Method for Network Traffic Generation based on User Behavior of Streaming Media

Shuai Zeng¹, Zhongying Wang²*, Guofeng Zhao³, Shaocong Gai⁴, Jie Duan⁵ and Yi Zhang⁶

¹-⁶The School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China
¹Zengshuai@cqupt.edu.cn, ²wzying_cqupt@189.cn, ³zhaogf@cqupt.edu.cn, ⁴zaihuishou_cong@163.com, ⁵duanjie@cqupt.edu.cn, ⁶cyzhangyi@126.com

Abstract

Existing traffic generation methods generate traffic by the flow generator structured and then sent data packets, which had two defects. Namely, the generated traffic is not enough authenticity, and the rate of flow is not high enough. For coping with the mentioned network traffic generation issues, based on user behavior of streaming media and streaming media server was presented. The test results show that comparing with D-ITG, the size of traffic produced by this method has been obviously improved, and its principle indicates that the flow is highly authentic. Rather than following the traditional thinking of sending packet by traffic generator itself, the proposed method provides an innovation scheme which decreases the difficulty of the realization of the traffic generator and has recommendable value for the researches and development of traffic generator.

Keywords: traffic generator, traffic authenticity, streaming media, user behavior

1. Introduction

Traffic Generator is a tool that generates quantitatively network traffic in the network, which plays an important role in many fields, like research in network performance, study on quality of service about network, test and deploy of new protocols, network security [1]. Along with the development of the Internet, the structure of network is increasingly complex and the scale is increasing, the existing traffic generation methods are facing serious challenge in two aspects. One is generation capability of high rate traffic, another is authenticity of traffic. So, how to make the network traffic generated by traffic generator to close to the characteristics of actual network traffic, and the traffic is big enough, has become one of the important research direction.

At present, there are two traffic generation methods which could improve the traffic generation rate in order to meet the demand of high-speed network and interconnection equipment performance test are improve performance of hardware and reduce consumption of system. The former uses better processing chip to increase the sending rate of packets. And the latter improve the traffic rate by enhancing the software performance or decrease the complexity of traffic model. For example, Spirent Avalanche [2] based on traffic model runs on a special hardware platform (Spirent hardware platform), which can be able to generate traffic about 80 Gbps. Ixia IxLoad [3] (Ixia hardware platform) can simulate millions of sessions. But
these traffic generators must be run on specific hardware devices which are expensive, so that they have not been widely used. In addition, TCPivo [4, 5] based on method of traffic playback uses software way to generate traffic. This traffic generator employs low delay priority scheduling mechanism to support a higher rate of data packet replay. But traffic generated is constant because of restrained of log file, so the method is lack of flexibility.

On the other hand, in order to improve the authenticity, there are two methods used commonly, the method based on traffic replay [6, 7] and traffic model [8, 9]. The principle of former is that capturing and recording traffic of real network, then playback it to test network. Although this method can reproduce the state of a period of network, traffic generated is constant because of restraint of log file, such as TCPReplay [10, 11], TCPopera [12]. The other method based on traffic model builds the corresponding mathematical model by analyzing the behavior of network user and characteristics of traffic. Then traffic generator uses the mathematical model to structure and send packets, so the traffic’s characteristics conform to the characteristics of network. In general, the more traffic model has parameters, the more model is precise and the network traffic produced is authentic, which will result in the model is more cumbersome and complex, resource consumption is increasing in the system and ability of Traffic generation will decline. For example: TG [13], RUDE/CRUDE [14, 15] and D-ITG [16, 17], etc.

All traffic generation methods described in above have one common characteristic that is the traffic generator structure and send packets through it to implement the purpose of generating traffic. There are two problems with this approach:

(1) Authenticity of the traffic generated isn’t enough. Although the existing traffic generator can simulate real network traffic greatly, but the authenticity of traffic still Don’t meet requirements, which the main reason is that packets is structured by traffic generator itself, rather than from the Internet server. The traffic produced by only sending packets don’t conform to the actual network, so the authenticity is not strong.

(2) The rate of traffic is not high. The method that traffic generator structures and sends packets through itself will consume lots of system resource, which leads to rates of traffic produced are not high enough, this can’t meet the demand of heavy traffic in gigabit or high-speed experiment network.

Aiming at the shortcomings of the existing traffic generation methods, this paper proposes a method that is based on user behavior and streaming media server to produce traffic, the theory of this method is that by simulating the behavior of the streaming media users to send the corresponding request to streaming media server in the real network, the streaming media servers response it, and then send video data to the client, so traffics are generated in the network. On the one hand, in our method, we visit to the real streaming media server to produce traffics, so the traffics are authentic; on the other hand, in this system, we only need to send a small amount of request packets to streaming media servers, which consumes a small quantity of system resource, and simulates a large number of users at the same time, so this method can generate high rate traffics.

The remainder of the paper is organized as follows. In Section 2, we describe the method for network traffic generation based on user behavior of streaming media in detail. In Section 3, we give the architecture design of this method. In Section 4, we verify this method, analyze and compare the result with other method. In Section 5, we summarize the paper and discuss the future work.
2. Method based on User Behavior of Streaming Media

There is lots of streaming media servers in real network, which can be a part of the traffic generation system. Traffic generator can get response and video data from these servers by sending requests, so we can use a small price to generate a lot of traffic in the network. According to this, we propose a new method which based on behavior of streaming media user and streaming media server to produce traffic. As shown in Figure 1.

![Figure 1. Method for Network Traffic Generation based on User Behavior of Streaming Media](image)

We have two reasons to choose the way to generate traffic through accessing to streaming media servers. The first one is that streaming media application has become one of the most popular applications [18], so it is representative. Second, traffic produced by watching streaming videos much higher than browsing web and SNS services, and video traffic is the main component of network traffic.

Two advantages of this method:

(1) The authenticity of traffic

Because the method is to simulate the behavior of the streaming media users to watch video, which visits streaming media server in the network, so the traffic produced by this method is consistent with users to watch streaming videos, so it is authentic.

(2) A lot of traffic

By this method, the traffic generator only need to send a few request packets streaming media servers, then the servers respond a lot of video data. Because of little system consumption, we can simulate a large number streaming media users to send requests concurrently, which brings about large amount of network traffic.

2.1. Behavior model of streaming media users

In order to ensure the traffic generated by this method conforms to the actual network traffic, we must ensure that behaviors of streaming media users to watch video which
are simulated are authentic (e.g., the operation of users, time of the operation, etc.). So we need to research and analyze the user behavior firstly, then build the behavior model of streaming media users.

The simulate process for behaviors of streaming media users as shown in Figure 2, which mainly contains the following steps:

**Figure 2. The Simulate Process for Behaviors of Streaming Media Users**

The first step, determine the number (we name it as N) of videos will be watched.

The second step, determine one video from N.

The third step, determine the operations. There are a lot of operations when Users are watching videos, such as Forward, Backward, Play, Pause, Stop and so on. We determine operations according to the probability of each operation. If operation is Stop, then going to step 5, otherwise, going to step 4.

The fourth step, determine time of the operation which is from previous step. According to the operation, we can acquire the corresponding execution time. If the time is over, going to step 3.

The fifth step, N must minus 1 each time. We are going to step 6 if N is greater than zero. Otherwise, the user has watched all videos, then exiting.

The sixth step, determine the interval between two videos. Generally, there is a time interval which is mainly used for looking for the next video between two
videos watched by users. To determine the interval which is as dormancy time, then going to step 2.

2.1.1. Determine the number of videos: The amount of videos watched which may be one or more in the process of view is discrepant for different users. We were able to use a random number of the distribution to represent the number videos that will be viewed by user, if you know the distribution and parameters which the number of videos watched obeys.

As the distribution of probability density is shown in Figure 3, the X-axis represents the value, Y axis represents the probability ($f(x), 0 \leq f(x) \leq 1$) of the value. According to the distribution and parameter, we can acquire a random number which represents the number of videos. The method contains the following two steps:

The first step, producing a random number $x$ between Min and Max (Min is the minimum value, Max is the maximum value), then obtaining the corresponding probability value - $f(x)$ through the distribution of probability density;

The second step, randomly generating a random number $i$ from 0 to $M$ (M is the maximum of probability density function), if $i \leq f(x)$, the random number $x$ is we need, which can represent the number of videos that will be watch by a user; Otherwise, turning to the first step, until the random Numbers produced conform to the requirement.

$$f(x) \leq x \leq M$$

![Figure 3. The Probability Density Function](image)

By the second step, the bigger the value of probability density $f(x)$ calculated by the random number $x$ elected, the greater the probability of $x$ is selected, which makes random numbers produced consistent with the specified distribution as a whole.

2.1.2. Determine one video to watch: In daily life, users will generally choose videos which they are interested in or are popular, and the popularity of different videos is variable, so that the probability of different videos watched are discrepant. In order to close to the actual situation, reflect the discrepant popularity of different video, we give a weight to each video. Supposing there are $n$ videos ($V_1, V_2, ..., V_n$), which are endowed with different weights ($W_1, W_2, ..., W_n$). The more popular the video is, the greater weight given is, so the probability of this video viewed is greater.

To determine one video from N to view, we use the following two steps:

The first step, choosing a video $V_i$ randomly (i from 1 to N);

The second step, obtaining the weight $W_i$ of $V_i$, and using $W_i/100$ as the probability of this video selected (for example, the weight of a video is 10, so the
probability to be watched is 10%). Then generating a random number x which is between 1 and 100, if x is less than or equal to \( W_i \), the corresponding video will be watched; Otherwise, turning to the first step until the video selected meets the demand.

The probability \( P_i \) of video \( V_i \) can be selected as shown in formula (1), which indicates that \( P_i \) is affected by \( W_i \). And the sum of probability of all videos is 1.

\[
P_i = \frac{W_i}{\sum_{j=1}^{n} W_j}, \quad \left( \sum_{i=1}^{n} P_i = 1 \right)
\]

2.1.3. Determine the Operations: In the process of users watching videos, the probability of different operations performed are different, so different operation will be given different execution probability, and the greater the probability is, the greater the chance of operation executed. Suppose there are n operations \( (C_1, C_2... C_n) \), which are given different execution probability \( (P_1, P_2... P_n) \), and the sum of all probability is 1. In order to determine the operation will be executed, we take the method described below:

The first step, choosing an operation \( C_i \) randomly from 1 to n;

The second step, we are obtaining the probability of \( P_i \). We are generating a random number x between 1 and 100. If x is less than or equal to \( P_i \) multiplied 100, we are performing the current operation \( C_i \); otherwise, we are turning to the first step, until the operation selected meets the demand.

2.2. Determine the Parameters of the Model

We have analyzed and processed the log file of one video on demand system, then extracted the related parameters of model of streaming media user behavior, which can help us determine the parameters of the model.

2.2.1. Introduction of Video on Demand System Log: We chose three years and four months logs from the video on demand system (February 17, 2010 to June 17, 2013), which contains 835501 records, the format of log from video on demand system as shown in Figure 4. A complete system log record includes fellow content: serial number, type of users, IP address of users, start time of watching video, name of video, duration of video, UNIX time, running time of servers, etc. We need to extract IP address of users, start time of watching video, duration of video. Due to the log file does not record every operation, only the overall playing information such as time of play and names of video, so we can only acquire the numbers of videos viewed and the time interval between videos.

\[
6952665,"server",3","2010-02-23 15:30:28.123000000","",-1.10035,"1575267","-1]\|1266910228"
\]

\[
\]

\[
\]

\[
\]

**Figure 4. The Format of Log from Video on Demand System**

2.2.2. Determine the Interval of Viewing between Videos: The method to extract the number of videos watched by users is calculating how many videos will be view by user in a process, as shown in Figure 5. Assuming that the processes of viewing are
regarded as two times if the intervals of videos are more than 1800 seconds and the formula (2) shows how to calculate the interval.

\[ T = T_1 - T_2 - T_3 \]  

(2)

And T is the interval time, T1 is the moment starting to play of current video, T2 is the moment starting to play of previous video, T3 is the length of previous video.

We adopt the KS test to verify the interval of videos from the video on demand system logs, and the result is shown in Figure 6, which follows the generalized Pareto distribution \((k = 1.4, \sigma = 3.53, \theta = -0.34)\) [19], and the formula (3) is the distribution of probability density function. Figure 6 illustrates the interval of 86.2\% in all processes is less than 60 seconds.

\[ f(x|k, \sigma, \theta) = \frac{1}{\sigma} \left(1 + \frac{k(x-\theta)}{\sigma}\right)^{-\frac{k+1}{k}} \]  

(3)

2.2.3. Determine the Amount of Videos: We use the KS test for the number of videos in one process of viewing from the video on demand system logs, and the results is shown in Figure 7, which follows the generalized Weibull distribution \((\alpha = 0.75, \beta = 1.3)\) [20], and the formula (4) is the distribution of probability density function. Figure 7 illustrates 92.9\% of user view less than or equal to five videos in a process.
The number of videos
Cumulative probability

Data of lof file
Weibull distribution

Figure 7. KS Test for the Number of Videos

\[ f(x | \alpha, \beta) = \frac{\beta}{\alpha} \left( \frac{x}{\alpha} \right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^\beta} \]  

(4)

3. Implementation

In order to achieve the traffic generation method based on behavior of streaming media users, we design and realize the traffic generation system, which contains four key modules: parameters configuration module, user behavior module, traffic request module, log module (as shown in Figure 8). This modular design has following advantage: firstly, the system is easy to maintain and add new function module. On the other hand, it is conducive to shorten the development cycle of software.

Figure 8. The Structure of Traffic Generation System based on Behavior of Streaming Media Users
3.1. Parameters Configuration Module

Developer can configure these parameters, such as the number of users and the user behavior model, according to the different scenarios and requirements. Configurable content mainly includes: the distribution and related parameters of each operation which contains play, pause, forward, and backward. The probability of each operation is performed; list of video, the number of mock users and the interval of each mock user join to this system.

3.2. User Behavior Module

User behavior module simulates the behavior of streaming media users. In our system, the model of user behavior is initialized by these parameters from parameters configuration module, which is used by user behavior module to determine the name of video be watching, operation, and time of the operation and so on. Then, these parameters are passed to traffic request module.

3.3. Traffic Request Module

The function of this module is sending request to streaming media servers. According to the parameters from user behavior module, traffic request module can send the corresponding request to servers. For example, operation of user simulated is playing, so the module sends playing request to streaming media server and start timing. Then the server responds the request and transmits video data.

3.4. Log Module

Log module is aiming to record the related data per second which includes generated traffic, number of mock users, round trip delay of streaming media server, response time of streaming media server and so on. Through analyzing the log file, developer would know the performance of this system and network.

4. Experiment

4.1. Experimental environment

To test the actual performance of our method based on the behavior of streaming media users and streaming media servers, this paper implements a high-performance network traffic generating system which adopts the above method, as shown in Figure 9. The network traffic generating system consists of streaming media servers, traffic generator, experimental network and the Internet. And the traffic generator is a host with Windows 7 operating system configured with 2 x 2.6 GHz Intel Pentium processor, 4 GB RAM.
4.2. The Test for Ability of Traffic Generation System

To test the ability to generate network traffic of our system, we use this system to simulate users to access to streaming media servers, and gradually increase the number of users, record the change of the traffic. In order to better show the ability to generate traffic of our system, we compared with D-ITG traffic generation system, the result is shown in Figure 10.

Figure 10 shows that the traffic produced by D-ITG is consistent with the theoretical value when traffic rate is less than 600 MB/s. However, D-ITG can’t meet the requirements when the traffic rate is required to greater than 600 MB/s. Thus, the ceiling of traffic rate generated by D-ITG is 600 MB/s in such a network environment. But, the actual traffics of our traffic generation system described in this paper keep consistent with the theoretical value. And the Comparison illustrates...
that in the same experimental environment, the traffic rate of our traffic generation system is higher than D-ITG system, and the ability to produce traffic can reach the transmission capacity of this network (1 GBPS), which meet the demand of gigabit experiment network.

4.3. The Test for the Sudden of Traffic

In recent years, many studies have shown that network traffic has characteristics with sudden and self-similarity. Sudden characteristic means network traffic can be rapidly growth in a very short time, then forming some shapes like spikes or saw tooth. Besides, this breaking phenomenon still exists in very wide time scales, and the shape of traffic in this scenario may be not smoother than the traffic shape of narrow time scales. The outstanding trait of sudden characteristic is that the emergency of network traffic doesn’t have fixed length of time, and it can’t be smoothed out, so the sudden of network traffic doesn’t decrease over sampling time increase [21].

In order to verify the traffic’s sudden produced by our traffic generation system, we use this traffic generation system to simulate to 100 users (whose keep performing playing operation with 10000 seconds) to access to streaming media server simultaneously, and then acquire the statistics of traffic. In order to compare, we also employ D-ITG produce 5000 packets per second in the same condition, and collect the traffic generated by this system. We choose 1 second, 2 seconds, 5 seconds, and 10 seconds as the interval of sample to analyze the statistics of traffic collected from two traffic generation systems. For example, if the interval is 10 seconds, so the 10000 seconds will be divided into 1000 time slots, and we gain the average of traffic from each slot. The result is shown in Figure 11 and Figure 12, the red line in figure is the average of traffic per second.

![Figure 11. Traffic Generated by D-ITG](image-url)
In Figure 11, the sudden of network traffic produced by D-ITG is decreased over the interval of sampling increase, and the shape of traffic becomes smooth when the sampling interval is 5 seconds, it even almost become a straight line when the sampling interval is 10 seconds. On the contrary, Figure 12 indicates that the traffic produced by our method based on behaviors of streaming media users have strong sudden in different sampling intervals, it’s sudden doesn’t reduce along with the sampling interval increases, which is consistent with the characteristics of the real network traffic.

Contrasting Figure 11 and Figure 12, we know that the traffic generated by our traffic generation system has strong sudden characteristic, but the sudden of D-ITG is weak when the sampling interval is greater.

4.4. The Test for the Self-similarity of Traffic

Self-similarity means that the part of things can reflect similar features with overall. Self-similarity shows that network traffic has the statistical similarity at all sampling interval or a wide range of time. The similarity shown by the actual network traffic is Long Range Dependence-LRD. H is one of the important parameters to describe the LRD of businesses in Hurst index. There are three kinds of Hurst index [22]:

1. If $H = 0.5$, which indicates that time series is random walk;
2. If $0.5 < H \leq 1$, which states that time series have long-term memory;
3. If $0 < H < 0.5$, which shows that time series isn’t persistent.

So, If $0.5 < H < 1$, the series has LRD, otherwise don’t. And the more $H$ is closed to 1, the stronger the self-similarity is.

Hurst index is the key indicator to evaluate the self-similarity of network traffic, which can be calculated by seven methods [23]: Aggregation Variance method, R/S method, Period gram method, Absolute Value method, Variance of residuals method, Abry-Veitch method, Whittle estimator method. In this paper, we use Time Variance method [24] and R/S method [25].

By using Time Variance method and R/S method, we obtain the Hurst index of network traffic produced by our traffic generation system, the result is shown in Figure 13 and Figure 14.
Figure 13. The Test for the Self-Similarity of Traffic of our Traffic Generation System

Figure 14. The Test for the Self-Similarity of Traffic of D-ITG

Figure 13 and Figure 14 shows the Hurst index of network traffic in our traffic generation system is between 0.9 and 1, but the Hurst index of D-ITG varies from 0.6 to 0.7, so the self-similarity of traffic which adopts the method based on behaviors of streaming media users is higher than D-ITG. And the self-similarity of traffic in our method is strong, which is consistent with the actual network traffic.

5. Conclusions and Future Work

This paper proposes a network traffic generation method based on behaviors streaming media users and verifies the effectiveness of the proposed method. The traffic generation method has two advantages: First, the traffic generated is authentic and controllable; second, the traffic rate is higher.

Future work will focus on the following:

(1) Further research for behaviors of streaming media users. Due to the limitation of the log file, we can only extract the number of videos viewed in one process and the interval between videos. So we will study other parameters in the model of user behavior in the future.

(2) This paper only involves the traffic of streaming media, so we will extend the system to be able to generate other type traffic, such as Web traffic and P2P traffic, etc.
References


Authors

Shuai Zeng, lecturer of The School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, his main research direction is the Future Network and SDN.
Zhongying Wang, Postgraduate of Chongqing University of Posts and Telecommunications, his main research direction is the Future Network and SDN.

Guofeng Zhao, professor of Chongqing University of Posts and Telecommunications, his main research direction is the Future Network and SDN.

Shaocong Gai, Postgraduate of Chongqing University of Posts and Telecommunications, his main research direction is the Future Network and SDN.

Jie Duan, lecturer of Chongqing University of Posts and Telecommunications, her main research direction is the Future Network and CCN.

Yi Zhang, professor of Chongqing University of Posts and Telecommunications, his main research direction is Network management, embedded network technology.