Time based Internet Traffic Policing and Shaping with Weibull Traffic Model

Mohd Azrul Bin Abdullah (2013798299)
EE7002A, Fakulti Kejuruteraan Elektrik
UiTM Shah Alam
Shah Alam Selangor, Malaysia

Abstract- Quality of Services for bandwidth management is a crucial task in computer network. Policing and Shaping algorithms are identified as one of the task in bandwidth control, especially in tele-traffic engineering and computer network. This thesis presents the analysis of internet traffic flows in a campus network. Real internet traffic for inbound and outbound throughput flow is collected and statistical analysis is measured to characterize the real live internet traffic parameters. Empirical cumulative distribution function (CDF) model is presented in evaluating the real traffic distribution. Anderson-Darling (AD) and Goodness of Fit test is used to identify the best fitted distribution model to the real data. Four traffic distribution which are normal, lognormal, Weibull and exponential distribution are fitted and derived. Analysis results present Weibull Distribution model is the best fitted model. Two important Weibull parameters which are shape and scale are measured. Based on the identified statistical parameters, a new Time Based Policing and Shaping algorithm is developed and simulated. Policing process drops traffic while shaping process delay traffic to the next time transmissions. Mathematical model to formulate the algorithms is derived. The new algorithms present the existence inbound traffic burst is policed at 1200 Mbyte which is maximum allowable bandwidth while for outbound traffic burst, the policing traffic is policed at 680 Mbyte utilizes the maximum allowable bandwidth. The result of implementing policing and shaping traffic shows the burst can be controlled and thus, reduce the traffic congestion in the network. Furthermore, the drop data from policing traffic process can be saved and transmitted at the next transmission. Besides the method of congestion control, varying the shape parameter of Weibull distribution also help to reduce the burst and improve the performance the internet traffic as a whole.

Index Terms— Internet traffic, statistical analysis, cumulative distribution function, maximum likelihood estimation, Anderson-Darling, normal, lognormal, Weibull, exponential

INTRODUCTION

Nowadays, network traffic has grown tremendously due to the increasing usage of internet access. The significant growth of traffic is mainly driven by data applications, voice and video application. Network traffic is the main area to be controlled and managed in order to preserve network performance. Therefore the study of internet traffic has become very important task. It is essential to understand the overview characteristic of internet traffic. There are numerous traffic models that are used widely for traffic modeling with different categories of traffic models. Each model varies significantly from the other and suitable for modeling different traffic characteristics. Normal model has shown in [1] that normal distribution can be directly linked to the presence or absence of extreme traffic burst. However, Normal distribution is not suitable to model the traffic demand in large-scale network [2]. They show that from network traffic characterization from a point to another point in the network is a lognormal distribution, which has a slower decay than a normal distribution. While in [3], lognormal distribution is able to accurately statistical models for flow size and flow duration of traffic application. The techniques and real traffic parameter evaluation, yields changes in network performance.

Observation of invariant heavy tails in access traffic patterns of individual users has motivated [4] to investigate traffic transformation or aggregation as it traverses from access to core network. It shows that the variation of shape parameter of the Weibull distribution can capture the transformation of internet traffic which consists of sessions, flows and packet at inter-arrival level. In [5] study how the superposition of heavy-tailed renewal streams models the scaling behavior of traffic at different access networks and tiers of Internet hierarchy.

However, in [6] numerical results of throughput of network show that the network with exponential distributions of link capacities do not able to accommodate much more traffic as it is able for short range data. Exponential distribution is more suitable for non-long-tailed data. This is a sharp contrast to commonly made modeling choices that exponential assumptions dominate and show only short-range dependence [7]. The distribution model has its own characteristic and the selection of the traffic model is not only considering on the type of traffic but also depends on the application.

As high utilization of internet access application, network traffic management introduced called scheduling methods which will control the network traffic and improve performance. A goal of scheduling algorithms is to provide best-effort traffic, which all active flows should obtain the same services and guarantee no delay. In [8] proposed to

use multiple correlated token bucket to shape the traffic. However, it is not suitable for huge traffic flows and will increase the complexity in traffic policing. Meanwhile, in [9] studied the performance of leaky bucket and its application in traffic policing. The studied shows that the non-work conserving constraint the regulated traffic and allow limitation while network utilization improves in work conversing scheme.

There are several statistical models of probability distributions and estimate parameters used to estimates the behavior of data population such as Anderson-Darling (AD) test and maximum likelihood (ML) estimation. The population distributions generally use maximum likelihood (ML) estimation to estimate the distribution parameters which represent statistically characteristics. Study shown in [10-13] the maximum likelihood (ML) estimation technique used to identify the best distribution model that fits to real traffic data which represents the maximum log-likelihood. The maximum likelihood (ML) estimation aims at determining the certain parameters based on distribution models to maximize the likelihood function. The maximum likelihood (ML) estimation is likelihood function $L(\theta)$ as a function of θ and find the value of θ that maximizes it as defined as equation 1.1

$$\ell(\theta|x) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x|\theta), \qquad (1.1)$$

Analysis using maximum likelihood (ML) estimation has been tested to fit the established time series to existing distribution, such as Exponential and Pareto distribution [14]. The analysis pointed out that parameters value could be modeled using Pareto and exponential distribution.

In this thesis, network traffic throughput will be analyzed on real data to represent the real distribution and characteristic. Besides using Maximum Likelihood (ML) Estimate to determine the closer distribution, Anderson-darling (AD) and Goodness of Fit (GoF) test is another way to estimate the parameter that the best distribution can be fitted to data. The traffic data is tested using Anderson-Darling (AD) statistical test to find the best distribution model that fitted to real traffic network data. After identified the best distribution model, the new algorithm is proposed based on statistical parameter result. The algorithm is then modelled in mathematical equation. The expectation of the new algorithm is improvement on bandwidth usage and data saving.

LITERATURE REVIEW

The internet traffic analysis is essential to understand the network capabilities and requirements in order to provide reliability guarantees. In the past years, numerous traffic models proposed for understanding and analyzing the traffic characteristics of networks but none of the traffic model can be used for modeling traffic effectively in the networks.

Traffic modelling is one of the important aspects in order to meet quality of service (QoS) requirements of services and efficient utilization of network. Traffic modelling is become necessary in network traffic by controlling the bandwidth utilization. Bandwidth management is used to prevent traffic congestion and avoid traffic slow down during data transmission.

As the network performance changes significantly, the network application has to be monitored and controlled in order to balance the network performance. It is also proven that bandwidth management is a dynamic approach that provides adaptability, feasibility and efficiency for real time network [15, 16]. By bandwidth management, inbound and outbound traffic can be classified into application and service type. Without bandwidth management, utilization of available bandwidth is not fully occupied by internet applications and it also prevents other applications from sharing the network [17]. Thus, the traffic scheduling method is introduced in the network in order to control the bandwidth utilization [18]. In addition, the study in [19] shows that the bandwidth utilization can be controlled and the packet data can be saved.

In network traffic, one of scheduling method is traffic policing and traffic shaping process. Before the data transmits to the network, the data is policed at certain threshold and shaped in form of regulated data. Policing and shaping is one of the bandwidth management techniques to prevent the traffic congestion in the network and improve the quality of service performance. Policing process benefitted to the bandwidth saving and also avoid delay in transmission but the drawback is some of the data are dropped. In [20], policing process will have consequences to drop the traffic data but with shaping, the burst data is shaped and transmit at next submission time to the network.

The drawback of traffic policing and traffic shaping can only decrease the data rates but have less impact on aggregate burst traffic behavior. The additional aggregation will maintain the burst within burst phenomena in real time traffic [21] and thus make reduction of data transfer rate. Therefore, the next generation of network traffic must deploy new aggregation models as data rates moves toward bigger network bandwidth at higher transmission speed. New aggregation methods must rearrange burst to minimize heavy tail and self-similarity without degradation of performance in the quality of service (QoS). As growing high throughput in the network, it may increase the transmission bustiness with huge overflows at intermediate router queues. Implementation of packet layer policing in network is one example of bandwidth policing [22]. The introduction of inter-packet delays to spread packets in time helps to reduce bustiness in the network.

2.1 DISTRIBUTION THEORY

In order to keep the performance in constant, evaluation on traffic models and parameters should be defined to quantify the model is in optimum approached. The parameter of models defined must be related to the actual performance measures which are to be predicted from the traffic model. Several distribution models are identified to model the internet traffic data.

a) Normal distribution models

The normal distribution model is most widely used distribution and is the most significant model used in statistics. It is also can be called the "bell curve," as the curve looks bell-shaped curve. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss. The cumulative distribution function (CDF) of normal distribution as equation 2.1.

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \qquad (2.1)$$

Where μ is the mean and σ is standard deviation. These parameters show the characteristic of normal distribution.

b) Lognormal Distribution

The lognormal distribution has certain similarities to normal distribution. The lognormal distribution is versatile model that can empirically fit many types of data. The lognormal distribution is used to model continuous random data when the normal distribution is skewed curve. The cumulative distribution function (CDF) of lognormal distribution with essential parameters μ mean and σ standard deviation is given by equation 2.2

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{x} \frac{\exp\left(-\frac{[\ln(t)-\mu]^2}{2\sigma^2}\right)}{t} d \qquad (2.2)$$

The parameter σ is known as shape parameter while μ is the scale parameter.

c) Weibull Distribution

The cumulative distribution function (CDF) of Weibull distribution is given by

$$f(x;\alpha,\beta) = 1 - e^{-(x/\beta)\alpha} \qquad (2.3)$$

where $\beta \geq 0$ is scale parameter and $\alpha > 0$ is shape parameter. For the Weibull distribution, the process is heavy-tailed when $\beta < 1$ but it has moment finite [23]. Thus it suitable for convergence modelling in heavy tailed multiplexing area [24]. The Weibull distribution is a kind of exponential and Rayleigh distribution. The advantages of Weibull distribution model used to model traffic with constant failure rate, decreasing failure rate or increasing failure rate. This flexibility is one reason for the wide use of the Weibull distribution in reliability.

d) Exponential distribution model

Exponential distribution is the probability distribution when modeling the time between independent events that happen at a constant average rate. The cumulative distribution function (CDF) of an exponential distribution is

$$f(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad (2.4)$$

with λ is scale parameter. The exponential distribution is also related to the Poisson distribution and it is widely used to model waiting times.

There is formal statistical technique for assessing the principal distribution of traffic data called the Goodness of Fit (GoF) tests. The Goodness of Fit (GoF) test produces the result of estimates parameters that are numerically convoluted and usually require specific software to perform the lengthy calculations. The Goodness of Fit (GoF) test is applied to measure the compatibility of the real data with theoretical probability distribution function or empirical distribution of data population. However [25], no goodness of fit (GoF) tests is considered the best for all the distributions model selected. Probably this is related to the fact that the different tests behave better or worse depending on the specific distributions.

But in the goodness of fit (GoF) test [26], Anderson-Darling (AD) statistical parameter is important to determine the fitted model as well as probability level (p-value). Anderson-Darling (AD) one of the most powerful Goodness of Fit (GoF) test as it is more complex due to introduction of a weight function in test statistic.

Goodness of fit (GoF) test is essentially based on either of two distribution which are the cumulative distribution function (CDF) and probability density function (PDF). In Anderson-Darling (AD) test [27], it uses the cumulative distribution function (CDF) approach and therefore belongs to the class of distance tests.

2.2 SCHEDULING TRAFFIC MODEL

In scheduling traffic area, few process need to carried out for data optimizing such as data classification, data shaping and policing, buffer and queue strategy before scheduled to transmit. Scheduling algorithm is used usually in the traffic shaping process where the algorithm is applied to regulate the average rate of data flows before entering into the network. Leaky bucket and token bucket are the kind of mechanism for traffic shaping.

The Leaky Bucket algorithm used as rate control in a network by averaging the data rate, while the Token Bucket algorithm permit the variation of output rate depending on the size of burst. The leaky bucket algorithm [28] has been selected by some international standard organizations as the scheme to determine specified parameters compliances. However, in [29], the leaky bucket policing algorithm is classified as single traffic class. Study in [8] shows the possibility solution to use multiple token bucket to regulate the traffic into independent classes. Beside of leaky bucket, one of mechanism used in the network traffic is token bucket concept [30] in policing process .

Token Bucket	Leaky Bucket
Token dependent	Token independent
Data can only transmitted when enough token	Data are transmitted continuously
Data sent at faster rate and then constant rate	Data sent at constant rate
Token is saved to transmit large bursts	Token is not saved

Table 1: Difference between Token Bucket and Leaky Bucket

METHODOLOGY

The internet traffic data flow was monitored in a campus environment. The area of organization campus was selected to characterize the behavior of internet flow at speed of 16 Mbps. Solarwind software is setup at gateway router for network traffic collection. Inbound and outbound throughput is collected in Mbit. The data were collected every 10 minutes inter-arrival times. The internet traffic data were collected every day for 7 days. The study was performed on data collection with 1108 sample size.

1.1 DISTRIBUTION MODELLING

In order to characterize the internet traffic, statistical analysis approach is used. The measurement data is analyzed to determine the most suitable analytic model. Basically, the analysis process will go few steps before making decision which statistical distribution can be modelled. Flow chart in figure 2 is the way the analysis has done.

The collected internet traffic data at campus organization is then performed the statistical test to characterize the behavior of data distribution. Based on the measured data, several statistical models of probability distributions and estimate parameters are analyzed. The distribution models come with estimate parameters which represent the characteristic of data population. For each distribution statistical parameter were analyzed to find the best model that represent the data.

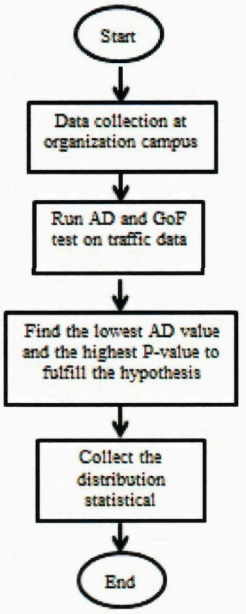


Figure 1: Flow chart of analysis step

The distribution of data plotted using an empirical cumulative distribution function (CDF). The graphs of empirical CDF that is used to evaluate the fit of a distribution to collected data, estimate percentiles, and compare different sample distributions. An empirical cumulative distribution function (CDF) plot performs a similar function as a probability plot.

However, unlike a probability plot, the empirical cumulative distribution function (CDF) plot has scales that are not transformed and the fitted distribution does not form a straight line. In addition, Empirical cumulative distribution function (CDF) is useful to approximate the theory cumulative distribution function (CDF) if the sample size is large.

The Anderson-Darling (AD) test is calculated from following equation 3.1

$$A^2 = -n - (1/n)\sum_{i=1}^{n}(2i-1)[\ln(w_i) + \ln(1-w_{n-i+1})] \quad (3.1)$$

where n is the sample size and w_i is the standard normal cdf $[(x-\mu)/\sigma]$. For Weibull distribution w_i is $1-\exp(-(x_i/\beta)^{\alpha})$.

The Anderson-Darling (AD) and Goodness of Fit (GoF) test use specific distribution in calculating critical values. This has become an advantage of Goodness of Fit (GoF) that more sensitive test is allowed and the disadvantage that the calculation of critical values for each distribution must be performed. The parameters of each data population were compared and the most suitable distribution is chosen. The distribution model with lower of Anderson-Darling (AD) value and higher probability value (p-value) is selected.

Once the analysis is done on the network traffic data, the process of controlling congestion has to take charge before the data can transmit to the network. One of the methods in congestion traffic control in network is traffic policing and traffic shaping.

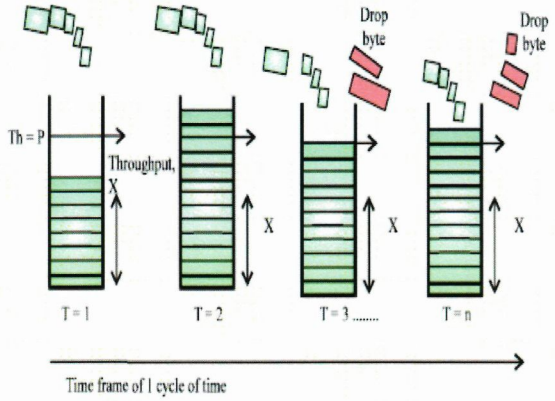


Figure 2: The Token Bucket Concept

The network traffic use token bucket mechanism in traffic congestion control. Figure 2 [11] represents the Token Bucket transitions of throughput, X on policy and threshold, Th. Furthermore, threshold Th is the maximum possible transmission rate in bytes/second or identified as Policy, P. The maximum burst time is the time where the rate of throughput, X is fully utilized. Bytes as token are discarded from the bucket if it beyond the threshold, Th or policy, P and per Bucket size, B_k. The incoming throughput is put in bucket according to identified policy condition P.

3.2 DEVELOPMENT MATHEMATICAL MODEL

The Goodness of fit (GoF) test has produced estimates parameter to represent the traffic characteristic. Based on information from goodness of fit (GoF) test, the distribution can be modelled by developing the mathematical equation for optimizing the traffic data. The information can be gathered as table 2.

Parameter	Symbol	Values
Time-based Traffic	X	0:00am < X < 11:50pm
Number of day	D	7
Network speed	S	16Mbps
Inter-Arrival Time	Ta	10 minutes
Time frame minimum	Tmin	0:00am
Time frame maksimum	Tmax	11:50pm
Weibull distribution	x	1 < x < 1108
Weibull Shape	α	2.33 and 1.9909
Weibull Scale	β	1058.523 and 379.8954
Number of Data minimum	i	1
Number of Data miximum	n	1108
Policing inbound threshold	Tht	1200Mbyte and 1050Mbyte
Policing outbound threshold	Tht	680Mbyte and 434Mbyte
Number of time	t	(1,2)

Table 2: Summary of variable Parameter

After identified the best distribution model based on statistical analysis, the internet traffic data input will go through the process of data classification, data shaping and policing, buffer and queue strategy before scheduled to transmit The process is called scheduling management. Normally in the internet traffic, scheduling algorithm is implemented where the policing and shaping algorithm is applied to regulate the data flows by averaging data rate before entering into the network.

The concept of policing and shaping in this process use token bucket methodology. Token bucket concept is used to realize the utilization of bandwidth by relocating the data. Essential of policing and shaping has made the development of new algorithm. The policing algorithm can be modelled in mathematical equation as Eq. 3.1.

$$B_p = \begin{cases} \sum_{i=1,t\in(1,2)}^{n,m} X_i - (X_i - Th_t), & X_i > Th_t \\ \sum_{i=1,t\in(1,2)}^{n,m} X_i, & X_i \leq Th_t \end{cases} \quad (3.1)$$

With B_p is bucket for policing.

In shaping process, the data is shaped according to how big the allocation of bandwidth at certain time. The

purpose of shaping is to optimize the traffic and increase the utilization of internet bandwidth. In order to achieve the benefit of shaping, the new algorithm is created. The new algorithm is created in mathematical equation as in Eq.3.2 with B_s is bucket shaping.

$$B_s = \begin{cases} X_i - (X_i - Th_t), & X_{i=1} > Th_{t=1} \\ \sum_{i=1,t\in(1,2)}^{n,m} X_i + (X_{i-1} - Th_t) - (X_i + (X_{i-1} - Th_t) - Th_t)', & X_{i>1} > Th_t \\ \sum_{i=1,j=1}^{n,m} X_i, & X_i \le Th_t \end{cases} \quad (3.2)$$

As the analysis of internet traffic data shown that the Weibull distribution is the best model fitted to data, the algorithm of Weibull policing and shaping is made. For Weibull distribution model, the empirical cumulative distribution function (CDF) mathematical equation as Eq.3.3:

$$x = 1 - e^{-(x/\beta)\alpha} \quad (3.3)$$

with condition $\beta \geq 0,\ 1 < \alpha < 3$.

$$B_{wp} = \begin{cases} \sum_{i=1,t\in(1,2)}^{n,m} 1 - e^{-(x_i/\beta)\alpha} - (1 - e^{-(x_i/\beta)\alpha} - Th_t)', & x_i > Th_t \\ \sum_{i=1,t\in(1,2)}^{n,m} 1 - e^{-(x_i/\beta)\alpha}, & x_i \le Th_t \end{cases} \quad (3.4)$$

$$B_{ws} = \begin{cases} 1 - e^{-(x_i/\beta)\alpha} - (1 - e^{-(x_i/\beta)\alpha} - Th_t), & x_{i=1} > Th_{j\in1,2} \\ \sum_{i=1,t\in(1,2)}^{n,m} 1 - e^{-(x_i/\beta)\alpha} - (1 - e^{-(x_i/\beta)\alpha} - Th_t)', & x_{i>1} > Th_t \\ \sum_{i=1,t\in(1,2)}^{n,m} 1 - e^{-(x_i/\beta)\alpha}, & x_i \le Th_t \end{cases} \quad (3.5)$$

With

$$x_i = 1 - e^{-(x_i/\beta)\alpha} + (1 - e^{-(x_{i-1}/\beta)\alpha} - Th_t), \quad x_{i>1} > Th_t$$

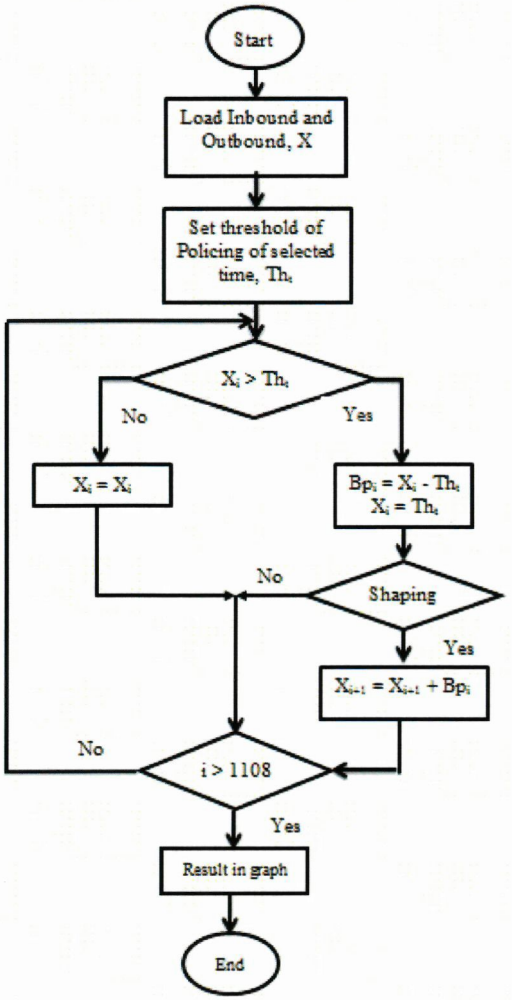


Figure 3: Flow chart of algorithm development

The mathematical model for Weibull policing and shaping as Eq. (3.4) and (3.5) with B_{wp} is Weibull bucket for policing and B_{ws} is Weibull bucket for shaping.

The important of this new algorithm is to manage traffic data at the highest performance of quality of service (QoS). This mathematical equation is then converted into programming language in order to handle the big internet traffic data automatically with no hassle. The programming is written using Matlab software. The programming of new algorithm can be represented as figure 3.

ANALYSIS AND RESULT

Internet traffic data collection of 7 days with 10 minutes interval is analyzed. Analysis has to be interpreted to understand the characteristic of network traffic in real time. The internet traffic data of 7 days plotted in times series plot to have general overview of time based internet traffic. The data is plotted according to time and date for 7 days.

Figure 4 shows heavy throughputs over time since day 1 to day 7. The trend of inbound data is increasing over time while slightly stable for outbound data. The inbound data above 1000 Mbyte represent burst traffic exist in the traffic. It can be identified as bottleneck of the real network. The data of internet traffic can be view comprehensively using scatterplot.

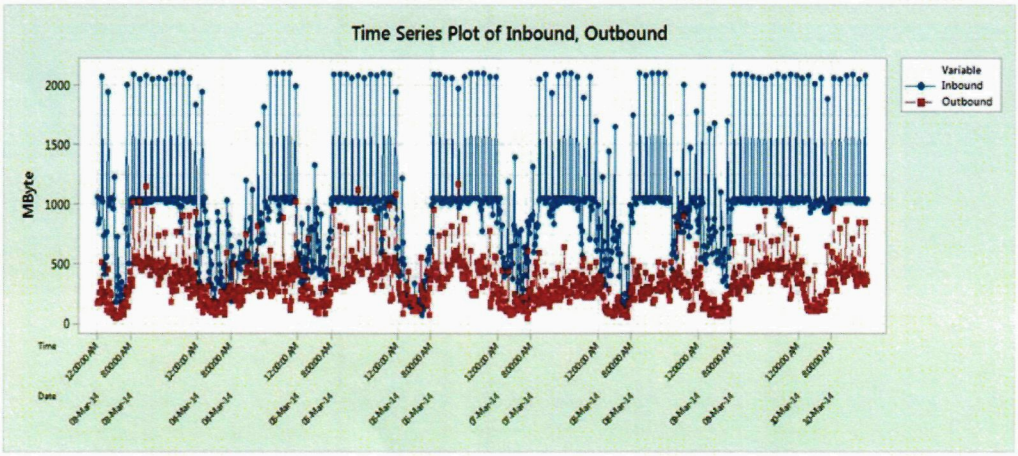


Figure 4: Time series plot of Inbound and Outbound

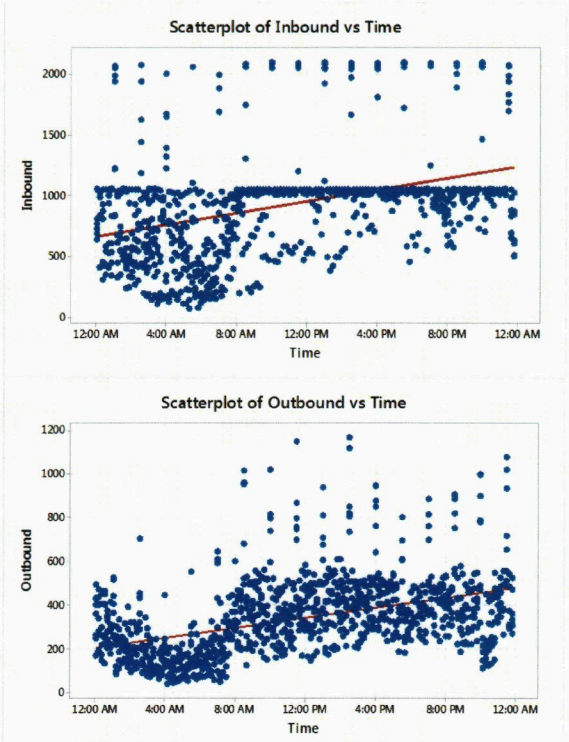


Figure 5: Scatterplot Inbound and Outbound Throughput

As figure 5, the inbound data distribution is concentrated at 1000 Mbyte during day time (8:00 AM to 12:00 AM). However, during between 12:00 AM to 8:00 AM, the throughputs are evenly distributed in area of 0 Mbyte to 100 Mbyte. For outbound throughputs, the distribution shows less throughputs during 12:00 AM to 8:00 AM while high throughput during 8:00 AM to 12:00 AM.

The inbound and outbound data can be adequately modeled by the selected distribution. For network traffic, four types of distribution selected are normal, lognormal, Weibull and exponential. Statistical analysis has to be interpreted to understand the characteristic of whole network traffic data.

Table 3 above is the statistics parameters of inbound and outbound internet traffic data. With the number of sample size N is 1108, the inbound data distributed at mean μ is 940.69 Mbyte with minimum data at 69.123 Mbyte and maximum is 2100.16 Mbyte. The standard deviation σ of inbound is 420.06. While the outbound data distributed at mean μ is 336.1Mbyte with deviation σ is 177.98. The lower value of population is 39.7602 Mbyte and maximum is 1165.58 Mbyte. From the statistical parameter, the inbound internet traffic data has wide variation compare to outbound data.

Descriptive Statistics						
	N	Mean (μ)	StDev (σ)	Median	Minimum	Maximum
Inbound	1108	940.688	420.062	1022.9	69.123	2100.16
Outbound	1108	336.111	177.984	322.086	39.7602	1165.58

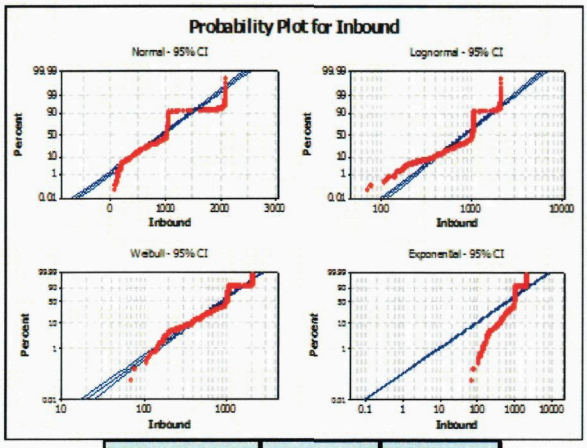
Table 3: Statistical parameters

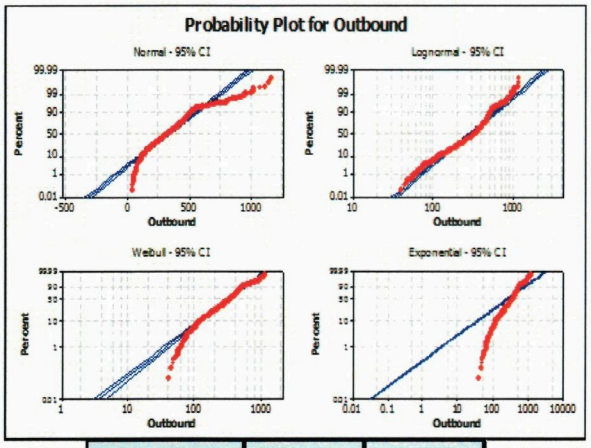
The collected internet traffic data is then tested with several analytical model distributions to measure compatibility of the real data with a particular distribution. The closest the distribution fits to the data, the statistical value will be smaller. Anderson-Darling (AD) statistic test is used to compare the fit of several distributions to select the best or to test whether samples of data come from a population with specified distribution.

The Anderson-Darling (AD) test is based on hypothesis as follow:

H_0 : The data follow the specified distribution
H_1 : The data do not follow the specified distribution

As probability value (p-value) for the Anderson-Darling (AD) test is lower than the chosen significance level α which in this case is 0.05, we can conclude that null hypotheses H_0 is rejected. It means the data do not follow the specified distribution. If probability value (p-value) of the Anderson-Darling (AD) test is higher than significance level, the analysis can be concluded that null hypotheses H_0 is accepted.





Distribution	AD	P-value
Normal	76.17	<0.005
Lognormal	87.901	<0.005
Weibull	73.197	<0.010
Exponential	195.344	<0.003

Distribution	AD	P-value
Normal	10.131	<0.005
Lognormal	14.77	<0.005
Weibull	4.479	<0.010
Exponential	118.548	<0.003

Figure 6: Goodness of Fit Test for Inbound

In Anderson-Darling (AD) test, selected four type of distribution test are Normal distribution, Lognormal distribution, Weibull distribution and Exponential distribution. As in figure 6 shows four types of distribution model have its own Anderson-Darling (AD) value and probability value (p-value).

Based on probability plot, the distribution with the smallest Anderson-Darling (AD) statistic with 73.197 has chosen the Weibull distribution is the closest model that suit to the data. Furthermore, the probability (p-value) of Weibull model with p-value <0.010 is higher than the other distribution models make the null hypothesis H_0 is accepted. While for outbound, Weibull model is the closest model fit to the outbound data with lowest Anderson-Darling (AD) value is 4.479 and the highest p-value <0.01 if compare to the other selected distribution models refer to figure 6.

Empirical cumulative distribution function (CDF) graphs are used to evaluate the fit of a distribution to the data or to compare different sample distributions. Each distribution provides the estimate parameter to understand the characteristic of data.

The analytic model distribution of Normal, Lognormal, Weibull and Exponential has fitted to the inbound data. Cumulative traffic data has reached bottleneck which is at 1000 Mbyte and it is about 40

percent of the internet traffic data. The internet traffic data has fully utilized the bandwidth of 1000 Mbyte about 40 percent to 90 percent of whole data. The burst exists at higher than 1000 Mbyte which allocate about 10 percent of whole inbound internet traffic usage as refer to figure 7.

While for outbound data, empirical cumulative distribution function (CDF) shows no threshold as bottleneck applied as refer to figure 8.

Inbound Estimate value		
Distribution	Type	Estimate
Normal	Location (μ)	940.688
	Dispersion (σ)	420.062
Lognormal	Scale (μ)	6.72441
	Shape (σ)	0.54934
Weibull	Scale (β)	1058.52
	Shape (α)	2.33
Exponential	Scale (θ)	940.688

Outbound Estimate value		
Distribution	Type	Estimate
Normal	Location (μ)	336.111
	Dispersion (σ)	177.984
Lognormal	Scale (μ)	5.66487
	Shape (σ)	0.58713
Weibull	Scale (β)	379.895
	Shape (α)	1.9909
Exponential	Scale (θ)	336.111

Table 4: Estimates Parameter for Inbound and Outbound

Table 4 shows the estimate parameters result of Goodnees of Fit (GoF) test for four types of distribution model. As the closest distribution fit to the internet traffic data, Weibull gives estimate parameters of scale and shape parameters for further analysis and improvement. Anderson- Darling (AD) and Goodness of Fit (GoF) test function describes for each set of distribution parameters, the chance that the true distribution has the parameters based on the data sample.

Based on characteristic of traffic data, optimization of throughput is needed by performing policing and shaping process. As the real traffic data were taken at 16Mbps, the threshold policing are set at 1200 Mbyte as maximum bandwidth of traffic for time 8 AM to 12 AM.

$$T_h \max = \frac{16 \; x \; 1000000 \; x \; 10 \; x \; 60}{8} = 1200 \; Mbyte$$

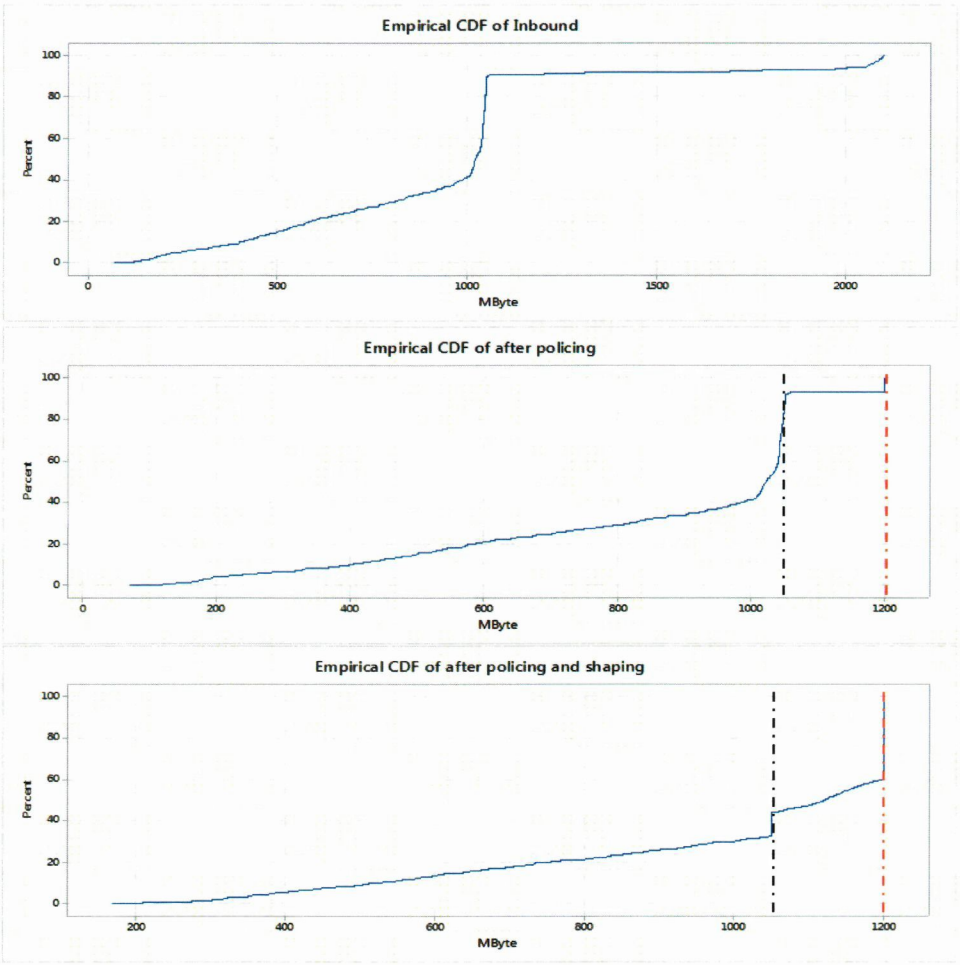


Figure 7: CDF of Inbound after Policing and Shaping

The threshold for time 12 AM to 8 AM is set 1050 Mbyte based on 95 percent throughput of original internet traffic population. The internet traffic data is policed using defined threshold according to time occurred. The internet traffic burst after policing at defined threshold is shaped according to mechanism of token bucket concept. This process called shaping process.

The policing process is done using mathematical model equation 3.1. Figure 7 shows the different of throughput after policing and shaping at threshold 1200 Mbyte (8 AM-12 Am) and 1050 Mbyte (12 AM-8 AM). The shaping process has reached 60 percent of throughput time at maximum allowable bandwidth 1200 Mbyte, while throughput for policing is at 90 percent total time to reach 1200 Mbyte refer to figure 7. It can be concluded that the shaping process is essential in term of data efficiency and quality of services QoS improvement.

The burst of inbound traffic after policing shows the bandwidth used up to 1000 Mbyte of whole usage of internet traffic. However, the burst of traffic data is kept for shaping process. The burst of traffic data after policing is then used to add on to the data of next time slot. The burst is kept in the bucket for next time process. Total burst exist in inbound traffic is 77799 Mbyte.

The algorithm shaping process is performed using mathematical equation 3.2. By allocating the bandwidth of next time data, the bandwidth can be fully utilized to maximum 1200 Mbyte with no drop data.

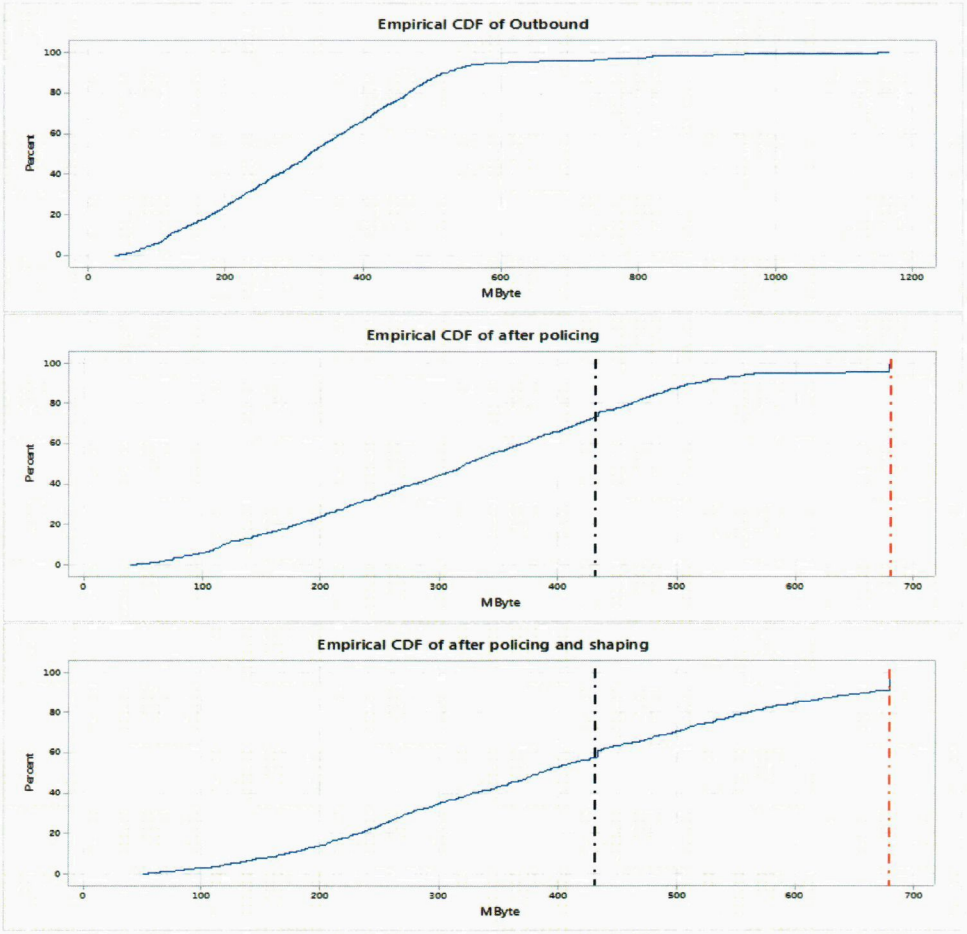


Figure 8: CDF plot of Outbound after Policing and Shaping

For outbound, the threshold of policing is set at 680 Mbyte for time 8 AM to 12 AM and 434 Mbyte for time 12 AM to 8 AM. The threshold is set based on 95 percent throughput of total outbound traffic.

Figure 8 shows the shaping has high throughput of 94.5 percent to reach 680 Mbyte while policing throughput is slightly less with 97.2 percent cumulative traffic to reach 680 Mbyte. The figure 8 shows the policing and shaping process can benefit the performance by fully utilizing the allocated bandwidth and improve the quality of service (QoS).

After policing, the burst data has used the bandwidth up to 500 Mbyte in one time. This burst can be kept in bucket by doing shaping. Total burst for outbound traffic is 10285 Mbyte as in figure 9.

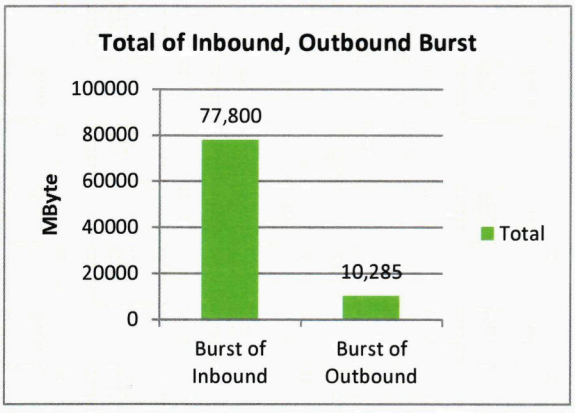


Figure 9: Total burst of Inbound and Outbound

From analysis of fit test, the closest distribution fit to real data is Weibull model. Weibull distribution has two important parameters which represent the characteristic

of distribution model. The parameters are shape parameter α and scale parameter β.

The random Weibull data created based on shape parameter (α=2.33) and scale parameter (β=1058.523). The Weibull data is created using Weibull mathematical equation 3.3. The Weibull data then is policed at 1200 Mbyte (8 AM-12 AM) and 1050 Mbyte (12 AM-8 AM) threshold. The data policed according to mathematical equation 3.3. The burst of Weibull after policing is then added with the data at the next slot for shaping process. The burst after policing is utilized up to 1600 Mbyte bandwidth. By shaping process, the traffic can be saved and fully utilized of the bandwidth.

In shaping process, the utilization of bandwidth is higher than just only policing. The network traffic has shaped and policed to make 65 percent usage to reach full utilization 1200 Mbyte while policing is at 82 percent usage to reach full utilization. It shows that the shaping after policing process has reach maximum utilization of bandwidth without compromising missing data. For outbound, Weibull distribution is generated with random data based on shape parameter (α=1.9909) and scale parameter (β=379.8954). The random Weibull data policed at 680 Mbyte and 434 Mbyte.

Time series of burst after policing from Weibull outbound shown has used up to 460 Mbyte bandwidth per time. The burst of Weibull outbound data is occupied in the next cycle in order to maximize the bandwidth utilization. That process is called shaping.

The cumulative traffic data increasing over time by comparing outbound traffic, traffic after policing and traffic after shaping and policing. The shaping after policing process gives improvement on utilization of bandwidth and performance of quality of service (QoS). The Goodness of Fit (GoF) test provides the estimate parameter value to do further analysis. From the test, Weibull model is selected to have nearest model fit to time based internet traffic data. The model comes with shape parameter and location parameter.

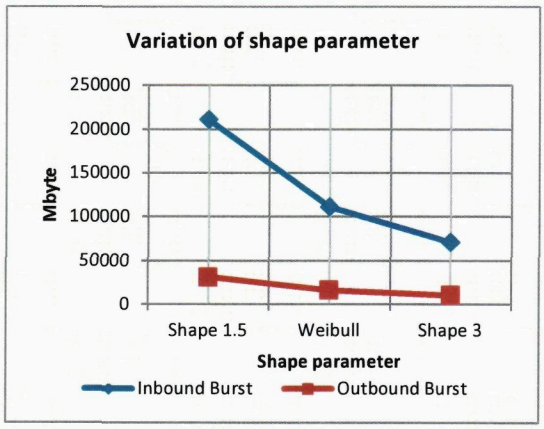


Figure 10: Total Inbound and Outbound Burst according to Shape variation

The total burst of inbound and outbound data is calculated according to variation of shape parameter α. By varying shape parameter, the total burst is also changing. Increasing number of shape parameter will reduce the total burst and higher throughput as shown in figure 10.

CONCLUSION

In this paper, statistical analysis for internet throughput presented. The distribution models were considered in this analysis is normal, lognormal, Weibull and exponential distribution model. The statistical parameter established such as mean, standard deviation and estimate parameter to characterize the behavior of data population. Empirical cumulative distribution function (CDF) is used to present the overview of data distribution and compare between samples. From analysis, Weibull model is found the best model fitted to data of inbound and outbound. The model is selected based on lower Anderson-Darling (AD) value and higher probability value (p-value). The probability values (p-value) show the higher the value concluded the null hypothesis is accepted. The analysis is important to understand the behavior of internet traffic throughput. In order to optimize the bandwidth usage, the network traffic data need to undergo policing and shaping process. The bandwidth of network usage can be saved and the available bandwidth is fully utilized. By varying the shape parameter of Weibull distribution, the burst data can be controlled. The analysis shows the higher shape parameter value, the lower burst of network throughput. This will help to increase the quality of service (QoS) in term of bandwidth utilization, speed and performance.

REFERENCE

[1] R. De O Schmidt, R. Sadre, N. Melnikov, J. Schonwalder, and A. Pras, "Linking network usage patterns to traffic Gaussianity fit," in *Networking Conference, 2014 IFIP*, 2014, pp. 1-9.

[2] K. Fukuda, "Towards Modeling of Traffic Demand of Node in Large Scale Network," in *Communications, 2008. ICC '08. IEEE International Conference on*, 2008, pp. 214-218.

[3] C. Baogang, X. Yong, H. Jinlong, and Z. Ling, "Modeling and Analysis Traffic Flows of Peer-to-Peer Application," in *Innovative Computing Information and Control, 2008. ICICIC '08. 3rd International Conference on*, 2008, pp. 383-383.

[4] M. A. Arfeen, K. Pawlikowski, D. McNickle, and A. Willig, "The role of the Weibull distribution in Internet traffic modeling," in *Teletraffic Congress (ITC), 2013 25th International*, 2013, pp. 1-8.

[5] M. A. Arfeen, K. Pawlikowski, A. Willig, and D. McNickle, "Internet traffic modelling: from superposition to scaling," *Networks, IET*, vol. 3, pp. 30-40, 2014.

[6] S. Hosoki, S. Arakawa, and M. Murata, "A Model of Link Capacities in ISP's Router-Level Topology," in *Autonomic and Autonomous Systems (ICAS), 2010 Sixth International Conference on*, 2010, pp. 162-167.

[7] J.-S. Park, J.-Y. Lee, and S.-B. Lee, "Internet traffic measurement and analysis in a high speed network environment: Workload and flow characteristics," *Communications and Networks, Journal of,* vol. 2, pp. 287-296, 2000.

[8] L. Tsern-Huei, "Correlated token bucket shapers for multiple traffic classes," in *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, 2004, pp. 4672-4676 Vol. 7.

[9] T. M. Au and H. Mehrpour, "Worst case analysis of leaky bucket regulators in broadband integrated networks," in *TENCON '94. IEEE Region 10's Ninth Annual International Conference. Theme: Frontiers of Computer Technology. Proceedings of 1994*, 1994, pp. 1095-1101 vol.2.

[10] M. Kassim, M. Ismail, and M. I. Yusof, "Adaptive throughput policy algorithm with weibull traffic model for campus IP-based network," *Information Technology Journal,* vol. 13, pp. 2632-2644, 2014.

[11] M. Kassim, M. Ismail, and M. I. Yusof, "A New Adaptive Throughput Policing and Shaping Algorithm on Campus IP-based Network," *ARPN Journal of Engineering and Applied Sciences* vol. 71, pp. 79-83, 2015.

[12] M. Kassim, M. Ismail, and M. I. Yusof, "A new adaptive throughput policy algorithm on campus ip-based network internet traffic," *Journal of Theoretical and Applied Information Technology,* vol. 71, pp. 205-214, 2015.

[13] M. Kassim, M. Ismail, and M. I. Yusof, "Statistical analysis and modeling of internet traffic IP-based network for tele-traffic engineering," *ARPN Journal of Engineering and Applied Sciences,* vol. 10, pp. 1505-1512, 2015.

[14] A. Tudjarov, D. Temkov, T. Janevski, and O. Firfov, "Empirical modeling of Internet traffic at middle-level burstiness," in *Electrotechnical Conference, 2004. MELECON 2004. Proceedings of the 12th IEEE Mediterranean*, 2004, pp. 535-538 Vol.2.

[15] S. Kashihara and M. Tsurusawa, "Dynamic Bandwidth Management System Using IP Flow Analysis for the QoS-Assured Network," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, 2010, pp. 1-5.

[16] A. Agbaria, G. Gershinsky, N. Naaman, and K. Shagin, "A bandwidth management approach for quality of service support in mobile ad hoc networks," in *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*, 2009, pp. 1-5.

[17] K. Ravindran, M. Rabby, and X. Liu, "Bandwidth measurement and management for end-to-end connectivity over IP networks," in *Communication Systems and Networks and Workshops, 2009. COMSNETS 2009. First International*, 2009, pp. 1-8.

[18] W. Fugui and P. Mohapatra, "An efficient bandwidth management scheme for real-time Internet applications," in *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, 2001, pp. 469-472.

[19] M. Kassim, M. Ismail, K. Jumari, and M. I. Yusof, "Bandwidth gain analysis for HTTP and HTTPs traffic on IP based network," in *Wireless Technology and Applications (ISWTA), 2012 IEEE Symposium on*, 2012, pp. 303-308.

[20] E. Vayias, J. Soldatos, and G. Kormentzas, "Traffic shaping based on an exponential token bucket for quantitative QoS: implementation and experiments on DiffServ routers," *Comput. Commun.,* vol. 29, pp. 781-797, 2006.

[21] D. S. Daian and D. H. Giura, "Traffic shaping and traffic policing impacts on aggregate traffic behaviour in high speed networks," in *Applied Computational Intelligence and Informatics (SACI), 2011 6th IEEE International Symposium on*, 2011, pp. 465-467.

[22] C. Caini and R. Firrincieli, "Packet spreading techniques to avoid bursty traffic in long RTT TCP connections [satellite link applications]," in *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*, 2004, pp. 2906-2910 Vol.5.

[23] F. Huebner, D. Liu, and J. M. Fernandez, "Queueing performance comparison of traffic models for Internet traffic," in *Global Telecommunications Conference, 1998. GLOBECOM 1998. The Bridge to Global Integration. IEEE*, 1998, pp. 471-476 vol.1.

[24] M. A. Arfeen, K. Pawlikowski, D. McNickle, and A. Willig, "Towards a combined traffic modeling framework for access and core networks," in *Telecommunication Networks and Applications Conference (ATNAC), 2012 Australasian*, 2012, pp. 1-7.

[25] S. Guatelli, B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon, and P. Viarengo, "Application of statistical methods for the comparison of data distributions," in *Nuclear Science Symposium Conference Record, 2004 IEEE*, 2004, pp. 2086-2090 Vol. 4.

[26] A. W. Azim, S. S. Khalid, and S. Abrar, "Analysis of modulation classification techniques using Goodness of Fit testing," in *Emerging Technologies (ICET), 2013 IEEE 9th International Conference on*, 2013, pp. 1-6.

[27] J. L. Romeu, "Anderson-Darling: A Goodness of Fit Test for Small Samples Assumptions," vol. 10, 2003.

[28] E. P. Rathgeb, "Modeling and performance comparison of policing mechanisms for ATM networks," *Selected Areas in Communications, IEEE Journal on,* vol. 9, pp. 325-334, 1991.

[29] "Traffic management specification," vol. Version 4.1, March 1999.

[30] Y. Dashdorj, N. Chuluunbaatar, B. Batzul, and S. Lee, "Characteristics of the token bucket parameters with self similar network traffic," in *Strategic Technology (IFOST), 2010 International Forum on*, 2010, pp. 198-202.