CBM: Online Strategies on Cost-Aware Buffer Management for Mobile Video Streaming

Jian He, Zheng Xue, Di Wu, *Member, IEEE*, Dapeng Oliver Wu, *Fellow, IEEE*, and Yonggang Wen, *Member, IEEE*

***Abstract*—Mobile video traffic, owing to the rapid adoption of smartphones and tablets, has been growing exponentially in recent years and started to dominate the mobile Internet. In reality, mobile video applications commonly adopt buffering techniques to handle bandwidth fluctuation and minimize the impact of stochastic wireless channels on user experiences. However, recent measurement work reveals that mobile users tend to abort more frequently than PC users during viewing videos. Such a high abortion rate results in a significant wastage of buffered video data, which is directly translated into monetary and energy cost for mobile users. In this paper, we propose an intelligent buffer management strategy called *CBM (Cost-aware Buffer Management)*, for mobile video streaming applications. Our purpose is to minimize cost induced by un-consumed video data while respecting certain user experience requirements. To this objective, we formulate the problem into a constrained stochastic optimization problem, and apply the Lyapunov optimization theory to derive the corresponding online strategy for cost minimization. Different from conventional heuristic-based strategies, our proposed CBM strategy can provide provably performance guarantee with explicit bounds. We also conduct extensive simulations to validate the effectiveness of our proposed strategy and our experimental results show that CBM achieves significant gains over existing schemes.**

***Index Terms*—Mobile video streaming, buffer management, energy consumption, bandwidth cost, Lyapunov optimization.**

I. Introduction

RECENT years have witnessed an exponential growth of the mobile Internet traffic, mostly driven by mobile video. Cisco predicted in [1] that mobile data traffic would increase by 18 times until 2016. Video traffic generated by mobile users will be the leading contributor to such explosive growth, reaching more than 70% of the total mobile data traffic with a compound annual growth rate (CAGR) of 78%. However, user experience of mobile video streaming applications is impacted by user mobility and the stochastic nature of wireless channels. A simple yet effective mechanism to mitigate the effect of channel fluctuation is to buffer a certain amount of video data on the mobile client, thus to reduce the possibility of playback freezing [2]. It has been verified that rate adaptive streaming with buffer support can significantly improve the QoE (Quality of Experience) of mobile video viewers [3].

However, the viewing patterns of mobile users will result in a wastage of system resources for mobile video streaming applications. Specifically, such wastage results from the un-consumed video data in the application buffer, which in turn is caused by random abortion of viewers. To reduce the possibility of buffer starvation during the playback period, mobile video applications (e.g., Youtube) tend to download video data aggressively in a periodical manner [4]. A recent measurement study [5] also suggested the viewing time of mobile users can be approximated by long-tail distribution. This can be translated into non-negligible random abortion probability, resulting in a significant fraction of video data unconsumed in the video playback buffer of mobile devices.

Moreover, this non-ideal buffer management would have direct impact on system design. *First, it would cause extra energy consumption for mobile devices*. In fact, it has been shown that video streaming service has been a new killer of battery energy of mobile devices [6]. The residual data buffered in the mobile client cost energy to receive from the source, including transferring data and keeping the WNIC (Wireless Network Interface Card) active [7]. *Second, the unconsumed video data in the buffer would induce a higher monetary cost than necessary for mobile users.* In reality, mobile users with 3G connectivity are charged by their data usage, depending on various pricing plans specified by ISPs [8]. The amount of unconsumed video data causes extra monetary cost for mobile users.

As a result, energy consumption and monetary cost induced by downloading un-consumed video data demand smarter buffer management strategies for mobile video applications to minimize wasted data while respecting a decent QoE for mobile video viewers. There exists a dedicate tradeoff between the minimization of wasted data and the improvement of user QoE. If always keeping the playback buffer at a rather low level to reduce the chance of data wastage, the playback buffer will be at a high risk of being drained out due to the time-varying nature of channel conditions, which directly results in frequent freezing and deteriorates the user QoE accordingly. However,

Manuscript received February 28, 2013; revised June 21, 2013 and August 21, 2013; accepted August 26, 2013. Date of publication October 09, 2013; date of current version December 12, 2013. This work was supported in part by the NSFC under Grant 61003242, Grant 61272397, the Fundamental Research Funds for the Central Universities under Grant 12LGPY53, Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S20120011187, the Program for New Century Excellent Talents in University under Grant NCET-11-0542, the US National Science Foundation under grant ECCS-1002214, CNS-1116970, the Joint Research Fund for Overseas Chinese Young Scholars under Grant 61228101, NTU SUG and MOE Tier 1 (RG 31/11). The corresponding author is Di Wu. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chia-Wen Lin.

J. He, Z. Xue, and D. Wu are with the Department of Computer Science, Sun Yat-sen University, Guangzhou, China (e-mail: hejian9@mail2.sysu.edu.cn; xuezh@mail2.sysu.edu.cn; wudi27@mail.sysu.edu.cn).

D. O. Wu is with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA (e-mail: wu@ece.ufl.edu).

Y. Wen is with the School of Computer Engineering, Nanyang Technological University, Singapore (e-mail: ygwen@ntu.edu.sg).

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/TMM.2013.2284894

if maintaining the playback buffer at a high level to mitigate the short-term dynamics of channel conditions and thus improve the user QoE, it will cause a significant amount of data wastage due to random abortion of viewers (e.g., video browsing). Previous work focused more on minimizing energy consumption of mobile applications by traffic shaping [9], proxying [10], etc. However, the problem of over-buffering in mobile video applications and its implications to user cost have not received enough attention in the field.

In this paper, we propose an intelligent cost-aware buffer management strategy, called *CBM (Cost-aware Buffer Management)*, for mobile video applications. Our design objective is to minimize sunk cost (including energy cost and monetary cost) induced by downloading unconsumed video data, while ensuring a decent viewing quality for mobile viewers. To the best of our knowledge, our work is the first to consider the over-buffering problem for mobile video streaming applications. Mathematically, the above-mentioned buffer management problem can be formulated as a constrained stochastic optimization problem, which can be solved efficiently by using the Lyapunov optimization theory [11]. By applying the Lyapunov optimization theory, the stochastic constraints can be transformed into queue stability constraints. Different from previous heuristic-based strategies without performance guarantee, our proposed **CBM** strategy can provide explicit performance bounds and approach the optimality with tunable distance. By extensive simulations, we show that our proposed **CBM** strategy significantly reduces the extra cost induced by aggressive downloading and achieves comparable viewing quality for mobile users.

The rest of the paper is organized as follows. Section II reviews previous work. Section III provides the formulation of the cost-aware buffer management problem. The design of online strategy is introduced in Section IV. Simulation results are given in Section V. Section VI concludes the paper and discusses some future work.

II. Related Work

Mobile video streaming, due to the growing popularity of mobile devices (e.g., iPhones and Android phones), has been gaining a lot of attention from the research community. Previous work on mobile video streaming can be mainly divided into two categories: (1) measurement studies of existing popular mobile video streaming applications, and (2) design of new streaming protocols with improved user experience and higher resource utilization.

Most previous measurement work aims to analyze various system properties of mobile video streaming applications. Balasubramanian *et al.* [7] investigated the component breakdown of energy consumption during data transfer, and proposed an energy model for mobile devices based on obtained results. Other measurement work (e.g., [4], [6], [12], [13]) also investigated energy consumption of mobile devices. Finamore *et al.* [5] and Li *et al.* [14] analyzed viewing behaviors of mobile users and frequent random abortion of mobile users was observed during video viewing. Liu *et al.* [15] measured several key properties of mobile video streaming systems, including distribution of mobile devices, video length distribution, etc.

Researchers also conducted quite a few studies on comparing and improving mobile video streaming protocols. To meet the stringent QoS requirements of wireless multimedia streaming, cross-layer optimization and its effectiveness have been studied in [16]. Liu *et al.* [4] compared the architecture and performance of four typical video streaming protocols, including RTSP streaming, Pseudo streaming, Chunk-based streaming and P2P streaming, via measurement studies. In [12], an energy-efficient streaming protocol was proposed to increase the probability that the WNIC can be switched to the power-saving mode. The idea is to utilize burst downloading [17] opportunistically.

The design of buffer management strategies is critical for mobile applications. The objectives of buffer management could be quite diverse, depending on application scenarios. In [18]–[21], different buffer management strategies were proposed to minimize energy consumption of mobile users. In [22], Mastronarde *et al.* proposed to use a reinforcement learning algorithm to learn the optimal buffer control policy at run-time and improve the multimedia application performance dramatically. The strategy proposed in [23] considers the insurance of QoS over stochastic wireless channels. Our work in this paper differs from previous work in that our research is motivated by the real problem unveiled in the measurement work [5], which observed that a significant fraction of buffered data is wasted with no gain. To this end, we propose an intelligent cost-aware online buffer management strategy to minimize data wastage, harnessing the power of random scheduling across channel conditioning, playback rates and user behaviors. The use of the Lyapunov optimization framework can avoid the limitation of inaccurate prediction (or learning) of channel conditions and simplify the algorithm design greatly. Lyapunov optimization-based algorithm only needs to observe channel conditions at the beginning of each time slot, while conventional prediction-based algorithms (such as MDP-based algorithms) require to make long-term prediction on channel conditions in order to approach optimality. Although learning-based algorithms [25] can mitigate the need of long-term prediction, the high computation complexity, induced by state update and stochastic approximation, makes them unsuitable to be implemented on mobile devices.

III. Problem Formulation

A. System Model

In this section, we consider a typical mobile video streaming system as illustrated in Fig. 1, in which mobile devices (e.g., smartphones, tablets) directly obtain video streams from the streaming server via their wireless interfaces. On each mobile device, a video playback buffer is maintained to achieve smooth playback and is controlled by a smart buffer manager. The buffer manager aims at requesting video chunks properly according to the observed status information (e.g., available download bandwidth, current buffer state, pricing plan, energy consumption, etc.). Our objective in this paper is to design an intelligent buffer management strategy to minimize the cost induced by unconsumed downloaded data while still respecting the QoE requirements of video playback.

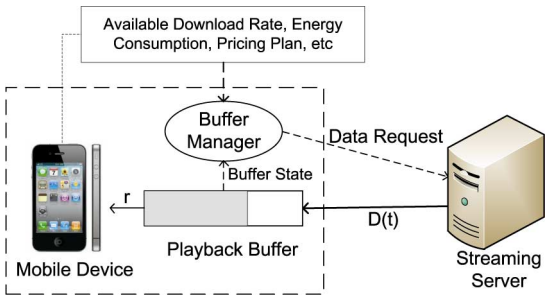


Fig. 1. A Generic Model of Mobile Video Streaming Service.

For a video streaming service, assume a video sequence is compressed with a constant bit rate r bits/sec;[1] the video sequence is divided into equal-size chunks with a size of $r \times T_0$ bits, where T_0 is the length of a time slot. The system is assumed to operate in discrete time with time slots $t \in \{0, 1, 2, \ldots, T\}$, where T is the maximum number of time slots needed to transmit the video sequence. The maximal download rate of a mobile device in slot t, denoted by $D(t)$, depends on the fading channel gain. We assume that the channel gain keeps constant within one time slot but can vary across time slots.

Define B as the maximum buffer size and $B(t)$ as the buffer size (in the unit of bytes) at the beginning of time slot t. The evolution of $B(t)$ is also a stochastic process with chunk arrival and departure. Suppose that a viewer starts its playback after pre-buffering is completed and continues to view the video until abortion. We assume a viewer can abort at the end of any time slot (say t) with a probability $p(t)$, which is determined by user viewing behaviors. At the time of abortion, all the unconsumed video data in the buffer is actually wasted with no gain.

Let $w(t)$ be the amount of wasted data when the viewer aborts its viewing at the end of time slot t, and $H(w(t))$ be the cost associated with $w(t)$. Note that the function $H(\cdot)$ is a generic cost function that can be defined according to user sensitivity to different types of cost. Such cost induced by wasted data (i.e., $H(w(t))$) is referred as the "*user-defined sunk cost*". Users can define the structure of $H(\cdot)$ to measure the cost incurred by wasted data. For example, $H(\cdot)$ can be defined as the battery energy consumption for downloaded bytes, or monetary cost charged for downloaded bytes, etc.

Meanwhile, we should also take user experiences into account when designing the buffer management strategy. In this paper, we mainly focus on the measure of freezing time, as it directly impacts user experiences. Freezing happens when the playback buffer runs empty. If the arrival rate of video chunks keeps being lower than the play-out rate of video chunks, the playback buffer will be drained out and a user will experience a period of freezing time. Denote $f(t)$ as the freezing time (in the unit of seconds) in one time slot t, and then the total freezing time can be given by $\sum_{t=0}^{T-1} f(t)$. Let η be the tolerance ratio of freezing time, which is defined as the maximum fraction of freezing time that can be tolerated by a viewer. It is essential to guarantee that the fraction of freezing time is less than the tolerance ratio η. To this objective,

[1]Usually, a streaming video (for transmission over networks) is compressed with a constant bit rate; for DVD, a video sequence is compressed with a variable bit rate.

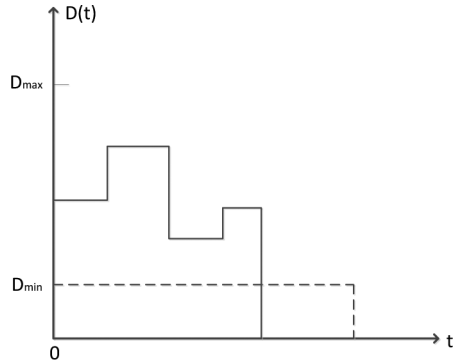


Fig. 2. Two downloading processes.

at the beginning of each slot t, the buffer manager needs to carefully determine the number of video chunks to be downloaded in that time slot, denoted by $n(t)$.

In addition, different from wireline environments, most data transmission protocols in the wireless environments have a power-saving mode. The WNIC consumes less power when switching from the "*active*" mode to the "*power-saving*" mode. According to the energy usage model in [7], the amount of energy usage can be defined as a linear function of the amount of downloaded data and the duration of the WNIC active period, namely,

$$\mathbb{E}(\tau, s) = \tau \cdot e_a + s \cdot e_d,$$

where $\mathbb{E}(\tau, s)$ is the amount of energy usage, τ is the duration of the WNIC active period, e_a is the average power consumption per unit time to keep the WNIC active, s is the amount of downloaded data, and e_d is the average power consumption per unit downloaded data. For a given amount of video data to be downloaded, it is preferable to quickly complete data download and switch the WNIC to the power-saving mode in order to save energy. Note that, there is no way to further reduce energy consumption if the WNIC is in the power-saving mode. Similar to previous work, we only consider energy consumption of the WNIC in the active mode.

To evaluate the efficiency of power saving, we define a *power-saving utility function* $U(\cdot)$ as below:

$$U(n(t), D(t)) = e_a \cdot \left(\frac{n(t) r T_0}{D_{min}} - \frac{n(t) r T_0}{D(t)}\right)$$

where D_{min} is the lower bound of download rate. $U(n(t), D(t))$ is the expected energy saving if downloading $n(t)$ video chunks with rate $D(t)$, which is compared with the case when downloading video data with the minimal download rate D_{min}.

To better understand the power-saving utility function, we plot two different downloading processes in Fig. 2. For the process drawn in the dashed line, the client downloads data with a constant minimal rate $D_{\min}$. In another process drawn in the solid line, the client's download rate fluctuates with time. Given the same amount of data to be downloaded, it is obvious that the former process (plotted in dashed lines) costs the WNIC to spend more time in the active state and consumes more energy accordingly. Thus, the former process can serve as a benchmark to evaluate the reduction of the WNIC active duration for other downloading processes. The direct implication is that more energy can be saved if we can minimize the download time.

B. Cost-Aware Buffer Management Problem

The update of the playback buffer can be defined as below:

$$B(t+1) = \max\{B(t) - r \cdot (T_0 - f(t)) + n(t) \cdot r \cdot T_0, 0\}.$$

In the above equation, the term $r \cdot (T_0 - f(t))$ stands for the amount of consumed video data in a time slot, in which r is the video playback rate. The term $n(t) \cdot r \cdot T_0$ is the amount of video data downloaded in a time slot, in which $n(t)$ is the number of video chunks downloaded in a time slot, r is the video encoding rate (which also equals to the playback rate) and $r \cdot T0$ is the size of a video chunk. In general, $w(t)$ can be expressed as a function of $n(t)$, namely, $w(t) = w(n(t))$.

Considering the above constraints, the cost-aware buffer management problem can be formulated into a stochastic optimization problem as below:

$$\begin{aligned} \min_{\{n(t)\}} \quad & \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} H(w(n(t))) \cdot p(t) \\ \text{s.t.} \quad & 0 \le B(t) \le B, \forall t \\ & n(t) \le \frac{D(t)}{r}, \forall t \\ & \lim_{T\to\infty} \frac{\sum_{t=0}^{T-1} f(t)}{T \cdot T_0} \le \eta \\ & \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} U(n(t), D(t)) \ge \alpha \cdot \bar{e} \cdot T_0 \end{aligned} \tag{1}$$

where the objective function is the time-average expected cost incurred by unconsumed data given that a viewer aborts at time slot t, and the expectation is calculated with respect to the probability distribution of $p(t)$. α represents a power-saving utility threshold. The first constraint indicates the buffer size should be no greater than its capacity at any time; the second constraint ensures that the number of requested chunks cannot be greater than the limit that a viewer can download in a time slot; the third constraint can guarantee the average fraction of freezing time is less than the tolerance ratio; the last constraint is to ensure the power-saving utility no less than a certain threshold α.

Before explicitly defining the freezing time at time slot t, we introduce three additional indicator functions as below:

$$1_N(t) = \begin{cases} 0, & n(t) = 0 \\ 1, & n(t) > 0 \end{cases}$$

$$1'_N(t) = \begin{cases} 0, & n(t) < 1 \\ 1, & n(t) \ge 1 \end{cases}$$

$$1_B(t) = \begin{cases} 1, & B(t) < rT_0 \\ 0, & B(t) \ge rT_0. \end{cases}$$

Lemma III.1: The freezing time in time slot t can be represented by the following expression:

$$\begin{aligned} f(t) = & 1_B(t) \cdot \max\{\frac{1'_N(t) r T_0}{D(t)} + \\ & \frac{(1 - 1'_N(t)) n(t) r T_0}{D(t)} - \frac{B(t)}{r}, 0\} + \\ & (1 - 1_N(t)) 1_B(t) (T_0 - \frac{B(t)}{r}). \end{aligned}$$

Proof: Please see Appendix A in our technical report [24] for the proof details. ■

The above-formulated problem differs from traditional convex optimization problems in that, our problem involves optimizing the time average of cost function subject to time average constraints.

IV. Design of Cost-Aware Online Buffer Management Strategies

To solve the constrained stochastic optimization in the above section, we exploit Lyapunov optimization theory to design online control strategies. A major benefit of Lyapunov optimization is that it does not require any priori knowledge about user behaviors and download rates. By taking actions to greedily minimize the drift-plus-penalty in each time slot, it can provide performance with explicit bounds.

A. Lyapunov Optimization

In the framework of Lyapunov optimization, the original stochastic optimization problem can be transformed to an optimization problem of minimizing the Lyapunov drift-plus-penalty. By using Lyapunov optimization, the time average constraints in Problem (1) can be transformed into a set of queue stability constraints [11].

Two virtual queues $F(t)$ and $G(t)$ are defined to transform the time average constraints on freezing time and power-saving utility into queue stability constraints. The updates of queues $F(t)$ and $G(t)$ are given by:

$$\begin{aligned} F(t+1) &= \max[F(t) + f(t) - \eta \cdot T_0, 0] \\ G(t+1) &= \max[G(t) - U(n(t), D(t)) + \alpha \cdot e_a \cdot T_0, 0]. \end{aligned}$$

Therefore, Problem (1) can be transformed into the following equivalent problem:

$$\begin{aligned} \min_{\{n(t)\}} \quad & \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} H(w(n(t))) \cdot p(t) \\ \text{s.t.} \quad & 0 \le B(t) \le B, \forall t \\ & n(t) \le \frac{D(t)}{r}, \forall t \\ & \lim_{t\to\infty} \frac{F(t)}{t} \le 0 \\ & \lim_{t\to\infty} \frac{G(t)}{t} \le 0. \end{aligned} \tag{2}$$

The last two constraints are used to guarantee the stability of virtual queues.

Denote the queue vector by $\mathbf{\Theta}(t) = (B(t), F(t), G(t))$, and define a quadratic Lyapunov function $L(\mathbf{\Theta}(t)) = 1/2[B(t)^2 + F(t)^2 + G(t)^2]$ to measure the size of the queue vector. The drift $\Delta(\mathbf{\Theta}(t))$ is defined as the expected change in the Lyapunov function over one time slot. Therefore, the solution to the original stochastic optimization problem can be approximately obtained by solving the problem of minimizing drift-plus-penalty $\Delta(\mathbf{\Theta}(t)) + VE\{O(t)|\mathbf{\Theta}(t)\}$ in each time slot, with an explicit approximation upper bound. The penalty $O(t)$ is the value of the objective function in the original

optimization problem (Note that $O(t) = H(w(t)) \cdot p(t)$ in our problem), and V is a tunable parameter that affects the performance of the online algorithm.

By exploiting Lyapunov optimization techniques, online strategy for Problem (1) can be obtained by solving the following problem in each time slot t:

$$\min_{\{n(t)\}} \quad \Delta(\mathbf{\Theta}(t)) + V \cdot H(w(n(t))) \cdot p(t) \tag{3}$$

$$\text{s.t.} \quad n(t) \leq \frac{D(t)}{r}, \forall t \tag{3a}$$

$$n(t) \leq \frac{B - \eta \cdot rT_0}{r \cdot T_0} \tag{3b}$$

$$G(t) - U(n(t), D(t)) + \alpha e_a T_0 \leq G \tag{3c}$$

$$F(t) + f(t) - \eta T_0 \leq F \tag{3d}$$

$$B(t) - r \cdot (T_0 - f(t)) + n(t) r \cdot T_0 \geq 0 \tag{3e}$$

where $\Delta(\mathbf{\Theta}(t)) = G(t) \cdot (-U(n(t), D(t))) + F(t) \cdot f(t) + B(t) \cdot (n(t) r T_0 + r f(t))$. In the above problem, Constraint (3b) is to ensure the buffer capacity constraint. Constraints (3c) and (3d) are used to ensure upper bounds of the occupancy of queues $F(t)$ and $G(t)$. The selection of F and G should guarantee that they are no less than the upper bounds of queues $F(t)$ and $G(t)$ under the optimal solution of Problem (1).

The main purpose of the transformation from Problem (1) to Problem (3) is to decrease the computation complexity of the optimization problem. By applying the Lyapunov optimization framework, the original complex optimization problem can be transformed into a simple optimization problem.

Before analyzing the performance bound of the online algorithm, we first prove the following lemma.

Lemma IV.1: The optimal strategy of Problem (3) can guarantee that

$$0 \leq B(t) \leq B, \quad \forall t. \tag{4}$$

Proof: Please see Appendix B in our technical report [24] for the proof details. ■

Denote the optimal solution of Problem (1) by $\hat{H} = 1/T \sum_{t=0}^{T-1} H(w(t))$ and we can obtain the performance bound with the following theorem.

Theorem IV.2: The time average cost incurred by the online algorithm derived by solving Problem (3) is within $[X + B \cdot (D_{max} T_0^2 + T_0) + (F + G\alpha e_a) T_0]/V$ of the optimal value, namely,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} H(w(n(t))) \leq \hat{H} + \frac{X + B \cdot (D_{max} T_0^2 + T_0)}{V} + \frac{(F + G\alpha e_a) T_0}{V} \tag{5}$$

if $D_{max} \leq B - \eta r \cdot T_0 / T_0$, where $X = 1/2(\eta \cdot T_0)^2 + 1/2(B - \eta \cdot r T_0)^2 + 1/2(\alpha e_a T_0)^2$, and D_{max} is the upper bound of download rate.

Proof: Please see Appendix C in our technical report [24] for the proof details. ■

Theorem IV.2 implies that the online algorithm can approximately approach the optimal solution of the original optimization problem within infinitely small distance. The distance to the optimality is determined by a tuning parameter V. For example, if $H(w(t))$ is a linear function of $w(t)$ and the viewer can tolerate the amount of wasted data less than $\hat{w}(t) + \epsilon$ at each time slot t, where $\hat{w}(t)$ is the amount of wasted data under the optimal solution of Problem (1), then the tunable parameter V can be derived as below:

$$V = \frac{X + B(D_{max} T_0^2 + T_0) + (F + G\alpha e_a T_0)}{H(\epsilon)}.$$

B. Cost Model

The cost function can be tailored to include different types of cost. Generally, there are two major types of cost for mobile users:

- *Bandwidth cost*, $C(w(t))$, which is the monetary cost incurred by the wasted data usage $w(t)$ at time slot t. For the usage-based data plan, the bandwidth cost can be defined as a linear function of $w(t)$, namely, $C(w(t)) = c_t \cdot w(t)$, where c_t is the price per unit data in time slot t. Normally, c_t is constant over time. However, under the time-dependent dynamic pricing strategy [26], c_t is constant within one time slot but varies over time slots. However, we can still obtain unit prices at the beginning of each time slot.
- *Energy cost*, $E(w(t))$, which is the energy consumption for downloading the wasted data $w(t)$. Based on the energy model stated in Section III, the energy consumption consists of the energy used to keep WNIC active and the energy used to download data $w(t)$. According to the measurement work in [7], $E(w(t))$ can be defined as $w(t)/D(t) e_a + w(t) e_d$, $w(t)/D(t)$ represents the duration of active period of WNIC when downloading wasted data $w(t)$ with a rate of $D(t)$ in time slot t. Recall that e_a is the power consumption per unit time to keep WNIC active and e_d is the power consumption per unit downloaded data.

Users can define the structure of $H(w(t))$ based on their own sensitivity to different types of cost. For instance, $H(w(t))$ can be defined simply as $C(w(t))$ or $E(w(t))$.

C. Online Strategy for Cost-Aware Buffer Management

The optimal online strategy can be derived by solving Problem (3). Denote the objective function in Problem (3) as $Y(t) = \Delta(\mathbf{\Theta}(t)) + V \cdot H(w(n(t))) p(t)$. The optimal online strategy is determined by the specific structure of the function $Y(t)$. The structure of $Y(t)$ can be further exploited to reduce the complexity when solving Problem (3). Assume that $Y(t)$ is a continuous function of $n(t)$, we can obtain the following easy-to-implement online strategy, called *Cost-aware Buffer Management Strategy (CBM)*, in which:

1) If $\partial Y(t)/\partial n(t) \geq 0$, then we choose the minimum $n(t)$ that satisfies the constraints in Problem (3).
2) If $\partial Y(t)/\partial n(t) < 0$, then we choose the maximum $n(t)$ that satisfies the constraints in Problem (3).
3) In other case, search all possible values of $n(t)$ that satisfy the constraints in Problem (3). As the feasible region of $n(t)$ is limited, the upper bound of time complexity is given $O(D(t)/r)$.

Fig. 3. Simulation Environment.

For the case that the derivative of $Y(t)$ with respect to $n(t)$ cannot be directly obtained in the range $[0, +\infty]$, we can divide the whole range into multiple subranges and calculate the derivative of $Y(t)$ in each subrange.

Intuitively, if the sizes of two virtual queues $F(t)$ and $B(t)$ are large, the viewer should choose smaller $n(t)$ to decrease the size of queues. If $F(t)$ and $B(t)$ are small, the viewer can choose a large $n(t)$. And if $G(t)$ is large, the viewer tends to choose a large $n(t)$ to guarantee the power-saving utility requirement. In the initialization step, both $F(t)$ and $G(t)$ are set as zero at time $t = 0$.

V. Experimental Evaluation

To evaluate the effectiveness of our proposed online strategy, we developed a discrete-event simulator to simulate the behavior of a video playback buffer of a mobile device under different buffer management strategies and evaluate the quality of user experience when viewing a video.

A. Simulation Setup

Fig. 3 briefly illustrates the simulation environment used in our experiments. To watch a video, a mobile device connects to a video streaming server through the intermediate network infrastructure. The access network can be WiFi, 3G/LTE, etc. Note that, from the perspective of a mobile client, the network side can be treated as a black box with any arbitrary structure. The information about network details is not required for the evaluation of our buffer management strategy. The network condition between the mobile device and the streaming server can be simply modeled as a stochastic process. By adjusting the stochastic process, we can capture the changes of different network parameters.

In our experiments, the video playback rate is set as 320 kbps and the video length is set as 200 seconds. The whole video is divided into equal-size chunks and each chunk has a length of 2 seconds. The maximum buffer capacity on the mobile device is set as the size of 20 video chunks, namely, $B = 1600\ KB$. Assume that the download capacity of a mobile device is bounded by the capacity of wireless channels in the last hop instead of the network core. Considering the fading effects of wireless channels, we adopt a Rayleigh distribution to approximate the stochastic changes of download rates as [27]. The fluctuation of download rates of the mobile device is plotted in Fig. 4. We use the trace-based probability of aborting in the viewing process from the measurement results in [5]. In default, the tolerance ratio of freezing time η is set as 0.03 and the amount of tolerable wasted data per time slot, ϵ, is set as 10 KB. All the parameter settings used in our experiments are summarized in Table I.

For comparison, we also investigate two other buffer management strategies that are being used in existing works, including:

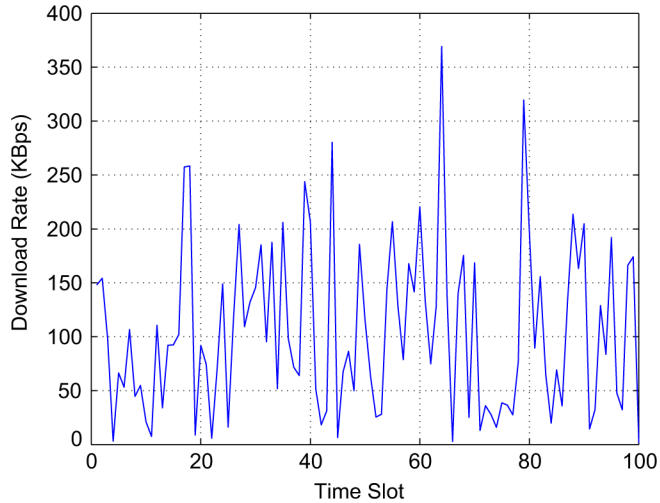


Fig. 4. Download rates of a mobile device (Rayleigh distribution with mean 50 KBps).

TABLE I
Simulation Parameter Settings

Download rate of a mobile device	Rayleigh distribution with a mean of 50KBps
Video playback rate	320kbps
Playback duration of one video chunk	2 seconds
Video length	200 seconds
Playback buffer size	1600KB
Freezing time tolerance η	0.03
Wasted data tolerance ϵ	10KB
Bandwidth price per unit data c_t	1.5 per KB
Average power consumption per unit time to keep WNIC active e_a	20 J/s
Average power consumption per unit downloaded data e_d	0.025 J/byte

- *Aggressive Buffer Management Strategy (ABM)*, in which a mobile video application downloads video data aggressively at each time slot until the buffer is full. *ABM* is commonly used by devices with unlimited power supply for video streaming services.
- *Periodical Buffer Management Strategy (PBM)*, in which a mobile video application requests a fixed number of video chunks from the server periodically and stops downloading if requested chunks have been completely downloaded before the end of the time slot or there is no room in the buffer. The measurement results in [4] show that *PBM* has been widely adopted by mobile devices. In our experiments, the default period under *PBM* is set as T_0 and the viewer requests two video chunks per period.

In our experiments, we evaluate the performance of our proposed algorithm under different types of cost functions. The main purpose to adopt different cost functions is to check the applicability of our buffer management strategy to different user requirements.

B. Buffer Status Analysis

We run each experiment, which lasts for 100 time slots (2 seconds per time slot), several times, and calculate the average results. In this experiment, we evaluate the effectiveness of our online algorithm **CBM** when users are sensitive to bandwidth

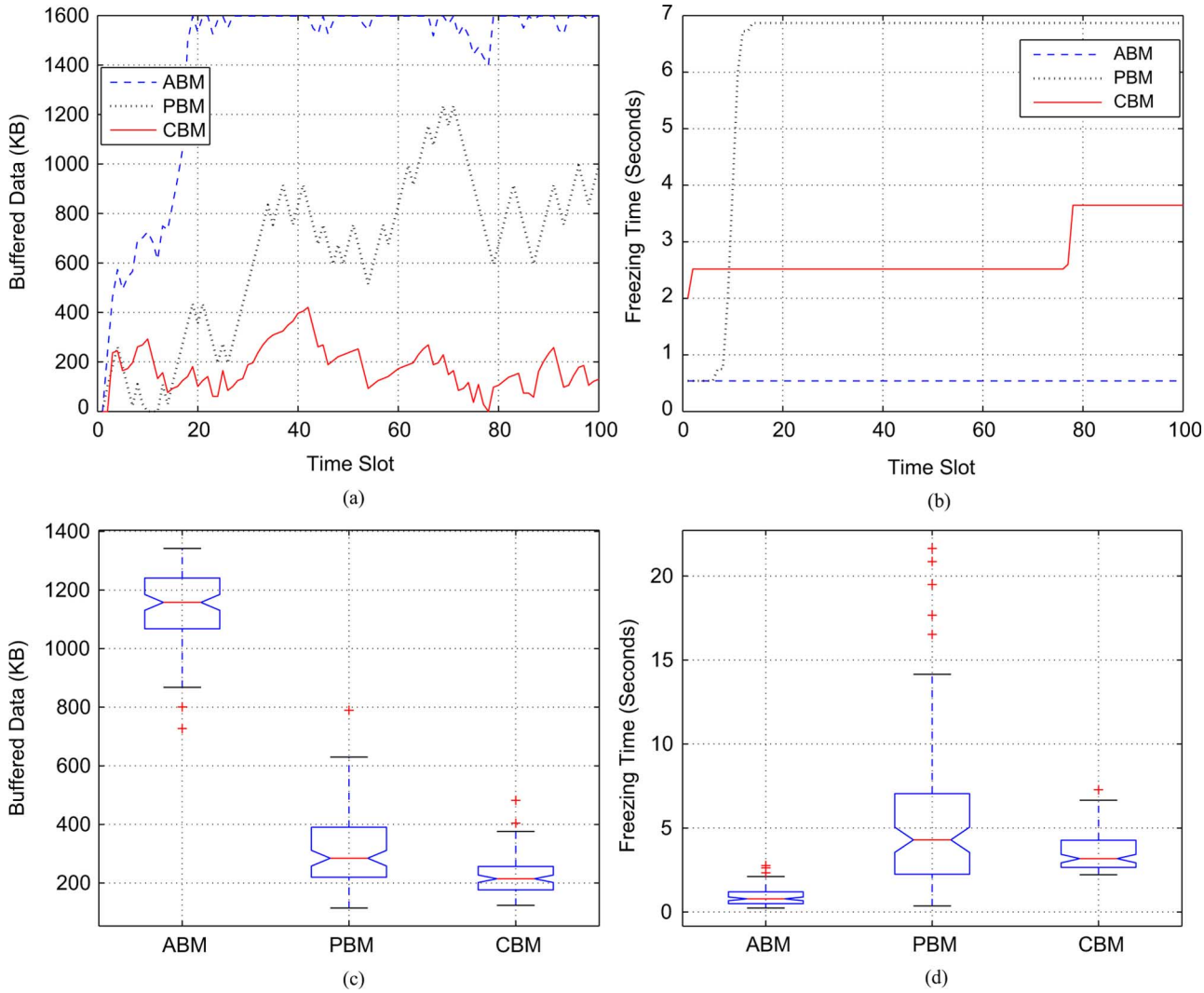


Fig. 5. Buffer state and cumulative freezing time under different strategies with $\mathrm{r} = 40$ KBps, $\eta = 0.03$, $\mathrm{B} = 1600$ KB, download rate follows a Rayleigh distribution with mean 50. Statistic information are obtained from 100 independent experiments with the same configurations. (a) Evolution of Buffer State. (b) Cumulative Freezing Time. (c) Statistics on the Amount of Buffered Data. (d) Statistics on the Cumulative Freezing Time.

cost. We adopt the simplest structure of the bandwidth cost function, namely, $H(w(t)) = w(t)$. In this case, the buffer state at the beginning of each time slot can be considered as the amount of wasted data if the viewer would abort in the coming time slot.

Fig. 5 illustrates the experimental results including buffer state and cumulative freezing time under default configurations. From Fig. 5(a), we can see that our **CBM** strategy keeps the buffer size less than 400 KB. As expected, the ABM strategy almost keeps the buffer full after the initial buffer filling up period, the PBM strategy keeps the buffer size high after the initial start-up period. From Fig. 5(b), the PBM strategy incurs the highest freezing time, and the freezing time of the ABM strategy only increases during the start-up period. Under our **CBM** strategy, it can be observed that the freezing time increases at around time slot 80. This happens because the channel condition in previous few time slots are bad, which is illustrated in Fig. 4, resulting in a temporary buffer drainage. This investigation suggests that our **CBM** strategy can simultaneously incur low buffer level and cumulative freezing time level.

The **CBM** strategy can adaptively make decisions according to buffer state and channel condition. To evaluate this, we investigate the decisions under four different cases:

1) When the channel condition is good (i.e., $D(t)$ is larger than the playback rate) and the amount of buffered data is large (i.e., no less than the size of two video chunks), the request is conserved. For example, the increase rate of the amount of buffered data at time slots 28 and 60 is only around 15 KB per second.
2) When the channel condition is bad and the amount of buffered data is large, the request is conserved. The decrease rate of the amount of buffered data at time slots 42 and 65 is around 30 KB per second. Since the consumption rate of buffered data equals to constant r, high decrease rate indicates little requested data.
3) When the channel condition is good and the amount of buffered data is small (i.e., less than the size of one video chunk), the request is aggressive. The increase rate of the amount of buffered data at time slots 24 and 86 is around 50 KB per second.

4) When the channel condition is bad and the amount of buffered data is small, the request is aggressive. The decrease rate of the amount of buffered data at time slots 23 and 85 is less than 10 KB per second.

Therefore, the **CBM** strategy can efficiently achieve trade-off between user-defined sunk cost and QoE, by adaptively controlling the buffer level according to channel condition and buffer state.

Moreover, our **CBM** strategy can ensure the buffer level larger than one video chunk over most of time slots, while the largest buffer level is much smaller than the other two strategies. The ratio between the cumulative freezing time and the length of the video resulted from our **CBM** strategy is less than 0.02, compared to 0.035 for the PBM strategy. Figs. 5(c) and 5(d) show statistic information obtained from 100 independent simulation experiments. The median of buffer size resulted from our **CBM** strategy is only 70% of that for the PBM strategy and the median of cumulative freezing time incurred by the **CBM** strategy is around 70% of that incurred by the PBM strategy.

C. Sensitivity Analysis

In this subsection, we investigate how the performance of our proposed **CBM** strategy varies as the system parameter changes and users are sensitive to energy cost. For this purpose, we adopt the energy cost as the cost function, namely, $H(w(t)) = E(w(t))$. Note that, in the energy usage model, e_d is set as 0.025 J/byte(which is adopted in[5]), and e_a is set as 20, which is larger than the value in [5], to represent the significant interests of mobile users in the power-saving utility.

First, we analyze the impact of different buffer management strategies on the amount of wasted data. Fig. 6(a) illustrates the relative ratio of the amount of wasted data under different buffer management strategies. The downloading rate has been normalized by the streaming rate. Note that, the downloading rate can be taken as an indicator of the underlying wireless channel quality. The **ABM** strategy is chosen as the baseline to evaluate the efficiency of other buffer management strategies. In the figure, **PBM-**k refers to the PBM strategy that requests k video chunks per period. We can observe that, the amount of data wasted by our **CBM** strategy is only around $15\% - 30\%$ of that of the **ABM** strategy. Compared with the **PBM** strategy, our **CBM** strategy can achieve a lower amount of wasted data than that of **PBM-2** and **PBM-3** . In the figure, **PBM-1** has the lowest amount of wasted data, however, Fig. 6(b) points out that the mean freezing time of **PBM-1** is the highest among all the strategies and its fraction of freezing time has exceeded the tolerance ratio η. In Fig. 6(b), the mean freezing time of the **ABM** strategy is the lowest due to its aggressive downloading behavior. The freezing time of our proposed **CBM** strategy is slightly higher than that of **ABM** , but the amount of wasted data is significantly decreased compared with **ABM** .

Fig. 7 further shows the mean energy cost induced by downloading wasted data under different buffer management strategies. Compared with the most aggressive **ABM** strategy, **CBM** can reduce the mean energy cost by around 70% on average. When compared with **PBM-2** and **PBM-3**, the percentage of energy cost reduction achieved by **CBM** is around 30%-40% in most cases.

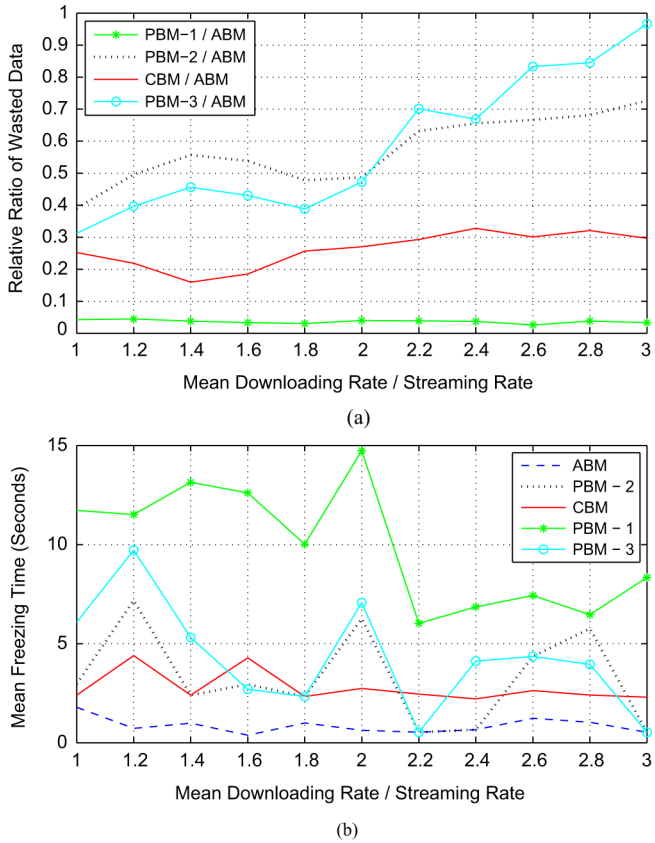


Fig. 6. Relative ratio of wasted data and mean freezing time under different ratios of download rate and streaming rate ($r = 40\ KBps$, $\eta = 0.03$)). (a) Relative Ratio of the Amount of Wasted Data. (b) Mean Freezing Time.

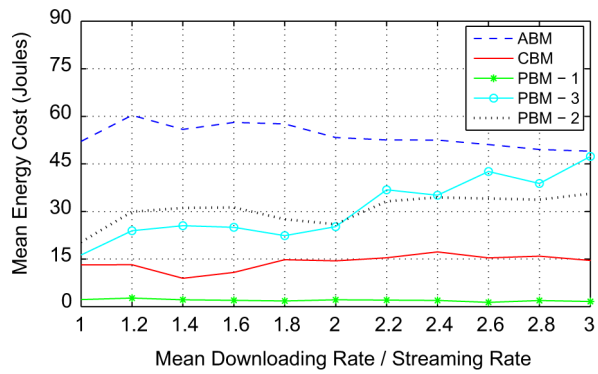


Fig. 7. Mean energy consumption induced by downloading wasted data under different ratios of download rate and streaming rate ($r = 40\ KBps$, $\eta = 0.03$).

Second, we analyze the performance of our **CBM** strategy under different freezing-time tolerance ratios (i.e., η). From Fig. 8, we can see that the mean energy cost decreases when increasing η. Moreover, we can observe that the convexity of the curve suggests that, the marginal gain in the energy cost diminishes as η increases. Specifically, as η increases from 0.01 to 0.03, the reduction of the mean energy cost is around 20%. If increasing η from 0.08 to 0.1, the reduction of the mean energy cost is only around 5%. Intuitively, more stringent requirements on freezing time will force mobile users to download more video data when the download rate is high, in order to mitigate the impacts of network condition fluctuation.

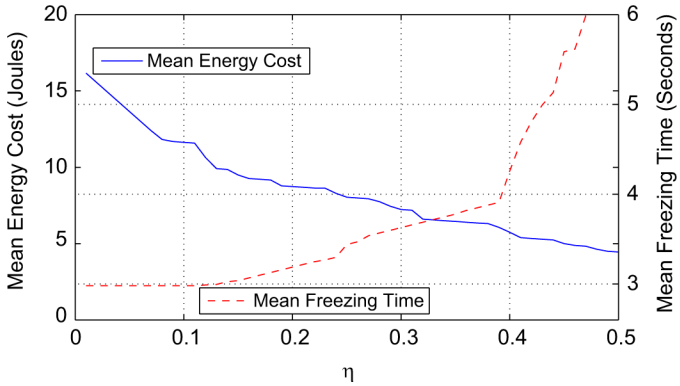


Fig. 8. Mean energy cost under different tolerance ratios of freezing time, η ($r = 40K\ Bps$, $B = 1600\ KB$).

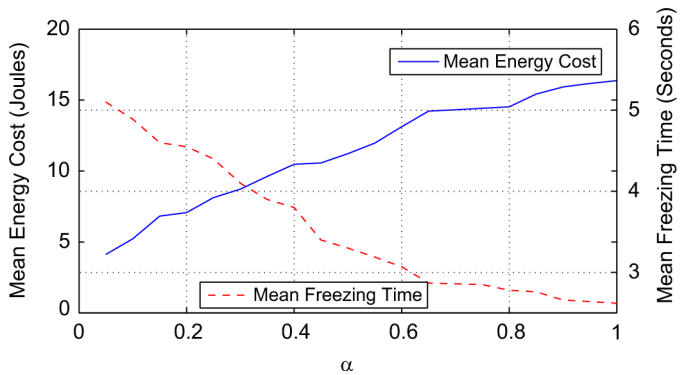


Fig. 9. Mean energy cost under different power-saving utility thresholds, α ($r = 40\ KBps$, $B = 1600\ KB$).

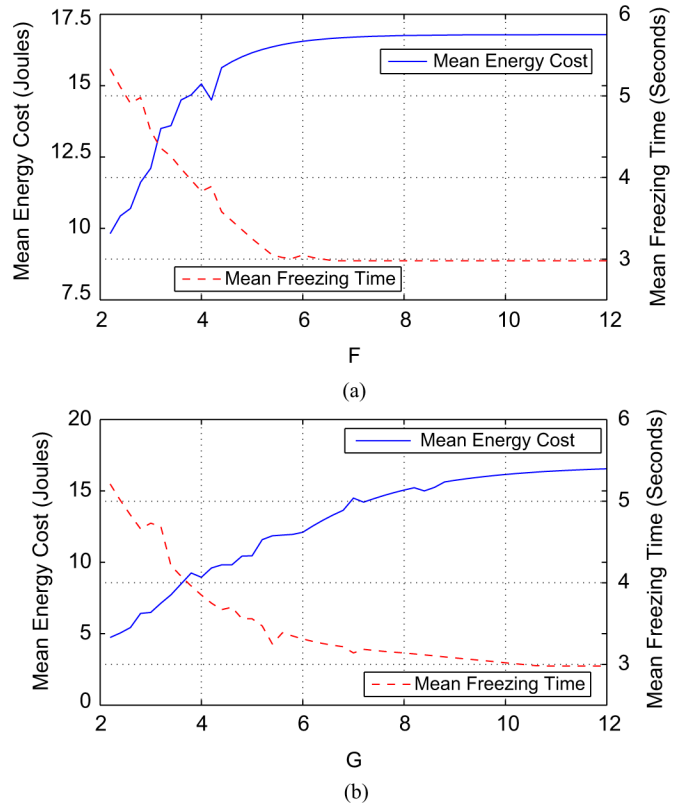


Fig. 10. Impacts of F and G on the mean energy cost under the CBM strategy ($\eta = 0.015$, $r = 40\ KBps$, $B = 1600\ KB$). (a) Mean energy cost under various F. (b) Mean energy cost under various G.

Finally, we investigate how our **CBM** strategy would behave under different values of α, which indicates the time average power-saving utility threshold during one time slot. Larger α drives mobile users to download more video data when the network condition is good, because mobile users can gain more utilities if downloading more video data when a higher download rate is possible. Fig. 9 illustrates the impacts of various α. The mean energy cost increases with α, and the underlying reason is that more video data will be requested and downloaded when the download rate is high.

D. Virtual Queues Analysis

In our online strategies, F and G are used as the bounds for the two virtual queues $F(t)$ and $G(t)$. The user-defined sunk cost function is defined as $H(w(t)) = E(w(t))$.

To validate whether the mean energy cost under the **CBM** strategy can approach to the optimal mean energy cost of Problem (2) with distance $H(\epsilon)$ (ϵ can be any small number), we need to confirm the existence of finite F and G. Fig. 10 shows the results under various values of F and G. The value of G is normalized by the playback rate of the video.

From the figure, we can observe that the mean energy cost is constant when F and G are larger than a certain threshold value. Therefore, the upper bounds on the occupancy of the two virtual queues under the optimal solution of Problem 2 are finite; Otherwise, the mean energy cost will increase with F and G all the time (proven in Theorem IV.2). We also observe that large values of F and G (i.e., $F \geq 3$ and $G \geq 10$ as illustrated in Fig. 10) can ensure that the **CBM** strategy approaches to the optimality with any small distance without violating the QoE constraints.

VI. Conclusions

In this paper, we investigated the over-buffering problem widely observed in mobile video streaming applications, by introducing an intelligent buffer management strategy. We formulated the buffer management problem as a constrained stochastic optimization problem, whose objective is to minimize the total sunk cost, including both energy consumption and bandwidth cost, under QoE and utility constraints. Leveraging the Lyaponov optimization framework, we derived an online cost-aware buffer management strategy **CBM** , which adaptively controls the buffer level in response to the stochastic channel condition and the buffer state. Our online algorithm has been verified via extensive simulations with real trace data. Compared to other buffer management strategies, our **CBM** strategy can significantly reduce the sunk cost due to unconsumed video data in the buffer, while providing a decent QoS measured by the percent of video freezing time. In our future work, we plan to extend our framework to address the same issue for adaptive bit rate (ABR) mobile streaming applications.

Acknowledgment

The authors would like to sincerely thank Dr. Jin Li at Microsoft Research for his insightful suggestions on our problem formulation. The corresponding author of this paper is Di Wu.

REFERENCES

[1] Cisco Visual Networking Index: Forecast and methodology, 2011–2016. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white paper c11-481360ns827 Networking Solutions White Paper.html.

[2] Y. Su, Y. Yang, M. Lu, and H. Chen, "Smooth control of adaptive media playout for video streaming," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1331–1339, 2009.

[3] C. Liu, I. Bouazizi, and M. Gabbouj, "Rate adaptation for adaptive HTTP streaming," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst.*, 2011.

[4] Y. Liu, L. Guo, F. Li, and S. Chen, "An empirical evaluation of battery power consumption for streaming data transmission to mobile devices," in *Proc. ACM Multimedia*.

[5] A. Finamore, M. Mellia, M. Munafó, R. Torres, and G. Rao, "Youtube everywhere: Impact of device and infrastructure synergies on user experience," in *Proc. ACM SIGCOMM Conf. Internet Measurement Conf. (IMC)*, 2011.

[6] A. Pathak, Y. Hu, and M. Zhang, "Fine grained energy accounting on smartphones with Eprof," in *Proc. EuroSys*.

[7] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: A measurement study and implications for network applications," in *Proc. ACM SIGCOMM Conf. Internet Measurement Conf. (IMC)*, 2009.

[8] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Pricing data: A look at past proposals, current plans, and future trends," Tech. Rep., Princeton Univ., 2012.

[9] J. Liu and L. Zhong, "Micro power management of active 802.11 interfaces," *Proc. Mobisys*, pp. 146–159, 2008.

[10] N. Ding, A. Pathak, D. Koutsonikolas, C. Shepard, Y. C. Hu, and L. Zhong, "Realizing the full potential of psm using proxying," in *Proc. Infocom*.

[11] M. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[12] Y. Liu, F. Li, L. Guo, and S. Chen, "Bluestreaming: towards power-efficient Internet P2P streaming to mobile devices," in *Proc. ACM Multimedia*.

[13] Y. Liu, F. Li, L. Guo, and S. Chen, "A measurement study of resource utilization in Internet mobile streaming," in *Proc. ACM NOSSDAV*.

[14] Y. Li, Y. Zhang, and R. Yuan, "Measurement and analysis of a large scale commercial mobile Internet TV system," in *Proc. ACM SIGCOMM Conf. Internet Measurement Conf. (IMC)*, 2011.

[15] Y. Liu, F. Li, L. Guo, B. Shen, and S. Chen, "A server's perspective of Internet streaming delivery to mobile devices," in *Proc. INFOCOM*.

[16] M. v. d. Schaar and S. Shankar, "Cross-layer wireless multimedia transmission: Challenges, principles, and new paradigms," *IEEE Wireless Commun. Mag*, vol. 12, no. 4, pp. 50–58, 2005.

[17] M. Hefeeda and C. Hsu, "On burst transmission scheduling in mobile TV broadcast networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, Apr. 2010.

[18] F. Tabrizi, J. Peters, and M. Hefeeda, "Dynamic control of receiver buffers in mobile video streaming systems," *IEEE Trans. Mobile Comput.*, vol. 12, no. 5, pp. 995–1008, May 2013.

[19] J. Adams, "Adaptive buffer power save mechanism for mobile multimedia streaming," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2007.

[20] X. Li, M. Dong, Z. Ma, and F. C. Fernandes, "Greentube: power optimization for mobile videostreaming via dynamic cache management," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 279–288.

[21] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, "Using crowd-sourced viewing statistics to save energy in wireless video streaming," in *Proc. ACM Mobicom*.

[22] N. Mastronarde and M. v. d. Schaar, "Online reinforcement learning for multimedia buffer control," in *Proc. ICASSP*.

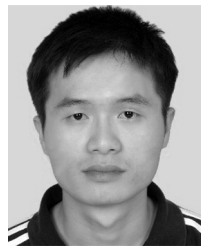
[23] G. Ji, B. Liang, and A. Saleh, "Buffer schemes for VBR video streaming over heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2009.

[24] J. He, Z. Xue, D. Wu, D. O. Wu, and Y. Wen, CBM: Online Strategies on Cost-aware Buffer Management for Mobile Video Streaming,CS-2013–0826, Sun Yat-sen University, Tech. Rep., 2013. [Online]. Available: http://sist.sysu.edu.cn/~dwu/cbm-tr.pdf.

[25] N. Salodkar, A. Bhorkar, A. Karandikar, and V. S. Borkar, "An on-line learning algorithm for energy efficient delay constrained scheduling over a fading channel," *IEEE J. Select. Areas Commun.*, vol. 26, no. 4, pp. 732–742, 2008.

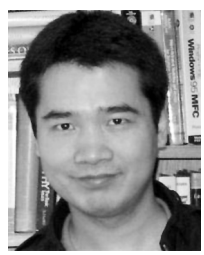
[26] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "TUBE: Time-dependent pricing for mobile data," in *Proc. ACM SIGCOMM*, 2012.

[27] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.



Jian He received the B.S. degree in Computer Science from Sun Yat-sen University in 2011. He is now a graduate student in the School of Information Science and Technology at Sun Yat-sen University. His research interests include content distribution networks, data center networking, green networking and network measurement.

Zheng Xue is a second-year graduate student in the Department of Computer Science, Sun Yat-sen University, Guangzhou, China. He received his B.S. degree from Sun Yat-sen University in 2012. His research interests include cloud computing, data center network, content distribution, network measurement, cloud-assisted mobile computing. His advisor is Prof. Di Wu.



Di Wu (M'06) received the B.S. degree from the University of Science and Technology of China in 2000, the M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2003, and the Ph.D. degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2007. From 2007 to 2009, he was a postdoctoral researcher in the Department of Computer Science and Engineering, Polytechnic Institute of NYU, advised by Prof. Keith W. Ross. He has been an Associate Professor in the Department of Computer Science, Sun Yat-Sen University, China, since July 2009. He was the winner of IEEE INFOCOM 2009 Best Paper Award, and is a member of the IEEE, the IEEE Computer Society, the ACM, and the Sigma Xi.

His research interests include multimedia communication, cloud computing, peer-to-peer networking, Internet measurement, and network security.

Dapeng Oliver Wu (S'98–M'04–SM'06–F'13) received B.E. in Electrical Engineering from Huazhong University of Science and Technology, Wuhan, China, in 1990, M.E. in Electrical Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1997, and Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, in 2003. Since 2003, he has been on the faculty of Electrical and Computer Engineering Department at University of Florida, Gainesville, FL, where he is a professor; previously, he was an assistant professor from 2003 to 2008, and an associate professor from 2008 to 2011. His research interests are in the areas of networking, communications, signal processing, computer vision, and machine learning. He received University of Florida Research Foundation Professorship Award in 2009, AFOSR Young Investigator Program (YIP) Award in 2009, ONR Young Investigator Program (YIP) Award in 2008, NSF CAREER award in 2007, the IEEE Circuits and Systems for Video Technology (CSVT) Transactions Best Paper Award for Year 2001, and the Best Paper Awards in IEEE GLOBECOM 2011 and International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine) 2006.

Currently, he serves as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, Journal of Visual Communication and Image Representation, and International Journal of Ad Hoc and Ubiquitous Computing. He is the founder of IEEE Transactions on Network Science and Engineering. He was the founding Editor-in-Chief of Journal of Advances

in Multimedia between 2006 and 2008, and an Associate Editor for IEEE Transactions on Wireless Communications and IEEE Transactions on Vehicular Technology between 2004 and 2007. He is also a guest-editor for IEEE Journal on Selected Areas in Communications (JSAC), Special Issue on Cross-layer Optimized Wireless Multimedia Communications. He has served as Technical Program Committee (TPC) Chair for IEEE INFOCOM 2012, and TPC chair for IEEE International Conference on Communications (ICC 2008), Signal Processing for Communications Symposium, and as a member of executive committee and/or technical program committee of over 80 conferences. He has served as Chair for the Award Committee, and Chair of Mobile and wireless multimedia Interest Group (MobIG), Technical Committee on Multimedia Communications, IEEE Communications Society. He was a member of Multimedia Signal Processing Technical Committee, IEEE Signal Processing Society from Jan. 1, 2009 to Dec. 31, 2012. He is an IEEE Fellow.

Yonggang Wen is an assistant professor with school of computer engineering at Nanyang Technological University, Singapore. He received his PhD degree in Electrical Engineering and Computer Science (minor in Western Literature) from Massachusetts Institute of Technology, Cambridge, USA. Previously he has worked in Cisco to lead product development in content delivery networks, which had a revenue impact of 3 Billion US dollars globally. Dr. Wen has published over 70 papers in top journals and prestigious conferences. His latest work in multi-screen cloud social TV has been featured by global media and has attracted much commercial attention. His research interests include cloud computing, green data center, big data analytics, multimedia network and mobile computing.