Enhancing Quality of Service by Exploiting Delay Tolerance in Multimedia Applications

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

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

ABSTRACT: *Delay-tolerant* multimedia applications, where clients are willing to wait for a specified time for the start of play back, fit the profile of many emerging applications, such as distance education, and corporate training. 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. This thesis deals with the issue of maximizing rates (quality) at the clients while satisfying their delay requirements, with minimal additional resources. To this end, we have developed an optimization-based approach to determine the rates at the clients through judicious placement of resources such as transcoders and caches. Simulation results demonstrate the usefulness of exploiting client delay tolerance specifications for delivering enhanced QoS with little or no additional resources. The final contribution of this thesis will be a tool that invokes appropriate transcoding and caching mechanisms to provide optimal quality to clients based on factors such as cost of additional resources and bandwidth availability.

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

1. INTRODUCTION

With the proliferation of world-wide computer networks, 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, and multinational corporations providing training to employees across cities. Heterogeneous architectures comprising of satellite, terrestrial links as well as the Internet are increasingly deployed for such applications. In these applications, a source disseminates multimedia contents that may be encoded at different rates to a set of geographically distributed clients through links of varying capacities and characteristics. In these applications the clients specify their QoS requirements in terms of a *minimum playout rate* they desire and a *maximum startup delay* they can tolerate. We define delay-tolerant applications as those where clients are willing to wait for a specified time for the start of play back.

***State of the art*:** A review of the existing mechanisms for effective and efficient delivery of multimedia in [2][5][6] 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 delays. A related problem has been addressed in the context of profit making e-commerce applications in broadcast environments [1]. Our focus, in contrast, is on exploiting explicit delay tolerance to maximize delivered multimedia quality in heterogeneous network environments.

Motivation: In delay-tolerant applications, there is no advantage to be gained by serving the clients before their scheduled playout time. However, there are several benefits to the Content Service Provider (CSP) (due to enhanced customer satisfaction) in offering a playout rate that is better than what a client demands, if this can be achieved without jeopardizing the ability to satisfy the minimum client requirements and *without additional resources*. Thus, the CSP's objective is not only to satisfy clients' requirements but also to provide the best possible playout rate, and to maximize the utilization of link capacities in the network.

Example: We motivate the context for solving this problem by considering an already deployed distance education model [4]:

- S is a source that captures classroom lectures and mixes with other multimedia inputs such as slides. S encodes the edited content at a high fidelity base rate r_{Base}. S provides several courses; a client can subscribe to one or more courses. S starts the streaming synchronously to all the clients of a course.
- Clients are typically institutions (regional colleges, corporations) which offer courses at scheduled times to registered participants; Clients can also be individual users. Each client specifies a minimum playout rate, i.e., the encoding rate that it requires. Each client also specifies a startup delay, i.e., the delay between the start of a lecture at the source and the start of the playout at the client.

The following system characteristics and parameters are known at S: (i) Topology of the dissemination network, including the physical capacities of the various links. (ii) r_{Base} the encoded rate of the stream (best possible rate at which clients can be served). (iii) Minimum rate requirement at client C_i, denoted as r_i and maximum delay requirement at client C_i, denoted as d_i.

In the figure below S serves five clients, C_1, C_2, C_3, C_4, and C_5 through the dissemination network, represented as a tree with link capacities in kbps given on every link. r_{Base} equals 512 kbps. Duration of the multimedia stream equals T. For this network, for different startup delay specifications (different multiples of T), Table 1 lists the average rate, minimum and maximum rates and the number of different rates (transcoders) required. These are generated by the optimization tool we have developed as part of this thesis.

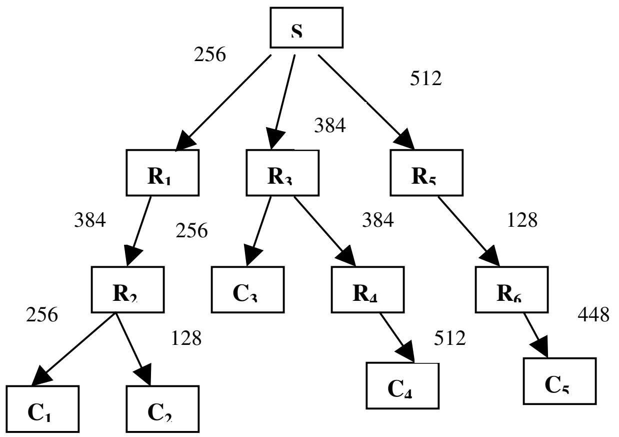


Table 1: Effect of exploiting startup delay on delivered QoS

Delay	Parameters	Transcoders (at any node)	Transcoders (only at S)
0	# of transcoders	**6**	**2**
	Average rate (kbps)	**230.4**	**170.5**
	Range (kbps)	**(128-384)**	**(128-234.25)**
0.5T	# of transcoders	**5**	**2**
	Average rate (kbps)	**313.6**	**268.8**
	Range (kbps)	**(160-512)**	**(192-384)**
T	# of transcoders	**4**	**1**
	Average rate (kbps)	**368.64**	**358.4**
	Range (kbps)	**(204.8-512)**	**(256-512)**
2T	# of transcoders	**4**	**3**
	Average rate (kbps)	**447.96**	**418.13**
	Range (kbps)	**(354.43-512)**	**(341.33-512)**
3T	# of transcoders	**2**	**1**
	Average rate (kbps)	**492.63**	**482.74**
	Range (kbps)	**(428.9 –512)**	**(438.8-512)**
4T	# of transcoders	**0**	**0**
	Average rate (kbps)	**512**	**512**
	Range (kbps)	**-**	**-**

By considering clients requiring zero delays, the first row in Table 1 focuses on minimizing the startup delay. This corresponds to the goals of existing research. The remaining rows in Table 1 demonstrate that when clients allow startup delays, the rates at which the multimedia content can be delivered to the clients can be improved: at a startup delay value 4T, all clients are served with the best possible rate (i.e., 512 kbps, the encoding rate of the file). While this result is intuitively obvious, it is also seen that this improved quality is achieved with fewer transcoders (i.e., less resources) than for the case with zero startup delay.

Closer scrutiny of the results reveals that a low capacity link *not* on the path from S to a client can affect its quality. For example, client C_2 having a low bandwidth link (128 kbps) affects the playout rate of C_1 and results in longer convergence time for the optimizer. Thus, using the optimization approach, a CSP can decide where to deploy additional resources for maximum impact on client QoS.

The rightmost column in Table 1 gives the results for the case when only the source can transcode. It is seen that in all but the 4T case, while a smaller number of transcoders suffice, a lower playout rate results compared to the case when transcoders are deployed anywhere in the network. Thus, we can see that several factors affect the rates at the clients in a delay tolerant multimedia application deployed over a heterogeneous network:

1. *Capabilities of the source*: Whether the source is capable of encoding at multiple rates (leading to replicated streams being disseminated by the source [3][6]).
2. *Capabilities at the relay nodes*: Whether transcoders [6] or caches [7] are available at the relay nodes.
3. *Link bandwidths along the path from the source to the client*: Whether the link characteristics are static or dynamic [2][6].

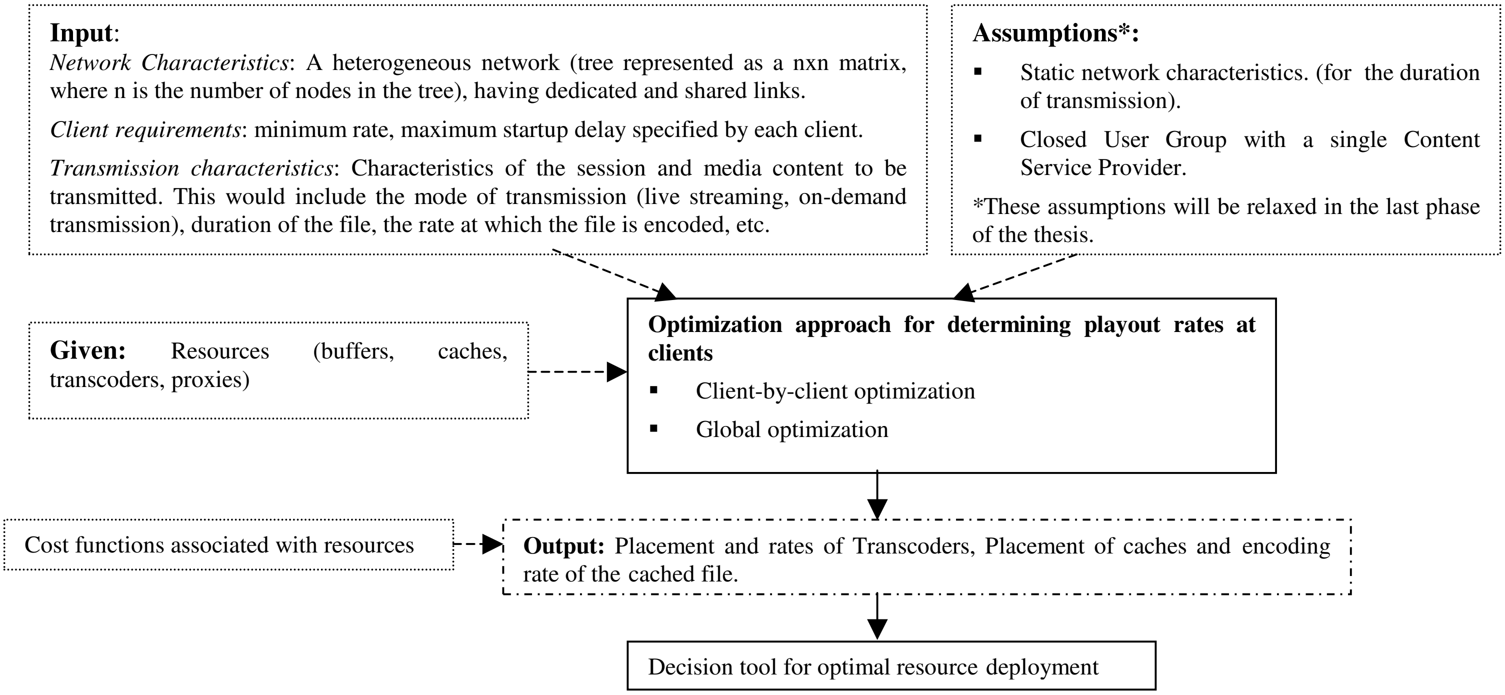
With the number of delay tolerant applications on the increase [8], it is important to investigate how the allowed delays can be exploited to deliver better quality multimedia content to the clients with little or no additional resources. This motivates the objectives and approach of our research, as elaborated in the next sections.

2. OBJECTIVES AND SCOPE

In the context of delay tolerant applications, our objective is to study the impact of a user specified *startup delay* on:

- *Playout rate (which defines the quality):* This involves understanding the buffer requirements at the nodes to maximize the playout rate while satisfying the delay requirement.
- *Resource deployment strategies of a CSP*: Considering transcoding and caching mechanisms, and costs associated with their deployment, ways by which such mechanisms can be deployed to optimally serve the clients have to be explored.

Our final goal is to develop a tool that determines the most suitable adaptive strategy based on optimal resource utilization that will invoke the appropriate mechanisms to provide optimal QoS to clients. We will also consider the effects of dynamically changing link capacities in some parts of the heterogeneous network and group membership. Scope of the thesis is summarized below:



Input:
Network Characteristics: A heterogeneous network (tree represented as a nxn matrix, where n is the number of nodes in the tree), having dedicated and shared links.

Client requirements: minimum rate, maximum startup delay specified by each client.

Transmission characteristics: Characteristics of the session and media content to be transmitted. This would include the mode of transmission (live streaming, on-demand transmission), duration of the file, the rate at which the file is encoded, etc.

Assumptions*:

- Static network characteristics. (for the duration of transmission).
- Closed User Group with a single Content Service Provider.

*These assumptions will be relaxed in the last phase of the thesis.

Given: Resources (buffers, caches, transcoders, proxies)

Optimization approach for determining playout rates at clients

- Client-by-client optimization
- Global optimization

Cost functions associated with resources

Output: Placement and rates of Transcoders, Placement of caches and encoding rate of the cached file.

Decision tool for optimal resource deployment

3. SUMMARY OF SOLUTION APPROACH

Work completed: *Static network characteristics, Transcoding only*		
	Description	*Details*
1	***Basic block analysis***: Analysis using a basic block consisting of a *source (S), relay(R)*, and a *client (C)*, the three types of nodes used to represent any heterogeneous network. 	The following three cases are considered: • Startup delay= Infinity; The best playout rate r_{Base} can be provided to the client. • Startup delay= zero; Playout rate at client depends on the weakest link in its path. • Startup delay= $d_{i.;}$ Playout rate= At least r_i as defined by client Ci.
2	***Client-by- client optimization***: Extension of basic block to path of client C_i, which specifies both r_i and d_i (Path from source to each client is considered in isolation).	Posed as an optimization problem, solved using Matlab's fmincon function; To handle large networks heuristics are also developed. Provides: • Feasibility check for each client to ensure that both requirements can be met, given the network constraints. • Upper bound of the optimal rate that can be realized at client C_i.
3	***Global optimization***: The entire network is considered. An additional constraint that accounts for multiple clients sharing links is added to the optimization problem.	• Finds optimal rates for each client • Provides placement and rates of transcoders to realize the optimal rates.
Roadmap for remaining work: *Adding caching capability to nodes, catering to changing network characteristics.*		
4	Analysis with other adaptive mechanisms such as caches and combination of transcoders and caches.	• Caching at relay nodes and caching strategies to provide optimal rates. • Viability of combining mechanisms (transcoding and caching) at different sub-trees.
5	Study of trade-off between costs incurred for deployment of transcoders/caches and rates provided to the clients.	• Refine global optimization to include minimization of number of transcoders/caches and develop approaches to factor costs involved in deploying resources and the additional processing delays.

6	Catering to: Links with dynamically varying capacities. Clients which dynamically join/leave the network.	• Explore how our optimization approach can be tailored to work with QoS mechanisms that use adaptive techniques based on bandwidth availability [2][6] in dynamic networks. • Study the impact on the choice of mechanisms and optimal rates at the client when clients exhibit dynamic behaviour.
7	Implementing the tool: Includes the design, protocol between the server and client modules.	• Centralized design: Having a server module that takes decisions on the mechanisms to be invoked and a client module that takes parameters from the server module and invokes the appropriate mechanisms. • Distributed approach: Here decisions are made at different levels of the tree. Additionally, nodes where decisions are made (strategic nodes from where sub-trees emanate) need a protocol to coordinate their decisions.

4. CONCLUSION

With our preliminary simulation experiments, we have shown that clients that can tolerate startup delays can be served with higher rates with little or no additional resources. Such applications fit the profile of many emerging applications such as distance education. We have formulated our solution as an optimization problem to serve clients in a heterogeneous network with optimal rates. 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) without incurring any additional costs for resources. In addition to the development of the optimization tool, our proposed work will provide a decision tool for optimal deployment of resources to best serve multimedia content to a set of delay tolerant clients in a heterogeneous network.

5. REFERENCES

1. C.C. Aggarwal, M.S. Squillante, J.L. Wolf, P.S. Yu, J. Sethuraman, Optimizing profits in the broadcast delivery of multimedia products, Fifth International workshop on Multimedia Information Systems, October 1999.
2. J. Liu, B. Li, Adaptive Video Multicast over the Internet, IEEE Multimedia, January-March 2003.
3. T. Kim, M. Ammar, A comparison of layering and stream replication video multicast schemes, NOSSDAV, June 2001.
4. S. Krithivasan, S. Iyer, To Beam or to Stream: Satellite-based vs. Streaming-based Infrastructure for Distance Education, EdMedia, June 2004.
5. S. Krithivasan, Mechanisms for Effective and Efficient Dissemination of Multimedia, Technical report –September 2004, URL: www.it.iitb.ac.in/~saras
6. X. Wang, H. Schulzrinne, Comparison of Adaptive Internet Multimedia Applications, IEICE Transaction Communication, VOL.ES2-B, NO.6, June 1999.
7. J. Wang, A Survey of Web Caching Schemes for the Internet, Cornell Network Research Group, 2001.
8. F. Warthman, Delay tolerant networks (DTNs): A tutorial, May 2003.
   http://www.ipnsig.org/reports/DTN_Tutorial11.pdf