Optimised Batch Patching with Classes of Service

Paul P. White
Visitor
Dept. Computer Science, UCL
Gower St.,London WC1E 6BT, UK
Tel: +44 7679 2000
p.white@cs.ucl.ac.uk

Jon Crowcroft
Professor
Dept. Computer Science, UCL
Gower St.,London WC1E 6BT, UK
Tel: +44 7679 7296
j.crowcroft@cs.ucl.ac.uk

ABSTRACT

In this paper we present a new technique called Optimised Batch Patching with Classes of Service(OBP with CoS) which can be used to leverage the benefits of multicast within the context of 'near' video-on-demand systems in IP networks. OBP with CoS builds on an earlier scheme, known as Optimised Patching, but is different in two respects. Firstly, in OBP with CoS, the server artificially delays requests in order to increase the probability of accumulating duplicates, thereby allowing greater exploitation of multicast compared to Optimised Patching albeit at the expense of higher latency until commencement of service. Secondly, the client is able to request a latency class of service which reflects how long the client is prepared to wait until playout of the video. The server scheduling algorithm attempts to minimise network bandwidth consumption within the constraints imposed by the class of each request. We present analysis and simulations of our scheme in order to validate its effectiveness.

Keywords

Multicast, video-on-demand, multimedia.

1. INTRODUCTION

Today IP networks are being used for many applications and services beyond conventional data transfer. One example is Video-On-Demand(VOD). VOD describes a system whereby a user is able to request any video from a library stored on one or more servers. In response to the request, the server will deliver the video as an isochronous stream for playout by the user. With 'true' VOD the delay between the user submitting a request and commencement of playout of the requested video should be small, typically a few seconds. An alternative is what could be referred to as 'near' VOD whereby the delay to commencement of playout is significant, typically a few minutes, but still acceptable to the user under certain conditions. For example, in the case of movies-on-demand, where the playout duration of a movie is say 90 minutes, a delay of several minutes might be acceptable to the user provided there is an appropriate cost reduction compared to 'true' VOD.

A major landmark in the recent evolution of IP networks is the introduction of multicast routing and forwarding techniques for group communication. When the same information needs to be sent to multiple recipients, a single multicast transmission using a protocol such as Protocol Independent Multicast(PIM)[4] will consume less network bandwidth and impose less of a load on the sender compared to a separate unicast transmission to each recipient.

Although many schemes have been proposed that attempt to exploit the benefits of multicast for VOD most of them are based on 'periodic broadcast' principles[1][3][5][7] whereby each video is split into portions which are continuously multicast over a number of multicast channels and as such are only bandwidth efficient provided request arrival rate is high.

A multicast scheme that is bandwidth-efficient over a wider range of request arrival rates is that of Patching[6] which facilitates 'true' VOD as well as 'near' VOD.

In the case of near VOD, each request incurs a delay at the server prior to commencement of service. In [6] this delay is only ever caused by bandwidth limitations on the server output link although, as we will explain later, an alternative is for delay to be artificially imposed by the server scheduling algorithm in order to increase the likelihood of obtaining duplicate requests thereby allowing further exploitation of multicast.

By the time service of a delayed request is about to commence the server may have accumulated additional requests for the same video. Each such group of identical requests is known as a batch. Following the service point of each request batch as determined by the server scheduling algorithm, the batch must then be served over 1 or 2 channels, that is multicast groups, either a so-called regular channel alone, or the combination of a regular channel and a so-called patching channel. A regular channel delivers a full video from start to finish while a patching channel delivers only the missing part of the video from the start until the point at which the clients of a batch join the regular channel. In [6] the multicast groups were realised at the application layer in order to minimise congestion on the internal server bus. However the technique would also work using network layer multicast which would result in network bandwidth savings in a multicast-capable IP network.

In [6], the preferred method of patching, known as grace patching, works as follows. At the service point of a batch, the server identifies the newest regular channel that is serving the requested video and equates the channel's age to a corresponding buffer space, e.g. to store 10 minutes of Mpeg video requires approximately 100 Mbytes of buffer space. If the corresponding buffer space exceeds the buffer capabilities of the clients in the batch then patching cannot be used for the batch in which case a new regular channel must be opened instead and its identity conveyed to the clients in the batch. Each client then joins the regular channel in order to receive the entire video.

Assuming the buffer space of the clients is sufficient then patching can be used as follows. The server opens a new channel known as a 'patching' channel which will deliver only the early part of the requested video that the clients will miss upon joining the video's newest regular channel. The server then conveys the identity of both patching channel and newest regular channel to each client in the batch. Each of these clients then joins both channels and buffers the regular channel while playing out the patching channel. When the patching channel has been