Predictive Network Traffic Engineering for Streaming Video Service

Haruhisa Hasegawa, Shouji Kouno, Akihiro Shiozu, Masaaki Sasaki, and Shinsuke Shimogawa
NTT Network Technology Laboratories
3-9-11 Midori-Cho Musashino-Shi, Tokyo 180-8585, Japan
E-mail: {hasegawa.haruhisa, kouno.shouji, shiozu.akihiro, sasaki.masaaki, shimogawa.shinsuke}@ lab.ntt.co.jp

***Abstract*—Next Generation Network services with Fiber-to-the-home have been spreading in Japan; therefore, the number of viewers using Video on Demand (VOD) services has also increased. Network operators are required to maintain service quality during peak hours, so they need to design bandwidth for VOD at those times. It is obvious that congestion occurs when large number of viewers watch VOD programs simultaneously. However, viewer behavior does not depend on only internal factors of VOD service but also external ones such as natural phenomena and social events. This makes it difficult for network operators to predict how traffic will increase and to design bandwidth adequately. We analyzed actual VOD traffic and clarified that it is effective to use a log-normal distribution model. We also predicted traffic distribution based on Bayes' method.**

Index Terms—VOD, Traffic management.

I. INTRODUCTION

Traffic volume of content delivery services has been rapidly increasing in telecommunication networks [1]. There are two major services for distributing content, multicast and unicast. Multicast is efficient for delivering a smaller number of programs simultaneously to a large number of viewers [2],[3]. In contrast, streaming services via unicast are obviously less efficient than via multicast because each viewer requires his/her own independent session. These sessions cannot be aggregated into one multicast tree. Recently, over-the-top (OTT) Video on Demand (VOD) services have been becoming increasingly popular [4]. At the same time, traffic has also been increasing rapidly since VOD services adopt unicast streaming over the network. As a result, telecom network operators face the problem of fitting the resources to the traffic characteristics of a VOD service. VOD traffic is usually transmitted with best-effort service over networks. This makes it is impossible for telecom carriers to guarantee the quality of service (QoS) of a VOD session by using bandwidth reservation technologies such as call admission control [5], [6]. If a large number of viewers are going to use a VOD service at the same time, traffic congestion will occur. This degrades the quality of experience since video streaming service is sensitive to QoS degradation, i.e., packet-loss and delay. To prevent degradation, it is necessary for telecom carriers to estimate the peak traffic of a VOD service and provide enough bandwidth for it.

II. TRAFFIC CHARACTERISTICS OF VIDEO STREAMING SERVICE

Subscriber behavior that causes congestion does not depend on only internal factors of a VOD service but also external ones such as natural phenomena and social events. For example, VOD traffic is strongly affected by TV broadcasting programs. TV broadcasting services have a huge number of viewers. If a TV program becomes uninteresting to viewers, most will quickly move to VOD services. On the other hand, telecom operators do not manage traffic or design network facilities based on TV programs. This makes it difficult for network operators to predict how traffic will increase and to design bandwidth adequately.

Network engineering based on traffic estimation has been extensively studied. The applications of traffic estimation are categorized into two types. One type is based on modeling short-term stochastic characteristics of traffic for queuing system design or traffic control. The other type is estimating traffic for relatively longer term forecast to design a network to fit future traffic. In the latter type, future traffic is predicted based on the historical traffic data. The Holt-Winters method [7] has been widely used for telephone traffic [8]. Since traffic characteristics are becoming complicated, estimating traffic data as a time series has been studied to identify its trend and daily, monthly, or yearly pattern by using time series models. There have also been many proposals to determine the parameters of each model [9]. As Internet traffic volume has been increasing, the major usage of subscribers has been constantly changing. Neural network techniques are used to solve this problem. However, a sudden traffic peak occasionally appears, and it is difficult to predict in a time series since such a peak is caused by external factors for network operation. To solve this problem, we analyzed actual traffic data of a VOD service and clarified the characteristics for designing network resources. We found that the use of a model based on log-normal distribution was effective. We explain how to estimate the upper bound of traffic by using predictive traffic distribution.

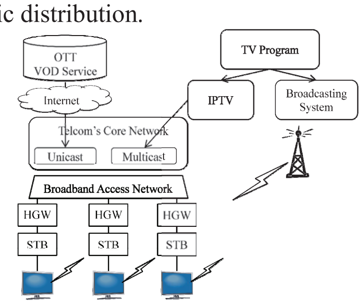

Fig. 1 System architecture of OTT-VOD services in Japan

A. Eco system of VOD service

A typical eco-system architecture of an OTT-VOD service in Japan is illustrated in Fig. 1. A TV program from a TV station is provided by both a broadcasting system and IP multicast service. Since VOD traffic is transmitted via unicast, the traffic also proportionally increases if the number of VOD viewers increases. This means that traffic volume on the network greatly depends on viewer behavior. In the next subsection, we discuss the analysis results of the relationship between viewer behavior and VOD traffic characteristics.

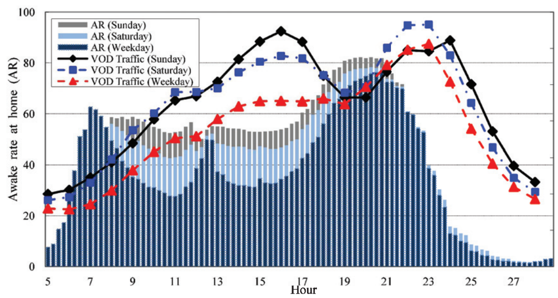

Fig. 2 Daily VOD traffic and awake rate at home (AR)Analysis of traffic based on viewer behavior

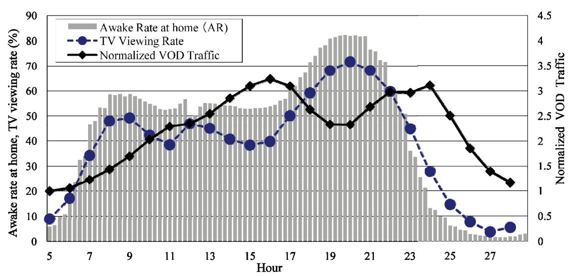

Fig. 3 VOD traffic, AR, and TV viewing rate on Sunday

B. Analysis of traffic based on viewer behavior

Figure 2 shows daily VOD traffic activity. We estimated VOD traffic of January 2010 at the interface between a telecom's IP network and a VOD service provider. The traffic volume was averaged hourly by day of the week and then normalized using the value at 5:00 on Monday. The rate of the population who are awake at home is shown as a histogram. We call this the awake rate at home (AR). We derived the AR based on the survey of [10]. The AR is calculated as the sum of the participation rates in home activities.

Weekday traffic had a single peak between 20:00-24:00. Weekend traffic had two peaks; at 14:00-17:00 and 21:00-25:00. The difference between Weekday and Weekend is clarified by the AR. The traffic between 13:00-18:00 depended on the AR. However, the peak traffic at night appeared later than the AR peak. It is not appropriate to explain a VOD peak simply based on the AR. VOD traffic and the TV viewing rate [12] on Sundays are shown in Figure 3. Since the most popular programs are broadcast around prime time (18:00-21:00), the rate of TV viewing greatly increases. As a result, VOD traffic is suppressed even if the AR is high.

Next, we discuss how social events or disasters affect VOD traffic. We show the daily VOD traffic per number of subscribers in Figure 4. Each plot indicates the maximum traffic, which is selected among hourly VOD traffic divided by the daily number of subscribers. There was a traffic peak on 2011 May 29, which was a holiday, when a powerful typhoon hit Japan. The TV programs broadcast, which had been scheduled for times with lower AR, were not as popular compared to those scheduled for prime time. However, the AR was actually higher than usual. As a result, a relatively large number of people tended to switch from TV programming to VOD service. A similar situation occurs on other holidays with bad weather. This phenomenon was most noticeable after the Great East Japan Earthquake on March 11, 2011. Almost all broadcasting programs were about the earthquake. However, it was not necessary to watch all day in areas far from the stricken areas. Therefore, VOD traffic increased to the highest level in these unaffected areas.

Based on the analysis, VOD traffic is highly dependent not only on the internal factors of VOD services but also external ones, such as natural phenomena, social events, and TV programs. It is quite difficult for operators to predict such external factors. Therefore, a technique for estimating the peak of VOD traffic in advance by taking into account traffic characteristics is required.

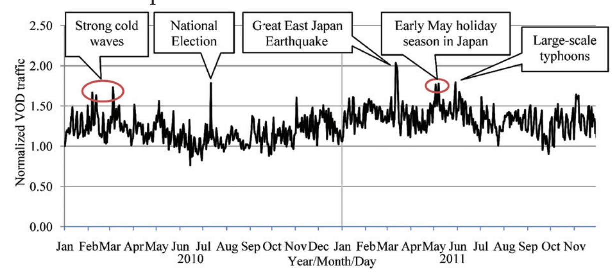

Fig. 4 Change in VOD traffic per number of subscribers

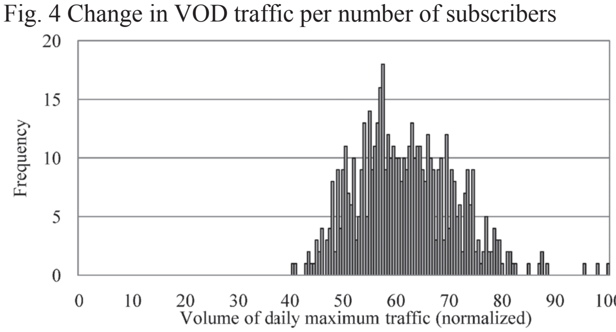

Fig. 5 Daily maximum traffic related to VOD service

III. MODELING IN TERMS OF PEAK TRAFFIC

Video streaming services are sensitive to QoS degradation. When traffic is transmitted with best effort QoS class in a congested network, some sessions may degrade. When a many VOD requests are concentrated, it means that service is attractive to most users and QoS reduction greatly deteriorates customer satisfaction. Network designers, therefore, need to evaluate not only the average trend in traffic but also the peak characteristics of VOD to manage related traffic. Figure 5 is a histogram of daily maximum traffic related to VOD of 2010-2011. The data were normalized by traffic at 22:00 on March

12, 2011 (the day after the Great East Japan Earthquake); the time at which highest traffic of all data was observed. The distribution is asymmetric with a longer tail on the right side. We estimated the distribution of traffic data by using Bayesian inference. The probability density function of traffic volume is written as $f(t|\theta)$. Here, t is traffic volume and θ is the m dimensional parameter vector of other observed events. The series of traffic data that have already been observed is written as $\{t_0, t_1, \cdots, t_{n-1}\}$. The elements of the series, t_i ($i = 0, 1, \cdots, n-1$) were assumed to be independent of each other and to have the same likelihood function.

$$f(t_0, t_1, \cdots, t_{n-1}|\theta) = \prod_{i=0}^{n-1} f(t_i|\theta)$$

Next, the probability density function of $\theta = \{\theta_1, \theta_2, \cdots, \theta_m\}$ is assumed to follow the function $\pi(\theta)$. Then, the posterior probability is derived with Bayesian theory. Using the posterior probability of $\pi(\theta \mid t_0, t_1, \cdots, t_{n-1})$, we can estimate the predictive distribution of traffic $w(t)$.

$$w(t) = \int_{-\infty}^{\infty} f(t|\theta)\pi(\theta|t_0, t_1, \cdots, t_{n-1})\, d\theta$$

Predictive traffic distribution depends on the type of likelihood distribution. We estimated five types of likelihood distribution: normal, gamma, Poisson, beta, and log-normal. We used the Bayesian information criterion (BIC) to estimate how the likelihood distribution fits the actual distribution [6]. A smaller BIC better fits actual traffic characteristics.

The results of the BIC and 99th percentile points are shown in Figures 6. (a) and (b) are related to holidays and all days, respectively. If a point is closer to the upper right than other ones, it is more suitable to fit actual traffic characteristics. The log-normal, gamma, and beta distributions are more suitable than normal and Poisson. This means that the distributions of the longer tail on the right side fit the actual traffic better than the distribution with symmetric tails. This analysis suggests that there are relatively fewer days with factors that would increase traffic. The distribution that expresses the peak against these factors is suitable for traffic management.

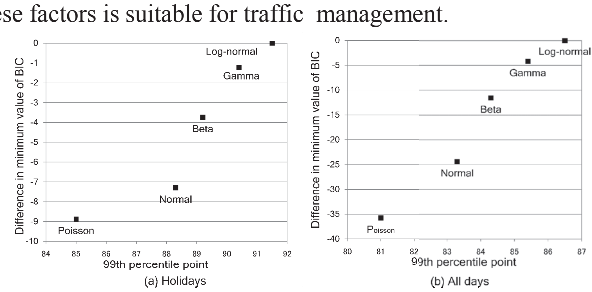

Fig. 6 Scatter diagrams of 99th percentile point and difference from minimum value of BIC

IV. Evaluation for Traffic Management

Based on the results discussed in the previous section, we used a log-normal distribution as a likelihood function. Assuming application to actual network operation, we compared predictive traffic distribution based on the traffic data of 2010-2011. Here, $\{x_j, y_j\}$ and $\{x'_i, y'_i\}$ are the points on the density function and histogram. We defined the distance d when x_j equals x'_i to measure the error between the predictive distribution and the actual histogram.

$$d := \sqrt{\sum_{i,j} \left| y_j - y'_i \right|^2 \Delta t}$$

It is important to design bandwidth against traffic near the peak value. Therefore, we summarize the evaluation of distances related to the 80th percentile for characteristics near the peak and whole region in Table 1. The log-normal distribution has the smallest distance near the peak, and the normal distribution has smaller distance values than the gamma and log-normal distributions. This is because the normal distribution is suitable in the region of the middle range, i.e., near the mode of the histogram. If it is necessary to design bandwidth concerning averaged traffic, a normal distribution should be adopted.

Table 1 Evaluation of distances between predictive distribution and actual histogram

	log-normal		gamma		beta		normal		Poisson	
	Distance	Rank	Distance	Rank	Distance	Rank	Distance	Rank	Distance	Rank
Near peak	0.00468	1	0.00484	2	0.00488	3	0.00524	4	0.00790	5
Wholeregion	0.13879	5	0.13790	3	0.13796	4	0.13562	2	0.07638	1

Table 2 Conditions for peak traffic characteristics

ID	Conditions
N1-1	Any hour when traffic is maximum at all hours of every observed day
N1-2	Any hour when traffic is maximum at all hours of Sundays, Saturdays, and holidays
P1-1	Higher traffic between 22:00 or 23:00 of Saturday and previous day before holiday
P1-2	Higher traffic between 22:00 or 23:00 of first weekend of month
P1-3	Higher traffic between 22:00 or 23:00 after first Tuesday of month

Table 3 95th and 99th percentiles of predictive traffic distribution to actual histogram

	Log-normal		Normal		Gamma		Beta		Poisson	
percentile	95%	99%	95%	99%	95%	99%	95%	99%	95%	99%
N1-1	77.5	**86.0**	76.5	83.0	77.0	84.5	77.0	85.0	73.0	78.0
N1-2	82.0	**91.0**	81.5	88.0	82.0	90.0	82.0	90.0	78.0	83.0
P1-1	82.5	**93.5**	79.0	86.0	81.0	90.0	81.0	90.0	74.0	79.0
P1-2	82.5	**94.5**	79.5	87.5	80.5	90.5	81.0	91.0	73.5	79.0
P1-3	80.5	**89.0**	79.5	86.0	80.0	88.5	79.5	87.5	75.5	81.0

We investigated to what degree each likelihood function would fit the actual histogram related to peak conditions. The conditions in which traffic tends to greatly increase are summarized from an analysis of actual past traffic characteristics in Table 2. From our experience, it is much more frequent for traffic to be highest between 22:00 and 23:00 than at other times. Condition P1-1 denotes Saturday and previous day before holidays. It is known that major VOD services send their subscribers a list of the latest content on the first weekend of the month. As a result, people tend to watch VOD programs more often on weekends. Therefore, we set condition P1-2 as denoting the first weekend of the month. We also found that the latest episodes of dramas are released every Tuesday; therefore, we set condition P1-3 as denoting the weekend after the first Tuesday of the month, which is more specific than P1-2.

The 95th and 99th percentiles of predictive traffic distribution to the actual histogram are listed in Table 3 for conditions in Table 2. The maximum values of all conditions for the same likelihood are underlined. The maximum values of all likelihood functions for the same condition are in bold. The 99th percentile of the log-normal distribution for condition P1-2 was the largest in the estimation. The second largest 99th percentile was also the log-normal distribution for condition P1-1. The 99th percentile of P1-3 was smaller than that of P1-2 for a log-normal distribution. On the other hand, the difference between P1-3 and P1-2 was relatively small for a normal distribution. Since P1-3 was more specific than P1-2, the data under P1-3 were extracted through screening. It is believed that the effect from the combination of other factors was almost cancelled, and the distribution became symmetric. It is not suitable for a log-normal distribution to symmetrically fit into a histogram because the mean is larger than the mode. On the contrary, since a normal distribution is suitable for symmetrically fitting into a histogram, it is considered to suit the condition, which is defined in order to extract the traffic distribution related to a specific factor such as P1-3.

The difference between the 95th and 99th percentiles of the log-normal distribution was larger than that in the normal distribution. The ranking in descending order concerning the difference between the 95th and 99th percentiles is as follows: log-normal, beta, gamma, normal, and Poisson. The 99th percentile of P1-2 was the largest among all conditions for the types of distribution that had longer tails on the right side, i.e., log-normal, beta, and gamma. In contrast, the 99th percentile of N1-2 is the largest for a normal distribution. Since N1-2 is a less specific condition than the conditions of the P series, the data have high values and the distribution has larger variation. As a result, the 99th percentile became high for N1-2. If the condition was more specific, the variation and 99th percentile became smaller. It is thought that approximating the entire histogram based on the normal distribution model is suitable if a specific peak condition is identified. In such a case, however, the histogram must be fit after screening data from empirical knowledge, such as in P1-3. On the other hand, when only raw traffic data without screening based on the empirical knowledge are available, it is appropriate to use a model with a distribution that has a longer tail on the right side.

If network operators design bandwidth by taking into account traffic in the peak range, a log-normal distribution should be adopted as the model. When an operator of a network applies the normal distribution, he/she tends to recognize the peak as an abnormal value. In fact, when the operator adopts a symmetric distribution, the bandwidth will become insufficient in peak traffic with a longer tail on the right side. Since VOD services have such traffic characteristics, designing bandwidth for a VOD service should be based on a log-normal distribution because sufficient QoS should be provided when many subscribers want to use the service.

V. Modeling Based on Log-Normal Distribution

The results discussed in the previous sections indicate that the log-normal distribution model is useful for estimating near-peak traffic. In fact, various factors may affect traffic volume. It is believed that many factors occur simultaneously and increase traffic at peak hours. Factor $\boldsymbol{X}$ is a set of n independent probability variables $\{x_1, x_2, \cdots, x_n\}$. Here, t_i is the traffic that factor x_i affects. Therefore, d_i, the increase in traffic by factor x_{i+1}, is expressed by the following equation

$$d_i = t_{i+1} - t_i$$

We assume that d_i, is proportional to function $g(s_i)$, which expresses a sensitivity to factors. The AR is set as s_i.

$$t_{i+1} = t_i + x_{i+1} \cdot g(s_i)$$

Therefore, the following equation is obtained.

$$x_1 + x_2 + \cdots + x_n = \sum_{i=0}^{n-1} \frac{t_{i+1} - t_i}{g(s_i)}$$

The change by each factor is assumed to be very small, and the following equation is therefore approximated.

$$x_1 + x_2 + \cdots + x_n = \int_{t_0}^{t_n} \frac{1}{g(s_i)} dt$$

The left side of this equation is approximated to a normal distribution by becoming n to infinity from the central limit theorem. Therefore, $log(t)$ follows a normal distribution. Thus, the characteristics of traffic affected by many factors fit well with a log-normal distribution.

VI. Conclusion

The normal distribution model often used in traffic management cannot express near-peak traffic characteristics, such as those of a VOD service. There may be many factors that increase traffic, but network operators are not able to identify such factors in actual network operation. For example, it is difficult for network operators to manage traffic depending on natural phenomena and TV programs in actual network operation. We analyzed actual VOD traffic and found that it is effective to use a model based on a log-normal distribution. Moreover, we were able to predict traffic distribution based on Bayes' method. Network operators can estimate the upper bound of traffic by using a predictive traffic distribution.

References

[1] Cisco, “Cisco Visual Networking Index: Forecast and Methodology, 2011-2016,” May 30, 2012.

[2] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, "Watching Television Over an IP Network," ACM IMC 08.

[3] Hikari TV http://www.hikaritv.net

[4] H. Yu, D. Zheng, B. Zhao, and W. Zheng, “Understanding User Behavior in Large-Scale Video-on-Demand Systems,” ACM EuroSys 06.

[5] H. Kasahara, J. Nishikido, K. Oda, K. Onishi, and Y. Kajiyama, “Network Core Technologies for a Next Generation Network,” NTT Technical Review, Vol. 5, No. 6, June 2007.

[6] ITU-T Recommendation Y.2111 "Resource and admission control functions in Next Generation Networks," 2006.9.

[7] P. R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," Management Science Vol. 6 1960.

[8] A. Passeron, L. Etev, "Forecasting model to set-up multi-season traffic matrices," International Teletraffic Congress 10 (ITC 10), 1983.

[9] C. D. Pack, B. A. Whitaker "Kalman filter model s for network forecasting," Bell System Technical Journal Vol. 61, 1982.

[10] http://www.stat.go.jp/english/data/shakai/index.htm

[11] H. Akaike, A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, Biometrika, 66(2): 237–242, 1979.

[12] Video Research Ltd. TV audience measurement http://www.video.jp