Real-time Simulated Traffic Generation and Verification Methods based on the Traffic Pattern Analysis

Hyuncheol Kim
Dept. of Computer Science
Namseoul Univ.
Cheonan, Korea
hckim@nsu.ac.kr

Wonhyuk Lee
Dept. of Computer Engineering
SKKU / KISTI
Suwon/Daejeon, Korea
livezone@kisti.re.kr

Jeom Goo Kim*
Dept. of Computer Science
Namseoul Univ.
Cheonan, Korea
jgoo@nsu.ac.kr

Abstract— **The explosive growth of Internet traffic has led to a dramatic increase in demand for data transmission capacity, which has spurred tremendous research activities in real-time traffic estimation and related methodologies. High-speed services and bulk transmission were available and various next-generation converged services are undergoing a change by various services. This paper presents improved real-time simulated traffic generation and verification schemes based on the actual traffic pattern analysis. For this, we analyzed traffic patterns of actual application system and generated simulated traffics. We also suggested scheme that verify similarity of simulated traffic and actual traffic.**

Keywords- Simulated Traffic Generation; Pattern Analysis; Traffic Verification

I. INTRODUCTION

With development of information age, the users are requesting faster and more diverse communication services, and for that purpose, several types of super speed transmission technologies and communication services based on various communication technologies are appearing on the stage one after another. As the services and speed provided by the communication networks are rapidly growing together with appearance of such new convergence services, service requirements of the users are very diversely changing from transmission of simple text information to image, voice and moving image information and required bandwidth is also fast increasing.

On the other hand, the communication service provides have conducted communication service demand forecast in order to satisfy such requirements of the users, but they cannot follow change of communication service paradigm upon popularization and super speed of the new type of convergence services. Accordingly, on order to organically provide the current communication services and various types of convergence services to be appeared in the future, it is urgently required to study on the simulated traffic generation and verification algorithm that may be effectively applied to both of the current communication services and new convergence services together with active response against change of the communication service paradigm.

This paper aims to suggest an algorithm that generates the same traffic under the network simulation environments as the actual traffic transferred from the communication networks. For that purpose, this paper has suggested a technique to analyze traffic generation pattern of the actual application system, generate simulated traffic based on it and verify the simulated traffic and the actual traffic.

In addition, this paper has suggested a method to forecast the traffic patterns of new services to be appeared in the future, in order to reflect the fast changing super speed network environments into the simulation environments.

II. REALATED STUDIES

In order to generate the simulated traffic, a study on the techniques to generate and verify the existing simulated traffic must be surely preceded. This section examines a representative modeling techniques to generate the existing simulated traffic and verify the simulated traffic and analyzes them.

A. Techniques to generate the Simulated Traffic

Traffic can be modeled in several patterns based on various characteristics. It can be largely divided into CBR (Continuous Bit Rate) and VBR(Variable Bit Rate), and VBR includes POISSON model, ON-OFF model, IPP (Interrupted Poisson Process) model, AR (Autoregressive) model, MMPP (Markov Modulated Poisson Process) model, etc. [1, 2, 3, 4].

Firstly, CBR is the simplest pattern in which traffic is generated at uniform intervals. CBR is a pattern that is generated from transmission of large file or the existing STM (Synchronous Transfer Mode) based traffic. At the moment, probability distribution function F(t) of the generation interval of cells is expressed as below.

$$F(t) = U(t - T_a) \tag{1}$$

Funding for this paper was provided by Namseoul university.

* Corresponding author

where U refers to a unit step function. The unit step function U(t) has 0 value on t<0, or 1 value on t≥0. In other words, CBR is generated only when t equals to Tc. Even though CBR model is the simplest traffic model, as it generates traffic at uniform intervals, it is difficult to reflect the various characteristics of traffic.

Poisson model has Poisson form of generation distribution in which probability that the number of cells entered during time t is n is expressed as (Equation (2)), and the distribution of cell generation intervals when cell generation rate of this traffic is λ is expressed as (Equation (3)) below.

$$P(n) = c^{\lambda t} \frac{(\lambda t)^n}{n!} \tag{2}$$

$$f(x) = \lambda e^{\lambda x} U(x) \tag{3}$$

$\lambda = \frac{1}{T_p}$ T_p : Average of Generation Intervals

In other words, on Poisson model, when n cells are generated in Poisson distribution within time t, time intervals of individual n cells has an exponential distribution. Poisson model is suitable for generating Poisson distribution traffic that is the representative characteristic of the internet traffic but it has a disadvantage that cannot generate the traffic having another characteristic of self-similarity.

ON-OFF model is a pattern to model traffic with large bustiness and has ON section and OFF section with exponential distribution. There are cells with uniform intervals generated from ON section but no cell from OFF section. ON-OFF model divides sections into ON section and OFF section and generates cells with uniform intervals on ON section as like CBR model. Thus it can be said that it is an advanced pattern of CBR model, but it cannot flexibly adjust intervals of traffic distribution for ON section.

IPP model is a special case of MMPP model and almost similar to ON-OFF model, of which generation of cells from ON section is not CBR distribution but Poisson distribution having λ=IPP(1/Tipp). That is, it has flexibility of ON-OFF model and it is possible to apply Poisson distribution to the traffic model of ON section. It has statistically Hyperexpornential distribution (H2).

$$\int H_2(t) = P\mu_1 e^{-\mu_1 t} + (1-P)\mu_2 e^{-\mu_2 t} \tag{4}$$

On the other hand, as the image signals have different patterns from bit rate generated depending upon image compression techniques (interframe differential pulse code modulation (DPCM), intraframe DPCM, conditional replenishment, motion compensation, transform coding (e.g. Discrete Cosine Transform, Run Length Coding, etc.), etc.) and scene types (video conference, sports, movie), type of camera, brightness of screen, camera moving (zooming, panning, etc.), it is very difficult to model it correctly. Out of them, those image signals with less change of screen such as video phone can be modeled with AR model.

MMPP model is a modeling of the most complicated pattern and sometimes used for modeling of several data sources, multiplied traffic of voice signals or image signals with fast changing scenes. 2-state MMPP can be expressed with summation of Poisson model and IPP model. In addition, as the multiplied s-state MMPP can be 2n-state MMPP model, more complicated traffic modeling may be available.

B. Simulated Traffic Verification Technique

The existing studies had been carried out on the basis of mathematics such as Poisson distribution or exponential distribution for network design rather than measurements. However, recent studies have showed that the actually measured traffic has different characteristics from those of the mathematical model.

Mathematical models are suitable for numerical analysis using mathematical analysis or computers, they may result in waste or short of resources when networks are designed based on them. Therefore, it is required to reflect the actual characteristics of traffic using the actually measured data when designing networks. For that purpose, self-similarity model has been raised and studies on it had been carried out [5, 6].

Self-similarity refers to a characteristic that new series have same self-correlation function as the original series when multiplying several series. That is, when a series of any data has a specific self-correlation function, new series generated by multiplying the data have also same self-correlation function as the original series. Representative methods include Time-Variance Plot technique to look into traffic reduction rate upon multiplication level of traffic, R/S plot technique using a characteristic that R/ S (Rescaled Range) becomes larger according to Power law with H.

R/S plot uses the characteristic that inclination becomes larger according to input load. On the equation, n is a natural number that means sampling interval of data points, R(n) and S(n) indicate statistical number of process and standard deviation of sample within adjusted scopes, respectively. R(n) is the value predefined by the user and S(n) is the actual network traffic. If average of the sample is $\overline{X}(n)$, equations for sample regulation process W_k and $R_{(n)}/_{S(n)}$ are defined as (Equation (5)) and (Equation (6)), respectively.

$$W_k = (X_1 + X_2 + ... + X_k) - k\overline{X}(n), (k \geq 1) \tag{5}$$

$$R_{(n)}/S_{(n)}=\max(0,W_1,W_2,\dots Wn)–\min(0,W_1,W_2,Wn)|/S_{(n)} \tag{6}$$

When expressing log[R(n)/S(n)] and n as log – log graph, a straight line of which inclination exists between 0.5 and 1 is appeared. On the equation for R(n)/S(n), as x increases, the inclination also increases. Therefore, inclination of R(n)/S(n) equation increases. At the moment, as the inclination is closer to 1, it has more correct self-similarity. From the actual internet traffic, as the input load (x) increases and then the strength of self-similarity also becomes larger, we can confirm that verification with RS plot well reflects characteristics of the actual traffic [7].

Time-Variance plot is a method to show self-similarity of the data to be observed in a graph. If we bind m data into one and regard the Random Process of the bundle as $X^{(m)}$ – ($Xk^{(m)}$: K= 1, 2, 3,…), then dispersion of $X_k^{(m)}(X_k^{(m)} - \frac{1}{m_{i-km}} \sum_{(m-u)}^{km} X_i)$ is defined as (Equation (7)).

$$Var(X^{(m)}) = \frac{1}{n}\sum_{k-1}^{n}(X^{(m)} - \overline{X^{(m)}})^2 = E[X^{(m)^2}] - \overline{X^{(m^2)}} \quad (7)$$

If this Random Process has self-similarity, dispersions from the equation above satisfy (Equation (8)).

$$Var(X^{(m)}) \approx \frac{Var(X)}{m^k} \quad (8)$$

III. SIMULATED TRAFFIC GENERATION AND VERIFICATION ALGORITHM BASED ON THE PATTEN ANALYSIS

A. Simulated Traffic Generation and Verification Module

Fig. 1 indicates overall structure of a real-time simulated traffic generator based on distribution function and parameters suggested by this paper. As shown on Fig. 1, this paper mainly consists of network traffic capture module, PDF (Probability Density Function)/ parameter extraction module, traffic generation module, comparison module and Hurst parameter generation module.

At first, Fig. 1 ① shows traffic capture module. It captures network packets to be measured on the traffic capture module and classifies them into flow information and packet header information. From the flow information, it is possible to understand data values such as how much traffic has been generated per a unit time, how much traffic is for a specific time band, etc., while, from the packet header information, it is possible to identify information on the packet itself such as origination and destination, length of packet, generated time, concerned operation system, etc. by analyzing field values contained in the header. This collected information is classified by each application system for comparison with the simulated traffic and required parameters are analyzed and managed separately.

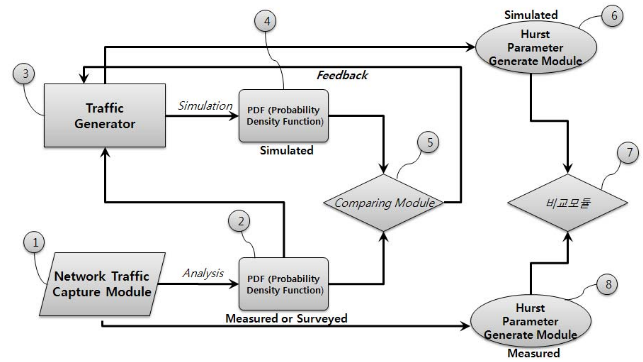

Figure 1. Configuration Diagram of Simulated Traffic Generation and Verification based on Pattern Analysis

Fig. 1 ② indicates PDF/ parameter extraction module, which in turn extracts Probability Density Function (PBF) of the relevant traffic and parameter values upon it through information acquired from the network traffic capture module. If it is required to compare traffic flow, it is possible to see what distribution the traffic amounts through packet capture shows using Q-Q Plot, and for detailed comparison group by the application system, it is possible to prepare traffic information separately using information contained in the packet header and find out distribution using Q-Q Plot. Q-Q Plot is a method to measure whether two distributions contained within the same comparison group are matched and often used as a method to measure which unspecific distribution is mostly similar to and suitable for which specific distribution. Through that, it is possible to find out the optimal probability density function, extract parameter values composing the distribution function and store distribution function-parameter pairs.

Fig. 1 ③ indicates traffic generation module, which classifies the traffics by the existing application systems and models them in order to generate the simulated traffic. During the processes to model the traffic by classifying traffics by the existing application systems as like that, it is required to enter the traffic distribution value suitable for each application system. If we model traffic through simulation, we can enter specific distribution and parameters for it. At the moment, it is possible to enter the measured probability density function and parameter values above. It is possible to generate the simulated traffic through the modeled traffic generator like that.

Fig. 1 ⑤ indicates a comparison module to compare parameters between actual traffic and simulated traffic. If the simulated traffic is generated through the traffic generator, it is possible to extract the probability density function and parameters through this traffic. If the modeling is successful and in correct state, the probability density function and parameters extracted through the simulated traffic should be matched with those extracted from the actual traffic and entered on modeling. If two values show different results, it is required to look into any design error in the traffic generator or pass through correction process through debugging process. If there is also error on the actually measured parameters and simulated traffic parameters through these processes, it is necessary to return to the first verification process and carry out analysis actual network traffic. If the probability density function and parameters measured first are wrong, it cannot be expected that the simulated traffic generated through input of the values would be similar to the actual traffic.

When the flow and parameter values compared through these processes are matched without big error, it may prove that design of the traffic generator and analysis on the parameters extracted from the actual traffic had been done without any abnormality.

Fig. 1 ⑥, ⑦ and ⑧ indicate Hurst parameter generation and comparison module. Hurst parameter generation and comparison module compares level of self-similarity of the actual traffic and simulated traffic through elf-similarity of the traffic. Level of elf-similarity can be identified by calculating Hurst parameter values through R/S Plot, VT Plot or Wavelet Method, etc. If Hurst parameter measured from the actual traffic shows a value between 0.5 and 1, then it can be said that it shows self-similarity and it is possible to identify what level of self-similarity is shown through the value. Since the simulated traffic generated by simulating the actual traffic should show self-similarity as like the actual traffic, it is possible to calculate the Hurst parameter value for the simulated traffic using the same method. If Hurst parameter value of the simulated traffic does not exist within the scope so that it does not show self-similarity, it becomes difficult to compare it with the actual traffic and two traffics can't be regarded as similar. Since two traffics should show same characteristics in order to demonstrate that they are similar, they should have common characteristics for self-similarity, and if both traffics have self-similarity, Hurst parameter value should also show a figure of match without any big error.

B. Similarity Verification Simulation

Before verifying the suggested traffic generation module, we perform a simulation for verification method of the traffic similarity suggested on this paper. Sequence of the simulation is shown on Fig. 2.

Firstly, it is required to collect actual traffic information as shown on Fig. 2. For that purpose, it is required to analyze information by collecting traffic generated from the concerned server using Packet Sniffers such as Tcpdump or Wireshark. As it is to simulate a methodology for measurement of the simulated traffic similarity, it is required to select specific application traffic, analyze flow information of the traffic and record the data into a file such as Excel.

Afterwards, as shown on Fig. 2, it is required to derive the optimal distribution and parameters and use Q-Q Plot to find out the optimal distribution of the actually measured traffic. If we indicate a specific distribution and the actually measured traffic distribution into Q-Q Plot, we can see whether two distributions are matched or not. If points on the Plot show a straight line shape, it means that two compared distributions are matched. With this method, we can find out the optimal distribution of the actually measured traffic from several standard distributions. In order to draw Q-Q Plot, this paper uses statistics programs such as SPSS to extract probability density function and parameter values depending upon it.

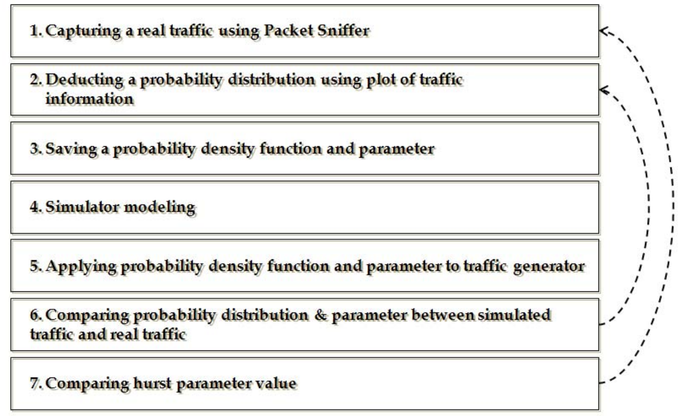

Fig. 2 Simulation Process through Probability Density Function and Self-similarity

Afterward, we carry out traffic modeling. We apply probability density function and parameters acquired by actually measuring one application system to the traffic modeling. At the moment, when it is required to design a simulated traffic generator to generate traffic generated from on application system through the simulator, it is possible to input the probability density function and parameters extracted above. When entering a specific distribution as like that, the simulator generates simulated traffics following the relevant distribution.

Finally, we compare the resulting values. It is possible to calculate the probability density function and parameter depending upon it from the generated simulated traffic as like the actually measured traffic. We also use Q-Q Plot for the case. We determine if the resulting value of the extracted simulated traffic as like that is matched with the probability density function and parameter values applied on traffic modeling. If the values are matched, we can assume that two traffics had been similarly generated. In order to verify if these two traffics have the same characteristics, we need to compare Hurt parameter values indicating level of self-similarity. We calculate the level of self-similarity on the actually measured traffic and simulated traffic with the Hurst parameter calculation method suggested on this paper and use it as verification of traffic characteristics. If we earned a result that two values are not matched each other or have contradict characteristics, we can't regard that two traffics are similar and we need to return to the processes to get the probability density function of the actually measured traffic and model the application system and carry out the simulation again.

C. Simulation Analysis

Fig. 3 indicates information collected to calculate Inter-arrival time. A column refers to the relative time values arrived by each packet and B column refers to Inter-arrival times calculated using differences of these times.

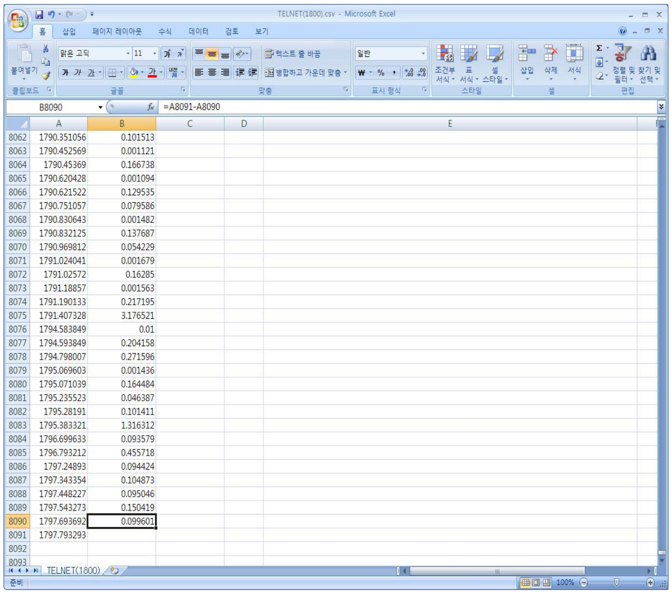

Fig. 3 Inter-arrival Time Calculation on the Extracted Traffic

For traffic modeling, we use OPNET Modeler. At the moment, the traffic modeling is simply organized so as to generate the simulated traffic through only distribution derived from the actual traffic and input of the parameter values.

This simulation uses R/S Plot that is most commonly used to measure level of self-similarity. Equation to be used for R/S Plot is calculated on the basis of packet arrival ratio of the traffic per hour.

Since the finally derived Hurst Parameter value 0.7304 together with results of some equations calculated during process of R/S Plot passing through these processes exist between 0.5 and 1, we can confirm that packet arrival ratio per hour of the traffic measured through this simulation has self-similarity.

IV. Conclusion

As the internet grows, increasingly many systems are connected to the network and network traffic shows gradually complicated aspect due to appearance of various services and rapid increase of user population.

Generally, traffic flows show different characteristics each other by application services, which can be modeled with mathematical distributions. Traffic generators of the most network simulation tools only one mathematical probability distribution such as exponential distribution or Poisson distribution to generate traffics for simulation in order to model such traffic flow. However, due to various services of the actual networks and traffic characteristics depending upon them, such a specific distribution can't completely reflect the characteristics of the actual traffic. Accordingly, if designing and implementing a network based on them, as it may result in waste or short of resources, it is required to conduct modeling by well identifying the characteristics and flow of the traffic actually generated by each application service.

This paper proposed an algorithm to generate the same pattern as the actual traffic transferred from the communication network under the network simulation environments. For that purpose, this paper proposed a technique to analyze traffic generation pattern of the actual application system, generate simulated traffic based on it, and verify similarity of the generated simulated traffic and actual traffic.

In addition, this paper proposed a traffic forecasting method of the new services to be appeared in the future in order to reflect fast changing super speed network environments into the simulation environments.

References

[1] A. Veres et al, "On the Propagation of Long-Range Dependence in the Internet", Computer Communication Review, Vol. 30, No. 4, Oct. 2000

[2] A. Veres, Zs. Kenesi, S. Molnar and G. Vattay, "On the Propagation of Long-Range Dependence in the Internet," SICCOMM. 2000.

[3] C. You and K. Cbandra, "Time Series Models for Internet Data Traffic," 24th conference on Local Computer Networks. 1999.

[4] Mark F. Crovella, Azer Bestavros, "Self-Similarity in World Wide Web Traffic Evidence and Possible Causes," IEEE/ACM Transactions on Networking. vol. s, No. 6, pp. 835-840, December. 2000.

[5] Y. Wang, CL. (Thou, CH. Huang. C.C. Hwang, Z.M. Yang, CC. Chiou, and CC. Lin(2003), "The design and implementation of the NCTUns 1.0 network simulator", Computer Networks: The International Journal of Computer and Telecommunications Networking Volume 42, Issue 2, pp 175-197.

[6] W. Willinger, M. Taqqu, K. Sherman, D. Wilson, "Sell-Similarity Through I-ugh-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," A IEEE/ACM Transactions on Networking. Vol. 5, No. 1, pp. 71-86, Jan. 14, 1997.

[7] Won-Ki Hong, S. U. Park, Y. M. Kang and J. T. Park, "Enterprise Network Traffic Monitoring, Analysis and Reporting Using Web Technology", Journal of Network and Systems Management, Plenum Press, March, pp. 89-111. 2001.