Maximum User Utility on Contents Delivery System with Multiple Priority Classes

Kyoko YAMORI[1], Yoshiaki TANAKA[1,2] and Haruo AKIMARU[3]
[1]Global Information and Telecommunication Institute, Waseda University
1-3-10 Nishi-Waseda, Shinjuku-ku, Tokyo, 169-0051 Japan
Tel : +81 3 5286 3831
[2]Advanced Research Institute for Science and Engineering, Waseda University
17 Kukui-cho, Shinjuku-ku, Tokyo, 162-0044 Japan
Tel : +81 3 3203 9434
[3]Professor Emeritus, Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, 441-8580 Japan.
Tel : +81 532 44 1551
e-mail : yamo@aoni.waseda.jp, ytanaka@waseda.jp, akimaru@beach.ocn.ne.jp

Abstract: For contents delivery systems, the service is considered in which the utility depends on each priority class. This paper deals with the multiple priority class of the contents delivery system from the viewpoint of the utility. The willingness to pay (WTP) is introduced as a measure of utility, and the optimum condition is analyzed to maximize the total user's utility. For the system with multiple priority classes, the optimum condition is given in terms of the traffic load, waiting time for service for each priority class. Systems with the priority classes, 1, 2 and 3 are analyzed, and the effect of the number of priority classes is examined.

1. Introduction

For contents delivery systems, such as video-on-demand service, the utility depends on the waiting time until the service is available. There have been a number of studies concerning the utility for such systems[1],[2].

This paper deals with the contents delivery systems with multiple priority classes from the viewpoint of the utility. The formula for estimating the mean waiting time is given for service available in packet transmission systems. The willingness to pay (WTP) is introduced as a measure of utility, and for its quantitative estimation formula is proposed. The objective function is defined as the total user's utility, and the optimum condition is analyzed to maximize the objective function. The optimum condition is given in terms of the traffic load, and the utility for each priority class. Numerical examples are shown to analyze the effect of the parameters. For the systems with multiple priority classes, 1, 2 and 3, the optimum conditions are analyzed, and the maximized total user's utilities are compared to examine the effect of the number of priority classes.

Section 2 describes the traffic model to evaluate the waiting time for service, and introduces the formula for the willingness-to-pay. Section 3 defines the objective function as the total user's utility, and analyzes the optimum condition to maximize the objective function. Section 4 discusses the number of priority classes. Section 5 summarizes the results obtained and gives concluding remarks.

2. Modeling

2.1 Traffic model

Consider a contents delivery system, for example, on a LAN, which is regarded as a single server model. Figure 1 shows the model with Q priority classes. Assume here the pre-emptive priority model, in which the higher priority class packets preempt the lower priority packets in service, if any. The motivation of using this priority model is that it is simple and convenient for analysis, because a lower priority class traffic has no effect on a higher priority class performance. Due to this simplicity, the optimum condition is given analytically, using which the effects of the parameters, such as one for the WTP formula, are examined.

Assume that requests of the data delivery occur at random (Poisson distribution) and that sufficiently large (infinite) buffers are provided. The contents data are transmitted in packet form for which the packet size is exponentially distributed. Then, the mean waiting time T_i for service available (all the packets to be sent) for priority class i is given by (See Appendix A)

$$T_i = \frac{h}{\left(1-\sum_{j=1}^{i-1}\rho_j\right)\left(1-\sum_{j=1}^{i}\rho_j\right)} \quad (i=1,\cdots,Q). \tag{1}$$

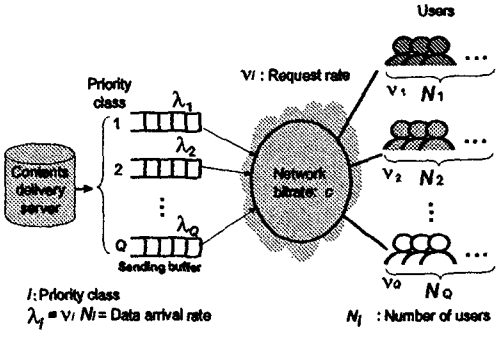

Figure 1. Preemptive priority traffic mode.

where h is the net mean transmission time of the data, and ρ_j the traffic load of the priority class j.

Letting H be the mean data volume, c the transmission bitrate, and if the overhead by the packet header is negligibly small, h is approximated by

$$h = H/c. \tag{2}$$

Denoting by N_i and ν_i the number of users and user's request rate (number of requests per unit time), respectively, for priority class i, we have

$$\rho_i = N_i \nu_i h. \tag{3}$$

2.2 Measure of utility

Let us introduce "Willingness-to-Pay: WTP" as a measure of utility for service [4]-[7]. The WTP is one of the methods of social sciences, and means a price of "How much willing to pay" for a certain service. In this paper, the WTP is used to estimate the utility of service.

Denote by U the WTP and by T the mean waiting time until the service is available. It is assumed that the decreasing rate of WTP against the mean waiting time, $-dU/dT$, is proportional to WTP, U.

In this case, we have the formula,

$$U = De^{-kT}. \tag{4}$$

The parameter D is a scale factor and may be selected arbitrarily. The parameter $k > 0$, with dimension [time^{-1}], is statistically estimated by opinion tests [8]. It is reported that the formula (4) presents a good agreement with measured data with about $k = 0.2$/sec [9].

3. Optimum Condition

3.1 Optimum condition

Define the objective function V, "number of users $\times$ request rate $\times$ utility" which represents the total user's utility as follows:

$$V = \sum_{i=1}^{Q} N_i \nu_i U_i = \frac{1}{h} \sum_{i=1}^{Q} \rho_i U_i. \tag{5}$$

In this section, the optimum condition for $Q = 2$ is examined. From (1) the mean waiting times T_1 and T_2, respectively, for the priority and non-priority classes are given by

$$\begin{aligned} T_1 &= h/(1-\rho_1) \\ T_2 &= h/[(1-\rho_1)(1-\rho)] \end{aligned} \tag{6}$$

where ρ_1 and ρ_2 are, respectively, the traffic loads of priority and non-priority classes, and $\rho = \rho_1 + \rho_2$ total traffic load. From (4) the utilities U_1 and U_2 of respective classes are given by

$$U_i = De^{-kT_i} \quad (i = 1, 2). \tag{7}$$

Putting

$$f_i = \rho_i U_i, \quad (i = 1, 2) \tag{8}$$

we have

$$V = \frac{1}{h}(f_1 + f_2). \tag{9}$$

If V is maximized, then,

$$\partial V/\partial \rho_i = 0, \quad (i = 1, 2) \tag{10}$$

from which, noting that f_1 is independent of ρ_2, we have

$$\begin{aligned} &\partial f_1/\partial \rho_1 + \partial f_2/\partial \rho_1 = 0 \\ &\partial f_2/\partial \rho_2 = 0. \end{aligned} \tag{11}$$

Using (6) and (7) in (8), from (11) we have the optimum condition (See Apendix B.),

$$\begin{aligned} &f_1\left(\frac{1}{\rho_1} - \frac{kh}{(1-\rho_1)^2}\right) - khf_2 \frac{2(1-\rho_1)-\rho_2}{[(1-\rho_1)(1-\rho)]^2} = 0 \\ &\frac{1}{\rho_2} - \frac{kh}{(1-\rho_1)(1-\rho)^2} = 0. \end{aligned} \tag{12}$$

For numerical calculation, solving the second equation in (12) for $0 < \rho < 1$, and using

$$\rho = 1 + \frac{kh}{2(1-\rho_1)} - \sqrt{kh + \left(\frac{kh}{2(1-\rho_1)}\right)^2} \tag{13}$$

in the first equation in (12), the optimum solutions are obtained by iteration.

3.2 Numerical examples

Table 1 shows calculated results of the optimum condition. For example, with mean data volume $H = 10$Mbit and the bitrate $c = 20$Mbps, from (2) the mean data transmission time is $h = 0.5$sec (packet header overhead neglected). With $k = 0.2$/sec, we have $kh = 0.1$, and the calculated results by (12) and (13) are shown in Table 1 (with the scale factor $D = 1$). It is shown that to obtain the maximum utility $V = 0.5251$, the total traffic load should be kept at $\rho = 0.7868$ with ρ_1 and ρ_2 as shown. The optimum condition may be attained by selection the number of users N_i in (3). The optimum waiting times and utilitis are $T_1 = 1.2796$sec, $T_2 = 6.0025$sec, $U_1 = 0.7742$ and $U_2 = 0.3010$.

Table 1 Optimum Condition for $Q = 2$

kh	ρ_1	ρ_2	ρ	V
1.0	0.2628	0.2071	0.4699	0.0837
0.1	0.6092	0.1776	0.7868	0.5251

4. Analysis of Number of Priority Classes

In this section, the cases of $Q = 1$ (without priority class) and $Q = 3$ are examined.

4.1 Non-priority control: $Q = 1$

In the case of non-priority control, $Q = 1$, letting the traffic load be ρ, from (1) and (4) the mean waiting time T and the utility U are given by

$$\begin{aligned} T &= h/(1-\rho) \\ U &= De^{-kT}. \end{aligned} \tag{14}$$

From (5), we have the objective function,

$$V_1 = \frac{\rho U}{h}. \tag{15}$$

If V_1 is maximized, then,

$$\frac{\partial V_1}{\partial \rho} = \frac{1}{h}\left(U + \rho \frac{\partial U}{\partial \rho}\right) = 0 \tag{16}$$

from which, we have the optimum condition,

$$\frac{1}{\rho} - \frac{kh}{(1-\rho)^2} = 0. \tag{17}$$

Solving (17) for $0 < \rho < 1$, we obtain the optimum traffic load,

$$\rho = 1 + \frac{kh}{2} - \sqrt{kh + \frac{(kh)^2}{4}}. \tag{18}$$

4.2 Three priority classes: $Q = 3$

In the case of $Q = 3$, from (1) and (4), the mean waiting time T_i, and the utility U_i for class i are given by

$$\begin{aligned} T_1 &= h/(1-\rho_1) \\ T_2 &= h/[(1-\rho_1)(1-\rho')] \\ T_3 &= h/[(1-\rho')(1-\rho)] \end{aligned} \tag{19}$$

$$U_i = De^{-kT_i}, \quad (i = 1, 2, 3)$$

where $\rho' = \rho_1 + \rho_2$ and $\rho = \rho_1 + \rho_2 + \rho_3$.

Using (8), the objective function is given by

$$V_3 = \frac{1}{h}(f_1 + f_2 + f_3). \tag{20}$$

If V_3 is maximized, then,

$$\frac{\partial V_3}{\partial \rho_i} = 0, \quad (i = 1, 2, 3) \tag{21}$$

From which, we have, in a similar manner as for $Q = 2$, the optimum condition,

$$\begin{aligned} &f_1\left(\frac{1}{\rho_1} - \frac{kh}{(1-\rho_1)^2}\right) - khf_2\frac{2(1-\rho_1)-\rho_2}{[(1-\rho_1)(1-\rho')]^2} \\ &\quad -khf_3\frac{2(1-\rho')-\rho_3}{[(1-\rho')(1-\rho)]^2} = 0 \\ &f_2\left(\frac{1}{\rho_2} - \frac{kh}{(1-\rho_1)(1-\rho')^2}\right) - khf_3\frac{2(1-\rho')-\rho_3}{[(1-\rho')(1-\rho)]^2} = 0 \\ &\frac{1}{\rho_3} - \frac{kh}{(1-\rho')(1-\rho)^2} = 0. \end{aligned} \tag{22}$$

For the numerical calculation, solving the third equation in (22) for $0 < \rho < 1$ and using

$$\rho = 1 + \frac{kh}{2(1-\rho')} - \sqrt{kh + \left(\frac{kh}{2(1-\rho')}\right)^2} \tag{23}$$

in the first and second equations in (22), we obtain the optimum solutions by iteration.

4.3 Numerical examples

The calculated examples are shown in Table 2. Figure 2 compares the maximum total user's utility with Q=1,2 and 3 for various kh.

Table 2 Optimum Condition for $Q = 1$ and $Q = 3$

Q	kh	ρ_1	ρ_2	ρ_3	ρ	V
1	1.0				0.3820	0.0757
	0.1				0.7298	0.5040
3	1.0	0.2014	0.1626	0.1501	0.5141	0.0862
	0.1	0.5255	0.1822	0.1038	0.8115	0.5311

As the number of priority classes increases, the maximum total utility is slightly increases, but a significant difference is not recognized. In general, the total utility increases as kh decreases. As the bitrate c increases, the mean data transmission time h decreases, and hence kh decreases with a given value of k. Since the transmission cost increases as the bitrate increases, there may be the optimum condition to maximize the utility per transmission cost.

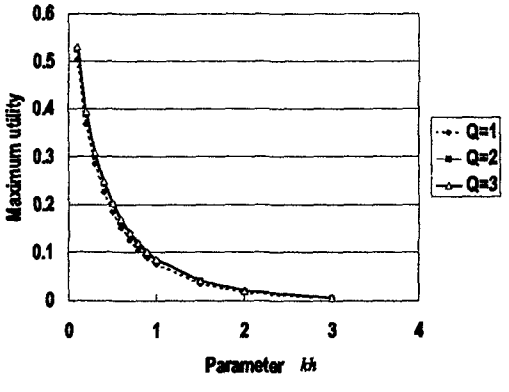

Figure 2. Maximum utility for optimum condition.

5. Conclusions

In this paper, we have examined the optimum condition to maximize the user's utility, and the analytical solution of the optimum condition has been given. From the numerical examples, the optimum condition in terms of the traffic load, waiting time for service and the utility for respective priority classes have been presented. The

effect of the number of priority classes has been examined. Although the total utility slightly increase as the number of priority classes increases, there is no significant difference as far as the maximum user's utility is concerned.

In this paper, it is assumed that the parameter k is not relevant to the value of D in the utility measure function. This point should be clarified by the opinion tests, etc. Even if more sophisticated formulas for the WTP are used, the framework of this paper may be applied in a similar manner.

The optimum design to maximize the utility per transmission cost is left for further study, considering the relation between the cost and the bitrate.

References

[1] R.J.La, and V.Anantharam,"Utility-Based Rate Control in the Internet for Elastic Traffic," IEEE Transactions on Networking, vol.10, no.2, pp.272-286, April 2002.

[2] Z.Cao, and E.W. Zegura,"Utility Max-Min: An Application-Oriented Bandwidth Allocation Scheme,"IEEE INFOCOM '99, New York, USA, pp.793-801, March 1999.

[3] H.Akimaru, and K.Kawashima, "Teletraffic - Theory and Applications, Second Edition," Springer-Verlag, 1999.

[4] R.L.Keeney, and H.Raifa, "Decision with Multiple Objectives, Preferences and Value Tradeoffs," John Wiley & Sons, 1976.

[5] R.C.Mitchell, and R.T.Carson, "Using Surveys to Value Public Goods, The Contingent Valuation Method," Resources for Future, 1989.

[6] J.A.Hausman, "Contingent Valuation - A Critical Assessment," North-Holland, 1993.

[7] D.J.Bjornstad, and J.R.Kahn, "The Contingent Valuation of Environmental Resources," Edward Elgar, 1996.

[8] K.Yamori, H.Akimaru, and M.R.Finley, "Dimensioning of Video Conferencing Systems in ATM Networks," International Conference on Computer Communications (ICCC'99), Tokyo, Japan, pp.294-299, September 1999.

[9] K.Nomura, K.Yamori, E.Takahashi, T.Miyoshi, and Y.Tanaka, "Waiting Time versus Utility to Download Images," 2001 Asia Pacific Symposium on Information and Telecommunication Technologies (APSITT2001), Kathmandu, Nepal/ Atami, Japan, pp.128-132, November 2001.

Appendix A Packet Transmission Modeling

Denoting by L_h the mean header size, and by L_p the mean payload size, the mean packet size is given by

$$L = L_h + L_p. \tag{A.1}$$

Letting H be the mean data volume to be transmitted, the number of packets for transmitting the data volume H is given by

$$N_p = \frac{H}{L_p}. \tag{A.2}$$

If the packets are originated randomly, and the packet size exponentially distributed, the mean system time (waiting and transmission) for class i packet is given by [3, p.83]

$$T_{pi} = \frac{h_p}{\left(1 - \sum_{j=1}^{i-1} \rho_j\right)\left(1 - \sum_{j=1}^{i} \rho_j\right)} \quad (i = 1, \cdots, Q). \tag{A.3}$$

where ρ_j is the traffic load of class j packet, and h_p is the mean packet transmission time given by

$$h_p = \frac{L}{c} \tag{A.4}$$

where c is the tramsmission bitrate.

Hence, the mean waiting time for class i service (all the packets to be sent) is given by

$$T_i = N_p T_{pi} \tag{A.5}$$

from which we have (1).

If the packet overhead α is defined as

$$\alpha = \frac{L_h}{L} \tag{A.6}$$

we have

$$h = \frac{H(1+\alpha)}{c}. \tag{A.7}$$

Usually, α is so small that h may be approximated by (2).

Appendix B Derivation of Equation (12)

Using (6) in (7), we have

$$\begin{aligned} f_1 &= D\rho_1 \exp\left(-\frac{kh}{1-\rho_1}\right) \\ f_2 &= D\rho_2 \exp\left(-\frac{kh}{(1-\rho_1)(1-\rho)}\right). \end{aligned} \tag{B.1}$$

Taking the logarithm yields

$$\begin{aligned} \log f_1 &= \log D + \log \rho_1 - \frac{kh}{1-\rho_1} \\ \log f_2 &= \log D + \log \rho_2 - \frac{kh}{(1-\rho_1)(1-\rho)}. \end{aligned} \tag{B.2}$$

Applying the formula for logarithmic derivative, we have

$$\begin{aligned} \frac{\partial f_1}{\partial \rho_1} &= f_1 \frac{\partial \log f_1}{\partial \rho_1} = f_1 \left(\frac{1}{\rho_1} - \frac{kh}{(1-\rho_1)^2}\right) \\ \frac{\partial f_2}{\partial \rho_1} &= f_2 \frac{\partial \log f_2}{\partial \rho_1} = -khf_2 \frac{2(1-\rho_1)-\rho_2}{[(1-\rho_1)(1-\rho)]^2} \\ \frac{\partial f_2}{\partial \rho_2} &= f_2 \frac{\partial \log f_2}{\partial \rho_2} = f_2 \left(\frac{1}{\rho_2} - \frac{kh}{(1-\rho_1)(1-\rho)^2}\right). \end{aligned} \tag{B.3}$$

Using (B.3) in (11), (12) follows.