Enhancing QoS for *Delay-tolerant* Multimedia Applications:

Resource utilization and Scheduling from a Service Provider's Perspective

Saraswathi Krithivasan
IIT Bombay
saras@it.iitb.ac.in

Advisor: Sridhar Iyer
IIT Bombay
sri@it.iitb.ac.in

ABSTRACT

Emerging applications such as distance education and corporate training are examples ***of delay tolerant*** multimedia applications where clients *request* the start of play back at a convenient time specified by $(t+d_i)$ where t is the current time and d_i is the *delay tolerance* acceptable to client i. Such applications typically involve a Closed User Group (CUG) network that exhibits heterogeneous characteristics, where a Content Service Provider (CSP) disseminates multimedia content to geographically dispersed clients. Our research deals with the issue of maximizing quality delivered at the clients while satisfying their delay tolerance with minimal additional resources. As a first step to this end, we have developed an optimization-based approach to determine the best quality that can be delivered to the clients through judicious placement of resources such as buffers and transcoders. Simulation results demonstrate the usefulness of exploiting client delay tolerance specifications for delivering enhanced Quality of Service (QoS) with little or no additional resources. By maximizing QoS to the clients with given resources, CSPs can (1) maximize the utilization of links (2) provide differentiated services to their clients, and (3) offer upgraded services to some clients (which may have a revenue implication). Our ongoing work deals with scheduling and admission control issues which would maximize profits for the CSPs while satisfying the admitted clients' requirements with optimal resource utilization.

Keywords Delay tolerant applications, Multimedia dissemination, Quality of Service (QoS), Transcoding, Caching, Heterogeneous networks, Distance Education.

1. INTRODUCTION

With the proliferation of networks connecting different parts of the world, several popular streaming media applications have emerged: Universities offering their courses to a set of global subscribers, service providers streaming movies requested by their clients, multinational corporations providing training to employees across cities are some examples. Typical characteristics of such applications include: a *source* that is responsible for the dissemination of contents; a set of geographically distributed *clients* connected through heterogeneous links of varying capacities and characteristics, and multimedia *contents* that can be encoded at different rates.

Let S be a Content Service Provider (CSP) that offers movies to a set of subscribed clients. While S may have several channels, at any point in time S streams a movie synchronously to a subset of subscribers requesting for that movie from a given channel. Suppose at time t, a client i demands uninterrupted and loss free play back and specifies a *minimum play back rate* r_i which defines its minimum required QoS and a *delay tolerance* d_i, which determines the start of the play back, $(t+ d_i)$. In the CSP's perspective following questions are important:

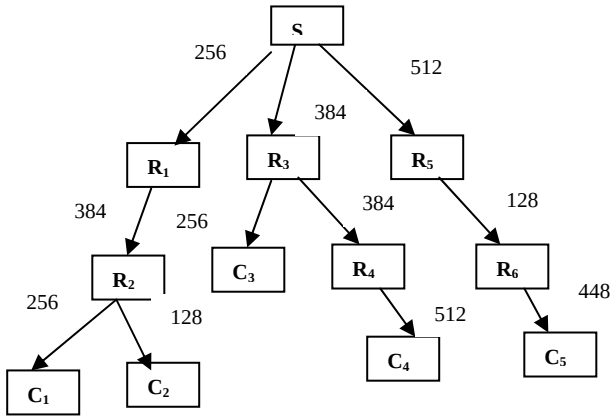
1.Using a single stream and by exploiting the delay tolerance of the clients what is the best quality that can be delivered to each client? 2. Assuming resources such as buffers and transcoders are available at (some of) the nodes in the network, how can such resources be optimally deployed to service the clients? 3. Given that some clients may have high bandwidth links that can support play back before their required start time, and most clients may be served at higher rates than their minimum requested rate (by exploiting their delay tolerance), how can the streaming schedule and admission control be designed such that requirements of all admitted clients remain satisfied and revenue is maximized?

A review of the existing mechanisms for effective and efficient delivery of multimedia in [3][4] indicates that existing work treats multimedia dissemination as real-time applications that can tolerate some transmission errors and explores ways to minimize the startup delay. In contrast, we focus on multimedia applications that can tolerate startup delays.

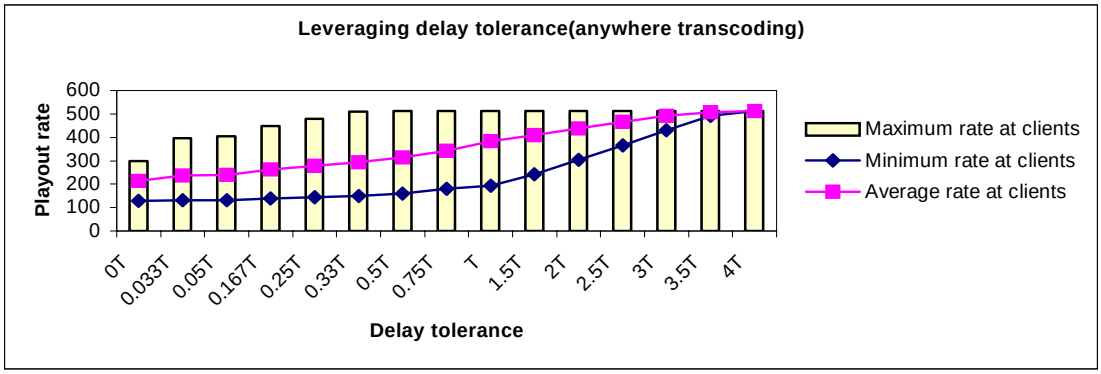
2. PROBLEM FORMULATION

2.1. Determining optimal quality deliverable to the clients

We use an optimization formulation to determine the best possible QoS deliverable to the clients. Through our studies, we have concluded that by exploiting the clients' delay tolerance, better quality is possible at the clients even with constrained links in the path from source to clients. We have presented our preliminary results at [2]. A simple network model in the figure below represents a given heterogeneous dissemination network as a tree, with the source as the root and clients as leaves. The intermediate nodes serve as *relay* nodes. We make the following assumptions: 1.The network is a Closed User Group network with a single service provider. 2. For the duration of a service, the network topology and link bandwidths do not change. Our optimization formulation results in the performance shown in the graph that follows.

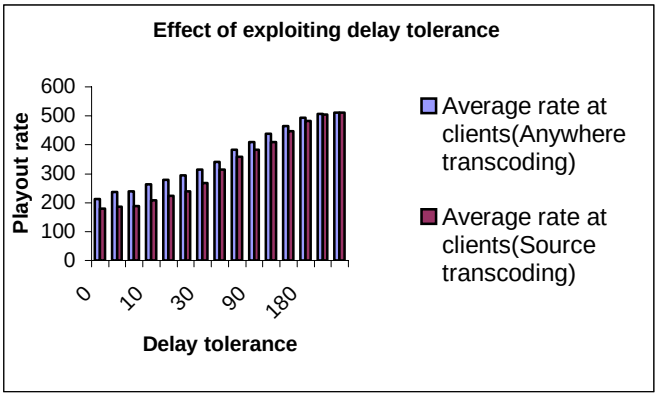


S: Source
$R_1, R_2, \ldots R_n$: relay nodes
$C_1, C_2, \ldots C_n$: clients
α= 512 kbps
T=1 hour



2.2. Optimal placement of resources

By changing the constraints in the optimization formulation, we can determine the placements for transcoders. We consider two cases (a) where the source provides multiple encoded streams and (b) where transcoders are deployed at intermediate nodes. Also, in the case where the entire content needs to be buffered at the client (as is likely under high bandwidth network links) to be played back at the specified time, we are also exploring the optimal distribution of caches within the network nodes. Such caches can then be used to serve future requests for the same multimedia content. Graph below shows the effect of placement of transcoders on the average playout rate (in kbps) at clients for



different delay tolerance values (in minutes).

2.3. Scheduling and admission control: Revenue model

In the first phase of our experiments, we considered a static model where the number of clients and the network topology remained static. In our on-going work, we consider dynamic arrivals, and hence changes in the network topology. Consider a static case where due to high bandwidth links or through the exploitation of delay tolerance, a number of clients exhibit *residual delay tolerance*, i.e., the clients are served with the best possible rate (the original encoding rate of the content) at a time earlier than their requested time. Let t_1 be the scheduled streaming time for servicing requests from clients c_1, c_2,…c_n. Let rd_1, rd_2,…, rd_n represent the residual delays for clients c_1, c_2…,c_n. Let us suppose that a subset of the clients have positive rd values while the others have zero as their rd values, indicating that these clients may not get the best possible quality. Let us assume that these clients however get a quality much better than the minimum quality they had specified.

One way to exploit the residual delay of the clients is to reschedule the streaming for a later time $t_1+\Gamma$, where Γ is the time by which the streaming can be postponed (hard deadline) without violating any of the admitted client requirements. We approach this problem in the following manner:

1. At t_1, we run a predictive tool which predicts the number of arrivals based on an arrival distribution and expected client requirements.
2. We define various price points for the enhanced quality of service when we start the stream at t_1 and additional revenue for each new admission if the stream were to be rescheduled at $t_1+\Gamma$. By considering the tradeoff between the two options, we recommend an appropriate alternative.
3. When the predictive tool favors rescheduling the streaming, we monitor each new arrival in the interval Γ and run the optimization tool to find the appropriate quality and placement of resources.

3. CONCLUSIONS AND FUTURE WORK

Delay tolerant applications cater to the clients' convenience while enhancing the quality of multimedia delivery. Our work explores the various options available for a CSP to utilize its resources optimally to achieve higher profits. The final contribution of this thesis would be a tool that invokes the appropriate adaptive mechanism/s [1] to provide appropriate quality to the clients given their requirements, while ensuring optimal use of resources and maximum revenue to the CSP. The challenges include design of the tool and protocols for communication between the server and client modules.

4. REFERENCES

1. J. Liu, B. Li, Adaptive Video Multicast over the Internet, IEEE Multimedia, January-March 2003.
2. S. Krithivasan, S.Iyer, Enhancing Quality of Service by Exploiting Delay Tolerance in Multimedia Applications, ACM Multimedia , Nov. 2005.
3. S. Krithivasan, Mechanisms for Effective and Efficient Dissemination of Multimedia, Technical report –September 2004, URL: www.it.iitb.ac.in/~saras
4. X. Wang, H. Schulzrinne, Comparison of Adaptive Internet Multimedia Applications, IEICE Transaction Communication, VOL.ES2-B, NO.6, June 1999.