Journal of Network and Computer Applications 35 (2012) 2016–2027

Contents lists available at SciVerse ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Content-aware rate allocation for efficient video streaming via dynamic network utility maximization

Mohammad H. Hajiesmaili [a,b,*], Ahmad Khonsari [a,b], Ali Sehati [a,b], Mohammad Sadegh Talebi [b]

[a] School of ECE, College of Engineering, University of Tehran, Tehran, Iran
[b] School of Computer Science, IPM, Tehran, Iran

ARTICLE INFO

Article history:
Received 7 May 2012
Accepted 31 July 2012
Available online 8 August 2012

Keywords:
Video streaming
Dynamic network utility maximization
Content-aware video utility model
Buffer underflow avoidance
Convex optimization

ABSTRACT

Nowadays it is vital to design robust mechanisms to provide QoS for multimedia applications as an integral part of the network traffic. The main goal of this paper is to provide an efficient rate control scheme to support content-aware video transmission mechanism with buffer underflow avoidance at the receiver in congested networks. Towards this, we introduce a content-aware time-varying utility function, in which the quality impact of video content is incorporated into its mathematical expression. Moreover, we analytically model the buffer requirements of video sources in two ways: first as constraints of the optimization problem to guarantee a minimum rate demand for each source, and second as a penalty function embedded as part of the objective function attempting to achieve the highest possible rate for each source. Then, using the proposed analytical model, we formulate a *dynamic network utility maximization* problem, which aims to maximize the aggregate hybrid objective function of sources subject to capacity and buffer constraints. Finally, using primal–dual method, we solve DNUM problem and propose a distributed algorithm called CA-DNUM that optimally allocates the shared bandwidth to video streams. The experimental results demonstrate the efficacy and performance improvement of the proposed content-aware rate allocation algorithm for video sources in different scenarios.

1. Introduction

1.1. Motivation

The recent advances in networking technologies make the multimedia applications such as on-demand streaming, IPTV, videoconferencing, and online video games, an integral part of the traffic passing through communication networks. In contrast to the conventional traffic like web browsing and file transfer, video transmission is characterized by its stringent Quality of Service (QoS) requirements, especially large bandwidth demand and continuous playback at receiver. To improve the satisfaction level of clients, multimedia networking has gained a lot of research interests, addressing many new challenges inherent to these applications. There is a large body of literature devoted to empirical and analytical study of multimedia communications.

Different from data traffic, video traffic is loss-tolerant, i.e., certain packets can be dropped when the network is congested. Hence, an appropriate key feature for video transmission applications is the adaptability of video quality on a per-user basis. Based on the simple idea of intelligently dropping video frames when facing congestion, several techniques have been proposed in the community for video adaptation. Scalable video coding (SVC) (Schwarz et al., 2007), transcoding (Ahmad et al., 2005; Xu et al., 2012), and video summarization (Li et al., 2005) are some of the main techniques that incorporate content adaptation into video-based transmission scenarios.

Traditionally, video encoding techniques and resource allocation in networking design scenarios are investigated separately in the area of multimedia and networking. As time went by, the idea of content-aware networking arose. The very idea of content-aware networking paradigm is to utilize the characteristics of the content for rate allocation in resource-limited scenarios. As such, this paradigm proves extremely promising for video transmission in congested networks, while it sounds inapplicable for conventional elastic traffic. This has led to emergence of several works from both multimedia and networking communities in recent years (Pandit et al., in press; Zhu et al., 2011; Li et al., 2009; Hong et al., 2010; Lu et al., 2005; Tan and Chou, 2012). The main motivation of content-awareness is based on the fact that parts of video content are not equal in terms of quality impact when facing loss. Put another way, different frames have different importance levels depending on how many frames have decoding

* Corresponding author at: School of ECE, College of Engineering, University of Tehran, Tehran, Iran. Tel.: +98 9125261811.
E-mail addresses: hajiesmaili@ipm.ir (M.H. Hajiesmaili), ak@ipm.ir (A. Khonsari), a.sehati@ipm.ir (A. Sehati), mstalebi@ipm.ir (M.S. Talebi).

1084-8045/$ - see front matter Crown Copyright © 2012 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2012.07.024

dependency on them. So, in order to provide better perceived visual quality for the end user, the basic idea is to solve content-aware resource allocation problem, which smartly allocates the resources according to the content importance for the users.

1.2. Dynamic Network Utility Maximization

Since the publication of seminal paper Kelly et al. (1998), *Network Utility Maximization (NUM)* has been used as a promising framework to analytically model the new wireline and wireless network resource allocation problems. NUM in its basic version (Low and Lapsley, 1999) concerns with maximizing the aggregate utility of users, using a fixed set of sources and predetermined paths, while satisfying the capacity constraints. An important issue in regard to utility-based optimization approaches like Kelly et al. (1998) and Low and Lapsley (1999) is that they solve the network utility maximization problem in a single period and fail to capture the temporal dynamics in modern networks. NUM responds to future changes in link capacities or flow utility functions by resolving the new problem and then adjusting flow rates accordingly.

Recently Trichakis et al. (2008) considered the explicit notion of time in NUM and called it *Dynamic Network Utility Maximization* (*DNUM*). The DNUM system model allows the flow utilities, link capacities, and source paths to vary over time. Moreover, they presented a multi-period variation of NUM, with constraints involving rate-sum guarantees over some prespecified periods, which are called *delivery contracts*.

Delivery contracts are conducive to express QoS requirements of multimedia applications in terms of the minimum aggregate rate requirements of each source during a specified time interval. Since delivery contracts are incorporated into the underlying network problem as constraints, the basic problem does not allow contract violations, which might make the problem infeasible. Thus, depending on the delivery contracts, the basic version of DNUM might be impractical in the realistic systems. Additionally, delivery contracts could be interpreted as buffer requirements of sessions, where the receiver of a streaming session employs buffers in order to cope with network-induced delay jitter. Thus, an issue of great importance is to ensure that receiver buffer does not run out of video frames during playback. As shown in this paper, we can characterize this requirement using delivery contracts, which will be discussed in the subsequent subsection.

1.3. Our contribution

Using the state-of-the-art content-aware networking paradigm, the main goal of this paper is to provide an efficient solution to support content-aware video transmission with continuous playback at the receiver in communication networks. Towards this, the main contributions of this paper are as follows:

- To consider characteristics of video content in bandwidth allocation, based on the simple idea of different importance levels of video frames, we introduce a content-aware time-varying utility model which intelligently tracks the quality impacts of video content.
- If users experience several freezes in video playback, they may give up the service. So, another contribution is to respect buffer underflow avoidance, where we employ the idea of delivery contracts. Since the delivery contracts are incorporated into the optimization problem as constraints, the basic version of DNUM does not allow contract violations. Herein, if the available bandwidth is less than the required amount to support the predetermined contracts, the optimization problem will be infeasible and it may have no optimal solution. To take one step further towards a more practical scheme, in this paper, we consider the idea of delivery contracts to model buffer requirements. The benefit of using the concept of delivery contracts is two-fold: first it strives to guarantee a minimum rate demand for each source and second it penalizes the objective when the allocated rate of a source is below its encoding rate.
  In the first case, delivery contracts appear as problem's constraints, while in the second case, it provides a subtractive term for the system utility to play its role as a penalty function.
- Using the proposed time-varying utility functions and analytical modeling of buffer requirements, we formulate rate allocation as a convex optimization problem which aims to maximize the aggregate hybrid objective function of sources subject to capacity and contract (buffer) constraints. By extending the basic version of DNUM so as to involve a penalty term, we could not employ the straightforward dual-based solution proposed in Trichakis et al. (2008) any longer. Hence, we solve the problem using primal–dual approach for which some iterative update equations are derived for the primal and dual optimization variables. Moreover, using the update equations we proposed a distributed algorithm called *Content-Aware Dynamic Network Utility Maximization* (*CA-DNUM*) which runs in predetermined time intervals and because of its awareness of content importance, it smartly allocates the transmission rate of each source. Finally, by experimental results we demonstrate the performance improvement achieved by our design.

This paper represents a follow-on work to our previous work Hajiesmaili et al. (2012). In this study, we employed the same utility model of Hajiesmaili et al. (2012). In Hajiesmaili et al. (2012), we have modeled buffer requirement of each video source as a linear inequality constraint of the underlying optimization problem. In this work, we have employed such a model for buffer requirements too. However, in this paper we have modified the objective function so as to include a penalty term, which is aware of buffer consumption of all video sources. The role of such a penalty term is to provide sources with rates near the video encoding rate. Also, the proposed algorithm in this work is armed with a content-aware frame dropping scheme to be implemented at the source. Moreover, some other practical issues have been discussed to provide guidelines for realistic implementation of CA-DNUM algorithm. In contrast to Hajiesmaili et al. (2012), our comprehensive simulation experiments in this paper provide a more clear comparison with similar design schemes, which help the reader judge the superiority of CA-DNUM with respect to others.

The rest of this paper is structured as follows. Section 2 is devoted to the related work. In Section 3 we give the system model. In addition, the content-aware time-varying video utility model and our proposed buffer underflow avoidance approach are also presented in this section. In Section 4, we propose the optimal solution of the optimization problem and CA-DNUM algorithm. Some practical notes are discussed in Section 5. Following this, the experimental results are provided in Section 6. Finally, the paper is concluded in Section 7.

2. Related work

There are several existing work on NUM-based resource allocation problems considering special characteristics of multimedia streaming applications (Chiang et al., 2007). Even though most of researchers have established NUM-based resource allocation for elastic traffic (like web and email) with concave and