Online Source Rate Control for Adaptive Video Streaming Over HSPA and LTE-Style Variable Bit Rate Downlink Channels

Jian Yang, *Member, IEEE*, Yongyi Ran, Shuangwu Chen, Weiping Li, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Online source rate control (RC) is designed for video streaming over high-speed packet access (HSPA) and Long-Term Evolution (LTE)-style variable bit rate (VBR) downlink channels. The problem is formulated as the adaptive adjustment of the operational mode of a video encoder based on the buffer overflow probability (BOP) feedback received from the radio link control (RLC) layer at the base station (BS). This allows us to maximize the attainable visual quality while keeping the transmitter BOP below a desired threshold and maintaining a video delay as low as possible. We derive an online measurement-based BOP estimation model for the RLC buffer, which is capable of operating with no prior knowledge of the channel variations and of the video characteristics. Based on this estimation model, an online adaptive RC algorithm is proposed to seamlessly adapt the bit stream to the characteristics of VBR channels. Our experiments are conducted in multiuser scenarios using VBR video encoding combined with adaptive modulation and coding (AMC) in the transceiver. The results demonstrate that the proposed source RC regime supports near-instantaneous yet smooth bit stream adaptability, which makes it useful for HSPA and LTE-style systems for the sake of accommodating unknown video traffic characteristics and dynamically fluctuating propagation conditions.**

***Index Terms*—Large deviation principle (LDP), rate control (RC), variable bit rate (VBR) channel.**

I. Introduction

ADVANCED transceiver techniques, including adaptive modulation and coding (AMC) and hybrid automatic repeat request (HARQ), are widely applied in wireless systems to combat the effects of both fading and interference [1], [2]. For instance, both the third-generation (3G) and high-speed packet access (HSPA) apply AMC and HARQ in the physical layer relying on the channel quality indicator's (CQI) feedback from the user equipment (UE) to achieve prompt link adaptation. The Long-Term Evolution (LTE) Advanced and WiMax standards also provide similar link adaptation based on AMC and short time-slot duration periods. Naturally, agile link adaptation may result in a time-varying effective throughput of the channel. Therefore, supporting near-instantaneous bit rate adaptivity is one of the most important features of video streaming applications.

Rate control (RC) aims to control the video bit rate to match the channel's achievable bit rate, given the prevalent channel quality to keep the video quality as high as possible [3]. Most existing studies of RC focus on allocating a bit rate budget to each group of pictures (GOP), frames, or macroblocks at the encoder. For instance, an efficient RC scheme was designed for MPEG-4 based on a quadratic rate–distortion (R–D) model invoked to maintain the target bit rate in [4]. In [5], the quadratic R–D model was also adopted to derive an adaptive RC for H.264 to meet the target bit rate. In [6], RC was designed for scalable H.264/Advanced Video Coding (AVC) encoding. H.264 reference software JM [7] has implemented the JVT-G012 [8] RC algorithm, which relies on a combination of the available channel bandwidth, the frame rate, the target buffer level, and the actual buffer fullness for determining the number of bits allocated for the current frame. A range of AMC-based schemes were conceived in [9] where no extra buffering was used by the RC. The design philosophy was to instruct the video encoder to produce the exact number of bits for each video frame, which was affordable for the near-instantaneous HSPA-style AMC transceiver mode that was periodically signaled back from the receiver to the transmitter. However, the aforementioned RC schemes rely on the knowledge of the target bit rate, which again requires the feedback of the expected bit rate based on the estimated channel quality. Unfortunately, the channel's bit rate fluctuations are not known *a priori* at the transmitter and the video encoder. Therefore, these RC schemes may impose substantial video quality fluctuations.

Recently, a control-theoretic approach-based RC regime has been proposed in [10] by jointly considering the encoder's RC and network congestion control. More specifically, an empirical R–D source model, a channel-induced distortion model, and their linearized models were applied to formulate a mathematically tractable system model. However, it is a challenge to formulate accurate models when streaming videos over

Manuscript received January 10, 2014; revised July 21, 2014 and November 3, 2014; accepted January 24, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61174062; by the State Key Program of the National Natural Science Foundation of China under Grant 61233003; and by the Fundamental Research Funds for the Central Universities. The work of L. Hanzo was supported by the European Research Council under an Advanced Fellow Grant. The review of this paper was coordinated by Prof. N. Arumugam.

J. Yang, Y. Ran, S. Chen, and W. Li are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: jianyang@ustc.edu.cn; yyran@mail.ustc.edu.cn; chensw@mail.ustc.edu.cn; wpli@ustc.edu.cn).

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2015.2398515

time-varying channels. Typically, intelligent packet scheduling is used at the base station (BS) by most of the existing solutions to achieve channel-quality-dependent adaptive video streaming over wireless networks. Cross-layer-adaptive scalable video streaming has been proposed by informing the media access control (MAC) layer of the amount of payload and the index of the modulation-and-encoding scheme to improve the attainable video quality [11]. An adaptive scalable video streaming strategy relaying on controlling the MAC buffer was presented in [12]. Both the packet deadlines and the channel characteristics were considered in video packet scheduling at the BS [13]. The cross-layer design-aided transmission of scalable video streams [14] involves packet scheduling combined with a video frame dropping strategy at the BS. Most of these papers discuss the channel-dependent adaptation of the MAC layer of the BS where no signaling is fed back to instruct the video source to adjust its bit rate. Dropping packets at the BS is a waste of resources both in the backhaul router and the core network. It was shown in [15] that packet dropping at the video source is more efficient than MAC layer dropping at the BS when we have to reduce the congestion in both the wired core network and in the wireless medium to the UE. By translating the CQI values to the number of enhancement layers scheduled for transmission, the video server facilitates adaptive video streaming [16]. However, the link throughput at the BS is determined not only by the CQI value but also by the MAC layer scheduling strategy of a multiuser scenario. Moreover, the bit rate of video clips having different amounts of motion activity may be very different even if they have the same number of enhancement layers [17]. Therefore, such CQI-mapping-based video source adaptation lacks robustness against the time-variant number of users and against the heterogeneous video bit rates. In [18], although a Markov decision process is invoked for dynamic video scheduling, its offline computation requires *a priori* knowledge of the channel dynamics. Hence, a heuristic online scheduling algorithm is derived based on the average channel throughput estimated with the aid of a first-order autoregressive model [AR(1)]. In the HSPA/LTE system, using a CQI-based link adaptation mechanism may lead to a burst-by-burst transmission block adaptation, and the linear estimator based AR(1) may not be appropriate for estimating the average channel throughput.

Against this background, we proposed a novel online source RC framework, where a buffer overflow probability (BOP) estimator is employed at the radio link control (RLC) layer of the BS, and the estimated BOP is signaled back to the video source to control its bit rate. The motivation of using a BOP-based feedback is that the BOP metric characterizes the degree of matching between the video bit rate of the source and the link throughput. Since the BOP is estimated before a buffer overflow is encountered, its feedback may assist the video source in promptly controlling the bit rate. In contrast to the methods found in the open literature [9]–[14], [16], [18], the main contributions of this paper are threefold.

- To enable the video encoder to generate a video stream that may be reliably delivered over variable bit rate (VBR) downlink channels, we formulate a constrained optimization problem subject to a constraint imposed on the transmission BOP to guarantee a low delay.
- A BOP estimation model based on large deviation principles (LDPs) [19] is proposed by monitoring the buffer fullness and its variation at the BS's RLC layer. The reason for applying the LDPs is because it accurately characterizes the probability of rare events, which assists us in achieving fine adjustment of the video source rate.
- We conceive an iterative RC algorithm to approach the most beneficial encoder rate, thus circumventing the difficulty of directly solving the related constrained optimization problem. The proposed RC algorithm allows the encoder to adjust its bit rate to that of the VBR channel without any *a priori* knowledge of both the channel quality variations and of the characteristics of the video source.

The remainder of this paper is organized as follows. Section II describes the system model, including the formulation of the source RC problem of video streaming over a VBR downlink channel. In Section III, we apply the LDP in [19] to derive a BOP estimation model, whereas an online measurement-based RC is proposed in Section IV. Our numerical simulation results are presented in Section V to characterize the attainable performance of the proposed algorithm. Finally, our conclusions are offered in Section VI.

II. System Model

A. System Overview

Fig. 1 shows a typical video streaming scenario over a wireless network. The main associated protocol stacks are also shown in the lower part of the figure. Naturally, a video streaming service involves not only maintaining the wireless connection between the mobile station (MS) and the base transceiver station (BTS) but supporting the public network connection (Internet) as well. The wired network provides high capacity and stability; hence, video streaming over the wired network has become a well-established service and has many successful applications, including video conferencing, surveillance systems, and Internet Protocol (IP) television. By contrast, video streaming over wireless networks faces unique challenges due to the time-varying nature of the wireless channel and owing to the scarcity of the system resources, which makes it difficult to guarantee any specific video quality of service (QoS). Hence, the wireless transmission in the video streaming service is likely to be a bottleneck, which is the focus of this paper.

Fig. 2 shows the basic video processing in the video network abstract layer (NAL) units of the 3GPP framework. A NAL unit may be encapsulated in a Real-Time Transport Protocol (RTP) data unit, and then, it may be transmitted over User Datagram Protocol (UDP)/IP. HTTP-based streaming protocols such as Apple HTTP Live Streaming (HLS) [20] are alternative media streaming communication protocols, which are capable of traversing any firewall or proxy server that lets through standard HTTP traffic, unlike UDP-based protocols such as RTP. 3GPP standardized an adaptive HTTP streaming protocol

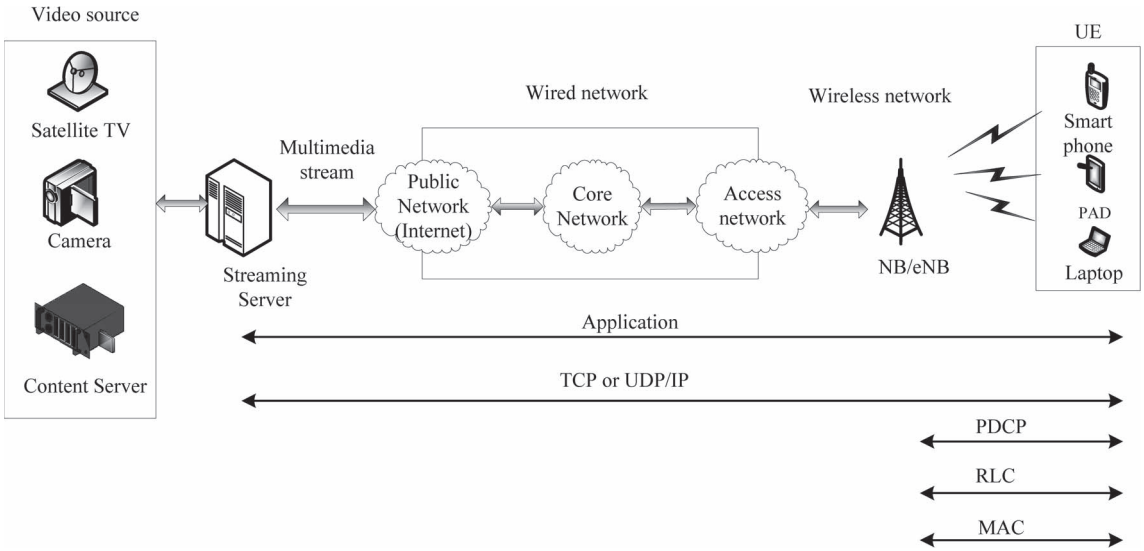


Fig. 1. Typical scenario for wireless video streaming in 3G Partnership Project (3GPP) framework.

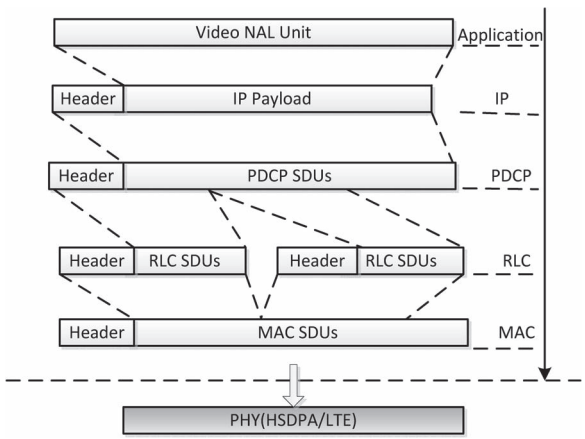


Fig. 2. Video data packetization at different layers in the 3GPP framework.

in Rel-9 [21]. Subsequently, we concentrate on the terminology used in HSPA NB/eNB, as shown in Fig. 2. The IP/UDP/RTP or IP/Transmission Control Protocol (TCP)/HTTP packet generated is encapsulated into a single Packet Data Convergence Protocol (PDCP) packet that becomes an RLC service data unit (SDU). Since a typical RLC SDU has a larger size than an RLC protocol data unit (PDU), it has to be segmented into smaller units. The length of the RLC PDU depends both on the selected bearer and on the AMC mode used. The RLC PDUs are forwarded to the MAC layer and are then encapsulated as MAC PDUs. The main functions of the PDCP layer are robust header compression and decompression, ciphering and deciphering, the transfer of data, and PDCP sequence number maintenance. The RLC layer in wireless systems is capable of operating in both an unacknowledged mode (UM) and acknowledged mode (AM), and both are capable of providing RLC PDU loss detection. However, while the UM is unidirectional and data delivery is not guaranteed, in the AM, automatic repeat request (ARQ) is applied for reliable data transmission. The main functions of the RLC layer include the transfer of upper layer PDUs; error correction relying on ARQ (only for AM); and the concatenation, segmentation, and reassembly of RLC SDUs, whereas the main functions of the MAC layer are resource scheduling and multiplexing/demultiplexing of MAC SDUs belonging to one or several logical channels into/from the relevant transport blocks (TBs). By comparing the functions of the RLC layer with those of the PDCP and MAC, the RLC is an appropriate layer for us to construct a model for characterizing the degree of matching between the network's throughput and the video source bit rate. Generally, the detection of a lost RLC PDU results in the loss of an entire PDCP packet; hence, the encapsulated IP and the NAL unit are lost. From the perspective of reacting to both the dynamics of the statistical fluctuation of the teletraffic and the variable channel conditions, agile bit rate adaptivity is one of the most important features for seamless video streaming over wireless systems. There are several ways of achieving bit rate adaptivity. For online encoding applications, the bit rate adaptivity can be achieved by controlling encoding parameters. For instance, H.264/AVC supports these features mainly by dynamically varying the quantizers but also by controlling temporal resolution. Scalable Video Coding (SVC) is another technique of implementing the bit rate adaptivity. It encodes the raw video clip into a base layer and a number of enhancement layers with different priorities. Naturally, the base layer has the highest priority since it contains the video bits with the highest importance, which can provide a minimum video quality. The enhancement layers with lower priorities may be progressively encoded to further refine the quality of the base-layer stream. This layered approach of the SVC codec allows an encoded stream to be flexibly prepared for meeting the bit rate constraint. In [22], the base-layer rate for a given video sequence is optimized to achieve the highest possible average perceived quality for heterogeneous clients, whereas in [23], a distribution regime was conceived for scalable videos, which guaranteed fairness for all end users. In Dynamic Adaptive Streaming over HTTP (DASH), the video content is partitioned into a sequence of small HTTP-based file

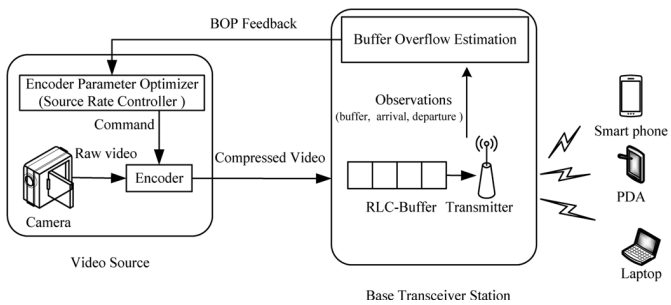


Fig. 3. Adaptive source RC framework for online video encoding.

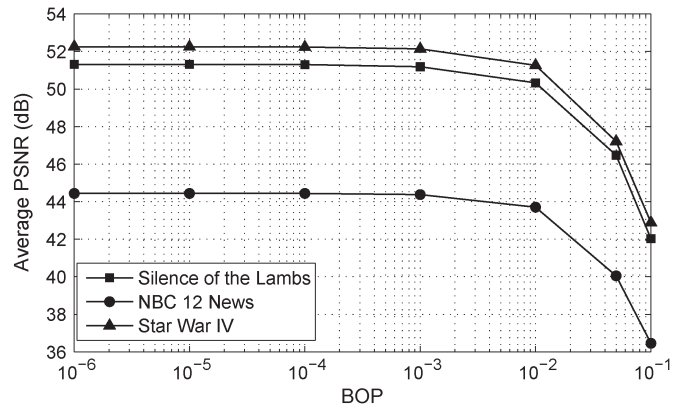


Fig. 4. Relation between BOP and received video quality by using three different video clips, namely, the Silence of the Lambs, NBC 12 News, and Star Wars IV. AQ1

segments, which are made available at a variety of different bit rates. More explicitly, segments encoded at different bit rates covering appropriately aligned short intervals of playback time are made available. Naturally, DASH also provides a flexible way of adjusting the video source bit rate. 3GPP has defined its Rel-10 version of DASH, termed as 3GP-DASH [24], which is a profile compatible with MPEG-DASH [25]. More work on HTTP-based adaptive video streaming can be found in [26]–[28].

Since the RLC buffer fullness indicates the degree of matching between the source video bit rate and the channel bit rate, we subsequently design a buffer fullness estimation scheme where the buffer fullness estimated at the RLC layer is fed back to the video source to adapt its bit rate. Without loss of generality, we consider a specific scenario of the online encoding RC relying on controlling encoding parameters, i.e., dynamically changing the quantization parameters (QPs). It is straightforward to extend the proposed method to both SVC streaming and to HTTP-DASH scenarios, as discussed in Section II-C.

B. Problem Formulation

The block diagram of a wireless video streaming system equipped with an encoder parameter optimizer conceived to control the source rate is shown in Fig. 3, which relies on a camera, a video encoder, and a BTS. The camera samples the video scene at certain frame scanning and forwards the frames to the encoder, which forwards the compressed video to the transmitter's buffer at the RLC layer for transmission. Here, we assume that the channel between the video source and the BTS is wired and reliable. The source RC problem is eliminating the congestion in the BTS while satisfying the delay constraints. Our basic philosophy is to feed back the mismatch between the current source rate and the channel's affordable throughput to the encoder parameter optimizer to adapt the source rate. If a mismatch does occur, the encoder parameter optimizer sends a command to the video encoder to adjust the video source rate. To quantitatively characterize this mismatch, we define BOP as the probability that the current wireless channel quality provides an insufficient throughput for the current video bit rate. A sufficiently low BOP indicates that we may increase the video bit rate for transmission over the wireless channel to achieve a higher video quality.

Let $\mathcal{M} \triangleq \{M_1, \ldots, M_L\}$ be the operational mode set of the video encoder. The video bit rate corresponding to the operational mode $M_i (i = 1, \ldots, L)$ is denoted by r_i. Let us assume that $r_1 < r_2 < \cdots < r_L$, which implies that a higher operational mode results in a higher video bit rate and hence achieves a better video quality. The operational mode is controlled by the encoder parameters. For instance, we can control the encoding mode by appropriately setting the I-P-B quantization mode or the target video bit rate parameter of the JM encoder. Here, we consider adapting the source rate to the channel quality by dynamically adjusting the operational mode (i.e., the encoding parameters) of the video encoder.

We used transmission slots of equal duration, which are normalized to unity. The RLC SDUs, which contain the video data generated by the video encoder, arrive at the RLC buffer, and they are then queued until they are transmitted. Since the data are processed at packet level instead of bit granularity, we define the length of the RLC buffer as the number of RLC SDUs. Let $A_n \in \mathcal{A} \triangleq \{0, \ldots, A\}$ denote the number of RLC SDU arrivals in slot n, where A is the maximum number of RLC SDU arrivals in a single slot, whereas $D_n \in \mathcal{D} \triangleq \{0, \ldots, D\}$ is defined as the number of RLC SDUs transmitted in slot n, where D is the maximum number of RLS SDUs transmitted in a single slot. It should be noted that an agile and sophisticated link adaptation mechanism based on AMC and HARQ is applied to maintain the required target bit error rate in most state-of-the-art wireless communication systems such as high-speed downlink packet access, 3G LTE, and WiMAX. Both AMC and HARQ rely on the CQI feedback received from the mobile terminals, which results in a burst-by-burst adaptive channel throughput. Hence, the downlink packet departure process D_n is assumed to be an independent identically distributed (i.i.d.) sequence. Let Q_n denote the buffer fullness expressed in terms of the number of packets at the end of slot n. The dynamics of the buffer fullness may be described as

$$Q_n = \max\left\{Q_{(n-1)} - D_n, 0\right\} + A_n. \tag{1}$$

According to Little's theorem [29], we have

$$\bar{Q} = \lambda \bar{D} \tag{2}$$

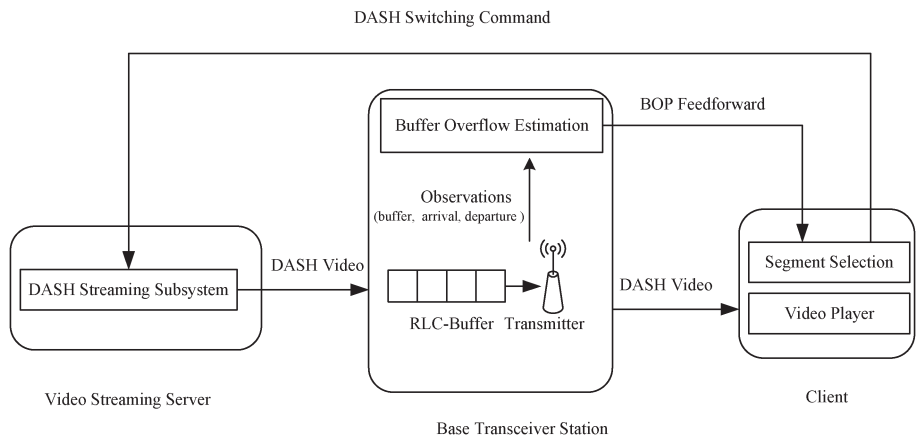


Fig. 5. Adaptive source RC framework for the DASH streaming application.

where $\bar{Q}$ is the average buffer fullness, $\bar{D}$ is the average delay, and λ is the packet arrival rate. This implies that, if λ is given, having a higher buffer fullness value implies a higher delay. Hence, we treat the delay as being synonymous with buffer fullness. Our objective is to select the best encoding mode for controlling the buffer fullness to satisfy the delay constraint of the specific video application. Let us formally define the BOP as

$$p_o = P(Q > B_h) \tag{3}$$

where Q is the buffer fullness and B_h is the buffer threshold. Higher BOP implies higher buffer fullness in the near future, in turn inducing a higher delay. To guarantee lip-synchronized interactive low-delay video transmission, the BOP should be kept low. Hence, the problem of RC may be reinterpreted as the problem of selecting the best encoding mode subject to a given BOP, which can be formulated as

$$\max_{m \in \mathcal{M}} \quad m \tag{4}$$

$$\text{s.t.} \qquad P(Q^m > B_h) < p_{\text{qos}} \tag{5}$$

where Q^m denotes the buffer fullness corresponding to the video encoding mode m, and p_{qos} is the maximum tolerable BOP.

It should be noted that the queue length $P(Q > x)$ in the scenario of an infinite buffer system is called an *overflow probability* or a *tail probability* [30], which has been used as a performance metric for an infinite buffer-aided system [31]–[33]. However, practical systems have a limited buffer capacity. When a video frame arrives at a full buffer, it gets dropped. Hence, the packet loss probability (PLP) $P_L(B_h)$ is invoked to characterize the performance of a finite-buffer-based system having a capacity B_h of video frames. For video streaming services, the PLP is an important QoS measure, but it is difficult to directly estimate the PLP of a general finite-buffer-based system. Fortunately, a theoretical justification was provided to approximate the PLP of a finite-buffer-based system with the aid of estimating the BOP of its infinite buffer counterpart [33]. Hence, we treat the BOP as being identical to the PLP, which implies that the video source rate optimization problem of a finite-buffer-based system having a capacity B_h of frames can be mapped to the mathematical model of (4) and (5). We have confirmed, by our simulation studies, the plausible fact that a high BOP may severely degrade the QoS, which is quantitatively characterized in Fig. 4 and motivates us to use the BOP for the feedback control.

C. Extensions for SVC and DASH Scenarios

The scheme shown in Fig. 3 is a point-to-point representation of our proposed framework, which may be also used for streaming to multiple users, as shown in our forthcoming SVC/HTTP-DASH streaming applications.

The framework detailed in Section II-B may be readily extended to SVC and HTTP-DASH scenarios. For an SVC scenario, a scalable video stream consists of a base layer l_1 and $(L-1)$ enhancement layers, i.e., $\{l_2, l_3, \ldots, l_L\}$. The operational mode $M_i (i = 1, \ldots, L)$ is defined as the scenario when the first i layers are selected for transmission. Thus, the source RC via adaptive enhancement layer selection can be also formulated as the problem [see (4) and (5)]. The corresponding video source in Fig. 3 may be also replaced by a video streaming server, which maintains the sessions corresponding to the multiple video users and receives the BOP feedback of a specific session for adjusting the number of the transmitted enhancement layers.

It should be noted that HTTP-DASH applied the pull-based streaming paradigm rather than the traditional push-based streaming paradigm relying on protocols such as the Real-Time Streaming Protocol. The client plays the central role of driving the video adaptation. To comply with the basic rule of DASH, the framework for the DASH streaming applications shown in Fig. 5 is based on the BOP feedforward to the client. The client may then use the received BOP to select the future video segments to be fetched. The DASH streaming server maintains multiple sessions for the sake of supporting video streaming to multiple users. In a DASH scenario, the operational mode set $\mathcal{M}$ is defined as the adaptation set, where the operational mode $M_i (i = 1, \ldots, L)$ denotes streaming the chunks corresponding to the ith quality level. Then, the source RC via HTTP-DASH can be similarly described as in (4) and (5) since a higher video quality implies a higher bit rate.

III. Large Deviation-Based Overflow Probability Estimation

Naturally, the BOP estimation is a key step to successfully solve the problem [see (4) and (5)]. To provide a high quality of experience for a video streaming service, an occurrence of buffer overflow is expected to be a rare event. The theory of large deviation provides a useful way of accurately characterizing the probability of rare events. Therefore, we present a large deviation-based BOP estimation model here. Although our practical buffer is of finite size, the BOP of the corresponding infinite-buffer-based system can be invoked to estimate the PLP, as shown in [31]–[33]. Hence, we subsequently aim to estimate the BOP of the corresponding infinite-buffer-based system. To derive an analytical model, the BOP is estimated based on the M/M/1 queueing model, which has been widely applied in wireless networks [34]–[37]. The buffer at the RLC is modeled as the M/M/1 queue, where λ RLC SDUs/slot and μ RLC SDUs/slot denote the Poissonian arrival rate and the AQ2 departure rate (or service rate), respectively. Let us now define the large deviation theory in [19] for estimating the probability of $P(Q_{n+N} > B_h|Q_n)$, where Q_n is the buffer length during the current time slot n. During the kth slot, the evolution of buffer length can be represented by

$$I_k = A_k - \min\left(Q_{(k-1)}, D_k\right). \tag{6}$$

Then, we have $Q_{n+N} = Q_n + \Sigma_{k=n+1}^{n+N} I_k$. Furthermore, $P(Q_{n+N} > B_h|Q_n)$ can be rewritten as

$$\begin{aligned} P(Q_{n+N} > B_h|Q_n) &= P\left(Q_n + \Sigma_{k=n+1}^{n+N} I_k > B_h\right) \\ &= P\left(\frac{1}{N}\Sigma_{k=n+1}^{n+N} I_k > a\right) \end{aligned} \tag{7}$$

where we have $a = (B_h - Q_n)/N$. Next, we apply the LDP to the sequence $I_{n+1}, I_{n+2}, \ldots$ to derive $P((1/n)\Sigma_{k=1}^{n} I_k > a)$. The LDP can be briefly described as follows [38].

Definition: A sequence $X_1, X_2, \ldots$ obeys the LDP associated with rate function $I(\cdot)$ if the following conditions apply.

1) For any closed set F, we have

$$\limsup_{N\to\infty} \frac{1}{N} \log P\left(\frac{1}{N}\sum_{i=1}^{N} X_i \in F\right) \leq -\inf_{a\in F} I(a). \tag{8}$$

2) For any open set G, we have

$$\liminf_{N\to\infty} \frac{1}{N} \log P\left(\frac{1}{N}\sum_{i=1}^{N} X_i \in G\right) \geq -\inf_{a\in G} I(a). \tag{9}$$

Obviously, $I_k \in \mathbb{R}\,(k \in [n+1, n+N])$ represents i.i.d. random variables with a mean of $E[I_k] = \lambda - \mu$ and moment-generating function (MGF) of $M(\theta) = E[e^{\theta I_k}]$ that is finite in a neighborhood of 0. According to *Cramér's theorem* [38], $I_{n+1}, I_{n+2}, \ldots$ obeys the LDP associated with the rate function $I(a) = \sup_{\theta>0}(\theta a - \log M(\theta))$, which implies that, for any $a > \lambda - \mu$, we have

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\frac{1}{N}\sum_{i=n+1}^{n+N} I_i > a\right) = -I(a). \tag{10}$$

It should be noted that (10) is logarithmically asymptotic. Hence, when N is large, the BOP can be approximated as

$$P\left(\frac{1}{N}\sum_{i=n+1}^{n+N} I_i > a\right) \approx \exp\left[-NI(a)\right]. \tag{11}$$

Since A_k obeys Poisson's distribution, we have

$$E[e^{\theta A_k}] = \sum_{\eta=0}^{\infty} e^{\theta\eta} \frac{\lambda^\eta}{\eta!} e^{-\lambda} = \exp(\lambda e^\theta - \lambda). \tag{12}$$

Similarly, we may arrive at $E[e^{-\theta D_k}] = \exp(\mu e^{-\theta} - \mu)$. Then, we may derive the MGF $M(\theta)$ as follows:

$$\begin{aligned} M(\theta) &= E[e^{\theta I_k}] \\ &= E[e^{\theta A_k}]E[e^{-\theta D_k}] \\ &= \exp(\lambda e^\theta + \mu e^{-\theta} - \lambda - \mu). \end{aligned} \tag{13}$$

Hence, the corresponding rate function $I(a)$ can be rewritten as

$$\begin{aligned} I(a) &= \sup_{\theta>0}(\theta a - \lambda e^\theta - \mu e^{-\theta} + \lambda + \mu) \\ &= (\theta a - \lambda e^\theta - \mu e^{-\theta} + \lambda + \mu)\big|_{\theta = \log\frac{a+\sqrt{a^2+4\lambda\mu}}{2\lambda}} \\ &= a \log \frac{a + \sqrt{a^2 + 4\lambda\mu}}{2\lambda} - \sqrt{a^2 + 4\lambda\mu} + \lambda + \mu. \end{aligned} \tag{14}$$

If λ and μ are given for any encoding mode $m \in \mathcal{M}$, the BOP $P(Q_{n+N} > B_h|Q_n)$ can be estimated by using (11) and (14). The problem [see (4) and (5)] can be then solved by calculating the BOPs corresponding to all encoding modes in $\mathcal{M}$. However, in practical situations, both λ and μ depend on the network's condition and on the source encoding mode, and they are not based on prior knowledge for each encoding mode.

An attractive technique is applying an iterative policy to solve the problem [see (4) and (5)]. Specifically, at the current mode update period, we can calculate the BOP corresponding to the current encoding mode by the online estimation of the arrival rate λ and the departure rate μ. Although λ and μ can be estimated by counting the number of RLC SDU arrivals and departures over a time period of T, they may change over time due to time-varying network conditions and dynamic encoding mode switching. The dynamic nature of λ and μ may be modeled using an autoregressive integrated moving average (ARIMA) process [39] as follows:

$$\left(1 - \sum_{i=1}^{p} a_i D^i\right)(1-D)^{d_1}\lambda(t) = \left(1 + \sum_{i=1}^{q} b_i D^i\right)\varepsilon(t)$$

$$\left(1 - \sum_{i=1}^{p} c_i D^i\right)(1-D)^{d_2}\mu(t) = \left(1 + \sum_{i=1}^{q} d_i D^i\right)\mu(t)$$

where D is the unit time-delay operator, and $\varepsilon(t)$ is the estimation error. Therefore, the task of estimating λ and μ becomes the task of estimating parameters p, q, d_1, d_2, a_i, b_i, c_i, and d_i for the ARIMA model [39]. Once λ and μ have been estimated, the corresponding BOP can be directly calculated by using (11) and (14). If the BOP estimate fails to satisfy the constraint (5), the source encoding rate should be decreased, thus reducing the BOP. Otherwise, if the BOP is much lower than p_{qos}, the video source bit rate may be increased, thus improving the video quality. Applying this iterative encoding mode update, we can iteratively solve the problem [see (4) and (5)] to find a new optimal encoding mode.

It should be noted that the aforementioned method relies on the classic M/M/1 queuing model, albeit the Poissonian assumption may be not the most realistic packet arrival and departure model. Moreover, our model of the BOP requires estimating both λ and μ. In the next section, we propose an online measurement-based RC, which no longer relies on the M/M/1 queuing model and on the estimation of the RLC SDU arrival rate λ and the departure rate μ.

IV. Online Encoding Mode Switching-Based Rate Control for Video Streaming

Let the index of the current time slot be n and the current buffer length be Q_n^m, whereas $m = M_i$ denotes the current video encoding mode. We estimate the BOP at the $(n+N)$th $(N > 0)$ slot under the assumption of maintaining the current encoding mode, where N is referred to as the prediction interval. Let $P_o^{n+N}(m)$ denote the overflow probability at slot $(n+N)$, which is defined as

$$P_o^{n+N}(m) = P\left(Q_{n+N}^m > B_h\right). \tag{15}$$

Subsequently, we derive a BOP estimation model based on the aforementioned LDP [19], and then propose an online adaptive source RC for video streaming over VBR channels.

A. BOP Estimation

When transmitting D_k RLC SDUs containing video data in slot k, the increased buffer length may be expressed as

$$I_k = A_k - \min\left(Q_{(k-1)}, D_k\right). \tag{16}$$

Since we have $0 \le A_k \le A$ and $0 \le D_k \le D$, the set of values for I_k is $\{-D, \ldots, 0, \ldots, A\}$. Let us introduce the variable $\pi_i = P(I_k = i)$ to denote the probability of having a buffer length increase of $I_k = i$. For a given wired network condition, A_k is determined by the video encoding mode since the encoded video frame size is affected by the encoding mode, and D_k is related to the channel bit rate. Their difference I_k shows the instantaneous mismatch between the video bit rate and the channel bit rate. Due to the time-variant RLC SDU arrivals and the fluctuating channel bit rate, the sequence $I_i(i = 1, 2, \ldots)$ may frequently alternate between negative and positive values.

The total increase in the buffer length during the interval spanning from slot n to slot $(n+N)$ is given by

$$I^{n+N} = \Sigma_{i=1}^{N} I_{n+i}. \tag{17}$$

Then, the buffer length Q_{n+N}^m at the end of the slot $(n+N)$ may be expressed as

$$Q_{n+N}^m = Q_n^m + I^{n+N}. \tag{18}$$

According to (6), the BOP at slot $(n+N)$ may be rewritten as

$$P_o^{n+N}(m) = P\left(Q_n^m + I^{n+N} > B_h\right). \tag{19}$$

Let us now define the expected value of the average buffer length increase in each of the future N slots as

$$m_o = E\left[\frac{\Sigma_{i=1}^{N} I_{n+i}}{N}\right] \tag{20}$$

where $E[\cdot]$ denotes the expectation operator.Furthermore, while keeping the current encoding mode fixed as $m = M_i$, the tolerable average buffer length increase during each slot is defined as

$$a_o = \frac{B_h - Q_n^m}{N}. \tag{21}$$

Having $m_o \ge a_o$ implies that there would be a high BOP after N time slots.

Let us now rewrite (19) as

$$\begin{aligned} P_o^{n+N}(m) &= P\left(Q_n^m + I^{n+N} > B_h\right) \\ &= P\left(I^{n+N}/N > \left(B_h - Q_n^m\right)/N\right) \\ &= P\left(\frac{\Sigma_{i=1}^{N} I_{n+i}}{N} > a_o\right). \end{aligned} \tag{22}$$

The term $\Sigma_{i=1}^{N} I_{n+i}/N$ in (22) represents the average buffer length change during a slot, which is jointly determined by the video encoded bit rate controlled by the encoding mode and the channel bit rate, whereas a_o is the tolerable average increase in the buffer length for each of the N future slots, as determined by the current buffer length. Therefore, the probability $P_o^{n+N}(m)$ in (22) may be viewed as an estimate of the remaining storage capacity in the buffer, which may be used to smoothen the fluctuation of the channel's affordable throughput in the future. If the probability $P_o^{n+N}(m)$ exceeds a predefined threshold value p_{qos}, we can decrease the encoding mode index to reduce the video bit rate, thus decreasing both the BOP and the buffer-induced delay.

Since $D_k(k = 0, 1, \ldots)$ are i.i.d. variables, $I_k(k = 1, 2, \ldots)$ are also i.i.d. random variables, which have a finite MGF of $M(\theta) = Ee^{\theta I_k}$ in the vicinity of 0. Then, provided that $a_o > m_o$ is satisfied, according to (11), for large N, we have

$$P_o^{n+N}(m) \approx \exp\left[-NL(a_o)\right] \tag{23}$$

where

$$L(a_o) = \sup_{\theta>0} \{a_o\theta - \log M(\theta)\} \tag{24}$$

$$\log M(\theta) = \log \left\{ \sum_{i=1-s_D}^{1} \pi_i \exp[i\theta] \right\}. \tag{25}$$

To calculate the approximate BOP of (23), the knowledge of a_o, m_o, and π_i is required. It is straightforward to calculate a_o according to (21). However, there is no prior knowledge about the histogram of I_k, and hence, we cannot derive analytical expressions for m_o and π_i. Therefore, we have to rely on their buffered history for estimating the current values of these parameters using the classic sliding-window-based method.

The observed buffer length increase/decrease sequence constitutes the input of $\{I_1, I_2, I_3, \ldots\}$. The sliding window takes into account the N_s most recent I_k values of $W_n = [I_n, I_{n-1}, \ldots, I_{n-N_s+1}]$.

For parameter m_o of (20), we use the sample mean as its estimate, i.e., we have

$$\hat{m}_o = \frac{\sum_{i=n-N_s+1}^{n} I_i}{N_s}. \tag{26}$$

Let $N_i (i \in \{-D, \ldots, 0, \ldots, A\})$ denote the number of events when $I_k = i$ appears in the sliding window, which is given by

$$N_i = \sum_{k=n-N_s+1}^{n} 1(I_k = i) \tag{27}$$

where $1(\cdot)$ is an indicator function. Then, the relative frequency of encountering $I_k = i$ may be estimated as

$$\hat{\pi}_i(n) = \frac{N_i}{N_s} \tag{28}$$

which can be applied in (25) to estimate the BOP.

B. Online Source RC Algorithm

The RC strategy advocated will be discussed in the context of three scenarios according to both the current buffer length Q_n and the average buffer length increase per slot $\hat{m}_o$ as follows.

1) $Q_n^m \geq B_h$: This implies that the current buffer length is above the threshold, which will impose an undesirable delay of the video frame. Therefore, we should reduce the video bit rate by decreasing the encoding mode index as

$$m = M_{\max\{i-1,1\}}. \tag{29}$$

2) $Q_n^m < B_h$ and $\hat{m}_o \geq a_o$: In this scenario, although the buffer length is under the threshold B_h, its average increase per slot $\hat{m}_o$ exceeds the tolerable average increase a_o of the buffer length per slot during the forthcoming N slots. This implies that, at the current buffer length increase rate, the buffer length will become higher than B_h after N slots. Therefore, in this case, the current video bit rate should be decreased to reduce the BOP. Then, we can adjust the encoding mode according to (29).

3) $Q_n^m < B_h$ and $\hat{m}_o < a_o$: In this case, the current buffer length is under the threshold B_h, and the average buffer length increase per slot $\hat{m}_o$ is within the range defined by the capacity a_o. Nevertheless, this does not imply that no buffer overflow will occur in the forthcoming N slots because $\hat{m}_o$ is the average buffer length increase per slot, which cannot directly characterize the buffer length increase in a certain time slot. Hence, there is still a chance of buffer overflow. Fortunately, since we have $a_o > \hat{m}_o$, buffer overflows remain a rare event. According to the large deviation-based probability estimation model of (23), for a sufficiently large N, the BOP may be approximated as

$$\hat{P}_o^{n+N}(m) = \exp\left[-NL(a_o)\right]. \tag{30}$$

An online measurement-based estimation method was presented in the previous section to calculate the BOP. Owing to the exponential decay of the estimated BOP probability with N, we can set N to a moderate value for the sake of acquiring an accurate BOP estimation instead of requiring a large N.

Our proposed RC algorithm aims to adjust the encoding mode to satisfy the BOP QoS requirement, i.e., p_{qos}. If we have $\hat{P}_o^{n+N}(m) \geq p_{\text{qos}}$, this implies that the current encoding mode index is too high to keep the BOP below p_{qos}. Therefore, we should decrease the encoding mode index to reduce the video bit rate, thus reducing the BOP. The future encoding mode index is adjusted as $m = M_{\max\{i-1,1\}}$.

By contrast, if we have $\hat{P}_o^{n+N}(m) < p_{\text{qos}}$, the currently affordable bit rate of the channel may be able to support a higher video bit rate. Hence, we should increase the encoding mode index to provide an improved video quality for the sake of fully exploiting the attainable bit rate of the channel. To achieve this, we define a threshold $p_T (< p_{\text{qos}})$ for the BOP. If $\hat{P}_o^{n+N}(m) < p_T$ is encountered consecutively $K > 0$ times, we will increase the encoding mode index according to

$$m = M_{\min\{i+1,L\}}. \tag{31}$$

The reason for requiring K consecutive threshold violation occurrences to trigger an encoding mode index adjustment is that this prevents frequent adjustments of the encoding mode, which would result in perceivable video quality fluctuations.

The RC regime is summarized in ***Algorithm 1***. Although the estimated average channel throughput was directly fed back to the BTS to control the source rate in [18], this technique does not characterize the burst-by-burst adaptive channel throughput on a sufficiently fine timescale, which, hence, fails to guarantee a low delay for the video packets. By contrast, ***Algorithm 1*** relies on the LDP of [19] to estimate the BOP, when the buffer overflow is a rare event, and applies the BOP constraint to trigger the source RC, thus achieving a low delay for video packets.

Algorithm 1 Online measurement-based adaptive rate control algorithm for streaming video over VBR channels.

Current encoding mode $M = M_i$
if $Q_n^M \geq B_h$ **then**
  $i \leftarrow \max(i-1, 1)$
else
  Calculate $\hat{m}_o$ and $\hat{a}_o$
  if $\hat{m}_o \geq \hat{a}_o$ **then**
    $i \leftarrow \max(i-1, 1)$
  else
    Calculate $\hat{P}_o^{n+N}(M)$
    if $\hat{P}_o^{n+N}(M) \geq p_{\text{qos}}$ **then**
      $i \leftarrow \max(i-1, 1)$
    else
      if $\hat{P}_o^{n+N}(M) < p_T$ consecutively happens K times **then**
        $i \leftarrow \min(i+1, L)$
      end if
    end if
  end if
end if

C. Discussion

Previously, the proposed solution was discussed in the context of a single user requesting a single video stream. However, it may be also extended to the scenario where multiple users having a different link quality desire the same video. A simple solution is for the video server to create a dedicated encoder instance for each encoding mode, where multiple video streams are generated by the encoders with the aid of different encoding modes. Then, based on the feedback triggered by the proposed method from the BTS, the video server may select an appropriate video stream for each user associated with a different link quality.

V. Performance Evaluation

Here, we characterize the performance of our online measurement-based adaptive RC (MBARC) algorithm. We first describe our simulation setup, including the network model and the video sequences employed. Then, the metrics used for performance evaluation are described. Finally, our simulation results are presented and analyzed. In the simulations, we also implemented the heuristic online adaptive RC (HOARC) algorithm in [18] and an offline method to provide pertinent performance comparisons to cutting-edge benchmarkers. The offline method simply relied on all the encoding modes and selected the mode having the best performance as the performance benchmark.

A. Simulation Setup

We consider an HSPA network [1] relying both on AMC and HARQ. We assume a UE (UE in HSPA parlance) belonging to category 10. According to the HSPA specifications [2], the CQI value ranges from 0 to 30, and the corresponding TB sizes are $0, 137, \ldots, 25558$ bits, respectively. In HSPA, there are 15 time-division multiplex slots per 10-ms frame. Three $10/15 = 2/3$ ms slots form a so-called transmission time interval (TTI) of 2-ms duration, where only one user is allowed to transmit with the aid of multiple spreading codes per TTI. Therefore, the time-varying number of users may result in a time-varying number of TTIs being assigned to each user, which, in turn, leads to a time-varying throughput for each user. Therefore, we simulated a multiuser scenario, where the maximum number of concurrently communicating users was set to $U = 8$, and the new user arrival process follows a Markov process with an arrival rate of $\lambda = 10^{-4}$ per TTI. The service rate was assumed to be $\upsilon = 1.5 \times 10^{-5}$ per TTI. A round robin scheme was applied to schedule the transmissions of the users. Then, our proposed strategy is applied for one of the users to implement its online source RC. Consider an i.i.d. Rayleigh channel for the target user, where the received SNR s is an exponentially distributed random variable described by the probability density function of $f(s) = (1/\gamma)e^{-(s/\gamma)}$ having an average of γ. We applied the SNR (in decibels)-to-CQI mapping in [40], i.e.,

$$\text{CQI} = \lfloor \text{SNR} + 4.5 \rfloor. \tag{32}$$

According to the specifications, the CQI reporting cycle is defined as 1, 2, 4, 5, 10, 20, 40, and 80 TTIs. In the simulations, we set the CQI reporting cycle to four TTIs.

TABLE I
Properties of the Video Sequences

Encoder	H.264 Full
Resolution	CIF(352×288): Silence of the Lambs NBC 12 News D−1(704×576): The Shawshank Redemption
Bit Rate	Variable Bit Rate (VBR)
Number of Frames	45,000
GoP Size	16
Frames Rate	30fps
Video Duration	25 minutes
No. B frames between I/P frames	1

B. Video Sequences Used for Performance Evaluations

Three different video sequences are used in our simulations, namely, the "Silence of the Lambs" clip, the "NBC 12 News" clip [17], and the "Shawshank Redemption" clip, scanned at 30 frames/s and encoded by the H.264 codec. The two former clips have a Common Intermediate Format (CIF) resolution, whereas the last clip has a D-1 resolution. The duration of the video sequence is 25 min, corresponding to 45 000 video frames, where a GOP is constituted by 16 frames. The properties of these video sequences are listed in Table I. Here, the encoder mode is denoted by (m_1, m_2, m_3), where $m_1, m_2,$ and m_3 represent the quantization scales for the I, P, and B frames, respectively. The operational modes are listed as follows: $M1 = (10, 10, 12)$, $M2 = (16, 16, 18)$, $M3 = (22, 22, 24)$, $M4 = (24, 24, 26)$, $M5 = (28, 28, 30)$, $M6 = (34, 34, 36)$, $M7 =$

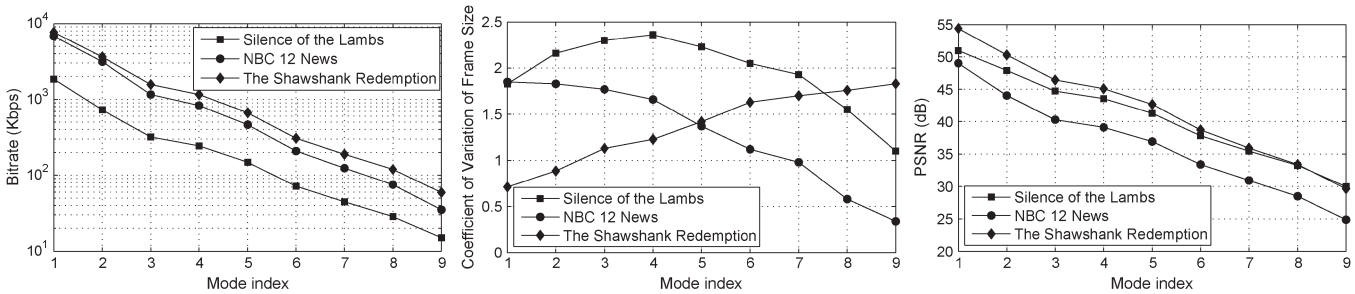


Fig. 6. Statistics of the video sequences. (a) Mean bit rate. (b) Coefficient of variation of frame size. (c) Mean PSNR.

$(38, 38, 40)$, $M8 = (42, 42, 44)$, and $M9 = (48, 48, 50)$. Fig. 6 characterizes the bit rate, the frame-size variation coefficient, and the peak SNR (PSNR) statistics of the video sequences in the different encoding modes, with the frame-size variation coefficient being defined in [41] as follows:

$$CoV_X^q = \frac{1}{\bar{X}_N^q}\sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1}\left(X_n^q - \bar{X}_N^q\right)^2} \tag{33}$$

where X_n^q denotes the frame size of video frame n encoded with the QP q, and $\bar{X}_N^q$ is the average frame size of the N video frames. In the first two simulations, the former two CIF clips were used for basic performance investigations, whereas in the rest of the simulations, the D-1 clip was used to characterize the achievable performance in the case of higher resolution videos.

C. Performance Metrics

In the simulations, each RLC SDU carried the bits of video frames. To satisfy the tolerable delay constraint of a specific video application, the RLC SDUs are dropped from the buffer when their delay exceeded the threshold of l milliseconds. To evaluate the quality of a received video sequence, an objective video quality evaluation model similar to [43] is adopted, which is defined by the rate of dropped video frames (DVFRs) that are undecodable at the MS, i.e., by

$$\text{DVFR} = 1 - \frac{N_{\text{dec}}}{N_{\text{total}-I} + N_{\text{total}-P} + N_{\text{total}-B}} \tag{34}$$

where N_{dec} is the total number of decodable frames, including all three types of frames. The decoding dependence values between different types of frames are also considered in counting the decodable frames. A lower DVFR value implies that a better quality is perceived by the recipient. Since the proposed adaptation method uses different bit rates, when configured for adapting the throughput and the quality, we also apply the average PSNR as a further performance metric.

All the results were averaged over 100 independent simulation runs. To characterize the performance improvement of our MBARC, we also conducted independent simulations for each encoder mode, and the best results were selected as performance benchmarks.

D. Simulation Results

1) Performance for Different Delay Thresholds: In the first experiment, the parameters of the proposed online MBARC algorithm were set as follows: length of the sliding window $N_s = 80$, prediction interval $N = 80$, buffer length threshold $B_h = 16$, $p_{\text{qos}} = 10^{-5}$, $p_T = 10^{-10}$, and $K = 32$. The initial mode index of the encoder is M_9, which has the lowest video bit rate. The MBARC was activated every eight video frame intervals. Its performance was investigated for the delay thresholds of {60, 100, 140, 180, and 220 ms}, which cover the delay requirements of various video applications. For example, the delay limit of 60 ms is applicable to lip-synchronized real-time interactive video conferencing applications, whereas the delay limit of 100 or 140 ms is applicable to wireless video surveillance, and finally, 180 and 220 ms are for digital television broadcast or video-on-demand services. Fig. 7(a) and (b) shows the DVFR and the average PSNR of Silence of the Lambs and NBC 12 News for different delay thresholds at the average channel SNR of $\gamma = 20$ dB. It can be observed in Fig. 7 that, as the encoder mode index increases, the DVFR decreases owing to the reduced source rate. For Silence of the Lambs, the DVFR of the operational mode $M6$ is similar to the DVFR of MBARC, but our MBARC improves the average PSNR by about 3 dB. For NBC 12 News, our MBARC and $M7$ have a similar DVFR, but MBARC improves the average PSNR by about 1 dB. This implies that the proposed MBARC is capable of adaptively adjusting the encoder mode when the source rate is temporarily higher than the affordable channel rate.

2) Performance for Different Channel SNRs: To investigate the proposed MBARC's source rate adaptation capability for different channel qualities, we conducted experiments at different average channel SNRs, namely, at $\gamma = \{12, 16, 20, 24,$ and 28 dB$\}$. The buffer length threshold was set to 28, whereas the remaining MBARC parameters were the same as in the first experiment. We set the maximum delay, which triggers dropping of the frames in the buffer to 200 ms. For each average channel SNR considered, we use an offline procedure to select the best encoding mode, which maximizes the average PSNR, while maintaining a DVFR similar to that of MBARC. The simulation results shown for the MBARC, HOARC, and offline mode selection method were plotted in Fig. 8(a) and (b). The results recorded in Fig. 8(b) for different average channel SNRs using the offline method were marked with (*Mode*). Fig. 8(b) demonstrates the average PSNR

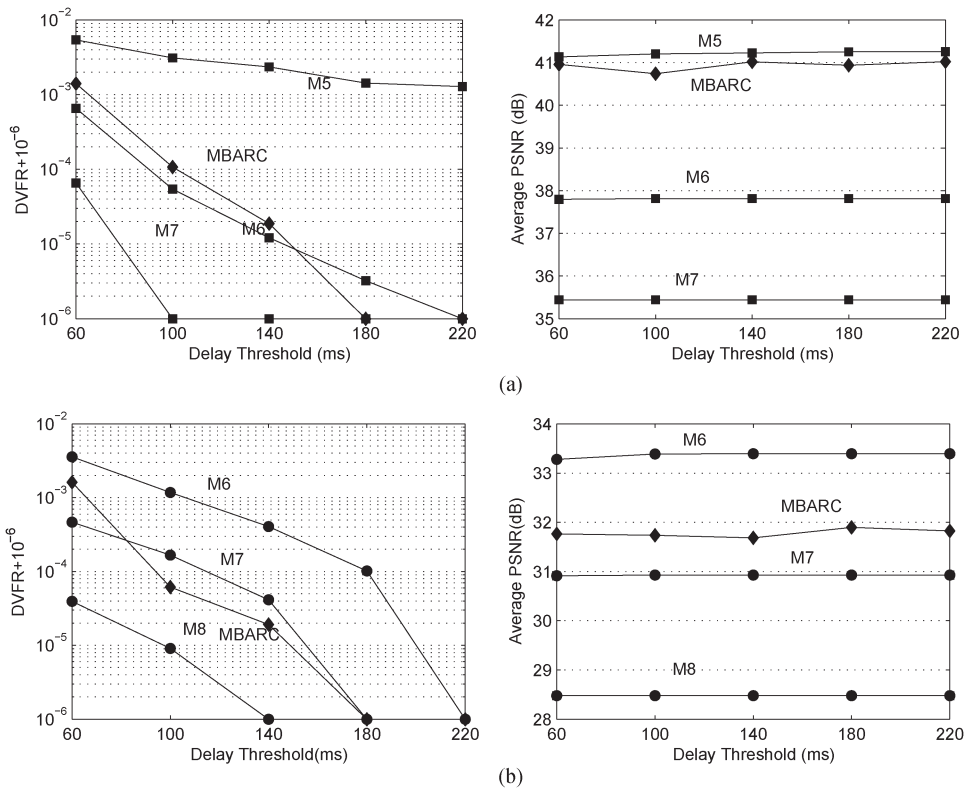


Fig. 7. (a) DVFR and Average PSNR of Silence of the Lambs for different delay thresholds with average channel SNR of 20 dB. (b) DVFR and Average PSNR of NBC 12 News for different delay thresholds with average channel SNR of 20 dB. (a) Silence of the Lambs. (b) NBC 12 News.

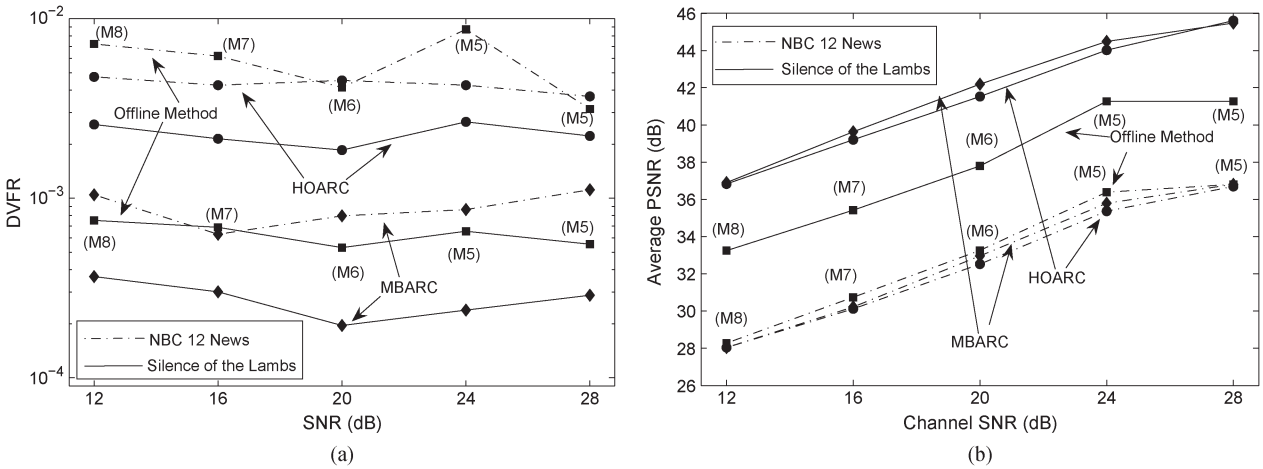


Fig. 8. (a) DVFR for different average channel SNRs of Rayleigh channel. (b) PSNR for different average channel SNRs of Rayleigh channel. The note in () represents the corresponding encoder mode for the offline method. (a) DVFR. (b) Average PSNR.

improvement of MBARC over the offline method, whereas its DVFR remains lower than that of the offline method. This demonstrates that MBARC is capable of adaptively accommodating different channel qualities by adjusting the encoding mode without any prior knowledge of the channel quality. Although Fig. 8(b) also shows that the HOARC technique in [18] is capable of adapting to time-variant channel conditions and approaching the PSNR performance of the proposed MBARC, it suffers from a much higher DVFR than that of the MBARC technique, as indicated in Fig. 8(a). The basic reason for this is elaborated as follows. The linear AR(1) of the HOARC technique may not be appropriate for estimating the average channel throughput since HSPA applies near-instantaneously adaptive burst-by-burst transmissions. Furthermore, the *average* channel throughput cannot fully characterize the *specific* near-instantaneously adaptive channel throughput at a certain transmission time instant, which, in turn, increases the delay of a specific packet and results in an increased probability of delay constraint violation. By contrast, the proposed MBARC technique is based on a BOP constraint, which may guarantee a

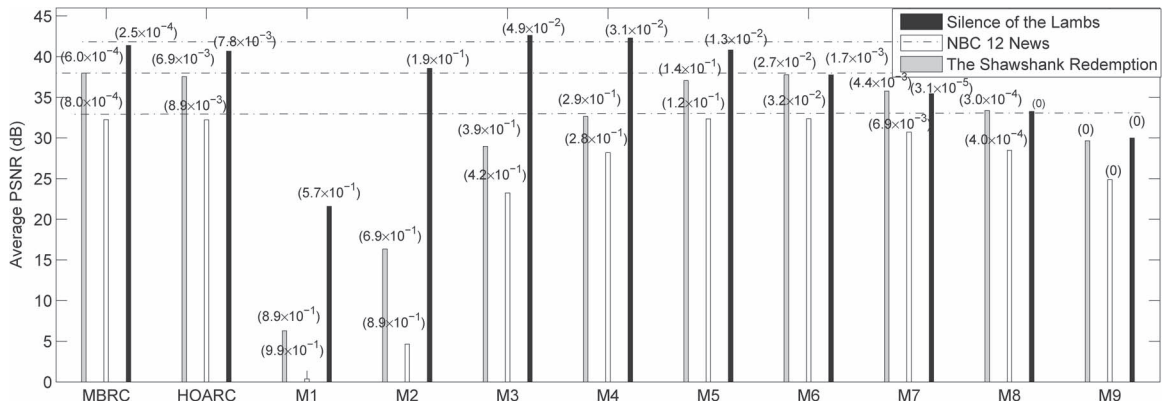


Fig. 9. Average PSNR for dynamic average channel SNRs for Rayleigh channel. The number in () represents the corresponding DVFR.

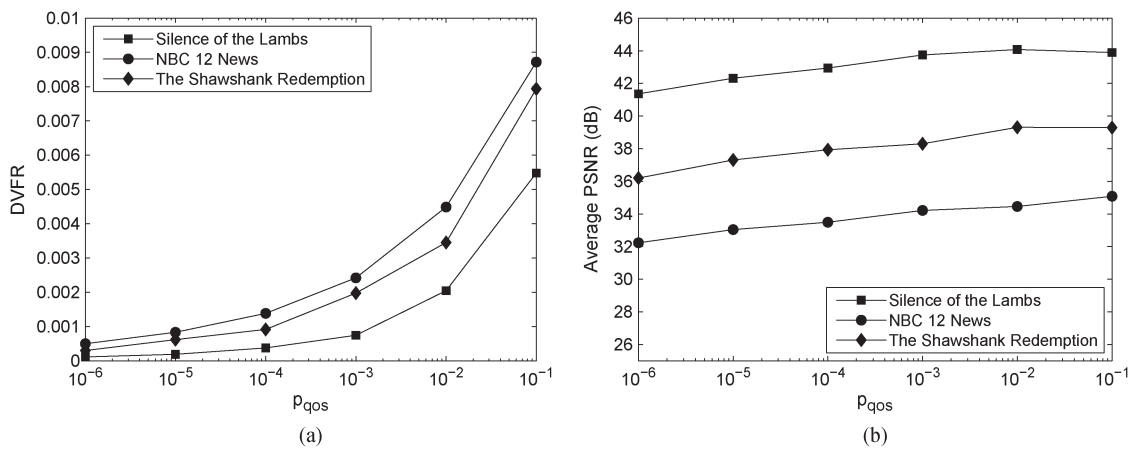


Fig. 10. (a) DVFR for different overflow probability thresholds $p_{\rm qos}$. (b) PSNR for different overflow probability thresholds $p_{\rm qos}$. (a) Silence of the Lambs. (b) NBC 12 News.

low delay for the video packets. The application of the LDP [19] assists MBARC in achieving an accurate estimation of BOP and, thus, improves the DVFR performance.

3) Performance for Dynamic Channel Quality: The previous investigations were conducted in a static environment by fixing the average channel SNR. Next, we investigate the performance of the MBARC in time-varying scenarios. The initial average channel SNR was set to 24 dB. At the instant of the 18 000th video frame interval, the average channel SNR was suddenly changed to 20 dB. Then, at the instant of the 36 000th frame interval, it was further reduced to 16 dB. All the MBARC parameters remained the same as in the previous experiment. Fig. 9 shows the corresponding simulation results, where each bar is marked with "(DVFR)" to characterize the corresponding DVFR. It is shown that the DVFR of our MBARC is lower than that of $M6$ for Silence of the Lambs and that it improves the PSNR by more than 2 dB in comparison with that of $M6$. For NBC 12 News, although $M8$ has a similar DVFR to that of our MBARC, its average PSNR is 2 dB lower than that of the proposed MBARC. This benefit is the result of the MBARC's adaptability, which adjusts the encoder mode online for VBR encoding to accommodate diverse dynamically fluctuating propagation environments. By contrast, for a fixed encoder mode, having a low channel SNR may inflict frame dropping events, resulting in low video quality, whereas at high SNRs, it fails to promptly decrease the video rate to improve the video quality in a timely manner. Similarly, MBARC is capable of adaptation and achieves an improved performance for the Shawshank Redemption clip having a higher resolution. Observe in Fig. 9 that the HOARC technique [18] approaches the PSNR performance of the proposed MBARC, but it exhibits a significantly higher DVFR than MBARC. This result also illustrates the benefit of the explicit BOP constraint in the problem formulation (4) and (5) and that of applying the LDP for accurately estimating BOP.

4) Performance Sensitivity of the Proposed Algorithm: The performance of the proposed method relies on parameters N, N_s, and K, as shown in **Algorithm 1**. Hence, this section investigates their effect on the performance of the proposed method. In the related experiments, one of these three parameters was set to different values, whereas the other parameters of our MBARC were the same as those in 2). The average SNR value of our Rayleigh channel model was fixed to 20 dB. The remaining parameters of the channel model were set as in 2).

Fig. 10 shows our simulation results of different BOP thresholds $p_{\rm qos}$ for the three clips. We may observe in Fig. 10 that a reduced overflow probability threshold provides a reduced DVFR. This is because a reduced probability threshold is capable of activating timely adjustments of the encoder mode to reduce the BOP. On the other hand, we can also observe in

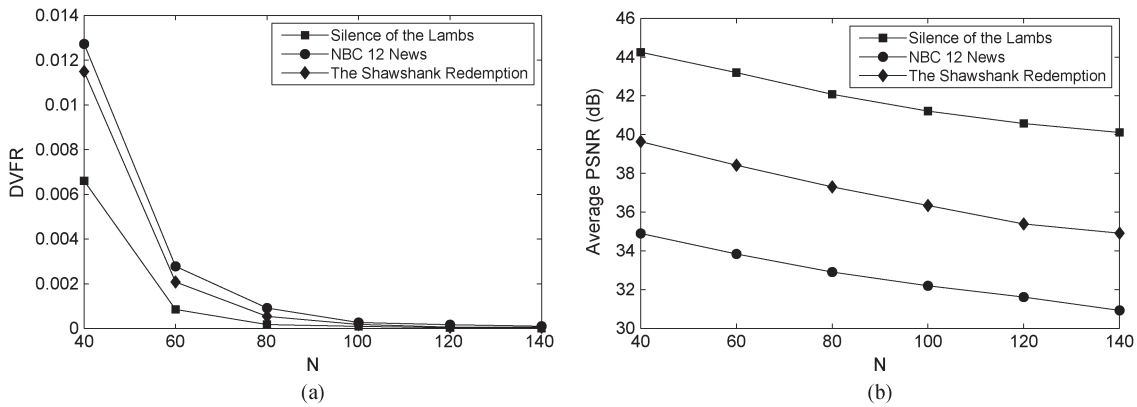


Fig. 11. (a) DVFR for different prediction intervals N. (b) PSNR for different predicting intervals N. (a) DVFR. (b) PSNR.

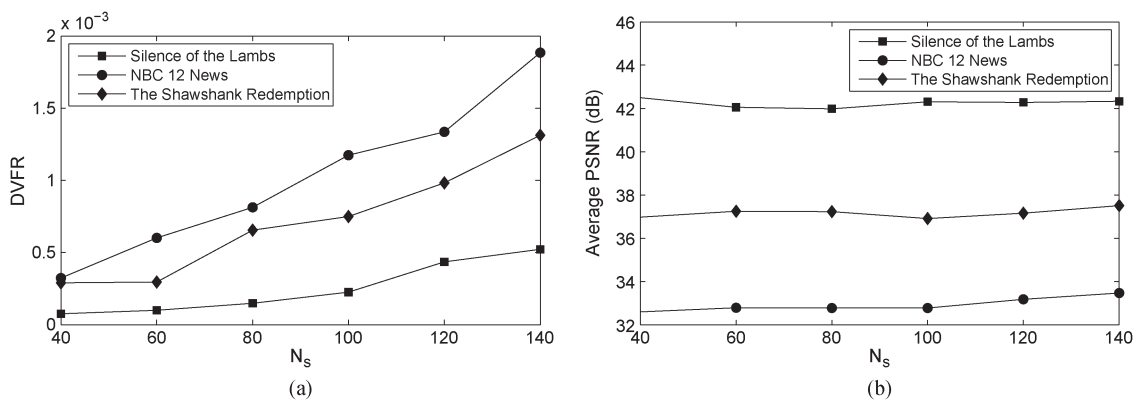


Fig. 12. (a) DVFR for different lengths of the sliding window N_s. (b) PSNR for different lengths of the sliding window N_s. (a) DVFR. (b) PSNR.

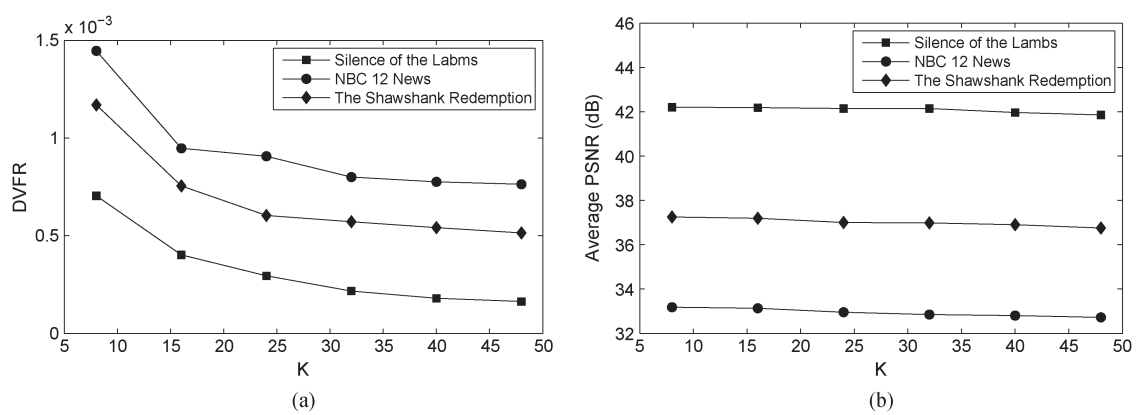


Fig. 13. (a) DVFR for different values of K. (b) PSNR for different values of K. (a) DVFR. (b) PSNR.

Fig. 10 that, as the BOP threshold decreases, the average PSNR increases. This is caused by the lower encoding mode index used to increase the video quality when the BOP threshold is higher. It should be noted that, although the DVFR is related to the BOP, their definitions are different. Hence, the value of DVFR is not the same as p_{qos}, as shown in Fig. 10(a). Fortunately, we may achieve an acceptable DVFR by setting p_{qos} to an appropriate value such as 10^{-5}.

Fig. 11 plots the simulation results for different prediction intervals N for the three clips. Fig. 11(a) shows that the DVFR is rapidly reduced upon increasing N, whereas Fig. 11(b) shows that the average PSNR is decreased as N is increased. The basic reason is elaborated as follow. According to (30), a large value of N leads to a more accurate BOP estimate, which assists the proposed method to trigger appropriate adjustments of the encoding mode and thus to reduce the DVFR. The result shows that $N \geq 80$ is appropriate for achieving an acceptable performance for the proposed method.

Fig. 12 shows the performance sensitivity of the proposed method to the sliding-window duration. It is shown in Fig. 12(a) that a larger value of N_s may result in a higher DVFR. This may be caused by the reduced sensitivity of the buffer variance when using a larger N_s for estimating $\hat{m}_o$ according to (26). Fig. 12(b) shows that N_s has only a modest effect on the PSNR.

Fig. 13 shows the sensitivity of the performance to parameter K. The proposed method triggers a video bit rate increase via adjusting the encoding mode only when the threshold violation of $\hat{P}_o^{n+N} < p_T$ is encountered consecutively K times. Therefore, having a larger K implies a more conservative adjustment policy, which leads to a lower DVFR and to a marginally reduced PSNR, which are shown in Fig. 13(a) and (b).

VI. Conclusion

An online adaptive source RC regime has been proposed for streaming videos in HSPA and LTE-like VBR downlink scenarios. Large deviation theory was invoked to derive the online measurement-based BOP model, applied at the RLC layer in the BS. The advantage of the resultant algorithm is that it exploits the online observations of buffer length and its variation for RC, requiring no prior knowledge of the channel variations and video characteristics. Our simulation results recorded in multiuser scenarios demonstrated that the proposed algorithm is capable of accommodating the channel quality variations, which may be beneficial in HSPA and LTE-style environments.

References

[1] L. Hanzo, J. Blogh, and S. Ni, *3G, HSPA and FDD Versus TDD Networking: Smart Antennas and Adaptive Modulation*. Piscataway, NJ, USA: Wiley/IEEE Press, 2008.

[2] Third-Generation Partnership Project (3GPP), "Physical layer procedures (FDD)," ETSI, Sophia Antipolis, France, 3GPP TS25.214 V6.2.0, 2004.

[3] M. Chen and A. Zakhor, "Rate control for streaming video over wireless," *IEEE Wireless Commun.*, vol. 12, no. 4, pp. 32–41, Aug. 2005.

[4] T. Chang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 246–250, Feb. 1997.

[5] Z. G. Li *et al.*, " A unified architecture for real-time video-coding systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 472–487, Jun. 2003.

[6] Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control of H.264/AVC scalable extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 116–121, Jan. 2008.

[7] "Joint Video Team Software JM 18.0." [Online]. Available: http://iphome.hhi.de/suehring/tml/

[8] Z. Li *et al.*, "Adaptive basic unit layer rate control for JVT," presented at the 7th Meeting, Pattaya II, Pattaya, Thailand, Mar. 2003, Paper JVTG012r1.

[9] L. Hanzo, P. Cherriman, and J. Streit, *Video Compression and Communications*. Hoboken, NJ, USA: Wiley, 2007.

[10] Y. Huang, S. Mao, and S. F. Midkiff, "A control-theoretic approach to rate control for streaming videos," *IEEE Trans. Multimedia*, vol. 11, no. 6, pp. 1072–1081, Oct. 2009.

[11] H. L. Lin, T. Y. Wu, and C. Y. Huang, "Cross layer adaptation with QoS guarantees for wireless scalable video streaming," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1349–1352, Sep. 2012.

[12] N. Changuel, N. Mastronarde, M. Van der Shaar, B. Sayadi, and M. Kieffer, "Adaptive scalable layer filtering process for video scheduling over wireless networks based on MAC buffer management," in *Proc. IEEE ICASSP*, May 2011, pp. 2352–2355.

[13] A. Dua, C. W. Chan, N. Bambos, and J. Apostolopoulos, "Channel, deadline, and distortion (CD^2) aware scheduling for video streams over wireless," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1001–1011, Mar. 2010.

[14] H. Zhang, Y. Zheng, M. A. Khojastepour, and S. Rangarajan, "Cross-layer optimization for streaming scalable video over fading wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 344–353, Apr. 2010.

[15] E. Piri, M. Uitto, J. Vehkapera, and T. Sutinen, "Dynamic cross-layer adaptation of scalable video in wireless networking," in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–5.

[16] R. Radhakrishnan and A. Nayak, "Cross layer design for efficient video streaming over LTE using scalable video coding," in *Proc. IEEE ICC*, Jun. 2012, pp. 6509–6513.

[17] Video Trace Library. [Online]. Available: http://trace.eas.asu.edu/

[18] C. Chen, R. W. Heath Jr., A. C. Bovik, and G. D. Veciana, "A Markov decision model for adaptive scheduling of stored scalable videos," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 23, pp. 1081–1095, Jun. 2013.

[19] A. Dembo, and O. Zeitouni, *Large Deviations Techniques and Applications*. 2nd ed. Berlin, Germany: Springer-Verlag, 1998.

[20] Apple HLS, "HTTP Live Streaming draft-pantos-http-live-streaming-08 (IETF draft)."

[21] Third-Generation Partnership Project (3GPP), "Transparent end-to-end Packet Switched Streaming Service (PSS); protocols and codecs," ETSI, Sophia-Antipolis, France, 3GPP TS 26.234, 2010.

[22] C. Hsu, and M. Hefeeda, "Partitioning of multiple fine-grained scalable video sequences concurrently streamed to heterogeneous clients," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 457–469, Apr. 2008.

[23] Z. Chen, M. Li, and Y. Tan, "Perception-aware multiple scalable video streaming over WLANs," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 675–678, Jul. 2010.

[24] Third-Generation Partnership Project (3GPP), "Transparent end-to-end Packet Switched Streaming Service (PSS); progressive download and dynamic adaptive streaming over HTTP (3G)," ETSI, Sophia-Antipolis, France, 3GPP TS 26.247, 2011.

[25] *Information Technology-Dynamic Adaptive Streaming Over HTTP (DASH)—Part 1: Media Presentation Description and Segment Formats*, ISO/IEC 23009-1:2012, 2012,

[26] G. Tian, and Y. Liu, "Towards agile and smooth video adaptation in dynamic HTTP streaming," in *Proc. ACM CoNEXT*, Dec. 2012, pp. 109–120.

[27] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE," in *Proc. ACM CoNEXT*, Dec. 2012, pp. 97–108.

[28] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," *Proc. ACM MOBICOM*, Sep. 2013, pp. 389–400.

[29] J. Little, "A proof of the queuing formula: $L = \lambda W$," *Oper. Res.*, vol. 9, no. 3, pp. 383–387, May 1961.

[30] A. Wilig, "A Short Introduction to Queueing Theory," 1999. [Online]. Available: http://www.cs.ucf.edu/lboloni/Teaching/EEL6785_Fall2010/slides/QueueingTheory.pdf AQ3

[31] F. Ishizaki, and F. Takine, "Loss probability in a finite discrete-time queue in terms of the steady state distribution of an infinite queue," *Queue Syst.*, vol. 31, no. 3/4, pp. 317–326, Jul. 1999.

[32] N. B. Shroff and M. Schwartz, "Improved loss calculations at an ATM multiplexer," *IEEE/ACM Trans. Netw.*, vol. 6, no. 4, pp. 411–421, Aug. 1998.

[33] H. S. Kim and N. B. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers," *IEEE/ACM Trans. Netw.*, vol. 9, no. 6, pp. 755–768, Dec. 2001.

[34] H. Bobarshad, M. van der Schaar, A. H. Aghvami, R. S. Dilmaghani, and M. R. Shikh-Bahaei, "Analytical modeling for delay-sensitive video over WLAN," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 401–414, Apr. 2012.

[35] S. T. Cheng, M. H. Tao, and C. Y. Wang, "Adaptive channel switching for centralized MAC protocols in multihop wireless networks," *IEEE Trans. Commun.*, vol. 58, no. 1, pp. 228–234, Jan. 2010.

[36] H. Bobarshad and M. Shikh-Bahaei, "M/M/1 queuing model for adaptive cross-layer error protection in WLANs," in *Proc. IEEE WCNC*, Nov. 2009, pp. 1–6.

[37] Y. Xu *et al.*, "Probabilistic analysis of buffer starvation in Markovian queues," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1826–1834.

[38] M. Mandjes, *Large Deviations for Gaussian Queues: Modelling Communications Networks*. Hoboken, NJ, USA: Wiley, 2007.

[39] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis Forecasting and Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.

[40] Motorola and Nokia, "Revised CQI proposal," 3GPP RAN WG1, Sophia, Antipolis, France, Tech. Rep. R1-02-0675, Apr. 2002.

[41] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, 4th Quart. 2012.

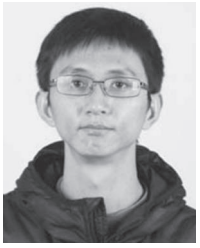
[42] Gabin *et al.*, "3GPP mobile multimedia streaming stands," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 134–138, Nov. 2010.

[43] C. H. Lin, C. H. Ke, C. K. Shieh, and N. K. Chilamkurti, "The packet loss effect on MPEG video transmission in wireless networks," in *Proc. IEEE AINA*, Apr. 2006, pp. 565–572.

Jian Yang (M'08) received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively.

From 2006 to 2008, he was a Postdoctoral Scholar with the Department of Electronic Engineering and Information Science, USTC. Since 2008, he has been an Associate Professor with the Department of Automation, USTC. His current research interests include distributed system design, modeling and optimization, multimedia over wired and wireless, and future networks.

Dr. Yang received the Lu Jia-Xi Young Talent Award from the Chinese Academy of Sciences in 2009.



Yongyi Ran received the B.S. and Ph.D. degrees in 2008 and 2014, respectively, from the University of Science and Technology of China, Hefei, China, where he is currently a Postdoctoral Researcher.

His research interests include cloud computing, service management, future networks, and stochastic optimization.

Shuangwu Chen received the B.S. degree in 2011 from the University of Science and Technology of China, Hefei, China, where he is currently working toward the Ph.D. degree with the School of Information Science and Technology.

His research interests include multimedia communications, future networks, and stochastic optimization.

Weiping Li (F'00) received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, China, in 1982 and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, USA, in 1983 and 1988, respectively, all in electrical engineering.

In 1987, he joined the Faculty of Lehigh University, Bethlehem, PA, USA, as an Assistant Professor with the Department of Electrical Engineering and Computer Science. In 1993, he was promoted to Associate Professor with Tenure. In 1998, he was promoted to Full Professor. From 1998 to 2010, he was with several high-technology companies in the Silicon Valley, where he had technical and management responsibilities. In March 2010, he returned to the USTC and is currently a Professor with the School of Information Science and Technology.

Prof. Li was elected Fellow of IEEE for contributions to image and video coding algorithms, standards, and implementations. He served as a member of Moving Picture Experts Group (MPEG) of the International Organization for Standardization (ISO) and an Editor of MPEG-4 International Standard. He served as a Founding Member of the Board of Directors of the MPEG-4 Industry Forum. His inventions on fine granularity scalable video coding and shape adaptive wavelet coding have been included in the MPEG-4 International Standard. As a Technical Adviser, he also made contributions to the Chinese Audio Video Coding Standard and its applications. He served as the Chair of several technical committees within the IEEE Circuits and Systems Society and at IEEE international conferences. He served as the Chair of the Best Student Paper Award Committee for the SPIE Visual Communications and Image Processing Conference. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He served as a Guest Editor for a special issue of the PROCEEDINGS OF THE IEEE. He has made many contributions to international standards. He received the Certificate of Appreciation from ISO and the International Electrotechnical Commission as a Project Editor in the development of International Standard in 2004, the Spira Award for Excellence in Teaching in 1992 from Lehigh University, and the first Guo Moruo Prize for Outstanding Student in 1980 from the USTC.

Lajos Hanzo (M'91–SM'92–F'04) received the Master's degree in electronics, the Ph.D. degree, and the *Doctor Honoris Causa* degree from the Technical University of Budapest, Budapest, Hungary, in 1976, 1983, and 2009 respectively, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2004.

During his career in telecommunications, he has held various research and academic posts in Hungary, Germany, and the U.K. Since 1986, he has been with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., where he holds the Chair in telecommunications. He was a Chaired Professor of Tsinghua University, Beijing, China. He is the coauthor of 20 John Wiley/IEEE Press books on mobile radio communications, totalling in excess of 10 000 pages, and has published more than 1400 research entries on IEEE Xplore. He is currently directing an academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) U.K., the European IST Program, and the Mobile Virtual Centre of Excellence, U.K. He is an enthusiastic supporter of industrial and academic liaison, and he offers a wide range of industrial courses.

Dr. Hanzo has acted as a Technical Program Committee Chair for IEEE conferences, presented keynote lectures, and has received a number of distinctions. He is the Governor of the IEEE Vehicular Technology Society and the Past Editor-in-Chief of the IEEE Press.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

AQ1 = Please check if “Star War IV” should be “Shawshank Redemption”
AQ2 = Please check if changes made in this sentence are appropriate.
AQ3 = Provided URL in Ref. [30] was not found. Please check.

END OF ALL QUERIES

Online Source Rate Control for Adaptive Video Streaming Over HSPA and LTE-Style Variable Bit Rate Downlink Channels

Jian Yang, *Member, IEEE*, Yongyi Ran, Shuangwu Chen, Weiping Li, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Online source rate control (RC) is designed for video streaming over high-speed packet access (HSPA) and Long-Term Evolution (LTE)-style variable bit rate (VBR) downlink channels. The problem is formulated as the adaptive adjustment of the operational mode of a video encoder based on the buffer overflow probability (BOP) feedback received from the radio link control (RLC) layer at the base station (BS). This allows us to maximize the attainable visual quality while keeping the transmitter BOP below a desired threshold and maintaining a video delay as low as possible. We derive an online measurement-based BOP estimation model for the RLC buffer, which is capable of operating with no prior knowledge of the channel variations and of the video characteristics. Based on this estimation model, an online adaptive RC algorithm is proposed to seamlessly adapt the bit stream to the characteristics of VBR channels. Our experiments are conducted in multiuser scenarios using VBR video encoding combined with adaptive modulation and coding (AMC) in the transceiver. The results demonstrate that the proposed source RC regime supports near-instantaneous yet smooth bit stream adaptability, which makes it useful for HSPA and LTE-style systems for the sake of accommodating unknown video traffic characteristics and dynamically fluctuating propagation conditions.**

***Index Terms*—Large deviation principle (LDP), rate control (RC), variable bit rate (VBR) channel.**

I. Introduction

ADVANCED transceiver techniques, including adaptive modulation and coding (AMC) and hybrid automatic repeat request (HARQ), are widely applied in wireless systems to combat the effects of both fading and interference [1], [2]. For instance, both the third-generation (3G) and high-speed packet access (HSPA) apply AMC and HARQ in the physical layer relying on the channel quality indicator's (CQI) feedback from the user equipment (UE) to achieve prompt link adaptation. The Long-Term Evolution (LTE) Advanced and WiMax standards also provide similar link adaptation based on AMC and short time-slot duration periods. Naturally, agile link adaptation may result in a time-varying effective throughput of the channel. Therefore, supporting near-instantaneous bit rate adaptivity is one of the most important features of video streaming applications.

Rate control (RC) aims to control the video bit rate to match the channel's achievable bit rate, given the prevalent channel quality to keep the video quality as high as possible [3]. Most existing studies of RC focus on allocating a bit rate budget to each group of pictures (GOP), frames, or macroblocks at the encoder. For instance, an efficient RC scheme was designed for MPEG-4 based on a quadratic rate–distortion (R–D) model invoked to maintain the target bit rate in [4]. In [5], the quadratic R–D model was also adopted to derive an adaptive RC for H.264 to meet the target bit rate. In [6], RC was designed for scalable H.264/Advanced Video Coding (AVC) encoding. H.264 reference software JM [7] has implemented the JVT-G012 [8] RC algorithm, which relies on a combination of the available channel bandwidth, the frame rate, the target buffer level, and the actual buffer fullness for determining the number of bits allocated for the current frame. A range of AMC-based schemes were conceived in [9] where no extra buffering was used by the RC. The design philosophy was to instruct the video encoder to produce the exact number of bits for each video frame, which was affordable for the near-instantaneous HSPA-style AMC transceiver mode that was periodically signaled back from the receiver to the transmitter. However, the aforementioned RC schemes rely on the knowledge of the target bit rate, which again requires the feedback of the expected bit rate based on the estimated channel quality. Unfortunately, the channel's bit rate fluctuations are not known *a priori* at the transmitter and the video encoder. Therefore, these RC schemes may impose substantial video quality fluctuations.

Recently, a control-theoretic approach-based RC regime has been proposed in [10] by jointly considering the encoder's RC and network congestion control. More specifically, an empirical R–D source model, a channel-induced distortion model, and their linearized models were applied to formulate a mathematically tractable system model. However, it is a challenge to formulate accurate models when streaming videos over

Manuscript received January 10, 2014; revised July 21, 2014 and November 3, 2014; accepted January 24, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61174062; by the State Key Program of the National Natural Science Foundation of China under Grant 61233003; and by the Fundamental Research Funds for the Central Universities. The work of L. Hanzo was supported by the European Research Council under an Advanced Fellow Grant. The review of this paper was coordinated by Prof. N. Arumugam.

J. Yang, Y. Ran, S. Chen, and W. Li are with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: jianyang@ustc.edu.cn; yyran@mail.ustc.edu.cn; chensw@mail.ustc.edu.cn; wpli@ustc.edu.cn).

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2015.2398515

time-varying channels. Typically, intelligent packet scheduling is used at the base station (BS) by most of the existing solutions to achieve channel-quality-dependent adaptive video streaming over wireless networks. Cross-layer-adaptive scalable video streaming has been proposed by informing the media access control (MAC) layer of the amount of payload and the index of the modulation-and-encoding scheme to improve the attainable video quality [11]. An adaptive scalable video streaming strategy relaying on controlling the MAC buffer was presented in [12]. Both the packet deadlines and the channel characteristics were considered in video packet scheduling at the BS [13]. The cross-layer design-aided transmission of scalable video streams [14] involves packet scheduling combined with a video frame dropping strategy at the BS. Most of these papers discuss the channel-dependent adaptation of the MAC layer of the BS where no signaling is fed back to instruct the video source to adjust its bit rate. Dropping packets at the BS is a waste of resources both in the backhaul router and the core network. It was shown in [15] that packet dropping at the video source is more efficient than MAC layer dropping at the BS when we have to reduce the congestion in both the wired core network and in the wireless medium to the UE. By translating the CQI values to the number of enhancement layers scheduled for transmission, the video server facilitates adaptive video streaming [16]. However, the link throughput at the BS is determined not only by the CQI value but also by the MAC layer scheduling strategy of a multiuser scenario. Moreover, the bit rate of video clips having different amounts of motion activity may be very different even if they have the same number of enhancement layers [17]. Therefore, such CQI-mapping-based video source adaptation lacks robustness against the time-variant number of users and against the heterogeneous video bit rates. In [18], although a Markov decision process is invoked for dynamic video scheduling, its offline computation requires *a priori* knowledge of the channel dynamics. Hence, a heuristic online scheduling algorithm is derived based on the average channel throughput estimated with the aid of a first-order autoregressive model [AR(1)]. In the HSPA/LTE system, using a CQI-based link adaptation mechanism may lead to a burst-by-burst transmission block adaptation, and the linear estimator based AR(1) may not be appropriate for estimating the average channel throughput.

Against this background, we proposed a novel online source RC framework, where a buffer overflow probability (BOP) estimator is employed at the radio link control (RLC) layer of the BS, and the estimated BOP is signaled back to the video source to control its bit rate. The motivation of using a BOP-based feedback is that the BOP metric characterizes the degree of matching between the video bit rate of the source and the link throughput. Since the BOP is estimated before a buffer overflow is encountered, its feedback may assist the video source in promptly controlling the bit rate. In contrast to the methods found in the open literature [9]–[14], [16], [18], the main contributions of this paper are threefold.

- To enable the video encoder to generate a video stream that may be reliably delivered over variable bit rate (VBR) downlink channels, we formulate a constrained optimization problem subject to a constraint imposed on the transmission BOP to guarantee a low delay.
- A BOP estimation model based on large deviation principles (LDPs) [19] is proposed by monitoring the buffer fullness and its variation at the BS's RLC layer. The reason for applying the LDPs is because it accurately characterizes the probability of rare events, which assists us in achieving fine adjustment of the video source rate.
- We conceive an iterative RC algorithm to approach the most beneficial encoder rate, thus circumventing the difficulty of directly solving the related constrained optimization problem. The proposed RC algorithm allows the encoder to adjust its bit rate to that of the VBR channel without any *a priori* knowledge of both the channel quality variations and of the characteristics of the video source.

The remainder of this paper is organized as follows. Section II describes the system model, including the formulation of the source RC problem of video streaming over a VBR downlink channel. In Section III, we apply the LDP in [19] to derive a BOP estimation model, whereas an online measurement-based RC is proposed in Section IV. Our numerical simulation results are presented in Section V to characterize the attainable performance of the proposed algorithm. Finally, our conclusions are offered in Section VI.

II. System Model

A. System Overview

Fig. 1 shows a typical video streaming scenario over a wireless network. The main associated protocol stacks are also shown in the lower part of the figure. Naturally, a video streaming service involves not only maintaining the wireless connection between the mobile station (MS) and the base transceiver station (BTS) but supporting the public network connection (Internet) as well. The wired network provides high capacity and stability; hence, video streaming over the wired network has become a well-established service and has many successful applications, including video conferencing, surveillance systems, and Internet Protocol (IP) television. By contrast, video streaming over wireless networks faces unique challenges due to the time-varying nature of the wireless channel and owing to the scarcity of the system resources, which makes it difficult to guarantee any specific video quality of service (QoS). Hence, the wireless transmission in the video streaming service is likely to be a bottleneck, which is the focus of this paper.

Fig. 2 shows the basic video processing in the video network abstract layer (NAL) units of the 3GPP framework. A NAL unit may be encapsulated in a Real-Time Transport Protocol (RTP) data unit, and then, it may be transmitted over User Datagram Protocol (UDP)/IP. HTTP-based streaming protocols such as Apple HTTP Live Streaming (HLS) [20] are alternative media streaming communication protocols, which are capable of traversing any firewall or proxy server that lets through standard HTTP traffic, unlike UDP-based protocols such as RTP. 3GPP standardized an adaptive HTTP streaming protocol

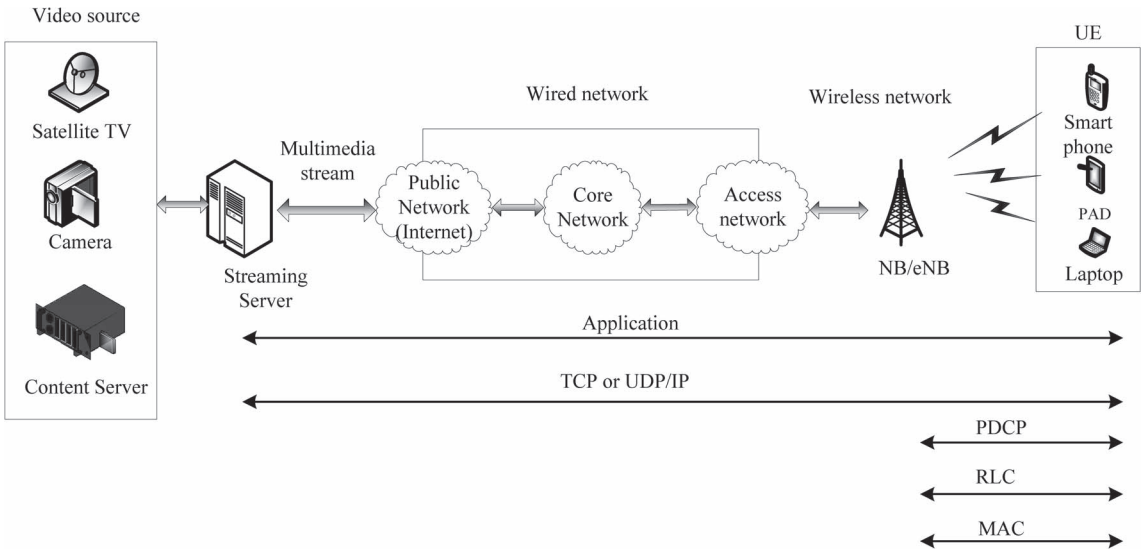


Fig. 1. Typical scenario for wireless video streaming in 3G Partnership Project (3GPP) framework.

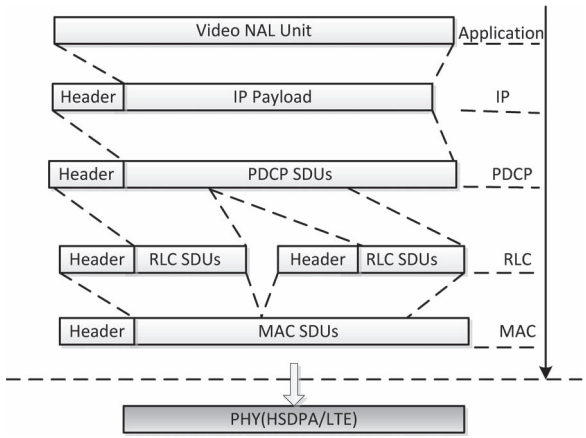


Fig. 2. Video data packetization at different layers in the 3GPP framework.

in Rel-9 [21]. Subsequently, we concentrate on the terminology used in HSPA NB/eNB, as shown in Fig. 2. The IP/UDP/RTP or IP/Transmission Control Protocol (TCP)/HTTP packet generated is encapsulated into a single Packet Data Convergence Protocol (PDCP) packet that becomes an RLC service data unit (SDU). Since a typical RLC SDU has a larger size than an RLC protocol data unit (PDU), it has to be segmented into smaller units. The length of the RLC PDU depends both on the selected bearer and on the AMC mode used. The RLC PDUs are forwarded to the MAC layer and are then encapsulated as MAC PDUs. The main functions of the PDCP layer are robust header compression and decompression, ciphering and deciphering, the transfer of data, and PDCP sequence number maintenance. The RLC layer in wireless systems is capable of operating in both an unacknowledged mode (UM) and acknowledged mode (AM), and both are capable of providing RLC PDU loss detection. However, while the UM is unidirectional and data delivery is not guaranteed, in the AM, automatic repeat request (ARQ) is applied for reliable data transmission. The main functions of the RLC layer include the transfer of upper layer PDUs; error correction relying on ARQ (only for AM); and the concatenation, segmentation, and reassembly of RLC SDUs, whereas the main functions of the MAC layer are resource scheduling and multiplexing/demultiplexing of MAC SDUs belonging to one or several logical channels into/from the relevant transport blocks (TBs). By comparing the functions of the RLC layer with those of the PDCP and MAC, the RLC is an appropriate layer for us to construct a model for characterizing the degree of matching between the network's throughput and the video source bit rate. Generally, the detection of a lost RLC PDU results in the loss of an entire PDCP packet; hence, the encapsulated IP and the NAL unit are lost. From the perspective of reacting to both the dynamics of the statistical fluctuation of the teletraffic and the variable channel conditions, agile bit rate adaptivity is one of the most important features for seamless video streaming over wireless systems. There are several ways of achieving bit rate adaptivity. For online encoding applications, the bit rate adaptivity can be achieved by controlling encoding parameters. For instance, H.264/AVC supports these features mainly by dynamically varying the quantizers but also by controlling temporal resolution. Scalable Video Coding (SVC) is another technique of implementing the bit rate adaptivity. It encodes the raw video clip into a base layer and a number of enhancement layers with different priorities. Naturally, the base layer has the highest priority since it contains the video bits with the highest importance, which can provide a minimum video quality. The enhancement layers with lower priorities may be progressively encoded to further refine the quality of the base-layer stream. This layered approach of the SVC codec allows an encoded stream to be flexibly prepared for meeting the bit rate constraint. In [22], the base-layer rate for a given video sequence is optimized to achieve the highest possible average perceived quality for heterogeneous clients, whereas in [23], a distribution regime was conceived for scalable videos, which guaranteed fairness for all end users. In Dynamic Adaptive Streaming over HTTP (DASH), the video content is partitioned into a sequence of small HTTP-based file

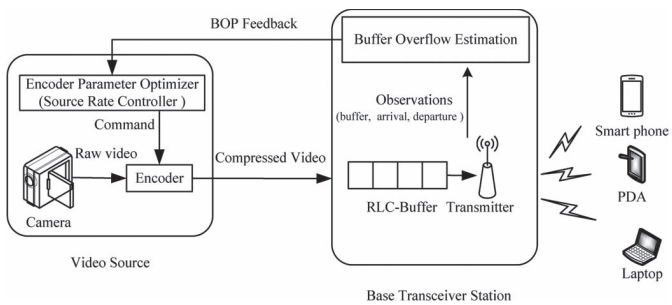


Fig. 3. Adaptive source RC framework for online video encoding.

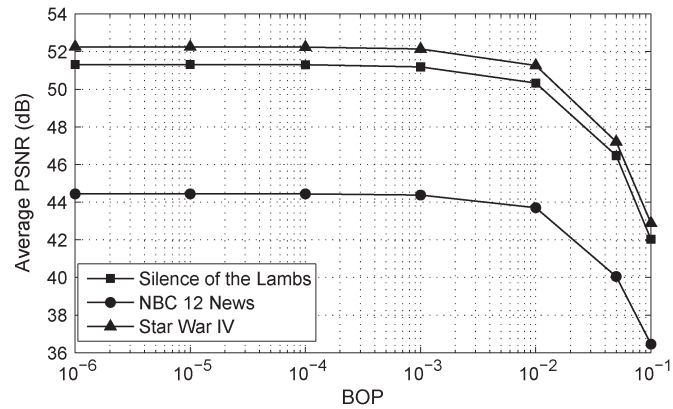


Fig. 4. Relation between BOP and received video quality by using three different video clips, namely, the Silence of the Lambs, NBC 12 News, and Star Wars IV. AQ1

segments, which are made available at a variety of different bit rates. More explicitly, segments encoded at different bit rates covering appropriately aligned short intervals of playback time are made available. Naturally, DASH also provides a flexible way of adjusting the video source bit rate. 3GPP has defined its Rel-10 version of DASH, termed as 3GP-DASH [24], which is a profile compatible with MPEG-DASH [25]. More work on HTTP-based adaptive video streaming can be found in [26]–[28].

Since the RLC buffer fullness indicates the degree of matching between the source video bit rate and the channel bit rate, we subsequently design a buffer fullness estimation scheme where the buffer fullness estimated at the RLC layer is fed back to the video source to adapt its bit rate. Without loss of generality, we consider a specific scenario of the online encoding RC relying on controlling encoding parameters, i.e., dynamically changing the quantization parameters (QPs). It is straightforward to extend the proposed method to both SVC streaming and to HTTP-DASH scenarios, as discussed in Section II-C.

B. Problem Formulation

The block diagram of a wireless video streaming system equipped with an encoder parameter optimizer conceived to control the source rate is shown in Fig. 3, which relies on a camera, a video encoder, and a BTS. The camera samples the video scene at certain frame scanning and forwards the frames to the encoder, which forwards the compressed video to the transmitter's buffer at the RLC layer for transmission. Here, we assume that the channel between the video source and the BTS is wired and reliable. The source RC problem is eliminating the congestion in the BTS while satisfying the delay constraints. Our basic philosophy is to feed back the mismatch between the current source rate and the channel's affordable throughput to the encoder parameter optimizer to adapt the source rate. If a mismatch does occur, the encoder parameter optimizer sends a command to the video encoder to adjust the video source rate. To quantitatively characterize this mismatch, we define BOP as the probability that the current wireless channel quality provides an insufficient throughput for the current video bit rate. A sufficiently low BOP indicates that we may increase the video bit rate for transmission over the wireless channel to achieve a higher video quality.

Let $\mathcal{M} \triangleq \{M_1, \ldots, M_L\}$ be the operational mode set of the video encoder. The video bit rate corresponding to the operational mode $M_i (i = 1, \ldots, L)$ is denoted by r_i. Let us assume that $r_1 < r_2 < \cdots < r_L$, which implies that a higher operational mode results in a higher video bit rate and hence achieves a better video quality. The operational mode is controlled by the encoder parameters. For instance, we can control the encoding mode by appropriately setting the I-P-B quantization mode or the target video bit rate parameter of the JM encoder. Here, we consider adapting the source rate to the channel quality by dynamically adjusting the operational mode (i.e., the encoding parameters) of the video encoder.

We used transmission slots of equal duration, which are normalized to unity. The RLC SDUs, which contain the video data generated by the video encoder, arrive at the RLC buffer, and they are then queued until they are transmitted. Since the data are processed at packet level instead of bit granularity, we define the length of the RLC buffer as the number of RLC SDUs. Let $A_n \in \mathcal{A} \triangleq \{0, \ldots, A\}$ denote the number of RLC SDU arrivals in slot n, where A is the maximum number of RLC SDU arrivals in a single slot, whereas $D_n \in \mathcal{D} \triangleq \{0, \ldots, D\}$ is defined as the number of RLC SDUs transmitted in slot n, where D is the maximum number of RLS SDUs transmitted in a single slot. It should be noted that an agile and sophisticated link adaptation mechanism based on AMC and HARQ is applied to maintain the required target bit error rate in most state-of-the-art wireless communication systems such as high-speed downlink packet access, 3G LTE, and WiMAX. Both AMC and HARQ rely on the CQI feedback received from the mobile terminals, which results in a burst-by-burst adaptive channel throughput. Hence, the downlink packet departure process D_n is assumed to be an independent identically distributed (i.i.d.) sequence. Let Q_n denote the buffer fullness expressed in terms of the number of packets at the end of slot n. The dynamics of the buffer fullness may be described as

$$Q_n = \max\left\{Q_{(n-1)} - D_n, 0\right\} + A_n. \tag{1}$$

According to Little's theorem [29], we have

$$\bar{Q} = \lambda \bar{D} \tag{2}$$

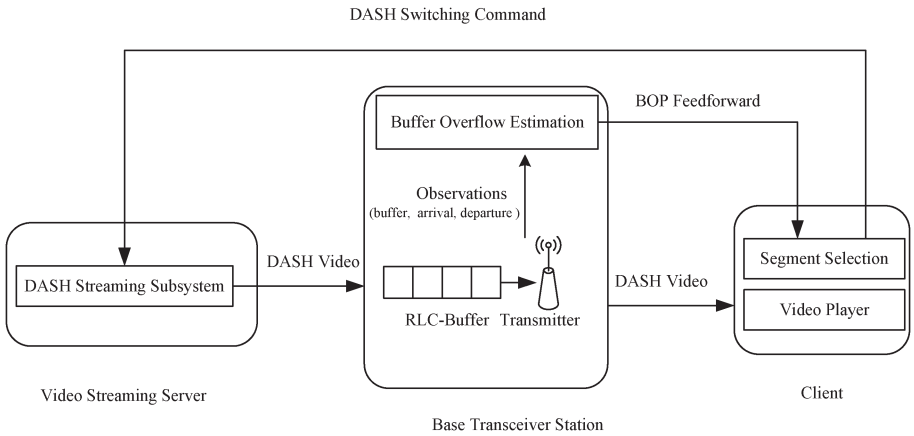


Fig. 5. Adaptive source RC framework for the DASH streaming application.

where $\bar{Q}$ is the average buffer fullness, $\bar{D}$ is the average delay, and λ is the packet arrival rate. This implies that, if λ is given, having a higher buffer fullness value implies a higher delay. Hence, we treat the delay as being synonymous with buffer fullness. Our objective is to select the best encoding mode for controlling the buffer fullness to satisfy the delay constraint of the specific video application. Let us formally define the BOP as

$$p_o = P(Q > B_h) \tag{3}$$

where Q is the buffer fullness and B_h is the buffer threshold. Higher BOP implies higher buffer fullness in the near future, in turn inducing a higher delay. To guarantee lip-synchronized interactive low-delay video transmission, the BOP should be kept low. Hence, the problem of RC may be reinterpreted as the problem of selecting the best encoding mode subject to a given BOP, which can be formulated as

$$\max_{m \in \mathcal{M}} \quad m \tag{4}$$

$$\text{s.t.} \qquad P(Q^m > B_h) < p_{\text{qos}} \tag{5}$$

where Q^m denotes the buffer fullness corresponding to the video encoding mode m, and p_{qos} is the maximum tolerable BOP.

It should be noted that the queue length $P(Q > x)$ in the scenario of an infinite buffer system is called an *overflow probability* or a *tail probability* [30], which has been used as a performance metric for an infinite buffer-aided system [31]–[33]. However, practical systems have a limited buffer capacity. When a video frame arrives at a full buffer, it gets dropped. Hence, the packet loss probability (PLP) $P_L(B_h)$ is invoked to characterize the performance of a finite-buffer-based system having a capacity B_h of video frames. For video streaming services, the PLP is an important QoS measure, but it is difficult to directly estimate the PLP of a general finite-buffer-based system. Fortunately, a theoretical justification was provided to approximate the PLP of a finite-buffer-based system with the aid of estimating the BOP of its infinite buffer counterpart [33]. Hence, we treat the BOP as being identical to the PLP, which implies that the video source rate optimization problem of a finite-buffer-based system having a capacity B_h of frames can be mapped to the mathematical model of (4) and (5). We have confirmed, by our simulation studies, the plausible fact that a high BOP may severely degrade the QoS, which is quantitatively characterized in Fig. 4 and motivates us to use the BOP for the feedback control.

C. Extensions for SVC and DASH Scenarios

The scheme shown in Fig. 3 is a point-to-point representation of our proposed framework, which may be also used for streaming to multiple users, as shown in our forthcoming SVC/HTTP-DASH streaming applications.

The framework detailed in Section II-B may be readily extended to SVC and HTTP-DASH scenarios. For an SVC scenario, a scalable video stream consists of a base layer l_1 and $(L-1)$ enhancement layers, i.e., $\{l_2, l_3, \ldots, l_L\}$. The operational mode $M_i (i = 1, \ldots, L)$ is defined as the scenario when the first i layers are selected for transmission. Thus, the source RC via adaptive enhancement layer selection can be also formulated as the problem [see (4) and (5)]. The corresponding video source in Fig. 3 may be also replaced by a video streaming server, which maintains the sessions corresponding to the multiple video users and receives the BOP feedback of a specific session for adjusting the number of the transmitted enhancement layers.

It should be noted that HTTP-DASH applied the pull-based streaming paradigm rather than the traditional push-based streaming paradigm relying on protocols such as the Real-Time Streaming Protocol. The client plays the central role of driving the video adaptation. To comply with the basic rule of DASH, the framework for the DASH streaming applications shown in Fig. 5 is based on the BOP feedforward to the client. The client may then use the received BOP to select the future video segments to be fetched. The DASH streaming server maintains multiple sessions for the sake of supporting video streaming to multiple users. In a DASH scenario, the operational mode set $\mathcal{M}$ is defined as the adaptation set, where the operational mode $M_i (i = 1, \ldots, L)$ denotes streaming the chunks corresponding to the ith quality level. Then, the source RC via HTTP-DASH can be similarly described as in (4) and (5) since a higher video quality implies a higher bit rate.

III. Large Deviation-Based Overflow Probability Estimation

Naturally, the BOP estimation is a key step to successfully solve the problem [see (4) and (5)]. To provide a high quality of experience for a video streaming service, an occurrence of buffer overflow is expected to be a rare event. The theory of large deviation provides a useful way of accurately characterizing the probability of rare events. Therefore, we present a large deviation-based BOP estimation model here. Although our practical buffer is of finite size, the BOP of the corresponding infinite-buffer-based system can be invoked to estimate the PLP, as shown in [31]–[33]. Hence, we subsequently aim to estimate the BOP of the corresponding infinite-buffer-based system. To derive an analytical model, the BOP is estimated based on the M/M/1 queueing model, which has been widely applied in wireless networks [34]–[37]. The buffer at the RLC is modeled as the M/M/1 queue, where λ RLC SDUs/slot and μ RLC SDUs/slot denote the Poissonian arrival rate and the departure rate (or service rate), respectively. Let us now define the large deviation theory in [19] for estimating the probability of $P(Q_{n+N} > B_h|Q_n)$, where Q_n is the buffer length during the current time slot n. During the kth slot, the evolution of buffer length can be represented by

$$I_k = A_k - \min\left(Q_{(k-1)}, D_k\right). \tag{6}$$

Then, we have $Q_{n+N} = Q_n + \Sigma_{k=n+1}^{n+N} I_k$. Furthermore, $P(Q_{n+N} > B_h|Q_n)$ can be rewritten as

$$\begin{aligned} P(Q_{n+N} > B_h|Q_n) &= P\left(Q_n + \Sigma_{k=n+1}^{n+N} I_k > B_h\right) \\ &= P\left(\frac{1}{N}\Sigma_{k=n+1}^{n+N} I_k > a\right) \end{aligned} \tag{7}$$

where we have $a = (B_h - Q_n)/N$. Next, we apply the LDP to the sequence $I_{n+1}, I_{n+2}, \ldots$ to derive $P((1/n)\Sigma_{k=1}^{n} I_k > a)$. The LDP can be briefly described as follows [38].

Definition: A sequence $X_1, X_2, \ldots$ obeys the LDP associated with rate function $I(\cdot)$ if the following conditions apply.

1) For any closed set F, we have

$$\limsup_{N\to\infty} \frac{1}{N} \log P\left(\frac{1}{N}\sum_{i=1}^{N} X_i \in F\right) \leq -\inf_{a\in F} I(a). \tag{8}$$

2) For any open set G, we have

$$\liminf_{N\to\infty} \frac{1}{N} \log P\left(\frac{1}{N}\sum_{i=1}^{N} X_i \in G\right) \geq -\inf_{a\in G} I(a). \tag{9}$$

Obviously, $I_k \in \mathbb{R}\,(k \in [n+1, n+N])$ represents i.i.d. random variables with a mean of $E[I_k] = \lambda - \mu$ and moment-generating function (MGF) of $M(\theta) = E[e^{\theta I_k}]$ that is finite in a neighborhood of 0. According to *Cramér's theorem* [38], $I_{n+1}, I_{n+2}, \ldots$ obeys the LDP associated with the rate function $I(a) = \sup_{\theta>0}(\theta a - \log M(\theta))$, which implies that, for any $a > \lambda - \mu$, we have

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\frac{1}{N}\sum_{i=n+1}^{n+N} I_i > a\right) = -I(a). \tag{10}$$

It should be noted that (10) is logarithmically asymptotic. Hence, when N is large, the BOP can be approximated as

$$P\left(\frac{1}{N}\sum_{i=n+1}^{n+N} I_i > a\right) \approx \exp\left[-NI(a)\right]. \tag{11}$$

Since A_k obeys Poisson's distribution, we have

$$E[e^{\theta A_k}] = \sum_{\eta=0}^{\infty} e^{\theta\eta}\frac{\lambda^\eta}{\eta!}e^{-\lambda} = \exp(\lambda e^\theta - \lambda). \tag{12}$$

Similarly, we may arrive at $E[e^{-\theta D_k}] = \exp(\mu e^{-\theta} - \mu)$. Then, we may derive the MGF $M(\theta)$ as follows:

$$\begin{aligned} M(\theta) &= E[e^{\theta I_k}] \\ &= E[e^{\theta A_k}]E[e^{-\theta D_k}] \\ &= \exp(\lambda e^\theta + \mu e^{-\theta} - \lambda - \mu). \end{aligned} \tag{13}$$

Hence, the corresponding rate function $I(a)$ can be rewritten as

$$\begin{aligned} I(a) &= \sup_{\theta>0}(\theta a - \lambda e^\theta - \mu e^{-\theta} + \lambda + \mu) \\ &= (\theta a - \lambda e^\theta - \mu e^{-\theta} + \lambda + \mu)\big|_{\theta=\log\frac{a+\sqrt{a^2+4\lambda\mu}}{2\lambda}} \\ &= a\log\frac{a+\sqrt{a^2+4\lambda\mu}}{2\lambda} - \sqrt{a^2+4\lambda\mu} + \lambda + \mu. \end{aligned} \tag{14}$$

If λ and μ are given for any encoding mode $m \in \mathcal{M}$, the BOP $P(Q_{n+N} > B_h|Q_n)$ can be estimated by using (11) and (14). The problem [see (4) and (5)] can be then solved by calculating the BOPs corresponding to all encoding modes in $\mathcal{M}$. However, in practical situations, both λ and μ depend on the network's condition and on the source encoding mode, and they are not based on prior knowledge for each encoding mode.

An attractive technique is applying an iterative policy to solve the problem [see (4) and (5)]. Specifically, at the current mode update period, we can calculate the BOP corresponding to the current encoding mode by the online estimation of the arrival rate λ and the departure rate μ. Although λ and μ can be estimated by counting the number of RLC SDU arrivals and departures over a time period of T, they may change over time due to time-varying network conditions and dynamic encoding mode switching. The dynamic nature of λ and μ may be modeled using an autoregressive integrated moving average (ARIMA) process [39] as follows:

$$\left(1 - \sum_{i=1}^{p} a_i D^i\right)(1-D)^{d_1}\lambda(t) = \left(1 + \sum_{i=1}^{q} b_i D^i\right)\varepsilon(t)$$

$$\left(1 - \sum_{i=1}^{p} c_i D^i\right)(1-D)^{d_2}\mu(t) = \left(1 + \sum_{i=1}^{q} d_i D^i\right)\mu(t)$$

where D is the unit time-delay operator, and $\varepsilon(t)$ is the estimation error. Therefore, the task of estimating λ and μ becomes the task of estimating parameters p, q, d_1, d_2, a_i, b_i, c_i, and d_i for the ARIMA model [39]. Once λ and μ have been estimated, the corresponding BOP can be directly calculated by using (11) and (14). If the BOP estimate fails to satisfy the constraint (5), the source encoding rate should be decreased, thus reducing the BOP. Otherwise, if the BOP is much lower than p_{qos}, the video source bit rate may be increased, thus improving the video quality. Applying this iterative encoding mode update, we can iteratively solve the problem [see (4) and (5)] to find a new optimal encoding mode.

It should be noted that the aforementioned method relies on the classic M/M/1 queuing model, albeit the Poissonian assumption may be not the most realistic packet arrival and departure model. Moreover, our model of the BOP requires estimating both λ and μ. In the next section, we propose an online measurement-based RC, which no longer relies on the M/M/1 queuing model and on the estimation of the RLC SDU arrival rate λ and the departure rate μ.

IV. Online Encoding Mode Switching-Based Rate Control for Video Streaming

Let the index of the current time slot be n and the current buffer length be Q_n^m, whereas $m = M_i$ denotes the current video encoding mode. We estimate the BOP at the $(n+N)$th $(N > 0)$ slot under the assumption of maintaining the current encoding mode, where N is referred to as the prediction interval. Let $P_o^{n+N}(m)$ denote the overflow probability at slot $(n+N)$, which is defined as

$$P_o^{n+N}(m) = P\left(Q_{n+N}^m > B_h\right). \tag{15}$$

Subsequently, we derive a BOP estimation model based on the aforementioned LDP [19], and then propose an online adaptive source RC for video streaming over VBR channels.

A. BOP Estimation

When transmitting D_k RLC SDUs containing video data in slot k, the increased buffer length may be expressed as

$$I_k = A_k - \min\left(Q_{(k-1)}, D_k\right). \tag{16}$$

Since we have $0 \le A_k \le A$ and $0 \le D_k \le D$, the set of values for I_k is $\{-D, \ldots, 0, \ldots, A\}$. Let us introduce the variable $\pi_i = P(I_k = i)$ to denote the probability of having a buffer length increase of $I_k = i$. For a given wired network condition, A_k is determined by the video encoding mode since the encoded video frame size is affected by the encoding mode, and D_k is related to the channel bit rate. Their difference I_k shows the instantaneous mismatch between the video bit rate and the channel bit rate. Due to the time-variant RLC SDU arrivals and the fluctuating channel bit rate, the sequence $I_i(i = 1, 2, \ldots)$ may frequently alternate between negative and positive values.

The total increase in the buffer length during the interval spanning from slot n to slot $(n+N)$ is given by

$$I^{n+N} = \Sigma_{i=1}^{N} I_{n+i}. \tag{17}$$

Then, the buffer length Q_{n+N}^m at the end of the slot $(n+N)$ may be expressed as

$$Q_{n+N}^m = Q_n^m + I^{n+N}. \tag{18}$$

According to (6), the BOP at slot $(n+N)$ may be rewritten as

$$P_o^{n+N}(m) = P\left(Q_n^m + I^{n+N} > B_h\right). \tag{19}$$

Let us now define the expected value of the average buffer length increase in each of the future N slots as

$$m_o = E\left[\frac{\Sigma_{i=1}^{N} I_{n+i}}{N}\right] \tag{20}$$

where $E[\cdot]$ denotes the expectation operator.Furthermore, while keeping the current encoding mode fixed as $m = M_i$, the tolerable average buffer length increase during each slot is defined as

$$a_o = \frac{B_h - Q_n^m}{N}. \tag{21}$$

Having $m_o \ge a_o$ implies that there would be a high BOP after N time slots.

Let us now rewrite (19) as

$$\begin{aligned} P_o^{n+N}(m) &= P\left(Q_n^m + I^{n+N} > B_h\right) \\ &= P\left(I^{n+N}/N > \left(B_h - Q_n^m\right)/N\right) \\ &= P\left(\frac{\Sigma_{i=1}^{N} I_{n+i}}{N} > a_o\right). \end{aligned} \tag{22}$$

The term $\Sigma_{i=1}^{N} I_{n+i}/N$ in (22) represents the average buffer length change during a slot, which is jointly determined by the video encoded bit rate controlled by the encoding mode and the channel bit rate, whereas a_o is the tolerable average increase in the buffer length for each of the N future slots, as determined by the current buffer length. Therefore, the probability $P_o^{n+N}(m)$ in (22) may be viewed as an estimate of the remaining storage capacity in the buffer, which may be used to smoothen the fluctuation of the channel's affordable throughput in the future. If the probability $P_o^{n+N}(m)$ exceeds a predefined threshold value p_{qos}, we can decrease the encoding mode index to reduce the video bit rate, thus decreasing both the BOP and the buffer-induced delay.

Since $D_k(k = 0, 1, \ldots)$ are i.i.d. variables, $I_k(k = 1, 2, \ldots)$ are also i.i.d. random variables, which have a finite MGF of $M(\theta) = Ee^{\theta I_k}$ in the vicinity of 0. Then, provided that $a_o > m_o$ is satisfied, according to (11), for large N, we have

$$P_o^{n+N}(m) \approx \exp\left[-NL(a_o)\right] \tag{23}$$

where

$$L(a_o) = \sup_{\theta>0} \{a_o\theta - \log M(\theta)\} \tag{24}$$

$$\log M(\theta) = \log \left\{ \sum_{i=1-s_D}^{1} \pi_i \exp[i\theta] \right\}. \tag{25}$$

To calculate the approximate BOP of (23), the knowledge of a_o, m_o, and π_i is required. It is straightforward to calculate a_o according to (21). However, there is no prior knowledge about the histogram of I_k, and hence, we cannot derive analytical expressions for m_o and π_i. Therefore, we have to rely on their buffered history for estimating the current values of these parameters using the classic sliding-window-based method.

The observed buffer length increase/decrease sequence constitutes the input of $\{I_1, I_2, I_3, \ldots\}$. The sliding window takes into account the N_s most recent I_k values of $W_n = [I_n, I_{n-1}, \ldots, I_{n-N_s+1}]$.

For parameter m_o of (20), we use the sample mean as its estimate, i.e., we have

$$\hat{m}_o = \frac{\sum_{i=n-N_s+1}^{n} I_i}{N_s}. \tag{26}$$

Let $N_i (i \in \{-D, \ldots, 0, \ldots, A\})$ denote the number of events when $I_k = i$ appears in the sliding window, which is given by

$$N_i = \sum_{k=n-N_s+1}^{n} 1(I_k = i) \tag{27}$$

where $1(\cdot)$ is an indicator function. Then, the relative frequency of encountering $I_k = i$ may be estimated as

$$\hat{\pi}_i(n) = \frac{N_i}{N_s} \tag{28}$$

which can be applied in (25) to estimate the BOP.

B. Online Source RC Algorithm

The RC strategy advocated will be discussed in the context of three scenarios according to both the current buffer length Q_n and the average buffer length increase per slot $\hat{m}_o$ as follows.

1) $Q_n^m \geq B_h$: This implies that the current buffer length is above the threshold, which will impose an undesirable delay of the video frame. Therefore, we should reduce the video bit rate by decreasing the encoding mode index as

$$m = M_{\max\{i-1,1\}}. \tag{29}$$

2) $Q_n^m < B_h$ and $\hat{m}_o \geq a_o$: In this scenario, although the buffer length is under the threshold B_h, its average increase per slot $\hat{m}_o$ exceeds the tolerable average increase a_o of the buffer length per slot during the forthcoming N slots. This implies that, at the current buffer length increase rate, the buffer length will become higher than B_h after N slots. Therefore, in this case, the current video bit rate should be decreased to reduce the BOP. Then, we can adjust the encoding mode according to (29).

3) $Q_n^m < B_h$ and $\hat{m}_o < a_o$: In this case, the current buffer length is under the threshold B_h, and the average buffer length increase per slot $\hat{m}_o$ is within the range defined by the capacity a_o. Nevertheless, this does not imply that no buffer overflow will occur in the forthcoming N slots because $\hat{m}_o$ is the average buffer length increase per slot, which cannot directly characterize the buffer length increase in a certain time slot. Hence, there is still a chance of buffer overflow. Fortunately, since we have $a_o > \hat{m}_o$, buffer overflows remain a rare event. According to the large deviation-based probability estimation model of (23), for a sufficiently large N, the BOP may be approximated as

$$\hat{P}_o^{n+N}(m) = \exp\left[-NL(a_o)\right]. \tag{30}$$

An online measurement-based estimation method was presented in the previous section to calculate the BOP. Owing to the exponential decay of the estimated BOP probability with N, we can set N to a moderate value for the sake of acquiring an accurate BOP estimation instead of requiring a large N.

Our proposed RC algorithm aims to adjust the encoding mode to satisfy the BOP QoS requirement, i.e., p_{qos}. If we have $\hat{P}_o^{n+N}(m) \geq p_{\text{qos}}$, this implies that the current encoding mode index is too high to keep the BOP below p_{qos}. Therefore, we should decrease the encoding mode index to reduce the video bit rate, thus reducing the BOP. The future encoding mode index is adjusted as $m = M_{\max\{i-1,1\}}$.

By contrast, if we have $\hat{P}_o^{n+N}(m) < p_{\text{qos}}$, the currently affordable bit rate of the channel may be able to support a higher video bit rate. Hence, we should increase the encoding mode index to provide an improved video quality for the sake of fully exploiting the attainable bit rate of the channel. To achieve this, we define a threshold $p_T (< p_{\text{qos}})$ for the BOP. If $\hat{P}_o^{n+N}(m) < p_T$ is encountered consecutively $K > 0$ times, we will increase the encoding mode index according to

$$m = M_{\min\{i+1,L\}}. \tag{31}$$

The reason for requiring K consecutive threshold violation occurrences to trigger an encoding mode index adjustment is that this prevents frequent adjustments of the encoding mode, which would result in perceivable video quality fluctuations.

The RC regime is summarized in ***Algorithm 1***. Although the estimated average channel throughput was directly fed back to the BTS to control the source rate in [18], this technique does not characterize the burst-by-burst adaptive channel throughput on a sufficiently fine timescale, which, hence, fails to guarantee a low delay for the video packets. By contrast, ***Algorithm 1*** relies on the LDP of [19] to estimate the BOP, when the buffer overflow is a rare event, and applies the BOP constraint to trigger the source RC, thus achieving a low delay for video packets.

Algorithm 1 Online measurement-based adaptive rate control algorithm for streaming video over VBR channels.

Current encoding mode $M = M_i$
if $Q_n^M \geq B_h$ **then**
  $i \leftarrow \max(i-1, 1)$
else
  Calculate $\hat{m}_o$ and $\hat{a}_o$
  if $\hat{m}_o \geq \hat{a}_o$ **then**
    $i \leftarrow \max(i-1, 1)$
  else
    Calculate $\hat{P}_o^{n+N}(M)$
    if $\hat{P}_o^{n+N}(M) \geq p_{\text{qos}}$ **then**
      $i \leftarrow \max(i-1, 1)$
    else
      if $\hat{P}_o^{n+N}(M) < p_T$ consecutively happens K times **then**
        $i \leftarrow \min(i+1, L)$
      end if
    end if
  end if
end if

C. Discussion

Previously, the proposed solution was discussed in the context of a single user requesting a single video stream. However, it may be also extended to the scenario where multiple users having a different link quality desire the same video. A simple solution is for the video server to create a dedicated encoder instance for each encoding mode, where multiple video streams are generated by the encoders with the aid of different encoding modes. Then, based on the feedback triggered by the proposed method from the BTS, the video server may select an appropriate video stream for each user associated with a different link quality.

V. Performance Evaluation

Here, we characterize the performance of our online measurement-based adaptive RC (MBARC) algorithm. We first describe our simulation setup, including the network model and the video sequences employed. Then, the metrics used for performance evaluation are described. Finally, our simulation results are presented and analyzed. In the simulations, we also implemented the heuristic online adaptive RC (HOARC) algorithm in [18] and an offline method to provide pertinent performance comparisons to cutting-edge benchmarkers. The offline method simply relied on all the encoding modes and selected the mode having the best performance as the performance benchmark.

A. Simulation Setup

We consider an HSPA network [1] relying both on AMC and HARQ. We assume a UE (UE in HSPA parlance) belonging to category 10. According to the HSPA specifications [2], the CQI value ranges from 0 to 30, and the corresponding TB sizes are $0, 137, \ldots, 25558$ bits, respectively. In HSPA, there are 15 time-division multiplex slots per 10-ms frame. Three $10/15 = 2/3$ ms slots form a so-called transmission time interval (TTI) of 2-ms duration, where only one user is allowed to transmit with the aid of multiple spreading codes per TTI. Therefore, the time-varying number of users may result in a time-varying number of TTIs being assigned to each user, which, in turn, leads to a time-varying throughput for each user. Therefore, we simulated a multiuser scenario, where the maximum number of concurrently communicating users was set to $U = 8$, and the new user arrival process follows a Markov process with an arrival rate of $\lambda = 10^{-4}$ per TTI. The service rate was assumed to be $\upsilon = 1.5 \times 10^{-5}$ per TTI. A round robin scheme was applied to schedule the transmissions of the users. Then, our proposed strategy is applied for one of the users to implement its online source RC. Consider an i.i.d. Rayleigh channel for the target user, where the received SNR s is an exponentially distributed random variable described by the probability density function of $f(s) = (1/\gamma)e^{-(s/\gamma)}$ having an average of γ. We applied the SNR (in decibels)-to-CQI mapping in [40], i.e.,

$$\text{CQI} = \lfloor \text{SNR} + 4.5 \rfloor. \tag{32}$$

According to the specifications, the CQI reporting cycle is defined as 1, 2, 4, 5, 10, 20, 40, and 80 TTIs. In the simulations, we set the CQI reporting cycle to four TTIs.

TABLE I
Properties of the Video Sequences

Encoder	H.264 Full
Resolution	CIF(352×288): Silence of the Lambs NBC 12 News D−1(704×576): The Shawshank Redemption
Bit Rate	Variable Bit Rate (VBR)
Number of Frames	45,000
GoP Size	16
Frames Rate	30fps
Video Duration	25 minutes
No. B frames between I/P frames	1

B. Video Sequences Used for Performance Evaluations

Three different video sequences are used in our simulations, namely, the "Silence of the Lambs" clip, the "NBC 12 News" clip [17], and the "Shawshank Redemption" clip, scanned at 30 frames/s and encoded by the H.264 codec. The two former clips have a Common Intermediate Format (CIF) resolution, whereas the last clip has a D-1 resolution. The duration of the video sequence is 25 min, corresponding to 45 000 video frames, where a GOP is constituted by 16 frames. The properties of these video sequences are listed in Table I. Here, the encoder mode is denoted by (m_1, m_2, m_3), where $m_1, m_2,$ and m_3 represent the quantization scales for the I, P, and B frames, respectively. The operational modes are listed as follows: $M1 = (10, 10, 12)$, $M2 = (16, 16, 18)$, $M3 = (22, 22, 24)$, $M4 = (24, 24, 26)$, $M5 = (28, 28, 30)$, $M6 = (34, 34, 36)$, $M7 =$

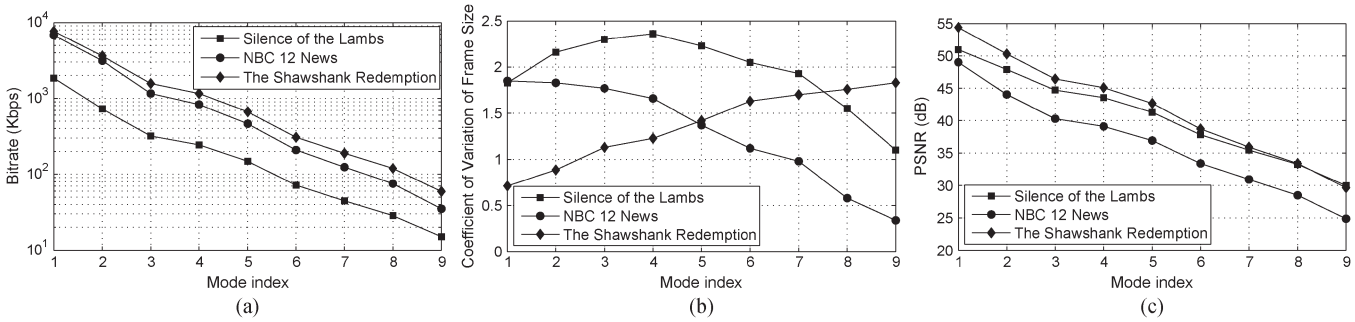


Fig. 6. Statistics of the video sequences. (a) Mean bit rate. (b) Coefficient of variation of frame size. (c) Mean PSNR.

$(38, 38, 40)$, $M8 = (42, 42, 44)$, and $M9 = (48, 48, 50)$. Fig. 6 characterizes the bit rate, the frame-size variation coefficient, and the peak SNR (PSNR) statistics of the video sequences in the different encoding modes, with the frame-size variation coefficient being defined in [41] as follows:

$$CoV_X^q = \frac{1}{\bar{X}_N^q}\sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1}\left(X_n^q - \bar{X}_N^q\right)^2} \tag{33}$$

where X_n^q denotes the frame size of video frame n encoded with the QP q, and $\bar{X}_N^q$ is the average frame size of the N video frames. In the first two simulations, the former two CIF clips were used for basic performance investigations, whereas in the rest of the simulations, the D-1 clip was used to characterize the achievable performance in the case of higher resolution videos.

C. Performance Metrics

In the simulations, each RLC SDU carried the bits of video frames. To satisfy the tolerable delay constraint of a specific video application, the RLC SDUs are dropped from the buffer when their delay exceeded the threshold of l milliseconds. To evaluate the quality of a received video sequence, an objective video quality evaluation model similar to [43] is adopted, which is defined by the rate of dropped video frames (DVFRs) that are undecodable at the MS, i.e., by

$$\text{DVFR} = 1 - \frac{N_{\text{dec}}}{N_{\text{total}-I} + N_{\text{total}-P} + N_{\text{total}-B}} \tag{34}$$

where N_{dec} is the total number of decodable frames, including all three types of frames. The decoding dependence values between different types of frames are also considered in counting the decodable frames. A lower DVFR value implies that a better quality is perceived by the recipient. Since the proposed adaptation method uses different bit rates, when configured for adapting the throughput and the quality, we also apply the average PSNR as a further performance metric.

All the results were averaged over 100 independent simulation runs. To characterize the performance improvement of our MBARC, we also conducted independent simulations for each encoder mode, and the best results were selected as performance benchmarks.

D. Simulation Results

1) Performance for Different Delay Thresholds: In the first experiment, the parameters of the proposed online MBARC algorithm were set as follows: length of the sliding window $N_s = 80$, prediction interval $N = 80$, buffer length threshold $B_h = 16$, $p_{\text{qos}} = 10^{-5}$, $p_T = 10^{-10}$, and $K = 32$. The initial mode index of the encoder is M_9, which has the lowest video bit rate. The MBARC was activated every eight video frame intervals. Its performance was investigated for the delay thresholds of {60, 100, 140, 180, and 220 ms}, which cover the delay requirements of various video applications. For example, the delay limit of 60 ms is applicable to lip-synchronized real-time interactive video conferencing applications, whereas the delay limit of 100 or 140 ms is applicable to wireless video surveillance, and finally, 180 and 220 ms are for digital television broadcast or video-on-demand services. Fig. 7(a) and (b) shows the DVFR and the average PSNR of Silence of the Lambs and NBC 12 News for different delay thresholds at the average channel SNR of $\gamma = 20$ dB. It can be observed in Fig. 7 that, as the encoder mode index increases, the DVFR decreases owing to the reduced source rate. For Silence of the Lambs, the DVFR of the operational mode $M6$ is similar to the DVFR of MBARC, but our MBARC improves the average PSNR by about 3 dB. For NBC 12 News, our MBARC and $M7$ have a similar DVFR, but MBARC improves the average PSNR by about 1 dB. This implies that the proposed MBARC is capable of adaptively adjusting the encoder mode when the source rate is temporarily higher than the affordable channel rate.

2) Performance for Different Channel SNRs: To investigate the proposed MBARC's source rate adaptation capability for different channel qualities, we conducted experiments at different average channel SNRs, namely, at $\gamma = \{12, 16, 20, 24,$ and 28 dB$\}$. The buffer length threshold was set to 28, whereas the remaining MBARC parameters were the same as in the first experiment. We set the maximum delay, which triggers dropping of the frames in the buffer to 200 ms. For each average channel SNR considered, we use an offline procedure to select the best encoding mode, which maximizes the average PSNR, while maintaining a DVFR similar to that of MBARC. The simulation results shown for the MBARC, HOARC, and offline mode selection method were plotted in Fig. 8(a) and (b). The results recorded in Fig. 8(b) for different average channel SNRs using the offline method were marked with (*Mode*). Fig. 8(b) demonstrates the average PSNR

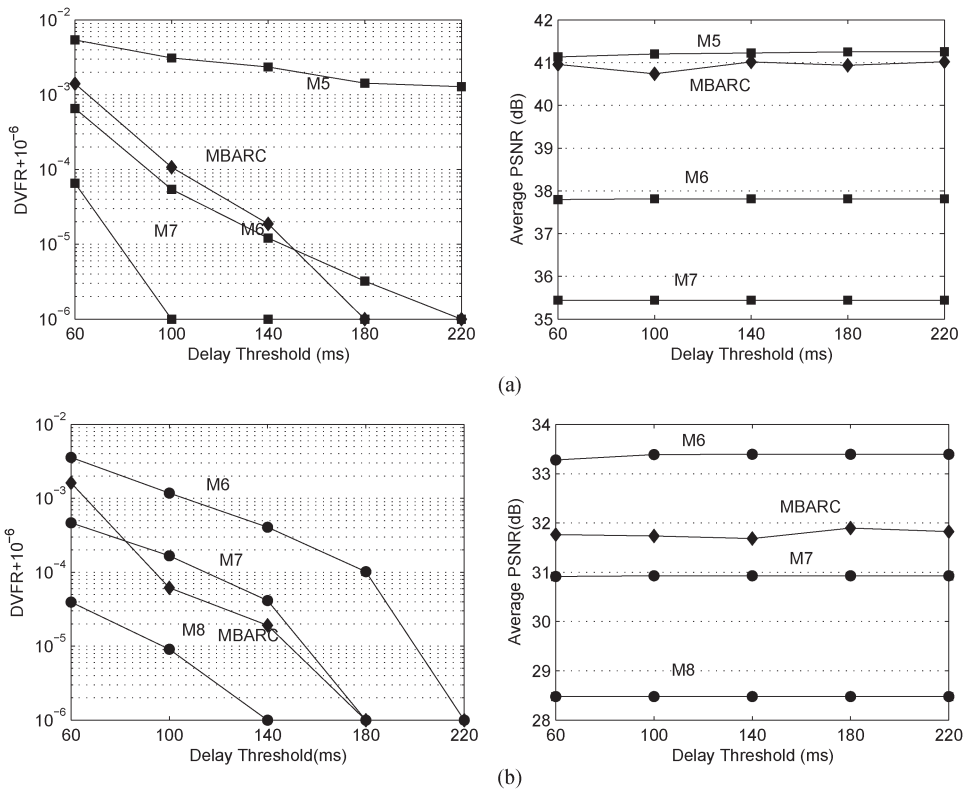


Fig. 7. (a) DVFR and Average PSNR of Silence of the Lambs for different delay thresholds with average channel SNR of 20 dB. (b) DVFR and Average PSNR of NBC 12 News for different delay thresholds with average channel SNR of 20 dB. (a) Silence of the Lambs. (b) NBC 12 News.

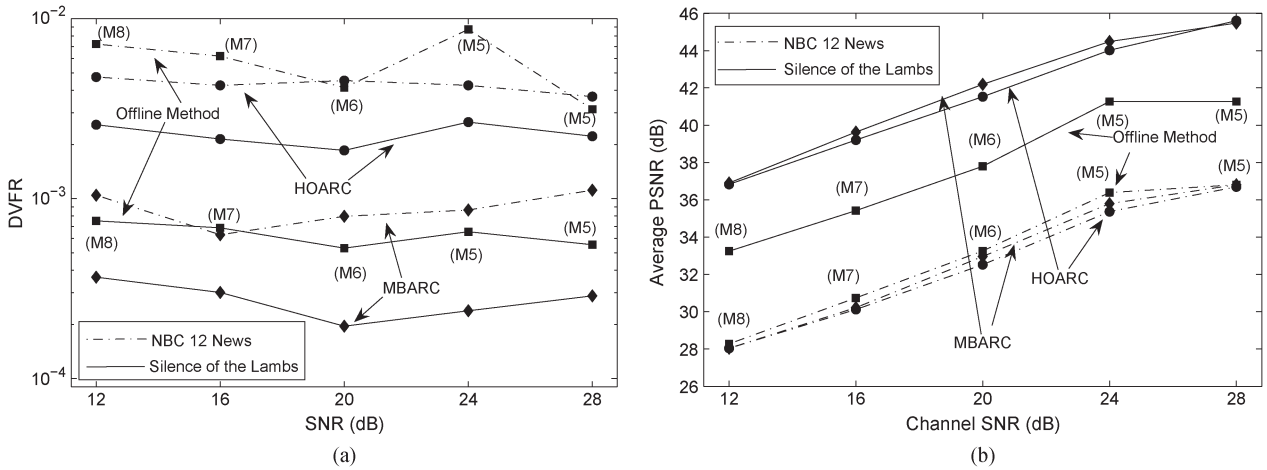


Fig. 8. (a) DVFR for different average channel SNRs of Rayleigh channel. (b) PSNR for different average channel SNRs of Rayleigh channel. The note in () represents the corresponding encoder mode for the offline method. (a) DVFR. (b) Average PSNR.

improvement of MBARC over the offline method, whereas its DVFR remains lower than that of the offline method. This demonstrates that MBARC is capable of adaptively accommodating different channel qualities by adjusting the encoding mode without any prior knowledge of the channel quality. Although Fig. 8(b) also shows that the HOARC technique in [18] is capable of adapting to time-variant channel conditions and approaching the PSNR performance of the proposed MBARC, it suffers from a much higher DVFR than that of the MBARC technique, as indicated in Fig. 8(a). The basic reason for this is elaborated as follows. The linear AR(1) of the HOARC technique may not be appropriate for estimating the average channel throughput since HSPA applies near-instantaneously adaptive burst-by-burst transmissions. Furthermore, the *average* channel throughput cannot fully characterize the *specific* near-instantaneously adaptive channel throughput at a certain transmission time instant, which, in turn, increases the delay of a specific packet and results in an increased probability of delay constraint violation. By contrast, the proposed MBARC technique is based on a BOP constraint, which may guarantee a

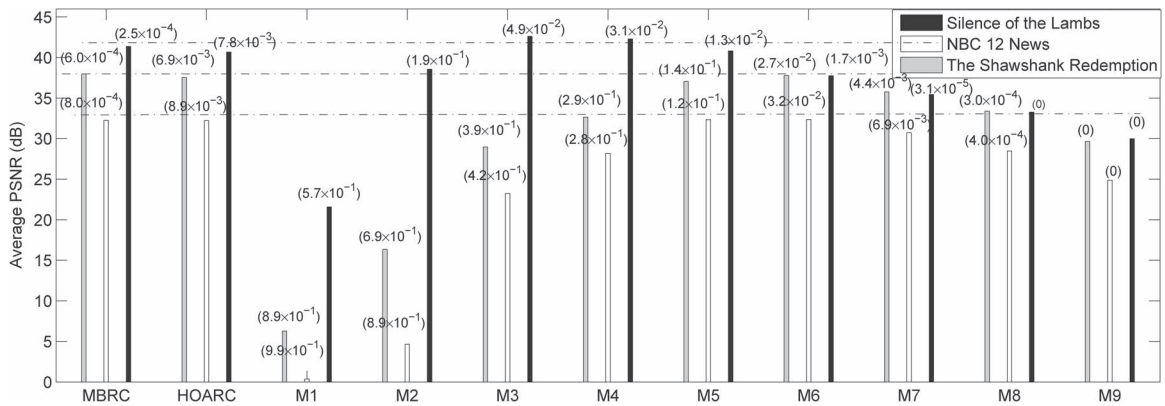


Fig. 9. Average PSNR for dynamic average channel SNRs for Rayleigh channel. The number in () represents the corresponding DVFR.

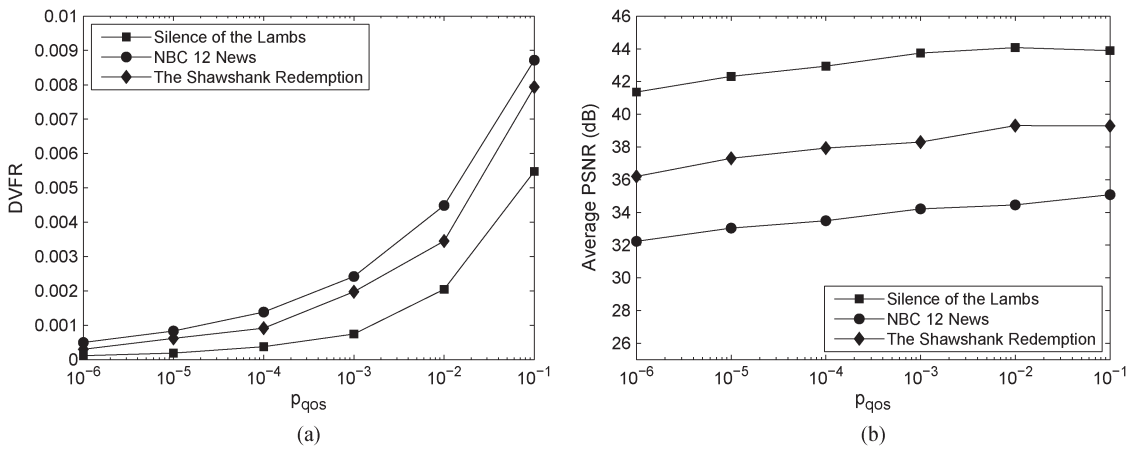


Fig. 10. (a) DVFR for different overflow probability thresholds $p_{\rm qos}$. (b) PSNR for different overflow probability thresholds $p_{\rm qos}$. (a) Silence of the Lambs. (b) NBC 12 News.

low delay for the video packets. The application of the LDP [19] assists MBARC in achieving an accurate estimation of BOP and, thus, improves the DVFR performance.

3) Performance for Dynamic Channel Quality: The previous investigations were conducted in a static environment by fixing the average channel SNR. Next, we investigate the performance of the MBARC in time-varying scenarios. The initial average channel SNR was set to 24 dB. At the instant of the 18 000th video frame interval, the average channel SNR was suddenly changed to 20 dB. Then, at the instant of the 36 000th frame interval, it was further reduced to 16 dB. All the MBARC parameters remained the same as in the previous experiment. Fig. 9 shows the corresponding simulation results, where each bar is marked with "(DVFR)" to characterize the corresponding DVFR. It is shown that the DVFR of our MBARC is lower than that of $M6$ for Silence of the Lambs and that it improves the PSNR by more than 2 dB in comparison with that of $M6$. For NBC 12 News, although $M8$ has a similar DVFR to that of our MBARC, its average PSNR is 2 dB lower than that of the proposed MBARC. This benefit is the result of the MBARC's adaptability, which adjusts the encoder mode online for VBR encoding to accommodate diverse dynamically fluctuating propagation environments. By contrast, for a fixed encoder mode, having a low channel SNR may inflict frame dropping events, resulting in low video quality, whereas at high SNRs, it fails to promptly decrease the video rate to improve the video quality in a timely manner. Similarly, MBARC is capable of adaptation and achieves an improved performance for the Shawshank Redemption clip having a higher resolution. Observe in Fig. 9 that the HOARC technique [18] approaches the PSNR performance of the proposed MBARC, but it exhibits a significantly higher DVFR than MBARC. This result also illustrates the benefit of the explicit BOP constraint in the problem formulation (4) and (5) and that of applying the LDP for accurately estimating BOP.

4) Performance Sensitivity of the Proposed Algorithm: The performance of the proposed method relies on parameters N, N_s, and K, as shown in **Algorithm 1**. Hence, this section investigates their effect on the performance of the proposed method. In the related experiments, one of these three parameters was set to different values, whereas the other parameters of our MBARC were the same as those in 2). The average SNR value of our Rayleigh channel model was fixed to 20 dB. The remaining parameters of the channel model were set as in 2).

Fig. 10 shows our simulation results of different BOP thresholds $p_{\rm qos}$ for the three clips. We may observe in Fig. 10 that a reduced overflow probability threshold provides a reduced DVFR. This is because a reduced probability threshold is capable of activating timely adjustments of the encoder mode to reduce the BOP. On the other hand, we can also observe in

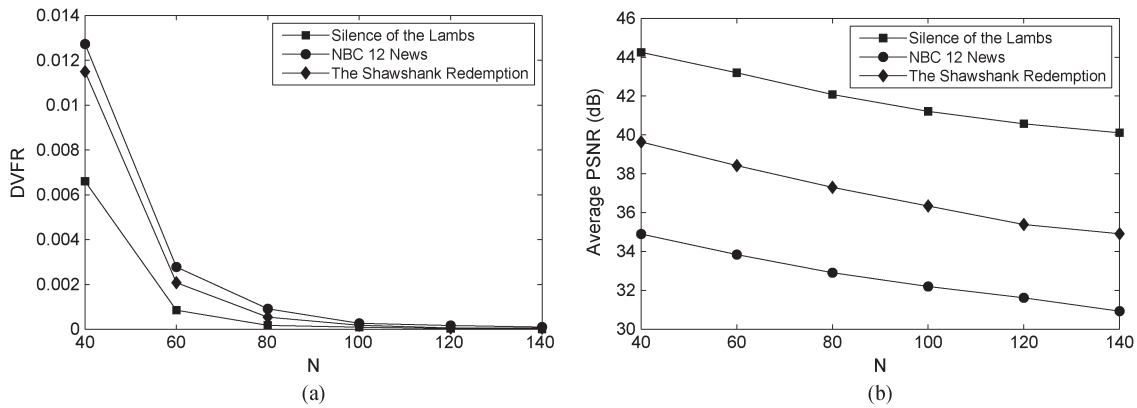


Fig. 11. (a) DVFR for different prediction intervals N. (b) PSNR for different predicting intervals N. (a) DVFR. (b) PSNR.

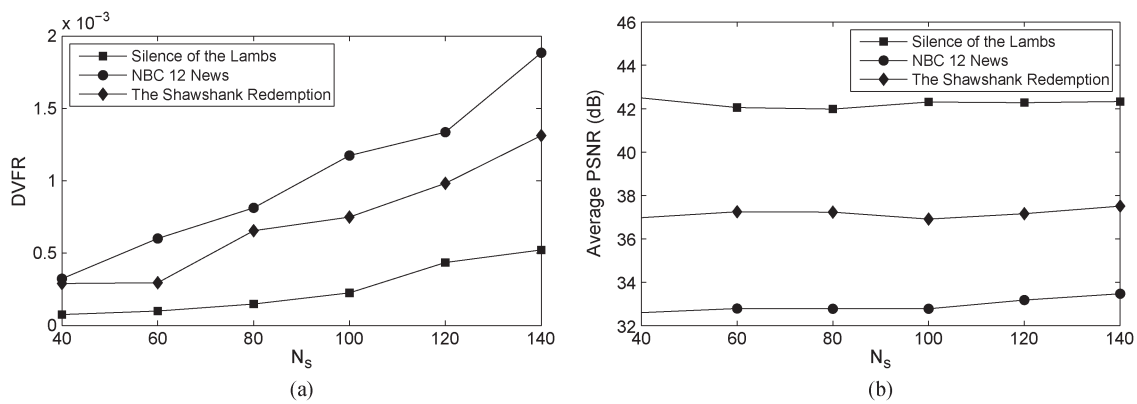


Fig. 12. (a) DVFR for different lengths of the sliding window N_s. (b) PSNR for different lengths of the sliding window N_s. (a) DVFR. (b) PSNR.

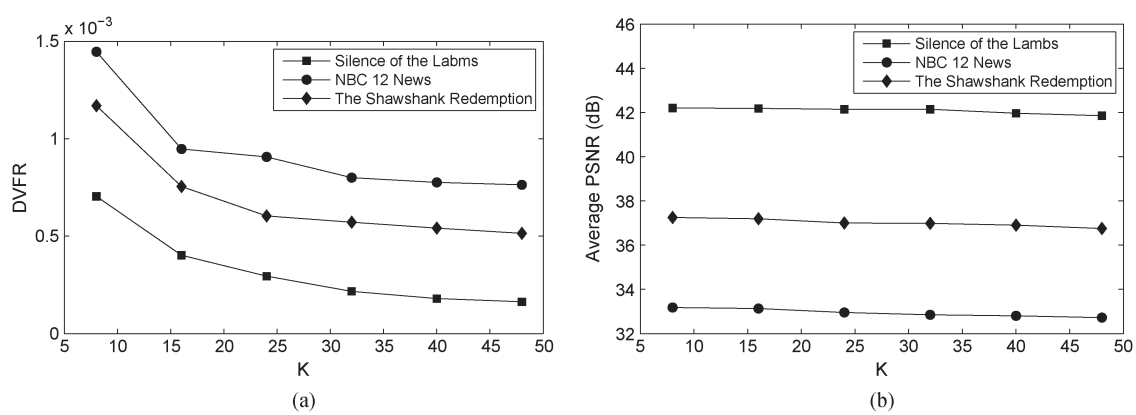


Fig. 13. (a) DVFR for different values of K. (b) PSNR for different values of K. (a) DVFR. (b) PSNR.

Fig. 10 that, as the BOP threshold decreases, the average PSNR increases. This is caused by the lower encoding mode index used to increase the video quality when the BOP threshold is higher. It should be noted that, although the DVFR is related to the BOP, their definitions are different. Hence, the value of DVFR is not the same as p_{qos}, as shown in Fig. 10(a). Fortunately, we may achieve an acceptable DVFR by setting p_{qos} to an appropriate value such as 10^{-5}.

Fig. 11 plots the simulation results for different prediction intervals N for the three clips. Fig. 11(a) shows that the DVFR is rapidly reduced upon increasing N, whereas Fig. 11(b) shows that the average PSNR is decreased as N is increased. The basic reason is elaborated as follow. According to (30), a large value of N leads to a more accurate BOP estimate, which assists the proposed method to trigger appropriate adjustments of the encoding mode and thus to reduce the DVFR. The result shows that $N \geq 80$ is appropriate for achieving an acceptable performance for the proposed method.

Fig. 12 shows the performance sensitivity of the proposed method to the sliding-window duration. It is shown in Fig. 12(a) that a larger value of N_s may result in a higher DVFR. This may be caused by the reduced sensitivity of the buffer variance when using a larger N_s for estimating $\hat{m}_o$ according to (26). Fig. 12(b) shows that N_s has only a modest effect on the PSNR.

Fig. 13 shows the sensitivity of the performance to parameter K. The proposed method triggers a video bit rate increase via adjusting the encoding mode only when the threshold violation of $\hat{P}_o^{n+N} < p_T$ is encountered consecutively K times. Therefore, having a larger K implies a more conservative adjustment policy, which leads to a lower DVFR and to a marginally reduced PSNR, which are shown in Fig. 13(a) and (b).

VI. Conclusion

An online adaptive source RC regime has been proposed for streaming videos in HSPA and LTE-like VBR downlink scenarios. Large deviation theory was invoked to derive the online measurement-based BOP model, applied at the RLC layer in the BS. The advantage of the resultant algorithm is that it exploits the online observations of buffer length and its variation for RC, requiring no prior knowledge of the channel variations and video characteristics. Our simulation results recorded in multiuser scenarios demonstrated that the proposed algorithm is capable of accommodating the channel quality variations, which may be beneficial in HSPA and LTE-style environments.

References

[1] L. Hanzo, J. Blogh, and S. Ni, *3G, HSPA and FDD Versus TDD Networking: Smart Antennas and Adaptive Modulation*. Piscataway, NJ, USA: Wiley/IEEE Press, 2008.

[2] Third-Generation Partnership Project (3GPP), "Physical layer procedures (FDD)," ETSI, Sophia Antipolis, France, 3GPP TS25.214 V6.2.0, 2004.

[3] M. Chen and A. Zakhor, "Rate control for streaming video over wireless," *IEEE Wireless Commun.*, vol. 12, no. 4, pp. 32–41, Aug. 2005.

[4] T. Chang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 246–250, Feb. 1997.

[5] Z. G. Li *et al.*, " A unified architecture for real-time video-coding systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 472–487, Jun. 2003.

[6] Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control of H.264/AVC scalable extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 116–121, Jan. 2008.

[7] "Joint Video Team Software JM 18.0." [Online]. Available: http://iphome.hhi.de/suehring/tml/

[8] Z. Li *et al.*, "Adaptive basic unit layer rate control for JVT," presented at the 7th Meeting, Pattaya II, Pattaya, Thailand, Mar. 2003, Paper JVTG012r1.

[9] L. Hanzo, P. Cherriman, and J. Streit, *Video Compression and Communications*. Hoboken, NJ, USA: Wiley, 2007.

[10] Y. Huang, S. Mao, and S. F. Midkiff, "A control-theoretic approach to rate control for streaming videos," *IEEE Trans. Multimedia*, vol. 11, no. 6, pp. 1072–1081, Oct. 2009.

[11] H. L. Lin, T. Y. Wu, and C. Y. Huang, "Cross layer adaptation with QoS guarantees for wireless scalable video streaming," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1349–1352, Sep. 2012.

[12] N. Changuel, N. Mastronarde, M. Van der Shaar, B. Sayadi, and M. Kieffer, "Adaptive scalable layer filtering process for video scheduling over wireless networks based on MAC buffer management," in *Proc. IEEE ICASSP*, May 2011, pp. 2352–2355.

[13] A. Dua, C. W. Chan, N. Bambos, and J. Apostolopoulos, "Channel, deadline, and distortion (CD^2) aware scheduling for video streams over wireless," *IEEE Trans. Wireless Commun.*, vol. 9, no. 3, pp. 1001–1011, Mar. 2010.

[14] H. Zhang, Y. Zheng, M. A. Khojastepour, and S. Rangarajan, "Cross-layer optimization for streaming scalable video over fading wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 344–353, Apr. 2010.

[15] E. Piri, M. Uitto, J. Vehkapera, and T. Sutinen, "Dynamic cross-layer adaptation of scalable video in wireless networking," in *Proc. IEEE GLOBECOM*, Dec. 2010, pp. 1–5.

[16] R. Radhakrishnan and A. Nayak, "Cross layer design for efficient video streaming over LTE using scalable video coding," in *Proc. IEEE ICC*, Jun. 2012, pp. 6509–6513.

[17] Video Trace Library. [Online]. Available: http://trace.eas.asu.edu/

[18] C. Chen, R. W. Heath Jr., A. C. Bovik, and G. D. Veciana, "A Markov decision model for adaptive scheduling of stored scalable videos," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 23, pp. 1081–1095, Jun. 2013.

[19] A. Dembo, and O. Zeitouni, *Large Deviations Techniques and Applications*. 2nd ed. Berlin, Germany: Springer-Verlag, 1998.

[20] Apple HLS, "HTTP Live Streaming draft-pantos-http-live-streaming-08 (IETF draft)."

[21] Third-Generation Partnership Project (3GPP), "Transparent end-to-end Packet Switched Streaming Service (PSS); protocols and codecs," ETSI, Sophia-Antipolis, France, 3GPP TS 26.234, 2010.

[22] C. Hsu, and M. Hefeeda, "Partitioning of multiple fine-grained scalable video sequences concurrently streamed to heterogeneous clients," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 457–469, Apr. 2008.

[23] Z. Chen, M. Li, and Y. Tan, "Perception-aware multiple scalable video streaming over WLANs," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 675–678, Jul. 2010.

[24] Third-Generation Partnership Project (3GPP), "Transparent end-to-end Packet Switched Streaming Service (PSS); progressive download and dynamic adaptive streaming over HTTP (3G)," ETSI, Sophia-Antipolis, France, 3GPP TS 26.247, 2011.

[25] *Information Technology-Dynamic Adaptive Streaming Over HTTP (DASH)—Part 1: Media Presentation Description and Segment Formats*, ISO/IEC 23009-1:2012, 2012,

[26] G. Tian, and Y. Liu, "Towards agile and smooth video adaptation in dynamic HTTP streaming," in *Proc. ACM CoNEXT*, Dec. 2012, pp. 109–120.

[27] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE," in *Proc. ACM CoNEXT*, Dec. 2012, pp. 97–108.

[28] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," *Proc. ACM MOBICOM*, Sep. 2013, pp. 389–400.

[29] J. Little, "A proof of the queuing formula: $L = \lambda W$," *Oper. Res.*, vol. 9, no. 3, pp. 383–387, May 1961.

[30] A. Wilig, "A Short Introduction to Queueing Theory," 1999. [Online]. Available: http://www.cs.ucf.edu/lboloni/Teaching/EEL6785_Fall2010/slides/QueueingTheory.pdf AQ3

[31] F. Ishizaki, and F. Takine, "Loss probability in a finite discrete-time queue in terms of the steady state distribution of an infinite queue," *Queue Syst.*, vol. 31, no. 3/4, pp. 317–326, Jul. 1999.

[32] N. B. Shroff and M. Schwartz, "Improved loss calculations at an ATM multiplexer," *IEEE/ACM Trans. Netw.*, vol. 6, no. 4, pp. 411–421, Aug. 1998.

[33] H. S. Kim and N. B. Shroff, "Loss probability calculations and asymptotic analysis for finite buffer multiplexers," *IEEE/ACM Trans. Netw.*, vol. 9, no. 6, pp. 755–768, Dec. 2001.

[34] H. Bobarshad, M. van der Schaar, A. H. Aghvami, R. S. Dilmaghani, and M. R. Shikh-Bahaei, "Analytical modeling for delay-sensitive video over WLAN," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 401–414, Apr. 2012.

[35] S. T. Cheng, M. H. Tao, and C. Y. Wang, "Adaptive channel switching for centralized MAC protocols in multihop wireless networks," *IEEE Trans. Commun.*, vol. 58, no. 1, pp. 228–234, Jan. 2010.

[36] H. Bobarshad and M. Shikh-Bahaei, "M/M/1 queuing model for adaptive cross-layer error protection in WLANs," in *Proc. IEEE WCNC*, Nov. 2009, pp. 1–6.

[37] Y. Xu *et al.*, "Probabilistic analysis of buffer starvation in Markovian queues," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1826–1834.

[38] M. Mandjes, *Large Deviations for Gaussian Queues: Modelling Communications Networks*. Hoboken, NJ, USA: Wiley, 2007.

[39] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis Forecasting and Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1994.

[40] Motorola and Nokia, "Revised CQI proposal," 3GPP RAN WG1, Sophia, Antipolis, France, Tech. Rep. R1-02-0675, Apr. 2002.

[41] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, 4th Quart. 2012.

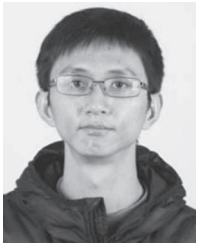
[42] Gabin *et al.*, "3GPP mobile multimedia streaming stands," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 134–138, Nov. 2010.

[43] C. H. Lin, C. H. Ke, C. K. Shieh, and N. K. Chilamkurti, "The packet loss effect on MPEG video transmission in wireless networks," in *Proc. IEEE AINA*, Apr. 2006, pp. 565–572.

Jian Yang (M'08) received the B.S. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2006, respectively.

From 2006 to 2008, he was a Postdoctoral Scholar with the Department of Electronic Engineering and Information Science, USTC. Since 2008, he has been an Associate Professor with the Department of Automation, USTC. His current research interests include distributed system design, modeling and optimization, multimedia over wired and wireless, and future networks.

Dr. Yang received the Lu Jia-Xi Young Talent Award from the Chinese Academy of Sciences in 2009.



Yongyi Ran received the B.S. and Ph.D. degrees in 2008 and 2014, respectively, from the University of Science and Technology of China, Hefei, China, where he is currently a Postdoctoral Researcher.

His research interests include cloud computing, service management, future networks, and stochastic optimization.

Shuangwu Chen received the B.S. degree in 2011 from the University of Science and Technology of China, Hefei, China, where he is currently working toward the Ph.D. degree with the School of Information Science and Technology.

His research interests include multimedia communications, future networks, and stochastic optimization.

Weiping Li (F'00) received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, China, in 1982 and the M.S. and Ph.D. degrees from Stanford University, Stanford, CA, USA, in 1983 and 1988, respectively, all in electrical engineering.

In 1987, he joined the Faculty of Lehigh University, Bethlehem, PA, USA, as an Assistant Professor with the Department of Electrical Engineering and Computer Science. In 1993, he was promoted to Associate Professor with Tenure. In 1998, he was promoted to Full Professor. From 1998 to 2010, he was with several high-technology companies in the Silicon Valley, where he had technical and management responsibilities. In March 2010, he returned to the USTC and is currently a Professor with the School of Information Science and Technology.

Prof. Li was elected Fellow of IEEE for contributions to image and video coding algorithms, standards, and implementations. He served as a member of Moving Picture Experts Group (MPEG) of the International Organization for Standardization (ISO) and an Editor of MPEG-4 International Standard. He served as a Founding Member of the Board of Directors of the MPEG-4 Industry Forum. His inventions on fine granularity scalable video coding and shape adaptive wavelet coding have been included in the MPEG-4 International Standard. As a Technical Adviser, he also made contributions to the Chinese Audio Video Coding Standard and its applications. He served as the Chair of several technical committees within the IEEE Circuits and Systems Society and at IEEE international conferences. He served as the Chair of the Best Student Paper Award Committee for the SPIE Visual Communications and Image Processing Conference. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He served as a Guest Editor for a special issue of the PROCEEDINGS OF THE IEEE. He has made many contributions to international standards. He received the Certificate of Appreciation from ISO and the International Electrotechnical Commission as a Project Editor in the development of International Standard in 2004, the Spira Award for Excellence in Teaching in 1992 from Lehigh University, and the first Guo Moruo Prize for Outstanding Student in 1980 from the USTC.

Lajos Hanzo (M'91–SM'92–F'04) received the Master's degree in electronics, the Ph.D. degree, and the *Doctor Honoris Causa* degree from the Technical University of Budapest, Budapest, Hungary, in 1976, 1983, and 2009 respectively, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2004.

During his career in telecommunications, he has held various research and academic posts in Hungary, Germany, and the U.K. Since 1986, he has been with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., where he holds the Chair in telecommunications. He was a Chaired Professor of Tsinghua University, Beijing, China. He is the coauthor of 20 John Wiley/IEEE Press books on mobile radio communications, totalling in excess of 10 000 pages, and has published more than 1400 research entries on IEEE Xplore. He is currently directing an academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) U.K., the European IST Program, and the Mobile Virtual Centre of Excellence, U.K. He is an enthusiastic supporter of industrial and academic liaison, and he offers a wide range of industrial courses.

Dr. Hanzo has acted as a Technical Program Committee Chair for IEEE conferences, presented keynote lectures, and has received a number of distinctions. He is the Governor of the IEEE Vehicular Technology Society and the Past Editor-in-Chief of the IEEE Press.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

AQ1 = Please check if “Star War IV” should be “Shawshank Redemption”
AQ2 = Please check if changes made in this sentence are appropriate.
AQ3 = Provided URL in Ref. [30] was not found. Please check.

END OF ALL QUERIES