Maximizing Revenue through Resource Provisioning and Scheduling in *Delay-Tolerant* Multimedia Applications: **A Service Provider's Perspective**

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

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

ABSTRACT

An emerging trend in multimedia applications such as distance education and corporate training is to service clients according to their convenience in terms of start time of the multimedia content and the quality of reception. Examples include clients requesting for a movie to start 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 network that exhibits heterogeneous characteristics, where a Content Service Provider (CSP) services requests from geographically dispersed clients synchronously. It is important for a CSP to utilize resources such as buffers, transcoders, and caches judiciously in order to minimize costs while serving clients with their required quality in order to maximize revenues. We approach the problem in the following steps: (i) Determine the optimal quality deliverable to the clients while satisfying their delay tolerance assuming static network characteristics for the duration of the play out (ii) Since the playback need to start at the requested time, determine the optimal placement of buffers, caches, and transcoders such that resource utilization is maximized and client device constraints are satisfied (iii) Use admission control and scheduling to consider the trade off in revenue if clients were admitted (dynamic arrivals) while satisfying the admitted clients' requirements by maximizing resource utilization. We have developed an optimization-based approach to determine the best quality that can be delivered to the clients using 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.

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

1. INTRODUCTION

Current research in multimedia dissemination focuses on providing clients with immediate service, i.e., minimize the start up delay. In networks with heterogeneous link characteristics achieving (close to) zero start up delay means that the quality is dictated by the weakest link in the client's path, assuming that the content can be transcoded at the appropriate rate. However, there are several popular streaming media applications where clients request for the service at a specified time based on their convenience. Examples include accessing course content in a distance education application, reserving in advance for streaming of a popular movie, and availing training material at convenience across cities in a corporate setting. Typical characteristics of such applications are: 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 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, $(t+ d_i)$. Such a scenario can be equated to advanced reservation for a movie that starts at the specified time convenient to the client, encoded at a rate that at minimum is of the client specified quality.

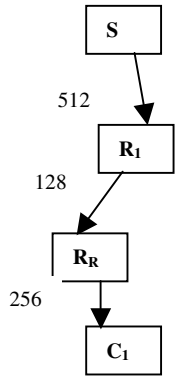
In the CSP's perspective following questions are important:

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 requirement of all admitted clients are met and revenue is maximized?

A review of the existing mechanisms for effective and efficient delivery of multimedia in [4][5] 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.

1.1.Motivating example

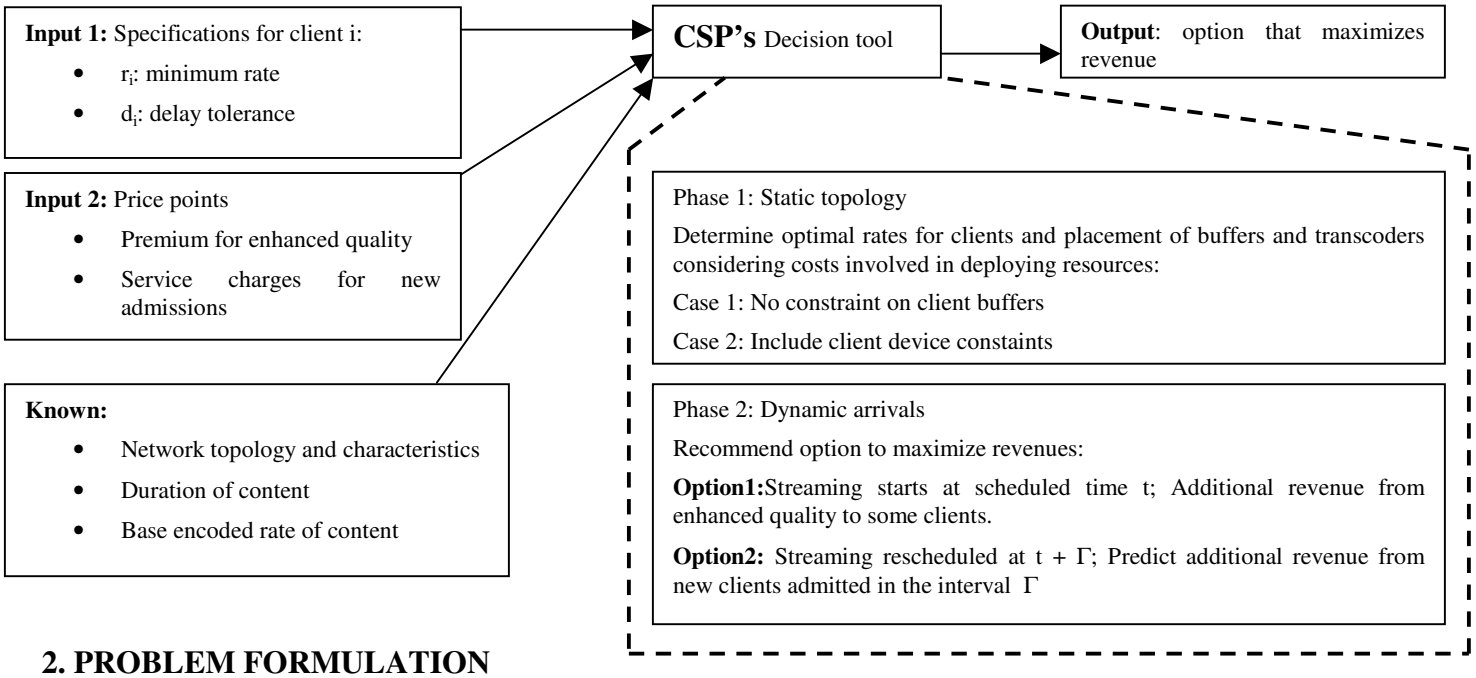
Consider a source S streaming multimedia content of duration T= 1 hr. encoded at 512 kbps to client C_1 connected through relay nodes R_1 and R_2 with links of capacities as shown in Figure 1. Table in figure 1 provides the rates at which the content can be served to C_1 for different values of delay tolerance d_1. By exploiting C_1's delay tolerance it can be provided better quality of reception using resources such as buffers and transcoders even though link capacities in its path are constrained [2][3].



Delay tolerance d_1	Rate at C_1 (kbps)
0	**128**
0.1T (6mins.)	**140.8**
0.25T(15 mins.)	**160**
0.5T(30 mins.)	**192**
T(1 hr.)	**256**

Figure 1: A motivating example

1.2. Scope of problem from the CSP's perspective



2. PROBLEM FORMULATION

We have formulated an optimization approach to determine the best possible rates at the clients and the placement of resources such as buffers and transcoders. We use a nonlinear least square function to determine the optimal rates that can flow across the links given the following constraints: (i) *Latency constraint*: this constraint will ensure that the cumulative delay incurred due to buffering in the network nodes is always bounded by the client specified delay tolerance (ii) *Rate constraint*: we use this constraint to ensure that the encoded rate is always >= the client specified minimum rate and <= the base encoded rate of the content, the best possible quality (iii) *Transcoding constraint*: this constraint will ensure that the content can be transcoded only to a lower rate. We present our experimental results in Figure 3 using Matlab for a network having 12 nodes and 5 clients as shown in Figure 2.

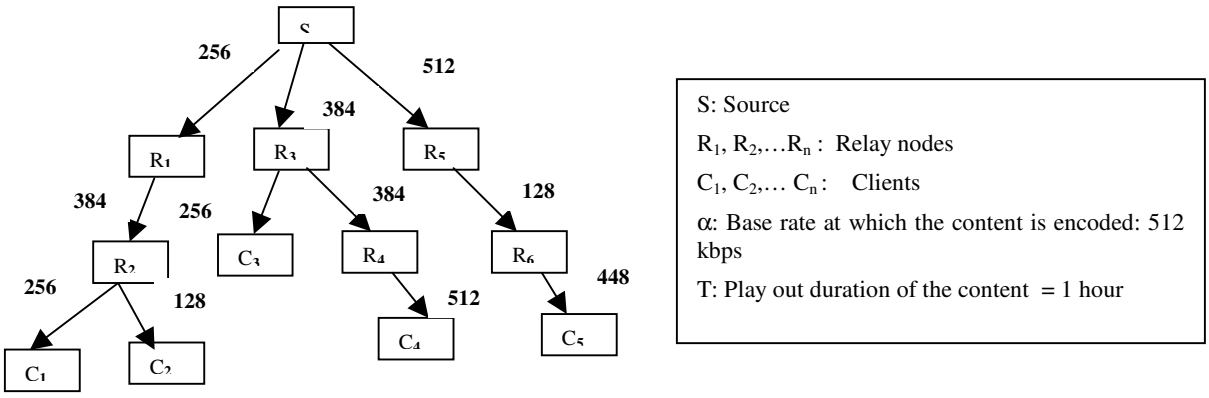


Figure 2: A dissemination tree

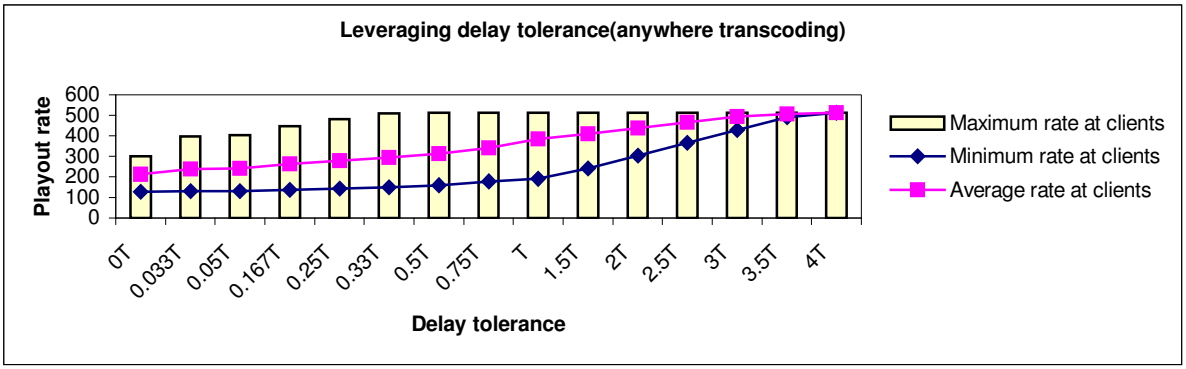


Figure 3: Improved quality through exploiting client delay tolerance

2.1. Optimal placement of resources

The study of placement of resources from a service provider's perspective is important due to the following reasons: (i) to minimize cost of deployment and maximize utilization of deployed resources and (ii) to account for client device constraints. The second point is very relevant to the current trend where the end user devices may have memory constraints and relay nodes supported by the service provider may be capable of caching. Such caches can then be used to serve future requests for the same multimedia content. We have studied placement of transcoders by changing the constraint specifications in the optimization formulation. We consider the cases where the source provides multiple encoded streams and the case where transcoders are deployed at intermediate nodes. Figure 4 shows the effect of placement of transcoders on the average rate (in kbps) for clients with different delay tolerance values (in minutes).

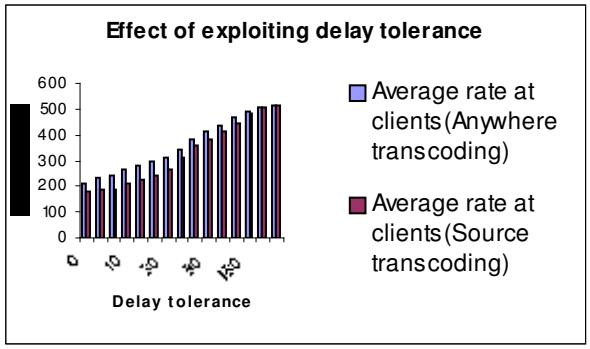


Figure 4: Placement of transcoders: Effect on quality

2.2. Scheduling and admission control: Revenue model

In the first phase of our research, 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 change in the network topology. Consider a static case where due to high bandwidth links or through 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_2 where $t_2=t_1+\Gamma$, 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, considering that maximizing revenue is the objective in the service provider's perspective:

1. At t_1, we run a predictive tool that predicts the number of arrivals and client requirements based on simple distributions.

2. We define various price points for the enhanced quality of service when we start the stream at t_1. In other words, when a client is served with quality better than its minimum required quality, a small premium is charged to the client. Another option is to reschedule the steaming at $t_1+\Gamma$ and admit new requests for the service during the interval Γ. New arrivals bring additional revenue proportional to the number of new requests. By considering the tradeoff between the two options, we recommend an appropriate alternative. In either case, all admitted clients would be served at their specified time with at least their minimum specified quality.

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 to maximize revenues for the service provider.

3. CONCLUSIONS

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. Final contribution of this work would be a tool that invokes suitable adaptive mechanisms [1] to provide appropriate quality to the clients given their requirements, while ensuring optimal use of resources and maximum revenue to the CSP.

4. REFERENCES

1.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, S.Iyer, Enhancing QoS for Delay-tolerant Multimedia Applications: Resource utilization and Scheduling from a Service Provider's Perspective, accepted in Student workshop, Infocomm 2006.

4. S. Krithivasan, Mechanisms for Effective and Efficient Dissemination of Multimedia, Technical report –September 2004, URL: www.it.iitb.ac.in/~sarask

5. X. Wang, H. Schulzrinne, Comparison of Adaptive Internet Multimedia Applications, IEICE Transaction Communication, VOL.ES2-B, NO.6, June 1999.