Efficient Contents Delivery Method with Scheduled Unicast and Multicast

Kazuki IRIE, Yutaka ARAKAWA, Satoru OKAMOTO, and Naoaki YAMANAKA
Dept. of Information and Computer Science,
Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, Japan 223-8522
Email: irie@yamanaka.ics.keio.ac.jp

Abstract—We propose the efficient contents delivery method with scheduled unicast to multicast to reduce the load of contents server and network congestion. In our proposed method, the timing of transmitting contents is arranged when many users access the contents in short period of time. Furthermore, the method of transmitting contents is switched from unicast to multicast.

We evaluate the performance by computer simulation in terms of the unicast holding time, the threshold of switching from unicast to multicast, and the multicast delivery interval with the parameters of the content size and the access traffic. The simulation results show that the unicast holding is effective for the relatively-little contents and average download time can be reduced in our proposed method. The optimum switching time and the multicast delivery interval are also derived for a typical model.

I. INTRODUCTION

With the growth of the Internet and the diffusion of Internet-capable mobile devices, a lot of people have come to use services such as web browsing, E-mail and so on routinely. In these days, high speed networking infrastructure based on ADSL (Asymmetric Digital Subscriber Line), and FTTH (Fiber To The Home) has improved the access speed [1]. Through the high-speed network infrastructure, the rich contents such as movies and music transmitted over the Internet have been increasing. The Peer to Peer applications, such as Winny [2] and BitTorrent [3], have also accelerated the exchange of large-size files over the Internet. With the start of Terrestrial Digital Broadcasting, larger-size files such as HDTV (High Definition TV) movies will be exchanged in the future. Therefore, the delivery of large-size files over the Internet draws increasing attention.

In Client-Server system, where the requested content is delivered from a content provider server called the origin server to the client PC which requested the content, the origin server may be heavily-loaded when many users access the content at the same time. Techniques to reduce the origin server load in case a lot of user access in short period of time, such as mirroring and proxy caching, have been proposed [4]. The network configuration in which some Surrogate Servers (cache servers) having all or a part of contents of the origin server are arranged in geographically-distributed manner has been applied for Contents Delivery Network (CDN) [5].

However, as the number of users who download large-size files is increasing, the network bandwidth and the processing capabilities of contents servers are inadequate. Transmitting popular contents will cause the heavy server load and network congestion.

In this paper, we propose a new content delivery system based on scheduled unicast and multicast to mitigate the server load and the network load. In our proposed method, the timing of transmitting contents is arranged when many users access the contents in short period. Furthermore, if the number of access continues to increase, the transmitting method is switched from unicast to multicast.

In Section 2, we present and analyze the conventional method for reducing the origin server load and the issue of contents delivery. In Section 3, we describe the outline of our proposed method and the processing flow including scheduling and unicast/multicast switching in detail. In Section 4, we describe the parameters which are used for system design and computer simulation. In Section 5, we evaluate the performance by computer simulation in terms of the unicast holding time, the threshold of switching from unicast to multicast and the multicast delivery interval with the parameters of the content size and the access traffic. Conclusions are summarized in Section 6.

II. ISSUE ON CONTENTS DELIVERY

As previously mentioned, because of the increase of rich contents such as music or movie files having large content size, network traffic and server load have a tendency to increase when the server transmits the contents. Figure 1 shows the relation between the content and download time of rich contents.

Several techniques for reducing server load delivering contents have been proposed. For example, two techniques are presented as follows.

• Proxy Caching [4]

This technique uses the server called Proxy Cache Server which is set up on LAN network to which users directly connect. When the user requests the content on the original server, the proxy cache server accepts the request and accesses the original server instead of the user.

In the case of the first access from the user, the contents is delivered from the original server, however, after second time the content would delivered from the proxy server, resulting in the reduction of the original server load.

• Mirroring

This technique is that all contents on the original server are copied to the other servers. These servers are set up on the network, and the user requests are distributed to servers. These servers assume the role of the original server.

CDN employs these techniques. In CDN, the network configuration in which some Surrogate Servers (cache
servers) having all or a part of contents of the origin server are arranged is applied in geographically-distributed manner. The techniques that the user requests are transferred to some servers effectively are researched and applied in CDN [6], [7]. Figure 2 shows the example of configuration of CDN network. As each Surrogate Server is limited in the processing capacity, Surrogate Server cannot deal with a lot of user access. In this case, it can be solved by adding new Surrogate Servers, but this means a vast amount of resource is consumed. In particular, the delivery of the large-size contents needs a lot of time. Therefore, the number of concurrent connection in the server tends to increase.

So we consider that each content server needs to be applied the techniques that mitigate the server load. Note that it is possible to combine the above-mentioned load distribution techniques and our proposed method effectively.

III. PROPOSAL OF SCHEDULING AND MULTICAST

In our system, three transmitting modes are adaptively changed. These are the unicast delivery, scheduled unicast and multicast. Figure 3 shows that the outline of the proposed scheme and Figure 4 shows how to switch three modes. Its details are described as follows.

When the number of access is lower than the unicast delay control threshold, the transmitting mode is the unicast delivery which is typical transmitting method. The advantage of this method is simplicity and stability. However, the network resource bottleneck may occur around the contents server when the number of concurrent access increases. So if the number of access in a unit of time exceeds the unicast delay control threshold, the transmitting mode will be changed to the scheduled unicast mode in which the timing of starting unicast delivery is delayed. As a result, the number of simultaneous access will be reduced. We define the amount of delay in this mode as a holding time. This operation would improve the download time for each user and decrease stress of the user caused by a long download time during congestion. By delaying the download timing, the growth in the download time of the already accessing users can be mitigated. In the case that the number of access continues to increase and even the load in the server wouldn't be mitigated by the setup of the holding time, the transmitting mode is switched from unicast to multicast and the number of simultaneous access is drastically reduced. In multicast mode, multicast delivery interval is assigned when user first request arrive the contents server. After waiting multicast delivery interval, requested contents are delivered to users who request

---

**Figure 1. Content and Download time**

**Figure 2. Configuration of CDN (Three Surrogate Servers)**

**Figure 3. Principle of the Proposed Scheme**

**Figure 4. Processing Flow in Server**
TABLE 1
SIMULATION CONDITION

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>10,000</td>
</tr>
<tr>
<td>Number of files</td>
<td>10 kinds</td>
</tr>
<tr>
<td>File size (HDTV)</td>
<td>22.5GB (120 minutes)</td>
</tr>
<tr>
<td></td>
<td>11.25GB (60 minutes)</td>
</tr>
<tr>
<td>Maximum bandwidth of the server</td>
<td>100Mbps</td>
</tr>
<tr>
<td>Maximum number of access to server</td>
<td>10,000</td>
</tr>
<tr>
<td>Maximum unicast bandwidth per user</td>
<td>30Mbps</td>
</tr>
<tr>
<td>Maximum multicast bandwidth per user</td>
<td>15Mbps</td>
</tr>
<tr>
<td>Zipf parameter</td>
<td>1</td>
</tr>
</tbody>
</table>

the same contents during the multicast delivery interval. This means that users who request contents during the multicast delivery interval are forced to wait because the requested contents are not delivered instantly. In the case of multicast mode, if the users who request contents are allowed to connect to the contents server freely in real time, the network congestion and/or the server overload would occur. As a result, the download time will increase. On the other hand, the waiting time for the next multicast delivery would make the download time increase a little, but it is relatively small compared to the download time when the content size is large.

The multicast has been used for delivery of the real-time streaming data to a number of users simultaneously. Therefore, assuming multicast for file download, the content delivery scheme is used to consider the conditions of each user terminal and network. More specifically, in order to transmit the content to the users of the same group at the same time, the download speed should be set to the network speed of the lowest capability user on the same group. Some retransmission scheme may be required in the case transmission error or packet loss occurs because the acknowledgment of receive is not used in multicast mode. Therefore reliable multicast schemes guaranteeing to receive the packet or the sequence of the received packets are proposed [8].

IV. SYSTEM PARAMETER DESIGN

The optimization of the parameters such as the holding time in the unicast delivery and transmission interval in the multicast delivery is required for the proposed system. The specific parameters are:

- the threshold value of access users at which the holding time is configured
- the holding time
- the threshold value of access users at which the delivery mode is switched from unicast to multicast
- the interval of multicast

The holding time and the interval of multicast should be determined based on the content size which affect the download time. The relation between the parameters and the download performance are shown with the results of computer simulation below.

V. SIMULATION

A. Simulation Condition

The simulation was conducted with the conditions listed in Table 1. The file size of the contents was determined considering the large size contents such as HDTV video file.

The probability of each user accessing the content file is assumed to follow Zipf distribution. With Zipf distribution, the probability of access to the i-th popular content, $P_N(i)$, among the all N-contents in the server is given by the following equation [9].

$$P_N(i) = \frac{1}{\sum_{j=1}^{N} \frac{1}{j^\beta}}$$  \hspace{1cm}(1)

B. Holding time and Download time

We have studied about the holding time from accept of user’s request to the start of transmission in scheduled unicast delivery mode in terms of content size. We assume two cases in scheduled unicast delivery mode. One is the case that the number of access per minute (hereinafter called “AN”) is 5 and the contents size is that of 120-minute HDTV file. The other one is the case that AN=8 and 60-minute HDTV file.

The relation between the holding time and download time or the maximum number of access in previously-mentioned two conditions is shown in Figure 5.

The unicast delay control threshold $\beta$ is set 50 for the both cases. The download time = holding time + content acquisition time Figure 5 shows that as the holding time in unicast scheduled delivery mode increases, the maximum number of access decreases without substantial increase of download time for both cases of 120-minute HDTV and 60-minute HDTV files. This is because the number of simultaneous access is kept by putting the user requesting the content download on waiting.

Compared to the case the holding time is not applied, the download time is increased by about 2.8% and 6.3% for each case of 120-minute HDTV and 60-minute HDTV file when the holding time is set to 60 minutes. This is because the holding time is included. However, as the download time of already delivering users are reduced, the increase of the download time is kept smaller than the assigned holding time.

The maximum number of access is found to be improved by 12% and 32% for each case of 120-minute HDTV and 60-minute HDTV files. This improvement is remarkable for the case of 60-minute HDTV file. This is because the effect can not be expected without the longer holding time for the large content as the download time increases for the large size file.

The holding time which is acceptable by many users should be evaluated in the future. It is assumed that the less 10% of download time for the holding time subjectively and that the holding time less than 30 minutes and 15 minutes are acceptable for the case of 120-minute HDTV file and 60-minute HDTV file, respectively. Therefore, considering that the effect of the holding time is small with the 120-minute HDTV file, the unicast delay control threshold and the holding time is set to 50 minutes and 15 minutes, respectively.

C. Threshold of Switching from Unicast to Multicast

The contents are delivered to all users in multicast mode when $\alpha$ is set to 0. In this case, the number of
simultaneous access for the server can be kept small, however the download time increases compared to that in unicast mode. The contents are delivered to all users in unicast mode when $\alpha$ is set to the maximum capacity of the server. In this case, the contents may be downloaded with the maximum throughput, however, the overload of the server or session loss will occur when the number of simultaneous access increases.

We have researched the optimum value of $\alpha$ which reduces the maximum number of server access and the download time. The relation among $\alpha$, download time and maximum number of access is investigated. The content size is assumed to 120-minute HDTV and 60-minute HDTV file for the case of large and small content size, respectively.

The relation between $\alpha$ and average download time or the maximum number of access for the case of AN=5 and 120-minute HDTV file size is shown in Figure 6. The relation for the case of AN=8 and 60-minute HDTV file size is also shown in the same figure. The figure shows that the average download time reaches the minimum value when the threshold $\alpha$ is set to 400 for both cases of 120-minute and 60-minute HDTV file size.

The optimum value of the threshold of switching from unicast to multicast, $\alpha$ is found to be 400 with this simulation condition.

D. Multicast Interval

We have investigated the relation between the download time and the multicast interval which is assigned against the download requests from users.

Figure 7 shows the relation between the average download time and the maximum number of access when the multicast delivery interval is changed for the case of 120-minute HDTV file with AN=5 and for the case of 60-minute HDTV file with AN=8. Figure 7 shows that the maximum number of access and the average download time result in minimum value when the multicast delivery interval is around 15 minutes.

E. Number of Access and Download time

The relation between the average download time and the AN with the range of 1 to 10 is shown in Figure 8. The changes in the maximum number of access are shown in Figure 9. The content size is assumed to be that of 120-minute HDTV file. The comparative performance was evaluated for the cases:

1) all contents are delivered by unicast
2) all contents are delivered by multicast
3) the contents are delivered by scheduled unicast with multicast switching (the proposed method)

The following parameters are used for the simulation: $\alpha$=400, $\beta$=50, unicast holding time =15 minutes and multicast interval = 15 minutes.

It is found that the download time is kept short in the unicast delivery mode in the case of small AN, because the sufficient network resource can be allocated for each user. However, with the increase of the access users, the number of simultaneous access increases rapidly, resulting in an increase in download time. Figure 9 shows that
the rejection of the request (session loss) happens when the number of access exceeds 4 per minute because the maximum number of simultaneous access goes beyond the capacity of the server.

In the multicast delivery mode, the number of simultaneous access increases is kept small even in the case AN increases. However, as the available transmission bandwidth in the multicast delivery mode is smaller than that in the unicast mode, the average download time is longer than that in the unicast mode.

In the proposed method, the average download time can be kept short by using the unicast delivery mode as far as possible and also with the use of scheduled unicast delivery in case that AN is small. When the AN increases, the delivery mode is switched to multicast mode to reduce the number of simultaneous access, and the multicast delivery interval is assigned to prevent the user in already delivering in unicast mode from the degradation of the download time.

In terms of the number of access, the maximum number of access can be kept under 60% of that in the server by switching to multicast mode resulting in the reduction of the number of simultaneous connections.

The simulation results show that the proposed method gives the best performance for increase of AN meaning that the load of the server and the network increases.

VI. CONCLUSION

We have proposed the efficient contents delivery method with scheduled unicast and multicast to reduce the server load and network congestion. In our proposed method, three transmitting modes, which are unicast delivery, scheduled unicast and multicast, are adaptively changed with an increase in the number of simultaneous access. In the scheduled unicast mode, timing of delivery contents is arranged when many users access the contents in short period. Furthermore, multicast delivery interval is assigned in the multicast mode to mitigate the number of the simultaneous connection.

We have evaluated the performance by computer simulation in terms of the unicast holding time, the threshold of switching from unicast to multicast and the multicast delivery interval with the parameters of the content size and the access traffic.

The simulation results showed that the unicast delay was effective for the relatively-little contents and average download time can be reduced. We have obtained the optimum the threshold of switching and the multicast delivery interval for a typical model.

The subjects of future investigation are dynamic control of the unicast delay, switching time from unicast to multicast and the multicast delivery interval.

ACKNOWLEDGMENT

This work was supported by Grant-in-Aid for Scientific Research(C) (19560395).

REFERENCES