Joint Rate Control and Scheduling for Video Streaming over OFDMA Networks

Peng Hui TAN, Jinghong ZHENG, Zhengguo LI, Sumei SUN

Institute for Infocomm Research, A*STAR
1 Fusionopolis Way, Connexis South Tower, Singapore 138632
Email: {phtan, jzheng, ezgli, sunsm}@i2r.a-star.edu.sg

Abstract—In this paper, we investigate joint rate control and packet scheduler for video streaming over downlink orthogonal frequency division multiple access (OFDMA) networks. We proposed algorithm which exploits rate based utility functions and the video packet’s importance information which is related to its position in the group of pictures. The time scale of rate control is larger than that of packet scheduler. The importance information for scheduling is embedded in the IP packets, making it more suitable for practical purposes. We extend the Lyapunov optimization technique for single time scale network to develop a low complexity algorithm, which does not require prior statistical information on the channel and queues. The algorithm also has a bounded worst case delay, which is useful for incorporating deadline constraints for video streaming. The performance of the proposed algorithm is evaluated by simulation in a LTE network. Simulation results show that our algorithm outperforms existing techniques.

Index Terms—Video Streaming, rate control, OFDMA scheduling, Lyapunov optimization

I. INTRODUCTION

Due to limited and time-varying capacity of wireless network, adaptive streaming is emerging for transmission of video streams over wireless networks. Video segments at multiple bitrates from video codec such as H.264/MPEG-4 SVC or AVC are available. To switch between different bitrates for better performance, one can use some rate control algorithms which are usually based on cross-layer design. The current end-system-centric approach for video streaming is HTTP-based adaptive streaming [1], where the end user chooses the video rate based on the estimated bandwidth and playout buffer length.

For the network-centric approach, the control is done at the lower layers of the transmission system to satisfy quality of service (QoS) parameters. However, the video application quality metric at the application (APP) layer requires an effective mapping in terms of the lower layer quality metrics. In this paper, we focus on network-centric approach for designing a cross-layer adaptive video streaming algorithm over an orthogonal frequency division multiple access (OFDMA) network, which is the downlink multiple access method for 3GPP Long Term Evolution (LTE) [2].

The video data are first packetized at the server before being injected to the base station servicing multiple clients over the (downlink) OFDMA packet network. Upon reception they are stored in the receiver buffer until the video playout is started. Packets which arrive after deadline are considered lost. Error concealment is used instead of retransmission from the server. The main contribution of this work is to design a joint rate control and scheduling algorithm that maximizes rate utilities and minimizes weighted number of bits dropped. The rate based utility function relates the rate decision to the video quality metric, whilst the weights quantify the importance level of the video packets. We start with a simple wireless network model with multiple queues. Extending the Lyapunov optimization technique developed in [3, 4], we exploit the time scale separation of the decisions to yield the algorithm for rate control, scheduling and bit dropping. The rate control algorithm operates on a large time scale, while the scheduling and bit dropping algorithms operate on a small time scale. The dropping of bits also leads to a bounded worst case deadline which is related to head-of-line (HoL) delay. This wireless model is then extended to the OFDMA network. With feedback of a scalar from the base station, it is possible for the user to use the video application quality metric at the at the application (APP) layer to make its own rate control decision. Multiple resource blocks (RB) scheduling with ARQ/HARQ error protection is also considered.

Before proceeding further, we should highlight the difference between our work and other existing studies. A typical assumption in most works [5–9] is that the decisions of the rate controller and scheduler are performed on the same time scale. In each time slot, single or several group of pictures (GoP) will be transmitted. However, in practice time scale of one or several GoPs typically span multiple scheduling time slots at the media access control (MAC) layer. Moreover, APP metrics are not available directly at the MAC layer where the scheduler resides. To address these limitations, in our work the APP metrics are mapped into the importance levels embedded into the IP packets. With packet inspection, the scheduler is able to obtain these information. In [10], the joint congestion control and scheduling has been designed using network utility maximization (NUM) approach. The two time scales modeling is presented, but only the rate constraints are applied as it was designed for data traffic. Lastly, despite both our work and [3, 11] use Lyapunov optimization technique for designing joint rate control and scheduling algorithm, our work includes the two time scales modeling and the use of packet’s importance level for scheduling.
The rest of the paper is organized as follows. We discuss
the system model used and the utility optimization problem
studied in this paper in Section II. In Section III, we develop
the rate control, scheduling and bit dropping algorithm based
on the Lyapunov optimization. Modifications of the algorithm
for OFDMA network are also discussed. The LTE simulation
scenario and the simulation results in Section IV. Concluding
remarks are listed in Section V.

II. WIRELESS NETWORK MODEL

For simplicity, we first consider a one-hop discrete time slots
wireless network. Data from $K$ queues are to be transmitted
to the $K$ end users by a scheduler at Media Access Control
(MAC) layer. Let $Q(t) = (Q_1(t), \ldots, Q_K(t))$ denote the
queue (in terms of bits) for time slot $t \in \{0, 1, 2, \ldots\}$. The vector of new arrivals is denoted by $a(t) = (a_1(t), \ldots, a_K(t))$
and $0 \le a_k(t) \le A^\text{max}_k, \forall k \in \{1, \ldots, K\}$. At time slot $t$, the scheduler determines the amount of bits in the queue to
schedule for transmission and drop. The dynamics of the queue
is modeled as follows:

$$Q_k(t + 1) = \max\{0, Q_k(t) - b_k(t) - d_k(t)\} + a_k(t),$$

where $b_k(t)$ and $d_k(t)$ are the amount of bits to be transmitted
and the amount of bits to be dropped, respectively.

The transmission vector $b(t) = (b_1(t), \ldots, b_K(t))$ is a
function of the current channel condition $S(t) = (S_1(t), \ldots, S_K(t))$ and the modulation and coding schemes
(MCS) chosen for error free transmission. Mathematically, it
is given by

$$b_k(t) = b_k(MCS_k(t), S(t)), \quad (2)$$

which is further bounded by $0 \le b_k(t) \le B^\text{max}_k(t) \le B^\text{max}_k$. Note that in our model, fragmentation of the data frame is
allowed.

The constraint for bit dropping decision $d_k(t)$, which is also
the amount of data dropped, made at every time slot is bounded by

$$0 \le d_k(t) \le D_k(t), \quad \forall k \in \{1, \ldots, K\}, (3)$$

where $D_k(t)$ is the amount of bits that can be dropped from queue $k$. The choice of $D_k(t)$ affects the performance of the algorithm. The way of choosing $D_k(t)$ is deferred until Section III.

The rate control decision $x_k(t)$, also the amount of bits injected to the scheduler, is made by the rate controller in
every $T > 1$ time slots (the larger time scale), i.e., $x_k(t) = x_k([t/T]/T)$. It corresponds to the video encoding rate for
single or multiple GoPs. In this work, we encode a single GoP for every $T$ time slots. The constraint imposed is given by

$$0 < X^{\text{min}} \le x_k(t) \le X^{\text{max}}. \quad (4)$$

Once rate control decision $x_k(t)$ is made, the GoPs are encoded and the bits are packetized and become the arrival of $Q_k(t)$. For integer $r \ge 0$ and $rT \le t \le rT + T - 1$, the arrival rate $a_k(t)$ is a function of rate control decision $x_k(t)$ and (binary) packet arrival indicator $J_k(t)$:

$$a_k(t) = a(x_k([t/T]/T), J_k(t)). \quad (5)$$

The arrival indicator $J_k(t)$ is a random process which indicates
if a new packet arrived at time slot $t$. We further assume that all the bits arrive within this period

$$x_k([t/T]/T) = \sum_{r=rT}^{rT+T-1} a_k(\tau) \quad (6)$$

Note that we have implicitly assuming that the starting of GoP is synchronized for all queues in the model. This makes
the analysis of the algorithm more concise. We will relax
this assumption in Section IV when obtaining the simulation
results.

A. Delay-Aware Virtual Queues

We introduce a modified virtual queue $Z_k(t)$ with initial value as $Z_k(0) = 0$ and operates at each time slot (smaller
time scale). This virtual queue update is given

$$\bar{Z}_k(t + 1) = \max\{0, Z_k(t) + c_k - b_k(t) - z_k(b_k(t))\} \quad (7a)$$

$$Z_k(t + 1) = \max\{0, \bar{Z}_k(t + 1) - [d_k(t) + z_k(d_k(t))]\} \quad (7b)$$

where $z_k(b_k(t))$ and $\bar{z}_k(d_k(t))$ are compensation to set $\bar{Z}_k(t + 1)$ and $Z_k(t + 1)$ to $\epsilon H_k(t + 1)$ when whole packet is
transmitted or dropped, respectively. Here, $H_k(t + 1)$ is the HoL delay of the packet and it increases by one unit at a time.
Note also that the transmission $b_k(t)$ is done before dropping $d_k(t)$. Since we do not consider the overflow of queues, the scheduling decisions come before the bit dropping decisions as it is a better strategy to serve and then drop the remaining bits. Moreover, if $\epsilon \ge A^\text{max}$, we have $Z_k(t) \le \epsilon H_k(t)$. The purpose of these virtual queues is to facilitate the incorporation into
the network model of deadline constraints related to $H_k(t)$. It will be clearer as we describe the dynamic algorithm in
Section III.

III. LYAPUNOV OPTIMIZATION AND DYNAMIC
ALGORITHM DESIGN

We formulation an optimization problem and solved it by
extending the Lyapunov optimization technique presented in
[3,4] for the single time scale. The optimization problem is given by:

$$\max_{\alpha(t) \in A_{\text{rate}}} \left\{ \sum_k \phi_k(y_k - u_k v_k) \right\} \quad (8a)$$

subject to

$$y_k \le x_k, \quad (8b)$$

$$u_k \le b_k + d_k \quad \text{and} \quad \epsilon_k \le \bar{b}_k + \bar{d}_k$$

where

$$\phi_k(y_k) = \frac{\epsilon_k}{T} \sum_{r=0}^{T/T-1} \phi_k(y_k(rT)), \quad x_k = \frac{T}{T} \sum_{r=0}^{T/T-1} x_k(rT), \quad u_k = \frac{T}{T} \sum_{r=0}^{T/T-1} v_k(\tau) \quad \text{and} \quad I_k(\tau) \text{ denotes the importance level} \quad (9)$$
of the HoL data frame. Variable $\pi_k$, $\overrightarrow{b}_k(t)$ and $\overrightarrow{\beta}_k$ are the time average of $a_k(t)$, $b_k(t)$ and $d_k(t)$, respectively.

We first introduce a virtual queue $G_k(t)$ with $G_k(0) = 0$. Since $y_k(t)$ is updated every $T$ time slots, the update of $G_k(t)$ is defined as:

$$G_k(rT + t) = \max[0, G_k(rT) + \delta (y_k(rT) - x_k(rT))],$$

(9)

for $r \geq 0$, $\delta > 0$ and $G_k(t) = G_k(t/T \cdot T)$. We design a drift-plus-penalty algorithm which minimizes a weighted sum of the $\overrightarrow{G}_k$.

Bounding (11), the control action for this upper bound, the following drift-plus-penalty algorithm according to (1), (7) and (9), respectively.

$$\max \left\{ \eta \overrightarrow{G}_k(t') x_k(t') - \frac{\sum_{\tau=(r-1)T}^{rT-1} Q_k(\tau) a_k(\tau)}{x_k((r-1)T)} \right\}$$

(15)

s.t. $X_k^{\min}(t') \leq x_k(t') \leq X_k^{\max}(t)$

- **Rate Control**: For each $t' = rT$ and $k \in \{1, \ldots, K\}$, observes $G_k(t')$ and chooses $x_k(t)$ so that:

$$\max \left\{ \eta \overrightarrow{G}_k(t') x_k(t') - \frac{\sum_{\tau=(r-1)T}^{rT-1} Q_k(\tau) a_k(\tau)}{x_k((r-1)T)} \right\}$$

(15)

• **Scheduling**: For each time slot $t$, choose $b_k(t)$ to

$$\max_{0 \leq b_k(t) \leq B_k^{\max}(t)} \left\{ \sum_k [Q_k(t) + Z_k(t)] b_k(t) \right\}$$

(16)

• **Bit Dropping**: For each time slot $t$, choose $d_k(t)$ to

$$\max_{0 \leq d_k(t) \leq D_k(t)} \left\{ [Q_k(t) + Z_k(t) - V u_k I_k^{\max}(t)] d_k(t) \right\}$$

(17)

With the above dynamic algorithm, we can ensure that the data frames are dropped with a deadline given in the following lemma [12].

**Lemma 1**: Suppose that $L_k(t) \geq D_k(t) \geq \epsilon_k$, where $L_k(t)$ is the minimum number of bits to be dropped such that the HoL delay of the subsequent packet has been decreased by more than one time slot. Then the deadline for the data frames is given by the maximum value of $Z_k^{\max} = V u_k I_k^{\max} \geq Z_k(t)$. The proof is shown via induction. By definition, we have $\epsilon_k(H_k(t) - 1) \leq Z_k(t) \leq \epsilon_k H_k(t)$. If deadline constraints based on $H(t)$ are present, we can adjust $V$, $u_k$ and $I_k^{\max}$ according to Lemma 1. Furthermore, it can be shown that the drift-plus-penalty algorithm described comes within $O(1/V)$ of the utility of a genie-aided $T'$-slot lookahead algorithm with an average delay constraint of $O(V)$. The proofs make use of the techniques described in [3, 4], while incorporating the two time scales decision making introduced in our network.

**A. Extension to layered OFDMA Network**

We now discuss the implementation of the drift-plus-penalty algorithm (14)-(17) for video streaming over downlink OFDMA network. Each queue in the network corresponds to the queue at the MAC layer for a user in the OFDMA network. The rate control decision determines the video encoding rate for each GoP at the APP layer, while the scheduling and bit dropping decisions are performed at the MAC layer.

To minimize message passing across layers, we choose to perform the rate control decisions at the MAC layer of the base station and then pass the decisions to the video encoders at the APP layer. The packet importance information is placed in the Type of Service (ToS) field in the IP header by the video server. The rate control decision is encoded by the base station into the Explicit Congestion Notification (ECN) field, which is also located in the IP header. With this ECN information, the user selects the video rates via sending HTTP or RTCP messages to the video server. However, this also implies that we cannot use the video application quality for $\phi_k(y_k(t))$ as
this information is not available at the MAC layer. Hence, we
stick to the log based utility function of rate:
\[ \phi_k(y) = \log(1 + y/1000). \] (18)
Furthermore, we do not know the bit sizes \( X_k^{\text{min}}(t) \) and \( X_k^{\text{max}}(t) \) of the GoP for each available rate when solving (15). To overcome this problem, we replace it with average number of bits generated for a GoP interval.

In each time slot, there are multiple resource blocks (RB) to be allocated for an OFDMA network and error protection via ARQ and HARQ needs to be considered. For finite queues, the optimization problem in (16) is NP-hard [13]. To reduce complexity, we consider RB allocation in a pre-determined order, which will be described later. For each time slot \( t \) and RB \( n \), the value of \( b_{k,n}(t) \) is obtained by solving
\[ \max_{b_{k,n}(t)} \{ |Q_k(t) + Z_k(t)| b_{k,n}(t) \}, \] (19)
where \( b_k(t) = \sum_n b_{k,n}(t) \). Bit dropping decisions and queue updates are the same as before. We label this controller as VLO and the HO scheduler, and LO scheduler when packet’s importance information is not used.\(^1\) As usual, the scheduling criterion includes the delay of packet and the channel condition experienced by the user. The MaxWeight scheduler in [14] uses \( Q_k(t) \) to reflect the delay of the packet, while Maximum-Largest Weighted Delay First (M-LWDF) in [15] uses the HoL delay. For the LO-type scheduler (LO and VLO), it uses \( Q_k(t) \) and \( Z_k(t) \) to reflect the delay of the packet. Furthermore, the LO-type scheduler is content aware as it uses embedded packet’s information \( I_k(t) \) to make scheduling decisions. If (14) and (15) are also implemented, we label them as VLO or LO rate controller.

The scheduling algorithm described in (19) is used only for new packet transmission at each RB. Other available transmission modes for the user are HARQ and ARQ retransmissions. A higher priority is given to the HARQ retransmission, followed by the ARQ retransmission and lastly new transmission [16]. Retransmission uses the same MCS and occupies the same number of RBs as before. The remaining RBs are then used by the new transmission.

**IV. Simulation Results**

We examine the performance of the joint rate control and scheduling algorithm. The simulation parameters of the LTE network are given in Table I and the encoding parameters related to the video quality are presented in Table II.

**A. Rate Control Performance**

For performance comparison on joint rate control and scheduling, we extend the NUM based rate controller proposed in [10], where rate control and scheduling are performed jointly over OFDMA network. However, in that formulation, only rate constraints were considered.

The common parameters for the LO-type rate controller are \( \eta = 1, \delta = 0.01, T = 10^3 \) and \( V = 8 \times 10^3 \). The rest of the parameters depend on the deadline constraint and the maximum importance level. For deadline of 100 ms (or time slots), \( V u_k = 100 \epsilon_k I_k^{\text{max}} = 8 \times 10^3 I_k^{\text{max}} \). Packets with lower importance level are dropped earlier. For the LO rate controller, we have \( I_k(t) = 1 \) and \( V u_k = 8 \times 10^3 \). The importance level assignment of the frames follows a linear decreasing function. The importance level of the I frame and the last P frame in the GoP are \( I_k^{\text{max}} \) and 1, respectively. Each packet from the same frame has the same importance level. We are interested in the performance of the cell-edged UEs. The comparison metrics used are the minimum and the ten percentile SSIM indices of the video stream, averaged over the 60 s transmission period. Two slow motion video sequences (Mother-daughter (MD) and News), one medium motion sequence (Foreman) and one fast motion sequence (Soccer) are used for the simulations.

Fig. 1 shows the empirical CDFs of the SSIM index for various rate controllers with Foreman video sequence, with \( I_k^{\text{max}} = 4 \). The rate controller has the option to encode the video GoP in either 150 kbps or 300 kbps. Two CDF curves (labeled as LO150 and LO300) are also shown for reference purposes. For these curves, the LO scheduler is used and all the UEs are transmitting at 150 kbps and 300 kbps, respectively. In this figure, we see that the LO-type rate controller outperforms the NUM controller at all percentile of SSIM index. Unfortunately, the packet’s importance information does not help much in the performance of VLO controller for this video sequence. It is because the rate controller is trying to match the video encoding rates to the transmission rates attainable by the UEs. Hence, it reduces the chance of mismatch and the usefulness of video packet’s importance information.

---

**TABLE I: Simulation parameters of LTE network**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sessions</td>
<td>10 with simulation length of 60s each</td>
</tr>
<tr>
<td>Physical Detail</td>
<td></td>
</tr>
<tr>
<td>Carrier Frequency</td>
<td>2 GHz</td>
</tr>
<tr>
<td>Bandwidth for the DL</td>
<td>5 MHz</td>
</tr>
<tr>
<td>Symbol for TTI</td>
<td>14</td>
</tr>
<tr>
<td>TTI duration</td>
<td>1 ms</td>
</tr>
<tr>
<td>Subcarriers per RB</td>
<td>12</td>
</tr>
<tr>
<td>Subcarrier spacing</td>
<td>15 kHz</td>
</tr>
<tr>
<td>eNodeB transmission power</td>
<td>43 dBm</td>
</tr>
<tr>
<td>2 TX and 2 RX antennas with ZF detector</td>
<td></td>
</tr>
<tr>
<td>Modulation Scheme</td>
<td>all available</td>
</tr>
<tr>
<td>Target BLER</td>
<td>10%</td>
</tr>
<tr>
<td>Overhead</td>
<td></td>
</tr>
<tr>
<td>PDCCH</td>
<td>3 OFDM symbols; CRC: 3 bytes</td>
</tr>
<tr>
<td>ARQ</td>
<td></td>
</tr>
<tr>
<td>No. of RLC ARQ retransmissions</td>
<td>0</td>
</tr>
<tr>
<td>No. of HARQ retransmissions</td>
<td>4</td>
</tr>
<tr>
<td>CQI</td>
<td></td>
</tr>
<tr>
<td>No. of RBs per CQI</td>
<td>1; Measured period: 2 ms</td>
</tr>
<tr>
<td>Traffic Model</td>
<td>H.264 Video traffic: 20 UEs with 150/300kbps; Best effort (BE) traffic: 10 UEs</td>
</tr>
</tbody>
</table>

**TABLE II: Selected encoding parameter of H.264 video**

<table>
<thead>
<tr>
<th>Encoding Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile</td>
<td>Baseline</td>
</tr>
<tr>
<td>Target bitrates</td>
<td>150/300 kbps</td>
</tr>
<tr>
<td>Frame rate</td>
<td>30 fps</td>
</tr>
<tr>
<td>GoP size</td>
<td>30</td>
</tr>
<tr>
<td>Sequence type</td>
<td>IPP..P</td>
</tr>
</tbody>
</table>
achieved by the LO scheduler with UEs transmitting at 150 kbps. Similarly, the maximum SSIM indices achieved by the LO-type rate controller are also close to the maximum SSIM index achieved by the LO scheduler with UEs transmitting at 300 kbps. Although the LO-type rate controllers have better SSIM index than the NUM controller, it has the worse BE throughput performance. It is because they are able to allocate more RBs to the video traffic.

Table III summaries the minimum and the ten percentile SSIM indices of various rate controllers and video sequences. It can be seen that the LO-type rate controller performs better than the NUM controller. The gain in ten percentile SSIM index of LO over NUM controller are 0.0098, 0.0220, 0.0070 and 0.0137 for MD, News, Foreman and Soccer, respectively. On average, an increase of 0.0134 in ten percentile SSIM index is achieved. For the minimum SSIM index, it achieves a higher average gain of 0.0292. It should also be noted that the linear decreasing packet’s importance information does not increase the ten percentile SSIM index of the rate controller by much.

V. CONCLUSION

We have considered the problem of joint rate control and scheduling for video streaming in the downlink of OFDMA network. The rate control decisions are determined on a larger time scale while the scheduling and bit dropping decisions are made on a smaller time scale. Our algorithm maximizes the time average of the weighted bits dropped. The weight accounts for the video packet’s importance information. Using the Lyapunov optimization technique, we have derived a simple way to implement the algorithm. No priori statistical information of the random events on channel and queuing states is required. Numerical simulations over LTE networks have been presented to show the effectiveness of the proposed algorithm. The importance level assignment used is a linear decreasing function of the frame position in the GoP. It is found that with rate control, the benefit of video importance information vanishes as the video encoding rate to the UE matches its transmission rate.

REFERENCES