An Approach to Select a Best Suitable Video Server

Falko Dressler
falko.dressler@rrze.uni-erlangen.de / fd@acm.org
http://bsd.rrze.uni-erlangen.de/~fd/
+49 9131 85-27802

University of Erlangen-Nuremberg
Regional Computing Center (RRZE) /
Department of Computer Science (Operating Systems)
Martensstr. 1, 91058 Erlangen
Germany

Abstract - **Today, many institutions begin to employ services to distribute multimedia content to their customers. Especially in the areas of tele-learning and tele-medicine, the requirements on the availability and the achieved transmission quality are very high. Therefore, multiple video servers for a particular service are distributed over the internet in order to allow the clients to choose the 'best' one. The term 'best' can be defined in various manners. Typically, the client uses a try-and-error based selection mechanism to choose a best fitting server. Additionally, methods have been developed to choose an optimum server based on its current load.**

The presented approach goes a little further. Quality of service (QoS) based mechanisms are employed for the selection process. In order to find the server which fits best for a particular service a quick check is started before the connection to the selected server is set up. This check includes the test of the availability of the server and the particular service as well as the measurement of the currently available quality of service from the server towards the client. This is done by initiating a packet stream for testing reasons only. The client analyzes the stream to calculate values such as the delay, the delay variation (the jitter), and the packet loss ratio. Based on the results of these measurements the 'best' server can be chosen.

I Introduction

Video streamings are widely employed in the internet. They are used in the fields of tele-medicine, tele-learning and home entertainment. Typically, such multimedia applications have high resource requirements. For example, a continuously throughput of about 5 Mbps is necessary for an assumed MPEG2 video broadcast in DVD quality.

Additionally, quality of service (QoS) parameters such as a maximum packet loss ratio, a maximum delay, and a maximum variance of the delay (jitter) have to be guaranteed.

The approach of most service providers is to operate more than a single video server for the same content. The end user, here called client, is requested to choose one of these servers for the particular transmission. Several selection criteria can be defined such as the slightest load of a server or an administratively controlled preference.

The problem description is shown in Fig. 1. Three server systems are deployed over the internet and the client has to choose a best fitting server depending on the quality demands of the forthcoming transmission.

New approaches are in progress, which allow a more meaningful selection. For example, the research area of peer-to-peer networks is working on such principles [5].

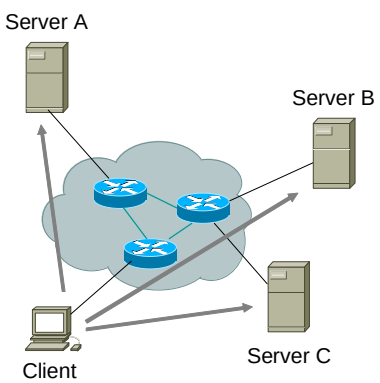

Fig. 1. Problem Description

IP multicast is often employed for the multimedia transmissions to save resources in the network and at the server [9]. Depending on the working principle of IP multicast, the selection becomes more difficult [1].

New mechanisms have been developed for using multicast techniques for video-on-demand-services [4].

The goal of this paper is to introduce an approach to select a best suitable video server based on the estimation of the available quality of service within the network from the server towards the client.

The sections are organized as follows: Section II introduces some basic classical selection criteria to be used for choosing a particular server. The new approach provided in this paper is presented in the following two sections, whereas new quality of service based selection criteria are discussed in section III, and a measurement methodology is described in section IV. An overview to a tool to measure the quality of service of a network connection and even of a multicast service, the multicast quality monitor, is provided in section V. Section VI discusses some application scenarios and a concluding summary of the work is provided in section VII.

II Classical Selection Criteria

There are numerous selection criteria to be considered for choosing a particular video server for an optimum distribution of some multimedia content. We divide the criteria into two main classes: The classical criteria which are already employed in several productive environments and the new quality of service based criteria.

The first selection criterion to be mentioned is the administrative preference. Numerous approaches have been investigated. Here, an overview of some of the most common definitions is provided. The mechanisms can be divided into two groups: centrally coordinated and dynamic selection methods.

A. Centrally coordinated selection

The concept of the centrally coordinated selection is to have a single server which only is responsible for the selection of a best fitting video source. The advantage of this method is the possibility to incorporate much more information into the selection process. Unfortunately, a new single point of failure is employed as well. Fig. 2. shows this mechanism.

First, the client is querying a selection server (1), which secondly responds with some information about the video server which should be preferably used (2). Finally, the client can connect to the selected video server in ortder to initiate the demanded transmission (3).

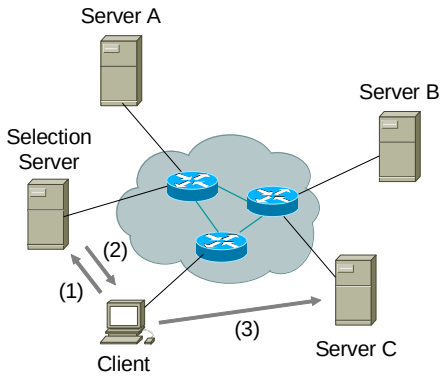

Fig. 2. Centrally coordinated Selection

Using such a central intelligence, the following mechanisms may be deployed:

- The access to single servers may be restricted to a limited number of end users, maybe to reduce the load of the most important servers.
- The knowledge about the behavior of the utilization degree may be used to create a list of favorite servers.
- A round-robin discipline may be employed to distribute the load equally to all the available servers.

B. Dynamic selection

The idea of a dynamic selection is to prevent the required existence of a single - potentially failing - selection server. A typical example for such a method is the usage of the DNS (domain name service) to create a round-robin schedule.

Another concept is to use IP multicast for this task. A well-known multicast group is used send query messages asking for available servers providing the requested content. Typically, this mechanism is used if the distribution of the data from the server towards the clients is already employing multicast.

A second selection criterion, which can be employed together with a centrally coordinated selection as well as with a dynamic mechanism to choose the best fitting servers, is the test of the connectivity.

The client can employ a try-and-error mechanism to connect to a server. It the first connection setup fails, a second and maybe a third try can be initiated in order to find an available server which supports the requirements of the client.

III New Quality of Service based Criteria

Beside the discussed classical selection criteria, new challenges appeared together with the strongly increasing demands for higher transmission qualities. Support for high quality video transmissions is necessary, therefore, the following parameter have to be included into the selection process: the server load and the transmission quality.

A. Server load

The first resource to measure is the current load of a particular server. If it exceeds some specific threshold value, the server became inoperative. Thus, the server has to deny new inquiries if the load reaches some soft-limit.

B. Transmission quality

This value describes the available quality of service in the network for a specific session. In particular, each service may define some minimum requirements which are needed for a suitable transmission. The most important quality of service parameters describing the network quality are discussed in the following.

- The end-to-end delay specifies the time it is required for the transfer of a single packet from the source to the destination. It is also called the transmission time or the latency. For unidirectional transmissions such as the mentioned video broadcasts, the absolute delay is mostly unimportant.
- Nevertheless, the variance of the delay, which is also known as the jitter, has a strong impact on the transmission quality. A typical multimedia stream requires a continuously stream of data which is decoded and played-back at the receiver. If the packets do not arrive continuously, a playout-buffer is required to absorb the effects introduced by the jitter. A problem arises if the jitter is higher than allowed by the size of the playout-buffer. Then the late packets are useless and, therefore, they are dropped.
- The last QoS parameter, which should be mentioned here, is the packet loss ratio. Depending on the type of the transmitted content, a small amount of lost packets can be tolerated. Examples for algorithms which introduce some basic forward error corrections are described in [8]. Nevertheless, the packet loss ratio has a high impact on the achieved transmission quality.

IV Measurement Methodology

In the previous section, a number of criteria have been provided which build a basis for the selection of a particular server for a forthcoming transmission. The goal of this section is to introduce a measurement methodology for the values of the single mentioned parameters. This section is divided into three parts describing the mechanisms to build a list of potential servers, to check the availability of each server and, finally, to measure the available quality of service in the transport network from all of the disposable servers to the client.

A. Building of a List of potential Servers

In an initial step, a list of potential servers, i.e., of such servers containing the requested data for distribution, must be created. The ’standard’ mechanisms are used to complete such a list of available servers. For example, a selection server as previously described can be employed for this task or a request can be sent using IP multicast to query all active video servers for their readiness to distribute the requested data.

B. Detecting the Availability of each Server

A necessary second step is the test the availability of each server. In general, the currently deployed video server have built-in features to test their availability as well as their readiness to offer a requested service.

Each server on the list which has been created in the first step must be verified. The result of this step is a modified list of servers which are ready to accomplish the requested task. Some of the currently installed server environments are already able to finish this part of the selection process.

At this time, the client is able to connect to ’a’ server and to request the demanded service from it. Obviously, no selection criteria based on any quality of service measurement has been included. Therefore, only the standard best effort transmission scheme can be employed.

C. Quality of Service Measurement

The measurement of the available quality of service for a forthcoming transmission between a particular server and the client is a difficult task. A special protocol has to be defined to achieve the necessary information.

In order to provide a maximum independence from the employed server software, this approach envisions a separate measurement service. A special program must be installed on the server to perform this task. The client can

connect to this service in order to achieve more information about the quality which can be expected for the transmission. The basic behavior is shown in Fig. 3. First, a separate control process is queried (1) and secondly, the data transmission from the distribution process is started (2).

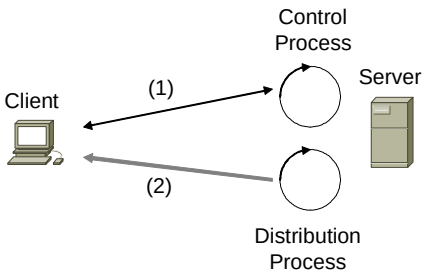

Fig. 3. Control and Distribution Processes

The first parameter, the server load, can be easily obtained using this control connection. Parameters such as the CPU load or the utilization of the network interface are good approximations for the overall server load.

To measure the available quality of service in the network, more sophisticated mechanisms are required. The proposed methodology is based on a short lasting simulated data transfer. The behavior of this transmission is adapted from the forthcoming multimedia distribution.

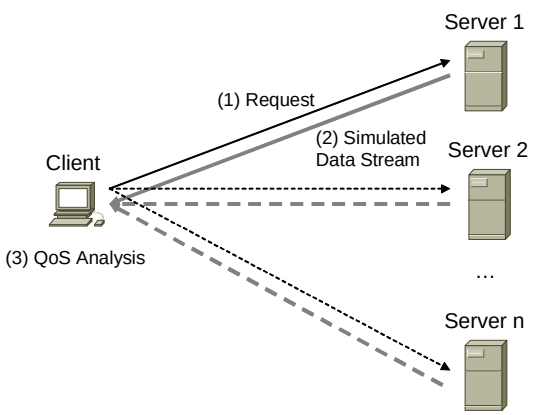

Fig. 4. QoS Measurement

Fig. 4. shows the principles of the QoS measurement. The client successively tests all the servers. First, a request message is sent to the server (1), requesting a simulated data stream. The server responds by initiating such a transmission (2) lasting for a short period of time. The behavior of this transmission depends on the type of the requested content. Finally, the client analyzes the received data in order to compute some quality of service parameters (3). This task is repeated for all available servers.

It is advised to use the same protocols for the measurement which will be used for the real transmission. Typically, the streaming is initiated using the real-time streaming protocol (RTSP, [7]). The data transfer itself is based on the real-time transport protocol (RTP, [6]). The advantage of RTP is the inclusion of elements such a sequence number and a timestamp in its header definition which can be used for the measurement and the analysis.

The proposed quality of service measurement methodology is based on the multicast quality monitor (MQM, [2], [3]). The MQM builds a framework for reliability and quality of service measurements in an IP multicast environment. The MQM is described in more detail in the next section.

Even if this framework was specified for application in an IP multicast network it can be employed in unicast environments as well. In general, the unicast case is a simplification the multicast case.

Some application scenarios are discussed succeeding the MQM section.

V Multicast Quality Monitor

The multicast quality monitor was designed to test the reliability of an IP multicast network and the transmission quality of multicast services. The primary goal was to estimate the available quality for forthcoming multicast services such as video conferences and TV broadcasts. The advantage of the MQM is its scalability. The single measurement methods are discussed in the following.

A. Reachability and Reliability

A new multicast ping mechanism was introduced with the multicast quality monitor. The concept is shown in Fig. 5. A measurement station or "probe" sends a MQM ping request message (1) to a preconfigured multicast group. The message is received (2) by the other probes which have been distributed over the network. They generate a response message (3). Finally, the originator of this mechanism is receiving the responses (4) and can analyze the results.

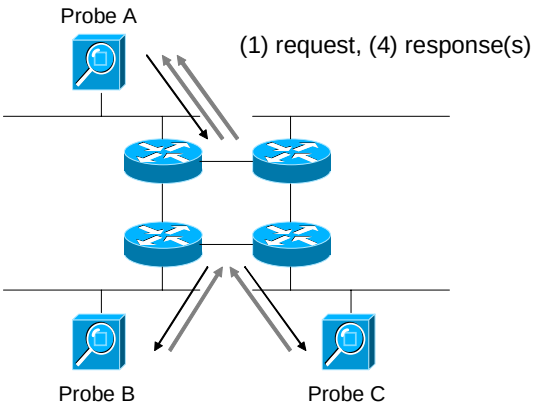

Fig. 5. MQM Ping Mechanism

Using the MQM ping mechanism, it is possible to test the reachability between several nodes in the network. The reliability can be calculated using the measured reachability information over a period of time.

To achieve a high scalability, the mechanism has been designed in a way that a complete reachability graph can be created using only two MQM ping requests.

The request-response cycle must be periodically repeated in order to refresh the forwarding state for the multicast group in all the involved network devices.

B. Quality of Service

The measurement of the available quality of service for a particular connection is divided into two different parts. The delay measurement, the one-way delay (OWD) as well as the round-trip time (RTT), is done using the MQM ping mechanism. The packet loss ratio and the jitter are measured using RTP streams.

- One-way delay and round-trip time
  Timestamp information is put into the header of each MQM ping packet. Using these timestamps, it is possible to calculate the OWD between the sender and the receiver and vice versa. The better the synchronization of the clocks of both computers, the better is the achieved result. The RTT can always be accurately measured because it depends only on the clock of a singe node.
- Packet loss ratio and jitter
  The concept of the MQM is to acquire as much information as possible by passively analyzing active RTP transmissions. The working principle of IP multicast allows one to join an active transmission without notifying the participants. Because most multimedia transmissions employ RTP to transfer their data, the capabilities of RTP can be used to calculate some QoS parameters, in our case the current packet loss ratio and the jitter (the variance of the delay). The scenario of passive QoS measurements is shown in Fig. 6.

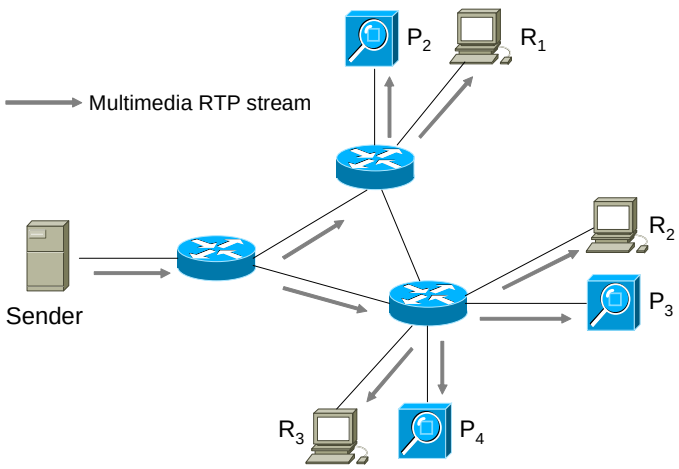

Fig. 6. Passive QoS Measurement

If no active session is available or if no session exists which shows a comparable behavior as the forthcoming transmission, an active measurement can be initiated by sending simulated RTP traffic.

In summary it can be concluded that the multicast quality monitor provides all the required capabilities which are necessary to estimate the available quality of service from a particular node (the server) to another one (the client). The mechanism can be employed for unicast transmissions as well.

VI Application Scenarios

Using the proposed methodology, a client has to measure to available quality of service in order to select a best suitable server. Fig. 7. shows a sample network consisting of three video servers, the client, and some network devices. After an initial query to create a list of potential servers, the client starts a measurement process for each available server.

In this example, all the servers have an instance of the multicast quality monitor implemented. Therefore, the client can use the MQM protocol to measure the delay, the jitter and the packet loss ratio.

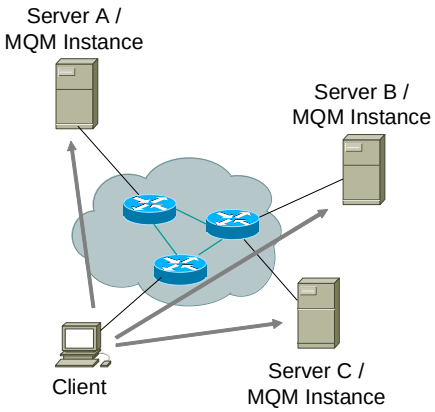

Fig. 7. Application of the MQM

Based on the demands of the forthcoming transmission, the best suitable server can be chosen based on the optimum quality. In some scenarios it also is a good approach to choose the first server which satisfies the requirements, therefore, a server is chosen with a sufficient quality.

Both scenarios have typical applications. For a high quality video transmission in the telemedicine, the optimum choice is searched. For home entertainment applications, such high demands are not necessary. Therefore, a first sufficiently working server can be selected.

VII Summary

The proposed approach to select a best suitable video server for forthcoming transmissions allows one to incorporate quality of service parameters into the decision process of selecting an available video server. The current techniques have been described and new methods have been discussed.

A new measurement methodology has been proposed which is based on the concepts of the multicast quality monitor. This methodology allows on to take decisions using information about the available QoS in the network and the availability of resources at the particular server.

Based on these several selection criteria, an optimum fitting server for the forthcoming transmission can be selected.

VIII References

[1] F. Dressler, "How to Measure Reliability and Quality of IP Multicast Services?" Proceedings of 2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE PACRIM'01), Volume 2, Victoria, B.C., Canada, August 2001, pp. 401-404.

[2] F. Dressler, "An Approach for QoS Measurements in IP Multicast Networks, MQM - Multicast Quality Monitor," Proceedings of Third International Network Conference (INC 2002), Plymouth, UK, July 2002.

[3] F. Dressler, "MQM - Multicast Quality Monitor," Proceedings of 10th International Conference on Telecommunication Systems, Modeling and Analysis (ICSTM10), Monterey, CA, USA, October 2002, pp. 671-678.

[4] K. Hua, Y. Cai, S. Sheu, "Patching: A Multicast Technique for True Video-on-Demand Services," ACM Multimedia, 1998.

[5] A. Kapur, R. Brooks, S. Rai, "Design, Performance and Dependability of a Peer-to-Peer Network supporting QoS for Mobile Code Applications," Proceedings of 10th International Conference on Telecommunication Systems, Modeling and Analysis (ICSTM10), Monterey, CA, USA, October 2002.

[6] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," RCF 1889, IETF, January 1996.

[7] H. Schulzrinne, A. Rao, R. Lanphier, "Real Time Streaming Protocol (RTSP)," RFC 2326, IETF, April 1998.

[8] T. Speakman, J. Crowcroft, J. Gemmell, D. Farinacci, S. Lin, D. Leshchiner, M. Luby, T. Montgomery, L. Rizzo, A. Tweedly, N. Bhaskar, R. Edmonstone, R. Sumansekera, L. Vicisano, "PGM Reliable Transport Protocol Specification," RFC 3208, IETF, December 2001.

[9] H. Tan, D. Eager, M. Vernon, H. Guo, "Quality of Service Evaluations of Multicast Streaming Protocols," Proceedings of ACM SIGMETRICS, 2002.