Influence of User's Behavior about Delay on Media Server

Hoon Lee* *Lifelong Member*

ABSTRACT

At present multimedia service composed of voice, data, and video service is prevalent in the Internet. As such, wide-scale penetration of Internet service imposes tremendous pressure to the network infrastructure such as the media servers and links as well as the nodes. In addition, users from a large-scale population require broad bandwidth and high level of QoS. This requires a network with reliable and scalable services to customers, which also necessitates a realistic method for the design of the a media server. In this work, we explore the influence of user's behavior about delay on the performance of media server that takes into account the system and user attributes in a realistic manner. By incorporating user's behavior about the delay-sensitivity, we present an analytic framework for the evaluation of the performance of the media server, via which we illustrate a meaningful intuition in the provision of Internet multimedia service.

Key Words : Internet Service, Media Server, Quality of Service, Grade of Service, Performance Analysis

I. Introduction

Advance of technology for the IP network with enhanced QoS(Quality of Service) functions gave birth to a fast growth of BcN(Broadband convergence Network). As such, rich multimedia services such as IPTV(Internet Protocol TV) as well as HSI(High-Speed Internet) and VoIP(Voice over IP) are provided by ISPs(Internet Service Providers).

On November 2008 TPS(Triple Play Service) was launched in Korea with addition of real-time IPTV service to the current VoIP and HSI services[1]. Contrary to the richness in the contents for the TPS and rapid penetration rate, customers are not sure about the quality of the TPS service. For example, when it comes to the traditional Internet service, people tolerate a certain amount of delay when they connect to a web site or send an e-mail. This made it possible to model a web or mail server by assuming a delay-insensitive user model that ignores the behavior of users about delay such as balking and reneging.

However, when it comes to IPTV service, one may feel that it is very nasty if the TV signal is not displayed immediately after one turns on the TV set or if one switches to a new channel. This means that the most important factor in the success of IPTV service is guarantee of QoS, especially the initial presentation delay or switching delay between different channels, which is comparable to the current cable TV service. The phenomenon like this can also happen in case of SIP(Session Initiation Protocol) server in the VoIP service and web server in the HSI service, too. One may have an experience to give up an access to the web server when the delay is too long before the response is arrived.

Recently, Lee et al. discovered the fact that customers begin to get nervous when the delay for the web access time is greater than 15 seconds[11].

We conjecture that user's dissatisfaction for the delay in the interactive service such as IPTV and

※ This research is financially supported by Changwon National University in 2009-2010.

* Dept. of Information & Communication Engineering, Changwon National University (hoony@cwnu.ac.kr)

논문번호 : KICS2010-05-211, 접수일자 : 2010년 5월 13일, 최종논문접수일자 : 2010년 7월 28일

on-line game will be more severe as compared with HSI service. This motivated us to explore the effect of user's delay-sensitivity to the performance of the media server for the Internet services such as IPTV. For the simplicity, let us call the server for the real-time services such as IPTV or VoIP service as a common name media server(MS).

In the year 2009 KCC(Korea Communications Commission) released a framework for the evaluation of QoE(Quality of Experience) for the IPTV service[2]. The QoE measures include channel switching time and video/audio qualities over the diverse load states of the IPTV system as well as the installation time for the set-top box. Among them the delay can mostly affect the QoE for the media server due to the real-time property of the applications.

Generally speaking, the packet delay can be defined for a flow, and it is identified and specified by a connection identification field, and network equipment or user terminal can control the QoS based on the specified value[13]. The former is called as a network-initiated QoS control, whereas the latter is called as a terminal-initiated one.

The installation time for the set-top box is determined by the physical characteristics of the user equipment. However, the QoE measures such as the channel switching time and experienced video/audio qualities over the diverse load states of the IPTV system depends on the state of the IP network as well as the server itself, where server is defined by an entity that arbitrates the reception of the request message for the IPTV service from the users and distribution of the requested program to the users.

Also, it is usually known that IPTV service requires sustained guarantee of bandwidth throughout the distribution of the program. This allows no opportunity for the oversubscription of the users, which is opposed to the traditional capacity dimensioning method for HSI.

In the previous work, we had presented a method for the design of the access and backbone network for the IPTV service in [4].

Noting that the bandwidth resource in the access and backbone part of the IP network is appropriately provisioned by ISP[3], the remaining problem lies in an appropriate model for the media server.

When it comes to the design of a media server, lots of problems have to be resolved such as the capacity, location, routing between servers and clients, link dimensioning, guarantee of QoS, etc.

We can find lots of works that promote the adoption of distributed server placement for the IPTV service[5]. If such a service environment is successfully implemented in the real-field, lots of problems have to be resolved such as the modeling of the user behavior concerning the user's tolerance about the delay, determination of the number of video contents to each server, optimal capacity of the link that takes into account the GoS(Grade of Service), etc[6].

Molina et al. proposed a general model for content delivery network which takes into account the distributed server location in [7]. They assumed two-level server model with one origin server and multiple surrogate servers, and they investigated the delay performance of the proposed model[7]. However, they did not consider the user's behavior about delay.

Agrawal et al. argue that a server for the large-scale IPTV network can be represented by an M/M/c/K queue[8]. This has an important implication, because the basic principle for the design of IPTV server lies in the flow-awareness, via which one can secure a reliable QoE to the IPTV service per customer. However, that work presents neither practical discussion about the system environment nor the behavior of users about delay.

Based upon our review from many literature other than those described above, we found that little work has been done in modeling the behavior of an MS which takes into account the user's reaction about the delay of the system even though it is a very important problem in the

design and operation of BcN. To the best of author's knowledge, we could find no results that deal with such a practical approach about evaluating the performance of MS regarding the delay and user's behavior. This motivated this work, and the novelty of this work lies at this point.

This work intends to propose a novel analytic method to evaluate the performance of an MS by taking into account the user behavior about incumbent delay in the MS.

This work is composed as follows: In Section II we present system architecture for the server of an IPTV service as an example. In Section III we propose a method to quantify the effect of user's behavior about the delay to the performance of an MS, which is the main contribution of this work. In Section IV a discussion about the performance of an MS is given by numerical experiment. Finally, in Section V, we summarize our work.

Ⅱ. System Model

There exist different types of media servers for the different types of services. IPTV service has a video server in the video distribution network, VoIP service has an SIP(Session Initiation Protocol) server, and HSI has a web server in the service provider's network, etc.

In order to focus on the delay-sensitivity in the access to the server, let us assume an IPTV service. Network architecture for the IPTV service varies depending on ISPs. Nevertheless, a typical IPTV network has three distinct areas: contents provider(composed of cluster of video servers), network provider(access and backbone network based on IP), and users who are connected via ISP.

Recently, total separation of storage from streaming is a trend for the large-scale video services, where the most popular content is cached at the edge of the network, whereas less popular content is transferred from the original storage, via which scaling problem can be resolved. This results in distributed server locations. In case video servers are located in a distributed manner, there exist problems such as the location and dimensioning of servers. We have investigated the server location and dimensioning problem in our previous study, which is described in [4]. Now our focus lies in the analysis of a media server by taking into account the behavior of users about delay in the server. We do not distinguish the type of MS, which can cover all the servers located at the cache as well as the main server.

Video contents that are stored at various video servers(especially, video storage servers) are first gathered at an MS, and they are transferred to an individual customer as a real-time stream when a request for video contents is issued by a user. It is usual that a video program from the video server to MS is transferred as unicast traffic. Multicasting is done after the video traffic arrives to an appropriate distribution node in the network. Our discussion here focuses on the initial phase of the video distribution in which a request message for the unicast of a video traffic is concerned.

In summary one can note that MS for the IPTV service acts as a broker for the process of a request message for a specific video program between the users and contents provider. That is to say, a single MS accommodates a very large number of customers for the process of request about the video contents. One can find that this is a typical client and server system, where MS is a server and users are clients.

Ⅲ. Media Server Model

An MS receives a message about the request of a TV program from the users, and it will translate the message and transfer it to a corresponding video server in the IPTV contents cluster. Then, the corresponding program is sent to the user.

Therefore, the MS is acting as a service broker for the IPTV service. As such tremendous amount of request messages that are generated from a large number of customers will arrive at an MS.

To provide a good-quality service to the users, the MS has to prepare a sufficient capacity for the process of the request messages from a large number of customers, otherwise the request message can not be served by MS. Therefore, we argue that a connection level GoS(Grade of Service) which is defined by CBP(Connection Blocking Probability) has to be defined for MS.

When it comes to a down-link(a link from MS to the network) it is usual that the streaming nature of video program requires sustained guarantee of bandwidth throughout the connection holding time, via which the packet level QoS for the video stream can be sustained.

It is usual that the down-link capacity between the users and IP network is appropriately provisioned by ISP before the service is introduced. See our previous work for this problem[4].

In this work let us investigate the up-link between the IP network and MS. Especially, we focus on the blocking probability of the request message at MS over which the requests for the video programs are served.

To model such a behavior of MS, let us assume a few conditions. First, let us assume that the request for video contents is generated by a group of users that are connected to an IP network following a Poisson process with mean rate λ. This is practical because the scale of users for a nation-wide IPTV service is sufficiently large. Second, when the request messages are accepted by MS, they are processed by a group of servers, where each server supports each video content with random processing time. Noting that the link capacity of MS is finite, we assume that at most c connections are served simultaneously. The buffer capacity for MS is assumed to be finite with size B.

Let us assume that the service time for the request message is exponentially distributed with mean service rate μ. The total system load is given by $\rho = \lambda/(c\mu)$.

Summarizing our argument about the system model, it is denoted by an $M/M/c/K$ queue, where $K=c+B$.

Now let us investigate the behavior of users for the IPTV services. As we have described above, users may experience unexpected delay for their requests for the IPTV program to be arrived at the MS, because MS has to support a large number of customers.

As one may have experienced in the real life, people can be patient or not for the unexpected delay. Some people waits without complaint, but some does not.

The behavior of customers in the IPTV service is considered to be the same. First, we can imagine that all the customers are patient, so that they are tolerant about the delay. In this case, all the request messages for the IPTV programs enter the MS if there is a vacant space in the buffer. Let us call this type of users as delay-insensitive or *persistent users*.

On the other hand, there may exist customers who are impatient. The impatient customers are concerned about the delay for their transactions. Using the auxiliary service that is provided by the network operator such as the notice of network state, customers can obtain an information about delay for their requests[9].

An impatient customer enter the system at first, and after that it leaves the system if it can not withstand the long delay. Such customers are called as delay-sensitive or *reneging users*. The behavior of reneging user is investigated in more detail later.

In the following subsections we will investigate the performance of MS for each type of user.

3.1 Delay-insensitive users

The requests for the IPTV program from the persistent users enter the MS if there is a vacant space in the buffer. Let $N(t)$ be the system state at time t, which is defined by the number of requests that are in the buffer and servers. Let $p_n^{DI}(t) = P\{N(t)=n\}, 0 \le n \le K$, be the probability that the system is in state n at time t. The probability that the system is in state n at equilibrium $(t \to \infty)$ is denoted by p_n^{DI}.

If one use the concept of birth and death process(BDP) to the above system, one can obtain a closed-form formula p_n^{DI} for the state of the system in equilibrium. A discussion about the details of the queuing analysis for the *M/M/c/K* queue is quite long, and it can be found at [10].

Noting that the mean arrival rate being λ and mean service rate of each server being μ, and the number of server is c, we can represent the mean rate of arrival and service at an MS as a function of the state of the system, which is given as follows:

$$\lambda_n = \lambda, 0 \le n \le K,$$
$$\mu_n = \begin{cases} n\mu, n < c, \\ c\mu, c \le n \le K. \end{cases} \quad (1)$$

Then, using the concept of BDP for the *M/M/c/K* queuing system, we obtain a formula for the equilibrium probability that the system is in state n, which is given as follows:

$$p_n^{DI} = \begin{cases} \left(\frac{\gamma^k}{n!}\right) p_0, 0 \le n < c, \\ \left(\frac{c^c \rho^n}{c!}\right) p_0, c \le n \le K, \end{cases} \quad (2)$$

where $\gamma = \frac{\lambda}{\mu}$ and $\rho = \frac{\gamma}{c}$, and p_0 is given as follows:

$$p_0 = \left(\sum_{n=0}^{c} \frac{\gamma^n}{n!} + \frac{c^c}{c!} \sum_{n=c+1}^{K} \rho^n \right)^{-1}.$$

Note that p_n^{DI} is the probability mass function(*pmf*) for the state of the system at equilibrium to be n. So, let us denote it by $f_{DI}(n)$.

On the other hand, the probability that a request message is not allowed to enter the MS due to buffer overflow is the probability that the state of the system is in state K, which is given as follows:

$$p_K^{DI} = \frac{c^c \rho^K}{c!} p_0. \quad (3)$$

Note that the above formula is the message blocking probability, because a newly arrived request message is blocked due to buffer overflow. So let us denote it by ϕ_{DI} .

3.2 Delay-sensitive users

When it comes to the MS with delay-sensitive users, a newly arrived customer joins the buffer if there is a vacant space. The mean arrival rate for the customers is independent of the system state, and it is given by λ.

The mean service rate of each server is μ and the number of server is c. As to the behavior of the customers, some customer determines to leave the system when a certain time is passed in waiting at the buffer. The time of the customer's departure varies from person to person, which is a function of the system state, and so it can be modeled as a random variable. Let us call this system as an *M/M/c/K-R* system, where R indicates *reneging*.

Noting that the state of the system for the *M/M/c/K-R* queue can be represented by the number of customers, it is realistic to define the reneging function as a function of the state of the queue.

There may exist lots of types of functions representing the behavior of users about the delay-sensitivity such as linear function, logistic function, and exponential function, etc. Among them, we conjecture that the exponential distribution is the most realistic one. Tezcan have assumed this kind of function, too[12]. So, let us assume the exponential function for the probability mass function of reneging where the rate of reneging is ν.

Note that, when there are n customers in the system, only n-c customers are located at the buffer and c customers are located at the server.

To model such a user's behavior about delay at the buffer, let us note that the total service rate of the system can be rearranged to the following equation:

$$\mu_n = \begin{cases} n\mu, 0 \le n < c, \\ c\mu + (n-c)\nu, c \le n \le K. \end{cases} \quad (4)$$

In order to avoid a confusion in the notation, let us denote that the mean offered load to the system is defined by $\rho = \frac{\gamma}{c}$, where $\gamma = \frac{\lambda}{\mu}$.

Now let us assume that the probability that the system is in state n at equilibrium is denoted by p_n^{DS}. Then, using the standard BDP model for the queuing system, the steady-state probability p_n^{DS} for the *M/M/c/K-R* system of which the state being n is given by

$$p_n^{DS} = \begin{cases} \left(\frac{\gamma^k}{n!}\right) p_0, 0 \le n < c, \\ \left(\frac{c^c}{c!}\rho^n \xi_{n-c}\right) p_0, c \le n \le K, \end{cases} \quad (5)$$

where $\xi_l = \left(\prod_{i=1}^{l}\left(1+\frac{i\nu}{c\mu}\right)\right)^{-1}$, and p_0 is given as follows:

$$p_0 = \left[\sum_{i=0}^{c}\frac{\gamma^i}{i!} + \frac{c^c}{c!}\sum_{j=c+1}^{K}\rho^j \xi_{j-c}\right]^{-1}.$$

Note that p_n^{DS} is the probability mass function for the state of the system at equilibrium to be n. So, let us denote it by $f_{DS}(n)$.

The blocking probability that a request message is not allowed to enter the buffer due to buffer overflow is the probability that the state of the system is in state K, which is given as follows:

$$p_K^{DS} = \frac{c^c}{c!}\rho^K \xi_{K-c} p_0. \quad (6)$$

Note that the above formula is the blocking probability of the request message for the *M/M/c/K-R* system. So let us denote it by ϕ_{DS} .

Ⅳ. Numerical Experiments

Note that our purpose lies in the evaluation of the performance for MS where the behavior of users for the delay in the system is acting as a parameter. As we have discussed in Section III, different user's behavior about delay incurs different performance in the system. As we have mentioned in the previous section, we have two different scenarios for the user behavior, the delay-insensitivity and delay-sensitivity. For each case, let us evaluate the *pmf* of the system state and the loss probability of the request message.

In order to investigate the implication of the proposed model for MS, let us assume a small scale network for the purpose of simplicity in computation. An extension to a large-scale network is trivial.

In our experiment, let us assume that the common parameters for MS are given as follows: the server capacity is 5, the queue size is 5, and the mean service rate is normalized to be 1. For the delay-sensitive users, the mean reneging rate is assumed to be 0.2, which is considered that the rate for the reneging is not so high. We conjecture that the performance of the MS depends on the offered load: the heavier the system is, the larger the influence of user's behavior of delay-sensitivity to the performance of the MS. Therefore, we assumed three cases: low, medium, and high load for the system.

Fig.1. illustrates the *pmf* of the state for the MS for the three cases of offered load. The *x*-axis illustrates the system state, whereas *y*-axis represents *pmf* of the system state. The left index indicates the offered load, whereas the right index represents the attributes of user. For example, (0.3,DI) represents the system that accommodates

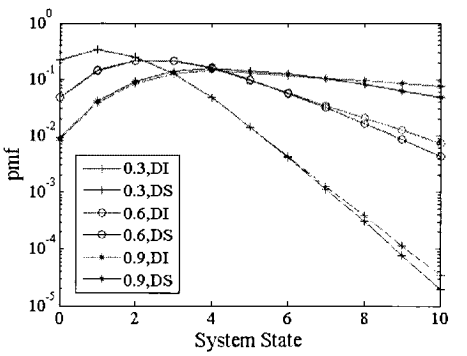


Fig.1. *pmf* for the MS

delay-insensitive users and offered load of 0.3.

Note that *pmf* is depicted as a log-scale, because it spans a large spectrum of values.

As we can see from Fig.1, the effect of user's sensitivity for the delay is not so evident about the system state for the low offered load of 0.3. This results from the fact that the incumbent delay is not so large, so that the number of customers who renege from the system is not large, either.

Now, when the offered load is 0.6, the system with moderate load, the *pmf* increases about a few tens to hundred times that of the low load system, which indicates that the effect of user's sensitivity for the delay is apparent for the moderate offered load.

When the offered load is 0.9, the system with high load, the effect of user's sensitivity for the delay is evident throughout all system state. This indicates that, as the offered load increases, the number of customers who renege from the system increases, too. Therefore, one can find that the effect of user's sensitivity for the incumbent delay is not negligible, and network operator has to take this fact into account in the design and operation of MS. Note also that there exists a cross-over point(system state=7 in this case) for the *pmf* between the system with delay-insensitive and delay-sensitive users. This indicates that the system with delay-sensitive users operates in relatively low state as compared with the system with delay-insensitive users, from which the probability of system overflow becomes lower. This comes at the cost of a priori reneging from the impatient customers.

In Fig.2 we illustrated the blocking probability of a connection(loss probability of request message) by varying the mean offered load to the system. The x-axis represents the mean offered load to the system, whereas y-axis represents the mean blocking probability for the request message.

As we can find from Fig.2, the blocking probability for the system that accommodates the delay-sensitive user is smaller than that for

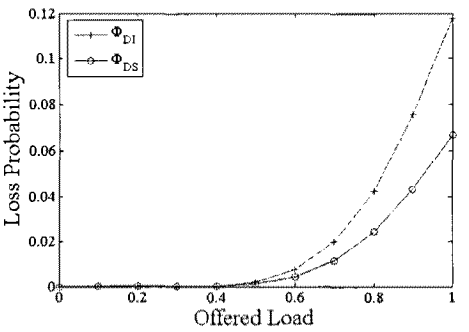


Fig.2. Loss probability

delay-insensitive user, which is obtained at the cost of a priori reneging due to impatience in the reneging user's delay-sensitivity.

Note also that the difference of loss probability for the request message becomes larger as the offered load increases. This indicates that the effect of impatience from the delay-sensitive customers becomes evident as the system becomes heavy.

On the other hand, note that the reneging behavior of users can contribute to the avoidance of system congestion, because the blocking probability for the system with reneging user is always smaller than that of persistent user. This is in line with the advocacy of the introduction of CAC(Call Admission Control) mechanism at the edge of the network for the avoidance of the degradation of the QoS for the incumbent users.

Note that the user's behavior is a reactive and self-control, whereas CAC is a preventive and forced-control. However, the effect is almost the same. This gives us a useful intuition for the design and control of network resources for the various types of services in the future Internet. Note that this intuition can be also useful in the operation of conventional Internet services such as web browsing, file transfer or e-mail servers, too.

V. Conclusions

In this work we proposed a method to analyze the effect of user's behavior about the delay to the performance of MS. Especially, we proposed

a mathematical model for evaluating the *pmf* of the system and blocking probability for the request message in MS by investigating the behavior of customer's delay-sensitivity.

To be more practical, we considered the behavior of users about the access of the IPTV service, which is incorporated into modeling the MS with delay-sensitive users. We categorized the users into delay-patient and delay-impatient users. For each type of user we evaluated the *pmf* of the system and probability of blocking for the request message. Especially the probability of blocking for the request message can be used as a measure of GoS for the IPTV service.

By carrying out numerical experiments, we illustrated the validity and practical implication of the proposition: First, we found that the performance of MS depends on the behavior of users toward the incumbent delay in waiting at the buffer of the MS. Second, we also found that the performance depends also on the offered load of the system. The heavier the system state is, the severer the reneging is.

From this observation we obtained an important intuition in the design of the IPTV service, which is the main contribution of this work.

There remain lots of future research areas related to this work. First, we are going to investigate the delay performance of the system such as mean delay that results from the system sojourn for a diverse set of user parameters such as the mean generation rate of the request message, the mean holding time, etc, because the characteristics of traffic for the future services will be different from the current services.

We are also going to investigate the user's behavior in more detail by accumulating the real-field data over the ongoing real-time services such as IPTV, VoIP, and videophone services.

We will also investigate the psychological aspects of the users by investigating whether users really renege in accessing the IPTV service when network is congested. This can be incorporated into the estimation of the more tangible GoS, and it can be utilized in the design of network for the future multimedia service.

References

[1] K. U. Ho, "Launching the commercial real-time IPTV service in Korea," *The Chosunilbo, November, 17, 2008.*

[2] "Evaluation of QoE for the IPTV service begins from 2009," *http://www.etnews.co.kr, November, 20, 2008.*

[3] M. Cha et al., "Case study: resilient backbone design for IPTV services", *IPTV Workshop, International WWW conference, May 23, 2006, Edinburgh, Scotland, UK.*

[4] Hoon Lee, "A practical network design for VoD services", *Vol.34, No.3, March, 2009, Journal of KICS.*

[5] Y. Takeda and Y. Kurihara, "Construction of fault-tolerant networks using server-load balancing," *Nikkei Communications, May 15, 2008.*

[6] G. O'Driscoll, Next generation IPTV services and technologies, *Wiley-Interscience, 2007.*

[7] B. Molina, C. E. Palau, and M. Esteve, "Modeling content delivery networks and their performance", *Computer Communications 27 (2004).*

[8] D. Agrawal, M. S. Beigi, C. Bisdikian, and K.-W. Lee, "Planning and managing the IPTV service deployment", *10th IFIP/IEEE Int'l Symposium on Integrated Network Management, Munich Germany May, 2007.*

[9] Y. Takizawa, "A review on the NTT FLET'S service for the last ten years", *Nikkei Communications 2009.11.15.*

[10] Hoon Lee, Queuing systems for Engineers, *Hongneung Publishing co., 2008, Korea.*

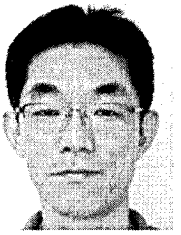
[11] Hoon Lee and Youngok Lee, "Evaluation of MOS for the access delay of Internet service" *Journal of KICS, Vol.34, No.9, September, 2009.*

[12] T. Tezcan and J. G. Dai, "Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic", *Operation Research Vol.58, No.1,*

Jan.-Feb., 2010.

[13] M. Alasti, B. Neekzad, J. Hui, and R. Vannithamby, "Quality of service in WiMAX and LTE networks", *IEEE Communication Magazine, May, 2010.*

Hoon Lee Lifelong Member



B. E. and M. E. from Kyungpook National University

Ph.D. from Tohoku University, Japan.

Feb. 1986~Feb. 2001 KT R&D Center

Mar. 2001~ Changwon National University

Mar. 2005~Aug. 2006 Visiting Scholar, University of Missouri-Kansas City, USA.

<Research fields> Network design and performance evaluation, Internet traffic engineering, QoS and charging