.

4.2: Analysis for Institutional Data (ID)

As mention in chapter 3 background analysis is important to analyze the ID because to find feasible method to predict the traffic distributions in a link. After getting the result from background analysis further analyzing has been done for ID. Results of the analysis was attached to appendix B. Table 4.21, Table 4.22, Table 4.23 and Table 4.24 are results taken under 1second, 6seconds, 24 seconds and 288 seconds respectively.

4.3: Mean error histograms

Institutional data is analyzed and the result is tabulated with its error. This error is interpreted for three types of distributions within 1,6,24 and 288 seconds. Table 4.26 is described the mean error for according to obtained results. Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4 are illustrated that mean error as histograms for all distributions.

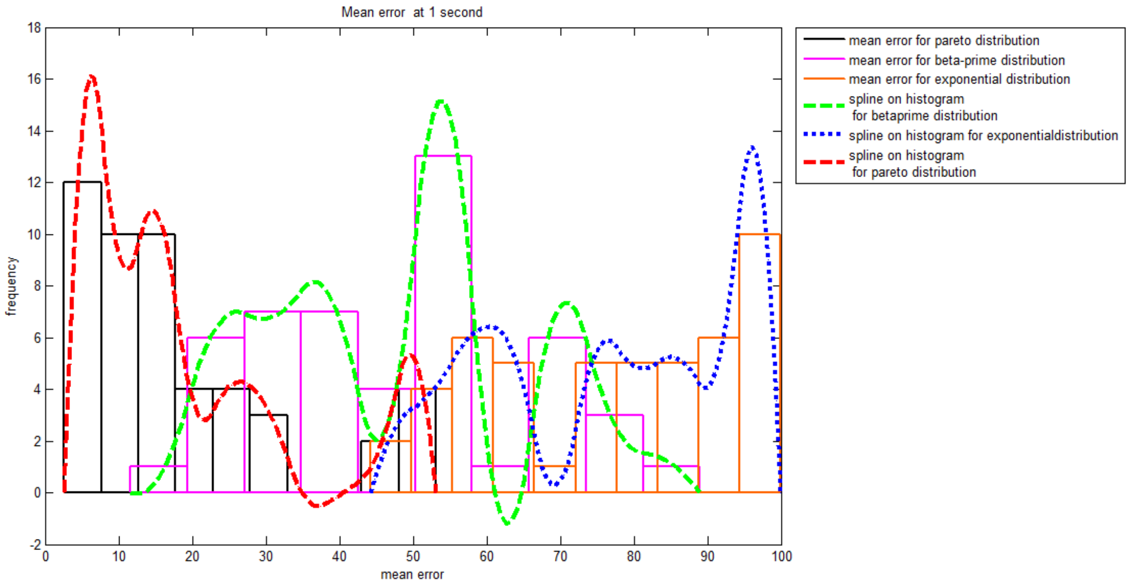


Figure 4.1: Mean error at 1 second

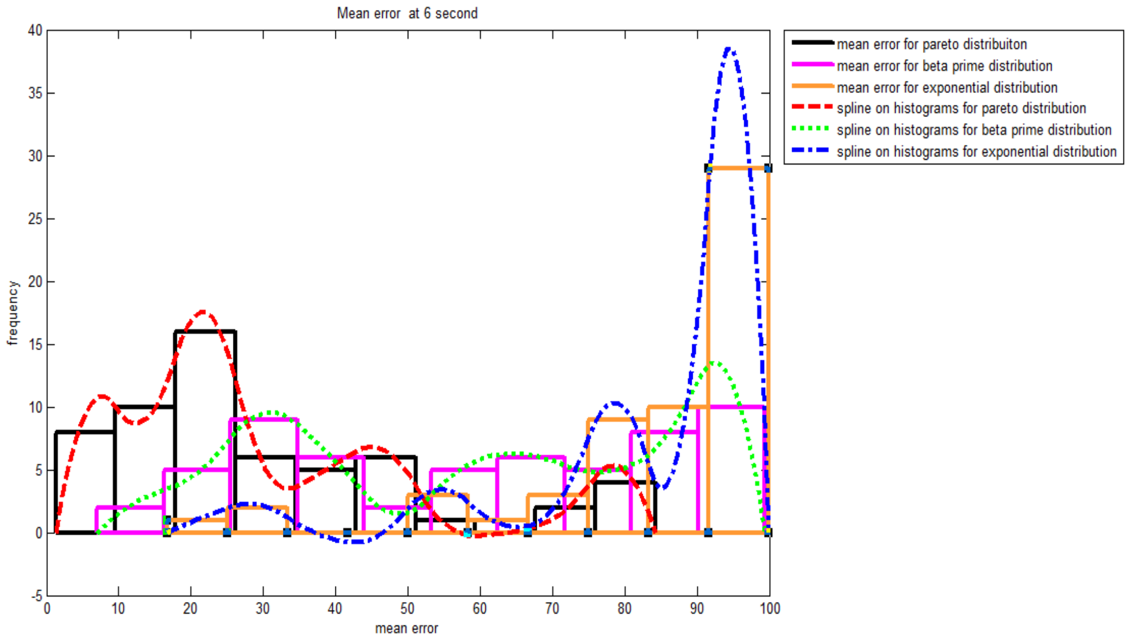


Figure 4.2: Mean error at 6 second

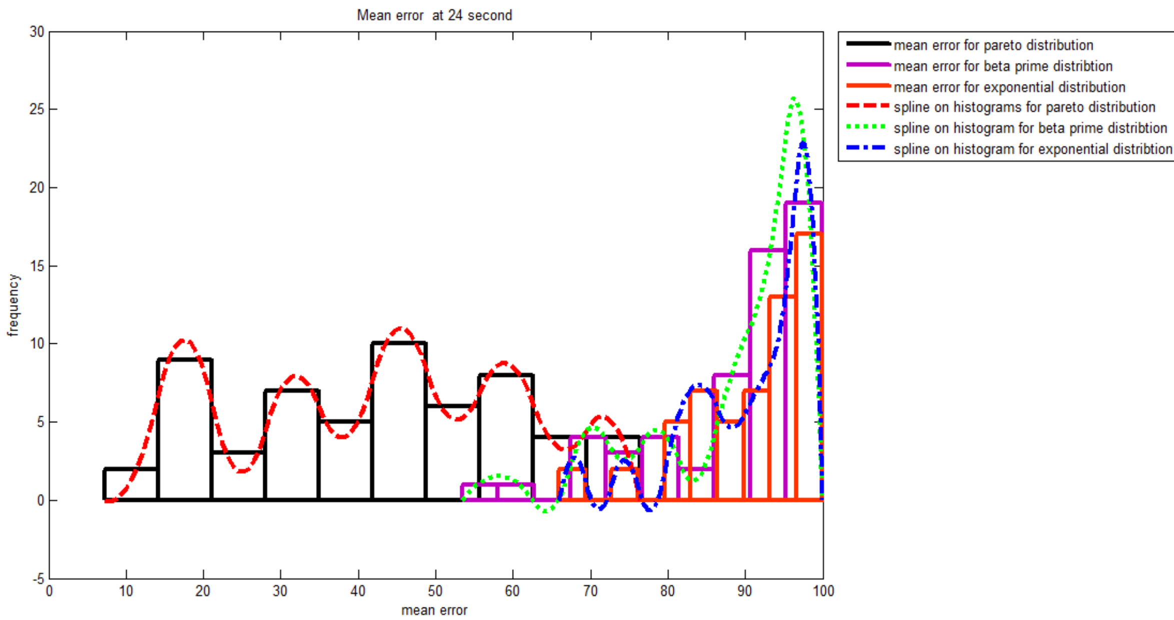


Figure 4.3: Mean error at 24 second

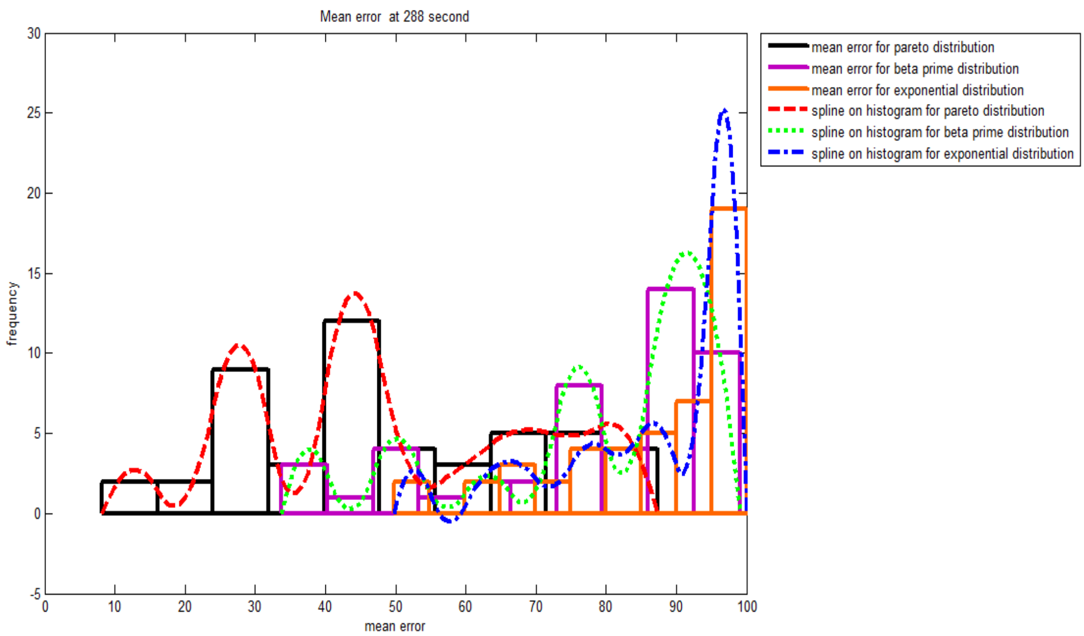


Figure 4.4: Mean error at 288 second

Mean error histograms represent how is that error for distributions model exist. According to the mean error histograms they emphasized that the best PDF as the Pareto distribution due to having of very small error. Also it interprets the error for

Exponential distribution is significantly high. Table 4.25 presents the mode and median of mean error for distribution.

Table 4.21: Mode and Median for distribution

Time Slot (seconds)	Pareto distribution		Beta-Prime Distribution		Exponential distribution	
	Mode	Median	Mode	Median	Mode	Median
1	5.15	27.8751	54.2	50.3201	97.2	72.0923
6	27.4	43.997	28.6	57.7496	95.8	58.3409
24	45.3	41.8671	97.6	76.6803	98.3	82.9694
288	43.8	47.7522	89.3	66.4012	97.5	74.8648

The results which is loaded in Table 4.26 parameters such as mode and median for desired PDFs.

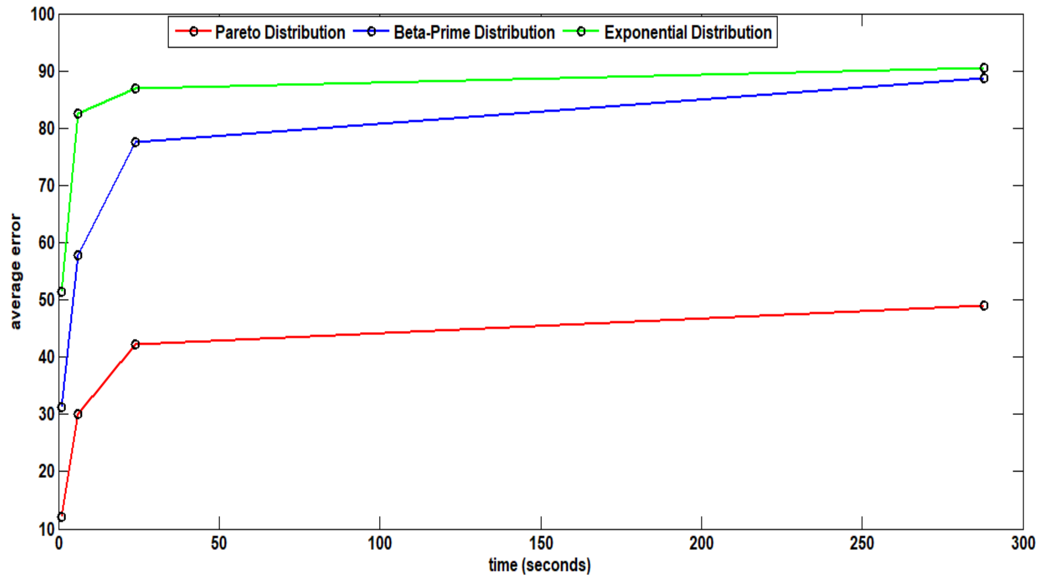


Figure 4.5: Mean error for three distributions within each time slots.

According to Figure 4.5 it shows that mean error is significantly less for Pareto distributing within the observed time period.

Furthermore, it has been analyzed the relationship between mean error and bandwidth. Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9 are described it.

At 1 Second

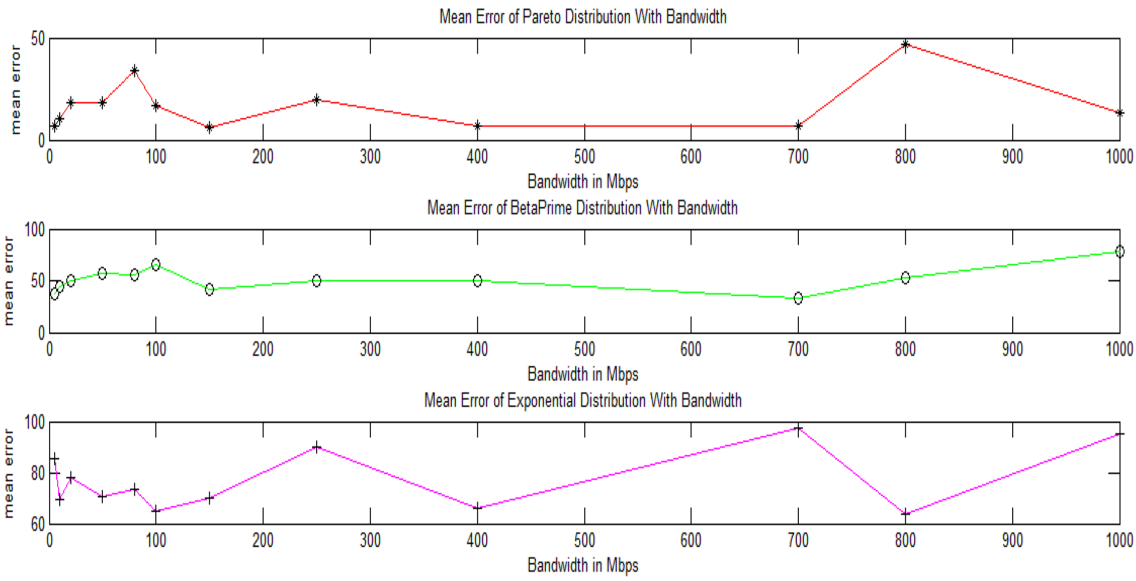


Figure 4.6: Mean error Vs bandwidth for all distributions within 1 second.

At 6 second

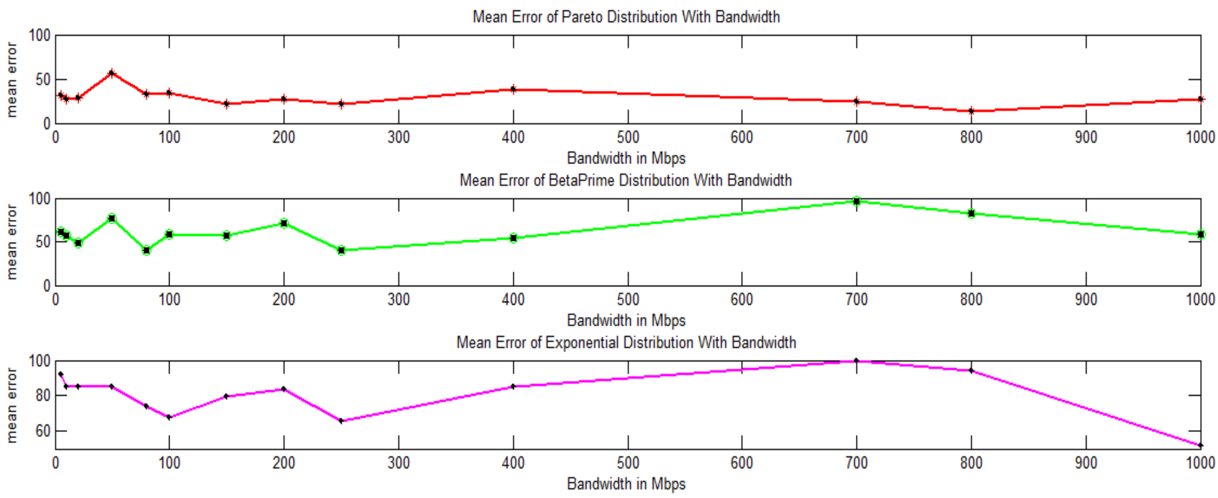


Figure 4.7: Mean error Vs bandwidth for all distributions within 6 second

At 24 second

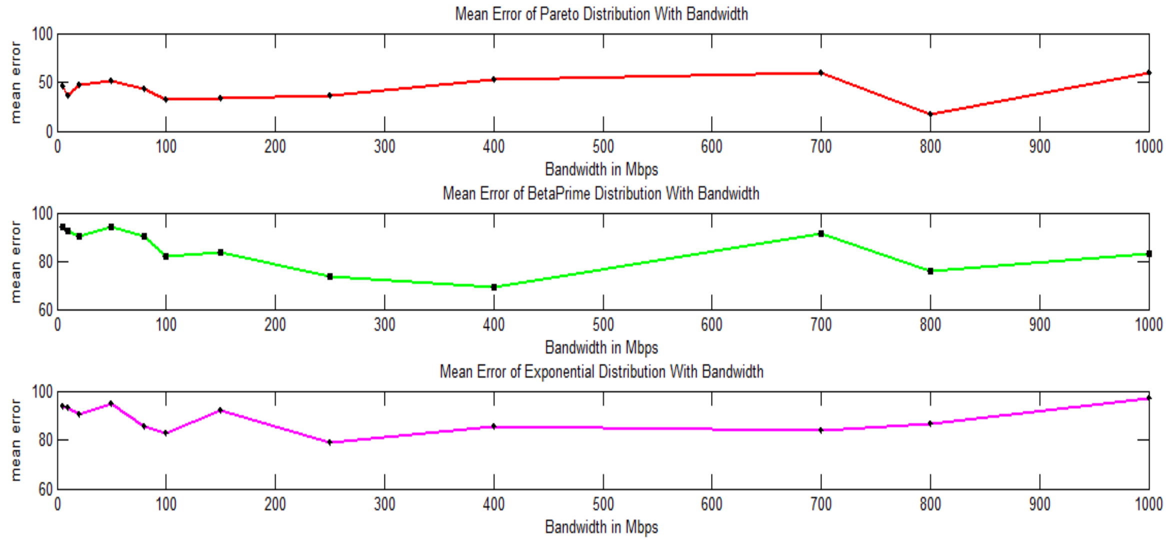


Figure 4.8: Mean error Vs bandwidth for all distributions within 24 second.

At 288 sconds

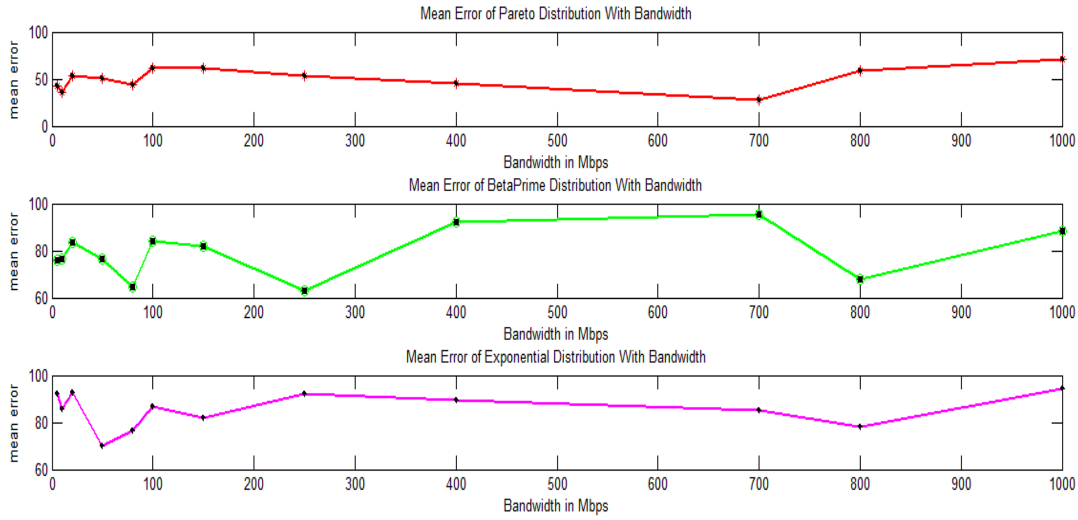


Figure 4.9: Mean error Vs bandwidth for all distributions within 288 second.

As per the figure 4.6, 4.7,4.8 and 4.9 indicates the minimum mean error in the Pareto Distribution for bandwidth.So all these analyzes are verified the best distributing fitting as the Pareto Distributing statistical model when compared with other distributions such as Beta-Prime distribution and Exponential distribution.

CHAPTER 05

CONCLUSIONS AND FUTURE WORKS

According to the results discussed in chapter 4 Pareto Distribution shows the best results for analyzing of traffic distribution compared to the other two distributions such as Beta-Prime Distribution and Exponential Distribution.

These results are gained by analyzing of two data sets. One is done as the background analysis for single machine. That data sets contain one hour traffics. Also uplink and downlink data are extracted form single machine data with one-hour traffic. Basically for the analysis of these data sources the speed of the network is considered as the 100Mbps. The best distribution as the Pareto Distribution is presented 29.6% average error. 82.6% and 95.1% average error is presented by Beta-Prime and Exponential Distribution respectively. So observed results are valid for this type of network.

Then analysis is done for Institutional data set which are collected from LEARN network. As mentioned in chapter 3 institutional data is taken under 1sec, 6sec, 24sec and 288 sec. According to the results shown in Table 4.26 and Figure 4.5 it clearly illustrates Pareto Distribution is the best distribution as it has minimum average error in different time scales compared to other two distributions.

According to [8] the Poisson process model was used for traffic modeling. But it ignored bursts completely due to focus on simplicity of analysis. But proposed statistical models for traffic distribution considered such characteristic of traffic distribution and derived the parameters on predicting traffic peaks. As reveal in [22] it has proved the model and estimation of packet traffic distribution for (Botswana International University of Science and Technology(BIUST) network based on Pareto distribution. Since that different networks provide traffic it's very essential to have various evaluation methods [39].

As a next step of this research observe the accuracy of these statistical models for predicting the traffic peak for high speed network like fiber link.

References

[1] “Cisco Visual Networking Index: Forecast and Methodology, 2011-2016 [Visual Networking Index (VNI)],” *cisco.* [online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns537/ns705/ns827/white_papaper_c11-481360_ns827_Networking_Solutions_White_paper.html. [Accessed:18-July-2018].

[2] T.E.Koch and E.Gelle, “On automating the network management in industrial automation systems,” *Proceedings of 12^{th} IEEE international conference on ECBS*, 123-128.

[3] B. Galloway and G.P.Hancke, “Introduction to industrial control networks,” *IEEE communications surveys and tutorials*, vol.15, No.2, pp. 860-880, May 2013.

[4] H. Abrahamsson. “Internet Traffic Management.”Malardalen University Press, Vasteras, 2008.

[5] Xu Lan, “Analysis and research of several network traffic prediction models”, *Chinese Automation Congress (CAC)*, pp.894-899, 2013.

[6] E.R.S.Castro, M.S. Alencar and I.E.Fonseca, “Probability density functions of the packet length for computer networks with bimodal traffic,” *International Journal of Computer Networks & Communication*, vol.5, no.3, pp.17-31, 2013.

[7] S.A.Musthaq and A.A.Rizvi, “Statistical analysis and mathematical modeling of network (segment) traffic,” *International Conference on Engineering Technologies*,Islamabad,2005.

[8] W.Feng, Y.Sun, Z.Zhou, Q.Rao, D.Chen, L.Yang and Y.Wang, “Study on multi-network traffic modeling in distribution communication network access service,” *International conference on advanced communication technology*, China, 2018.

[9] H. Zaho, “Multiscale analysis and prediction of network traffic,” *in Performance computing and communication conference, IEEE 28^{th} Internatonal. IEEE*, Dec.2009, pp.388-393.

[10] X.An, L.Qu and H.Yan, “A study based on self-similar network traffic model,” *Sixth International Conference on Intelligent Systems Design and Engineering Applications*, China,2015.

[11] R.Pries, F.Wamser, D.Staehle and P.Tran-Gia, “Traffic measurement and analysis of a broadband wireless Internet acess,” *IEEE 69th Vehicular Technology Conference*,Spain,2009

[12] W.John, and S.Tafvelin, “Analysis of internet backbone traffic and header anomalies observed,” *Proceeding of the 7th ACM SIGCOMM conference on Internet measurement*, New York,2007.

[13] R. Beverly and K.C. Claffy, “Wide-area IP multicast traffic characterization”, *Network IEEE*, vol.17, no.1, pp.8-15,2003.

[14] Michael Wilson, “A historical view of network traffic models”, [online]. Available:http://www.cse.wustl.edu/~jain/cse567-06/traffic_models2.htm. [Accessed:19-July-2018].

[15] U.Premarathne, U.Premarathne and K.samarasinghe, “Network traffic self-similarity measurements using classifier based hurst parameter estimation,” *ICIAfS*, 2010.

[16] V.Ndatinya, Z.Xiao, V.Maneppali, K.Meng and Y.Xiao, “Network forensics analysis using wireshark,” *Int.J.Sensor Networks*,vol.10,No.2,pp.91-106,2015.

[17] M.L.Mchugh, “The chi-square test of independence,”*Biochemia Medica 23*,143-149(2013).

[18] A.Dainotti, A.Pescape and H.Kim , “Traffic classification through joint distributions of packet-level statistics,” *In GLOBECOM*,pp.1-6,2011.

[19] S.Xu, *Proceedings of 2013 world agricultural outlook conference*, Springer, 2013,pp.19-21.

[20] H.T.Yura and S.G.Hanson, “Mean level signal crossing rate for an arbitrary stochastic process,” *Optics,image science and vision*, vol.27, pp. 797-804, Apr. 2010.

[21]“Math is fun”, [online]. Available: https://www.mathsisfun.com/numbers/percentage-error.html.[Accessed:24-July-2018].

[22] T.Solomon, A.M.Zungeru, R.Selvaraj and M.Mangwala, “A packet distribution traffic model for industrial application: A case of BIUST network”, *International journal of Information and Electronics Engineering,* vol.7, pp.136-140,2017.

[23] T.Bonald and M.Feuillet, “*Network performance analysis*”. Hoboken, NJ: Wiley, 2011.

[24] J.Zhang, Y.Xiang, Y.Wang, Yu Wang, W.Zhan, Y. Xiang and Y.Guan, “ Network traffic classification using correlation information,” IEEE transaction on parallel and distributed systems, vol.24,pp.104-117,2012.

[25] Y.Miao, Z.Ruan, L.Pan, J.Zhang, Y.Xiang and Y. Wang, “ Comprehensive analysis of network traffic data,” *In:2016 IEEE International Conference on Computer and Information Technology (CIT),* pp.423-430. IEEE 2016.

[26] X.Chen,J.Zhang,Y.Xiang,W.Zhou,”traffic identification semi-known network environment,” *in Proc. 16th IEEE conf. Computational Science and Engineering, 2013*,pp.572-579.

[27] A.Dainotti, A.Pescape, K.C.Claffy, “Issues and Future Directions in Traffic Classification,” *IEEE transactions on Network*,vol.26, No.1, pp.35-40, 2012.

[28] H.Akaike, “A new look at the statistical model identification,” *IEEE transactions on Automatic control,* vol.19,pp.716-723,1974.

[29] A.W.Moore, and D.Zuev, “Internal traffic classification using baysian analysis techniques,” presented at 5th Int. conf. on Measurement and Modeling of Computer Systems, Banff, Alberta, Canada, 2005.

[30] K.Thompson, G.J.miller, R.Wilder, “Wide area Internet Traffic Patterns and Characteristics,” *IEEE transactions on Network*,vol.11, No.6, pp.10-23, 1997.

[31] T.Karagannis, M.Molle, M.Faloutsos, A.Broido, “A nonstationary Poisson view of Internet traffic,” *IEEE INFOCOM 2004,Hongkong,China*,IEEE, 2004.

[32] V.Paxon, S.Floyd, “Wide Area Traffic: The Failure of Poisson modeling,” *IEEE Transactions on networking*, vol.3, No.3,pp.226-244, 1995.

[33] H.B.Mann and A.Wald, “ On the choice of the number of class intervals in the application of the chi-square test,” *IEEE transactions on Mathematical statistics*, vol.13, No.3, pp. 306-317, 1942.

[34] A.J.Rainal, “ Origin of Rice’s Formula,” IEEE transactions on Information Theory,” vol.34, No.6, pp.1383-1387, 1988.

[35] X.Zhang, D.Shasha, “Better Burst Defection,” Presented at 22nd International Conference on Data Engineering, Atlanta, USA, 2006.

[36] D.Dadabneh, M.st-Hilarie, C.Makaya, “Traffic model for long term evaluation network,” In.Proc. IEEE International Conference on mobile and wireless networking, 2013, pp.13-18.

[37] R.wald, T.M.Khoshqoftaar, R.Zuech and A.Napolitano, “Network Traffic Prediction Models For Long Term Predictions,” In.Proc.IEEE International Conference on Boformatics and Bioengineering’14, 2014,pp.362-368.

[38] Barath Kumar, Oliver Nigggemann and Juergen Jasperneite, “Statistical Models of Network Traffic,” *Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol.4, No.1 ,pp.177-185,2010.

[39] J.Markkula, J.Hagpola, “Impact of Smart Grid Traffic Peak Loads on Shared LTE Network Performance,” IEEE International Conference on Communications, Budapest, Hungary, 2013.

[40] A.Callado, C.Kamienski,G.Sszabo, “A survey on internet traffic identification*,” IEEE transaction on communication surveys and tutorials*, vol.11, No.3,pp.37-52,2009.

[41] Sung-Ho Yoon, J.S. Park, M.S.Kim, “ Behavior Signature for Big Data Traffic Classification,” IEEE International conference on big data smart omputing,Thailand, 2014.

[42] S.H. Low, F.Gaganini,J.C.Doyle, " Internet Congession Control," *IEEE transactions on control systems magazine,* vol.22, No.1, pp.28-43, 2002.

[43] F.Paganini, Z.Wang,J.Doyle, S.ow, " Congession Control for High Performance Stability and Fainness in General Networks, IEEE/ACM Transactions on Networking, vol.13, No.1,pp.43-56, 2005.

[44] D.H. Garcia, T. Hagakawa, " Using Congession Graphs to analyse the stability of network congession control, IEEE International Conference on Networking, Sensing and Control,2009,Japan,: IEEE,2009.

[45] K.Papagiannaki, "Long term forecasting of Internet backbone traffic," *IEEE transactions on neural networks,* vol.16, No.5,pp.1110-1124,2005.

[46] A.Sang, san-qi Li, "A predicatability analysis of network traffic," *IEEE transactions on Computer Networks*, vol.39, No.4, pp.329-345,2002.

[47] Q.He, C.Dovrolis,M.Ammar, "On the predictability of large transfer TCP Throughput," *in.Proc.5th IEEEconf.applications,technologies,architectureand protocols for computer communications*,2005,pp.145-156.

[48] M.F.Zhani, H. Elbiaze and F.Kamoun, "Analysis and prediction of real network traffic," IEEE transactions on Networks, vol.4, No.9,pp.855-865,2009.

[49] S.Basu, A.Mukherjee, S.Kilvanskey, "Time series models for internet traffic," presented at 96th conference on computer communications, San Francisco, USA,1996.

[50] C.Katris, S.Daskalaki, "Comparing forecasting approaches for internet traffic," *IEEE transactions on Expert Systems with Applications*,vol.42,No.21,pp.8172-8183,2015.