Seamless Dynamic Adaptive Streaming in LTE/Wi-Fi Integrated Network under Smartphone Resource Constraints

Jonghoe Koo, Juheon Yi, Joongheon Kim, *Senior Member, IEEE*, Mohammad Ashraful Hoque, and Sunghyun Choi, *Fellow, IEEE*

Abstract—Exploiting both LTE and Wi-Fi links simultaneously enhances the performance of video streaming services in a smartphone. However, it is challenging to achieve seamless and high quality video while saving battery energy and LTE data usage to prolong the usage time of a smartphone. In this paper, we propose REQUEST, a video chunk request policy for Dynamic Adaptive Streaming over HTTP (DASH) in a smartphone, which can utilize both LTE and Wi-Fi. REQUEST enables seamless DASH video streaming with near optimal video quality under given budgets of battery energy and LTE data usage. Through extensive simulation and measurement in a real environment, we demonstrate that REQUEST significantly outperforms other existing schemes in terms of average video bitrate, rebuffering, and resource waste.

Index Terms—DASH, adaptive video streaming, lyapunov optimization, smartphone, LTE/Wi-Fi networks

1 Introduction

Nowadays, smartphones utilize both LTE and Wi-Fi interfaces to enhance the performance of various applications. In particular, the performances of mobile applications in terms of Quality of Experience (QoE) have been modeled and actively discussed to be optimized [2]. Meanwhile, Cisco Visual Networking Index [3] suggests that video traffic will be 82 percent of all consumer Internet traffic by 2021, which is 73 percent higher and a threefold increment of video traffic from 2016 to 2021. Accordingly, video streaming is definitely one of the good candidates that can take advantage of utilizing both LTE and Wi-Fi wireless links towards better QoE provisioning [4], [5]. Among various video streaming standards and protocols, Dynamic Adaptive Streaming over HTTP (DASH) is one of the video streaming technologies which delivers video chunks containing video data with a specific playback time.

Mobile users can enjoy high-quality video streaming services over Wi-Fi wireless networks in high-bandwidth Wi-Fi hotspots, and watch video contents over LTE whenever LTE base stations can support reliable data communication even though there are no available Wi-Fi access points (APs). In a place where both LTE and Wi-Fi are available, higher quality videos can be supported or more resource efficient streaming is realized by judicially using both LTE and Wi-Fi wireless links, simultaneously.

Let us consider a video streaming scenario, where a user wants to enjoy content with long playback duration, e.g., a live soccer game or video on demand (VoD) contents from YouTube or Netflix. Since the content duration is quite long, Wi-Fi may be the desirable access network for the user. However, the user might be on move or take public transport while experiencing the content. Such mobility results in extremely fluctuating wireless link throughput or frequently disconnected Wi-Fi links from the nearby hotspots [6], [7]. This introduces significant video quality degradation and also causes frequent rebuffering (also called video stall or freezing), and hence, it is not desirable to watch the video over Wi-Fi link in this environment.

To tackle this problem, an intelligent video chunk requesting mechanism is desired which can utilize the unstable Wi-Fi links intelligently while guaranteeing the video quality requirements and uninterrupted playback without rebuffering in mobile environments. At the same time, battery energy and LTE data quota are important issues which are desired to be conserved as much as possible while ensuring satisfying video quality.

Accordingly, a chunk request technique should consider both enhancing video performances, i.e., video quality and rebuffering, and saving smartphone resources, i.e., battery energy and LTE data quota, to optimize QoE of video streaming. Besides, the problem is more complex in a multi-homing environment compared with a single link scenario [8], thus making it challenging to develop a well-designed DASH streaming client, and a solution is still due.

- *J. Koo is with Samsung Research, Samsung Electronics, Seoul 135-090, Korea. E-mail: jh89.koo@samsung.com.*
- *J. Yi and S. Choi are with the Department of ECE and INMC, Seoul National University, Seoul 08826, Korea. E-mail: jhyi16@mwnl.snu.ac.kr, schoi@snu.ac.kr.*
- *J. Kim is with the School of Computer Science and Engineering, Chung-Ang University, Seoul 156-756, Korea. E-mail: joongheon@cau.ac.kr.*
- *M. A. Hoque is with the Department of Computer Science, HIIT/University of Helsinki, Helsinki, Finland. E-mail: mohammad.a.hoque@helsinki.fi.*

Manuscript received 23 Dec. 2017; revised 9 June 2018; accepted 24 July 2018. Date of publication 6 Aug. 2018; date of current version 31 May 2019.
(Corresponding authors: Joongheon Kim and Sunghyun Choi.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TMC.2018.2863234

To provide an optimal video performance while satisfying resource usage constraints for battery energy and LTE data quota, we propose REQUEST, a bit*R*ate, *E*nergy, LTE data *Q*uota, and b*U*ff*E*r-aware video *ST*reaming, for DASH video over multi-homed smartphones. By utilizing Wi-Fi link as a supplementary link, REQUEST realizes a seamless DASH video streaming even in the environments where Wi-Fi link performance is not guaranteed. Also, REQUEST optimizes time-average video quality while satisfying time-average resource constraints by adopting Lyapunov optimization framework, and it is easily implemented in commercial smartphones as an application.

In addition, a control parameter V, which provides a tradeoff between a time-average video quality and time-average resource usages, is a critical factor to determine the performance of REQUEST. Since it is difficult to determine an optimal value of V, we adapt V in a heuristic manner for a better performance of REQUEST.

We claim the following major contributions:

- We propose a chunk request policy that achieves seamless playback of video using both Wi-Fi and LTE simultaneously even in situations where Wi-Fi is unstable.
- We formulate a Lyapunov optimization framework-based stochastic optimization problem to maximize time-average video quality under time-average energy and LTE data usage constraints and minimize rebuffering.
- We design REQUEST, an online video chunk request algorithm by using both LTE and Wi-Fi links, which provides a near-optimal solution of the Lyapunov optimization problem.
- We implement REQUEST by modifying ExoPlayer, Google's open-source DASH player for Android [9], and validate its performance in real-world scenarios.
- We propose a run time V (a control parameter for Lyapunov optimization) adaptation algorithm to provide a better performance of REQUEST and verify the enhancement through a trace-driven simulation.

Even though DASH has been studied for LTE data and energy optimization using both LTE and WiFi networks [5], [10], our work has the novelty as follows. We utilize the WiFi network as a supplementary network to enhance video quality and efficiently utilize the resources, i.e., battery energy and LTE data quota, while avoiding video rebuffering even when the WiFi link becomes unstable. We propose a new chunk request policy and formulate a stochastic optimization problem to maximize the time-average video quality under energy and LTE data quota constraints as mentioned earlier. Our solution is easy to implement on commercial smartphones and can be directly applied with already deployed network infrastructure.

The rest of the paper is organized as follows. In Section 2, we discuss issues arising when utilizing both LTE and Wi-Fi for DASH video streaming and related work. Section 3 describes the motivation of our work, and our proposed chunk request policy is presented in Section 4. We formulate the optimization problem in Section 5 and REQUEST algorithm is introduced in Section 6. We evaluate the performance of REQUEST in Section 7. In Section 8, we discuss the method to enhance the performance of REQUEST and conclude the paper in Section 9.

2 Background and Related Work

2.1 Background

2.1.1 Multi-Homing for DASH

DASH is a bitrate-adaptive streaming technique, which delivers video content over HTTP. Video content is encoded at a variety of bitrates in the video server. A DASH-enabled video streaming player downloads video chunks of a particular bitrate based on the experienced bandwidth for downloading earlier chunks. Initially, the player begins with lower or moderate quality to avoid longer start-up delay. Thanks to the development of techniques for simultaneous utilization of multiple network interfaces at mobile devices, especially, as application-level solutions [11], [12], both Wi-Fi and LTE interfaces can be utilized simultaneously to enhance video quality while saving battery energy and LTE data quota. By downloading video chunks with both LTE and Wi-Fi, the video player can maintain sufficient chunks in its buffer, thus avoiding rebuffering. However, chunks remaining in the buffer may cause possible data waste when the user stops watching the video in the middle of playback [8], [13].

Bandwidth Aggregation for Enhancing Video Quality. Traditional bandwidth aggregation in heterogeneous wireless network environment aims to support higher bitrates of video that cannot be sufficiently delivered by only one of the available networks [14], [15], [16], [17], [18], [19], [20].

Wi-Fi Offloading to Reduce LTE Data Usage. User may not fully utilize LTE link due to her/his remaining monthly LTE data quota or data plan. In this case, it is fairly useful to download video data with Wi-Fi as much as possible to save LTE data quota [7], [21].

Energy Consumption. The higher the video quality, the higher the energy consumption of smartphones, because the amount of data to be decoded and downloaded via Wi-Fi or LTE increases [5], [22]. The energy is also spent by the CPU for processing more packets [22].

Fast Prefetching. Prefetching video data, i.e., requesting more video chunks in advance, provides several advantages to a video client. The more video chunks to be consecutively requested, the less energy is spent for networking activities as network interfaces are able to stay in idle state for longer periods [8], [13], [23]. In addition, prefetching prevents rebuffering events because video buffer is filled with sufficiently many video chunks. However, a large amount of prefetching may waste as much energy and LTE data as the number of chunks remaining in the video buffer when the user stops watching before the video clip ends. Therefore, it is necessary to determine an optimal prefetching strategy considering the advantages of prefetching (prevention of rebuffering and energy efficiency) and disadvantages (waste of energy and LTE data due to user's leaving) to enhance QoE of video consumers.

2.1.2 Lyapunov Optimization Framework

Stochastic optimization aims at minimizing a time-average objective function subject to queue stability when the utility function and queue stability conditions are in tradeoff relationship. In addition, the stochastic optimization framework is able to utilize the concept of virtual queues for time-average constraint representation. In stochastic optimization formulation, Lyapunov function $L(t)$ is defined as the sum of squares of backlogs in actual and virtual queues on a slot (in a slotted system). Lyapunov drift $\Delta(t)$ is defined as the difference in the Lyapunov function per slot time. While pursuing the minimization of a time-average objective function, taking the minimum of the Lyapunov drift leads to the queue stability (i.e., main constraint), which is referred to as *drift-plus-penalty* minimization. By taking an action to greedily minimize the drift-plus-penalty every slot time, we can achieve a time-average utility deviating by at most $O(1/V)$ from optimality while satisfying time-average constraints and a time-average queue backlog bound of $O(V)$, where V is defined as a tradeoff factor between utility and queue stability. For further details about the theory of Lyapunov optimization, the book [24] can be referred to.

2.2 Related Work

Energy/Cellular Data Quota-Aware Video Streaming. Previous efforts [8], [25] control a video segment size to be prefetched in an HTTP-based video streaming service to improve the energy efficiency. *GreenTube* [25] aims to minimize an unnecessary active period of 3G/4G radios by scheduling each video segment downloading. *eSchedule* [8] utilizes crowd-sourced video viewing statistics and power models for energy efficient scheduling of video streaming. *QAVA* [10] manages the tradeoff between cellular data usage and video quality by predicting video client's usage behavior. *QAVA* automatically selects optimal video quality to enable users to keep under their data quota while maximizing video quality. Lee et al. [13] and Hu et al. [26] consider wastage of energy and LTE data quota incurring when users stop video streaming sessions early in their smartphones.

GreenBag [4] utilizes both LTE and Wi-Fi links to achieve a better quality of service (QoS) and energy efficiency. Go et al. [5] propose an energy-efficient HTTP adaptive streaming algorithm under a cellular data usage constraint. It considers the simultaneous usage of LTE and Wi-Fi for DASH video streaming on smartphones. It selects video bitrate and the number of chunks to request via each network in order to minimize a weighted sum of video distortion and energy consumption per video chunk.

Energy-quaLity aware Bandwidth Aggregation (ELBA) [19] and energy-distortion-aware MPTCP (EDAM) [20] have been proposed to provide energy-efficient and quality-guaranteed video streaming over heterogeneous wireless networks. ELBA models delay-constrained energy-quality tradeoff for multipath video transmission over heterogeneous wireless networks and enables energy-minimized quality-guaranteed streaming within an imposed delay by energy-minimized video rate adaptation, delay-constrained unequal protection, and quality-aware packet distribution. EDAM depends on utility maximization theory to minimize the energy consumption while achieving target quality for video streaming using Multipath TCP (MPTCP) over heterogeneous wireless networks.

Optimal Bitrate Selection of DASH Video. Video client selects a video chunk with an appropriate bitrate according to current network status [27], [28] or video player's buffer status [29]. Due to severe fluctuations of link throughput in a mobile environment, it is difficult for a link throughput-based bitrate adaptation to practically function, and for this reason, a buffer-based bitrate adaptation has been studied and practically used by real implementations [29], [30]. Kim et al. [31] designed a quality-aware device-to-device video scheduling and streaming technique for achieving time-average peak-signal-to-noise-ratio (PSNR) maximization subject to transmitter queue stability by selecting optimal chunk bitrates [32]. Since the baseline network architecture considered in this paper is completely different from the network in [31], a direct comparison is not possible.

3 Motivation

3.1 Wi-Fi Throughput Fluctuation Consideration

While a mobile device can utilize both Wi-Fi and LTE networks simultaneously, the Wi-Fi link may become unstable and its quality may fluctuate severely especially in a mobile or dense environment. For example, when a user moves around and goes out of the coverage of a Wi-Fi AP connected to the user's mobile device, the device cannot utilize Wi-Fi link until it finds an available Wi-Fi AP nearby and handover to a new Wi-Fi AP is successfully completed. It is also well known that Wi-Fi throughput degradation occurs in mobile environments due to poor performance of handover operations in commercial Wi-Fi devices [33], [34], [35]. In addition, if a Wi-Fi AP operates at 2.4 GHz, the Wi-Fi link quality may severely suffer from interference caused by other wireless devices operating in 2.4 GHz ISM band, such as Bluetooth, ZigBee, microwave ovens, and cordless phones [36], [37]. Furthermore, Wi-Fi at 5 GHz suffers from higher path loss than Wi-Fi at 2.4 GHz, thus increasing the possibility that mobile user experiences worse Wi-Fi link quality and goes out of Wi-Fi coverage. In contrast to Wi-Fi, a device may retain a seamless connection via the LTE network even though user moves fast, thanks to seamless handover between LTE base stations.[1] Likewise, in mobile and dense environments, Wi-Fi link may have more unstable quality and sometimes may be unavailable.

However, Wi-Fi can be still utilized for offloading purpose even in mobile environments [6], [7], and its offloading capability may increase in static environments or when a device associates to an IEEE 802.11ac-compliant Wi-Fi AP that can provide very high throughput. From this perspective, when a DASH client requests video chunks via both Wi-Fi and LTE in parallel in a multi-homed device, Wi-Fi link availability is like a double-edged sword. In other words, although the Wi-Fi link may enhance the video quality at low energy cost and reduce LTE data consumption, it can also increase the possibility of rebuffering events, as we cannot predict the exact bandwidth and availability of Wi-Fi in mobile and dense environments.

Therefore, it is challenging to design a chunk request policy for DASH by judicially utilizing Wi-Fi connectivity in addition to LTE to maximize the merit of utilizing Wi-Fi link while minimizing its side effects. In this work, we accept the challenge. We opportunistically request video chunks over Wi-Fi, thus reducing the side effects of unstable Wi-Fi links.

1. In this paper, it is assumed that seamless handover of LTE links is ensured. Other cellular links, e.g., 3G/HSDPA, may be applied instead of LTE.

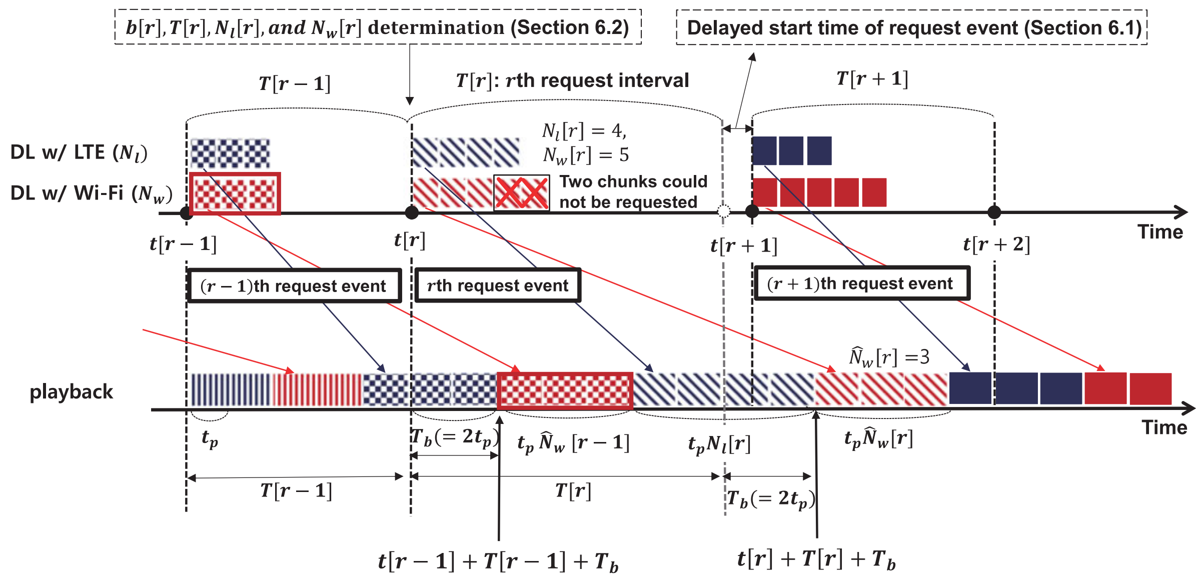


Fig. 1. An example of the proposed chunk request policy.

3.2 Resource Utilization Optimization

Mobile devices consume resources, i.e., battery energy and LTE data quota, during DASH video streaming over Wi-Fi and LTE networks. Since battery energy and remaining LTE data quota are usually limited, users will like to minimize the resource usage for DASH video streaming as much as possible so as to watch videos much longer or use the remaining resources for other applications. Assuming that a DASH client intelligently requests proper bitrate video chunks to balance the quality of video and resource usage. In this case, the video quality might be either increased in exchange for using more resources or decreased to conserve the resources. From this perspective, a problem can be formulated as optimizing video quality for DASH streaming given the constrained amount of resources in smartphones, i.e., requesting high quality video chunks with a given amount of battery energy and LTE data quota.

Unfortunately, the traditional optimization frameworks that have been used in the past studies do not support flexible resource utilization, e.g., sometimes allowing resource overuse to enhance video quality. For instance, a popular method to formulate optimization problem is to formulate a multi-attribute cost function with proper constraints based on a simple additive weighting (SAW) method. This approach is widely used for multi-attribute decision making (MADM) algorithms [5], [38], [39]. A challenge to optimize MADM-based cost function is to select appropriate weights for the attributes in the cost function to quantitatively balance them. In addition, even though the weights of the attributes determine the relative priority to provide a trade-off between attributes in a cost function, it is difficult to force an attribute to have a value within a certain range.

To overcome the limitation of MADM-based optimization, we adopt Lyapunov optimization framework to optimize time-average video quality while satisfying time-average resource constraints at the same time. By adopting time-average concepts, we occasionally allow resource overuse but eventually satisfy constraints.

4 Proposed Chunk Request Policy

Generally, before a DASH client sends a request for a video chunk, it determines an appropriate bitrate for the chunk according to the estimated link bandwidth [5] or the playback buffer status [29], [30]. In a multi-homed environment, a DASH client can further download a single chunk in parallel via multiple wireless interfaces by sending multiple HTTP range requests [40]. A number of proposed approaches follow this technique [4], [5]. However, considering a possible situation where Wi-Fi link is disconnected or extremely unstable during a chunk download, it is not efficient to divide one video chunk into two parts and receive them by both LTE and Wi-Fi separately. Accordingly, in this work, we use a single network to request and receive a single video chunk. This eliminates decoding failures due to partially received chunks, thus avoiding re-buffering and data wastage.

Therefore, we design our chunk request policy as illustrated in Fig. 1, where the notations in Fig. 1 are summarized in Table 1. We call an event of requesting a batch of video chunks a *request event*. A *request interval* $T[r]$ is defined as the time interval between the start time and the end time

TABLE 1
Important Notations

Symbol	Description
t_p	A fixed chunk playback time (sec)
T_b	Initial buffer-time (sec)
r	The index of request event
$t[r]$	Start time of the rth request event
$T[r]$	The rth request interval ($t[r+1] - t[r]$)
$b[r]$	Bitrate of chunks to be requested in the rth request event
$N_l[r]$	Number of chunks to request over LTE in the rth request event
$N_w[r]$	Number of chunks to request over Wi-Fi in the rth request event
$\hat{N}_w[r]$	Number of chunks received over Wi-Fi in the rth request event

of a *request event*. At the start time $t[r]$ of the rth *request event*, client determines the bitrate $b[r]$ of chunks,[2] *request interval* $T[r]$, and the numbers of chunks to request over LTE ($N_l[r]$) and Wi-Fi ($N_w[r]$) during $T[r]$. Ideally, all the chunks requested during $T[r]$, i.e., $N_l[r] + N_w[r]$ chunks, are expected to be successfully downloaded within $T[r]$. In this case, $t[r+1] = t[r] + T[r]$, i.e., the end time of the rth *request event* is equal to the start time of the $(r+1)$th *request event*.

Moreover, it is important to determine $T[r]$ judiciously to ensure that all the chunks are downloaded before they are played back for seamless video streaming. Especially, our policy tries to avoid rebuffering due to Wi-Fi throughput degradation. To enable this goal, the following features are introduced in our design:

i) We first set $T[r] = t_p(\hat{N}_w[r-1] + N_l[r])$, i.e., the playback time of $\hat{N}_w[r-1]$ chunks received over Wi-Fi in the $(r-1)$th *request event* and $N_l[r]$ chunks requested over LTE in the rth *request event*. By doing so, the chunks received over Wi-Fi accumulate in the buffer, which has the same effect as increasing the initial buffer-time at the start of the next *request event*. It also prevents rebuffering due to LTE throughput degradation that cannot be handled by the initial buffer.

ii) The video chunks requested over Wi-Fi in the rth *request event* should be played after T_b from the end time of the rth *request event*, where T_b is initial buffer-time.[3] For example, if $T_b = 2t_p$, $\hat{N}_w[r]$ chunks received over Wi-Fi in the rth *request event* are played after $t[r] + T[r] + 2t_p$ as illustrated in Fig. 1.

iii) If chunks are requested over both LTE and Wi-Fi, the video chunk requested over Wi-Fi is played right after the last video chunk requested over LTE in the same *request event*. In other words, the first sequence of $N_w[r]$ chunks is the right next to the last sequence of $N_l[r]$.

iv) If the client fails to request some chunks over Wi-Fi, those chunks are prioritized to request in the next *request event*. For example, as illustrated in Fig. 1, if two chunks could not be requested in the rth *request event*, the client starts requesting the chunks in order in the $(r+1)$th *request event* for the continuity of video.

Based on these features, we ensure that all chunks of a video can be eventually received and played even though some chunks could not be requested over Wi-Fi during a *request event*. However, in practice, chunks may not be received within a *request interval*, due to throughput degradation of either LTE or Wi-Fi. In this case, the start of the next *request event* can be delayed. The details are discussed in Section 6.1. In addition, to maximize the video quality while satisfying battery energy and LTE data quota constraints, we should properly determine b, T, N_l, and N_w for each *request event*. To achieve this, we formulate a stochastic optimization problem in the following section.

5 PROBLEM FORMULATION

Even if Wi-Fi and LTE links are sufficient to support the highest video quality, we cannot request high quality video if the battery energy and/or LTE data usage is limited. Thus, we have to accurately determine the bitrate of chunks and the number of requested chunks to maximize video quality while satisfying battery and LTE data usage constraints during the entire video playback.

Objective. The objective of our optimization problem is to maximize a time-average video utility. In this work, we define the time-average utility as a time-average logarithmic function of the video bitrate. The reason why we maximize the logarithmic function of bitrate instead of bitrate directly is to reflect the fact that the impact of increasing video quality on a user experience can be modeled by a concave function with diminishing returns [30].

Constraints. To guarantee longer battery life and use LTE data without exceeding budget, we assume that a mobile device is allowed to consume p_{av} (W) and consume LTE data with d_{av} (Mbps) on average during a video playback. In addition, p_{av} refers to the total power consumed by the display, CPU, and network interfaces including their tail energy.

In order to design an optimal chunk request policy for maximizing the time-average video quality with satisfying time-average resource constraints, it would require *a priori* statistical knowledge, such as the distribution of network bandwidth. It would also need a complex computational method, such as dynamic programming (DP) method for finding the optimal solution based on the *a priori* statistical knowledge [41]. However, it is practically difficult to obtain the exact distribution of network bandwidth to solve the problem with a DP method. Even if the distribution could be obtained, very large state space composed of request time, bitrate, number of chunks to request via networks would have to be constructed [30], [42].

In order to overcome this challenge, we apply Lyapunov optimization-based dynamic algorithm [42] that independently determines the number of chunks to request and their bitrate at the start of a *request event* by observing the chunk reception result of the previous *request event*. The performance of Lyapunov-based dynamic algorithm is close to that of the optimal solution by a DP algorithm with *a priori* statistical knowledge, but Lyapunov optimization algorithm does not require any *a priori* knowledge [42].

Renewal-Based Lyapunov Optimization. According to the chunk request policy proposed in Section 4, *request intervalT*$[r]$ is related to the number of requested chunks, and hence, it is time-varying. To reflect this, we adopt renewal-based Lyapunov optimization [24], [42].[4] We assume that the video playback time is infinite, i.e., $R \rightarrow \infty$,[5] for an approach similar to that in [30]. Then, the problem based on Lyapunov optimization framework to maximize time-average video utility while satisfying time-average energy and LTE data usage constraints is described as follows:

$$\begin{aligned} &\text{Maximize} && \overline{\log(b)T}/\overline{T}, \\ &\text{subject to} && \overline{e}_e \leq p_{av}\overline{T}, \quad \overline{e}_d \leq d_{av}\overline{T}, \end{aligned} \tag{1}$$

where $\overline{\log(b)T}, \overline{T}, \overline{e}_e$, and $\overline{e}_d$ are infinite horizon expectations, i.e.,

2. All the chunks requested in a *request event* have the same bitrate.

3. The initial buffer-time is the playback time of video chunks initially received before video player starts rendering the video.

4. A *request event* in Fig. 1 corresponds to a renewal frame in the renewal-based Lyapunov optimization.

5. R is the index of the last *request event* for a video.

$$\begin{aligned}&(\overline{\log(b)T}, \overline{T}, \overline{e}_e, \overline{e}_d)\\&\quad= \lim_{R\to\infty}\frac{1}{R}\sum_{r=0}^{R-1}(\log(b[r])T[r], T[r], e_e[r], e_d[r]),\end{aligned} \tag{2}$$

where $e_e[r]$ and $e_d[r]$ are the energy consumption and LTE data usage during the rth *request event*, respectively. Since *request interval* is time-varying, we give a time weight to $\log(b)$, thus maximizing $\overline{\log(b)T}/\overline{T}$, which denotes the time-average video utility.

Virtual Queues for Resource Constraints. We define virtual queues to solve our optimization problem based on a drift-plus-penalty ratio minimization algorithm, which greedily minimizes drift-plus-penalty ratio to achieve near-optimal time-average video utility while satisfying time-average resource constraints.

Let us first define a virtual queue for energy consumption $Q_e[r]$. The arrival and service process of $Q_e[r]$ at the rth *request event* are $e_e[r]$ and $p_{av}T[r]$, respectively. Then, the queue backlog of $Q_e[r]$ evolves as follows:

$$Q_e[r+1] = \max(Q_e[r] + e_e[r] - p_{av}T[r], 0). \tag{3}$$

If i) $Q_e[0] = 0$, ii) $Q_e[r]$ satisfies (3) for all $r \in \{0, 1, 2, \cdots\}$, and iii) $Q_e[r]$ is mean-rate stable (i.e., $lim_{R\to\infty}\frac{Q_e[R]}{R} = 0$) [24], then $\overline{e}_e \leq p_{av}\overline{T}$ for all integers $R > 0$.

Lemma 1. *If $Q_e[0] = 0$ and $Q_e[r]$ satisfies (3) for all $r \in \{0, 1, 2, \ldots\}$ and $Q_e[r]$ is mean rate stability [24], then for all integers $R > 0$:*

$$\overline{e}_e \leq p_{av}\overline{T}. \tag{4}$$

Proof. From (3) we have:

$$\begin{aligned}Q_e[r+1] &\geq Q_e[r] + e_e[r] - p_{av}T[r],\\ Q_e[r+1] - Q_e[r] &\geq e_e[r] - p_{av}T[r].\end{aligned} \tag{5}$$

Summing over $r \in \{0, 1, \ldots, R-1\}$ provides:

$$Q_e[R] - Q_e[0] \geq \sum_{r=0}^{R-1} e_e[r] - p_{av}\sum_{r=0}^{R-1} T[r]. \tag{6}$$

Dividing by R, using $Q_e[0] = 0$, and taking the limit $R \to \infty$, we obtain $\lim_{R\to\infty}\frac{Q_e[R]}{R} \geq \overline{e}_e - p_{av}\overline{T}$. Since $Q_e[r]$ is mean rate stability, i.e, $\lim_{R\to\infty}\frac{Q_e[R]}{R} = 0$, we obtain

$$\overline{e}_e[r] \leq p_{av}\overline{T}. \tag{7}$$

□

In addition, from (6), for all $R > 0$ such that $Q_e[R] = 0$, the average energy consumption $\frac{1}{R}\sum_{r=0}^{R-1} e_e[r]$ is less than or equal to $\frac{1}{R}p_{av}\sum_{r=0}^{R-1} T[r]$, i.e., the inequality constraint of (1) is satisfied.

The same approach can be applied to the LTE data usage virtual queue $Q_d[r]$, where the queue backlog evolution of $Q_d[r]$ is represented as follows:

$$Q_d[r+1] = \max(Q_d[r] + e_d[r] - d_{av}T[r], 0). \tag{8}$$

Drift-Plus-Penalty Ratio. The renewal-based Lyapunov optimization framework minimizes drift-plus-penalty ratio which is obtained by normalizing drift-plus-penalty by time [42]. Before defining the drift-plus-penalty ratio, we first define the Lyapunov function $L[r]$. In addition to Q_e and Q_d, we consider DASH client's video buffer $Q_r[r]$ which denotes the number of video chunks in video buffer at the rth *request event*. Then, $L[r]$ is defined by:

$$L[r] = \frac{1}{2}\left((Q_e[r])^2 + (Q_d[r])^2 + \left(\frac{Q_{\max}}{2} - Q_r[r]\right)^2\right), \tag{9}$$

where $Q_{\max}$ is defined as the maximum number of chunks stored in the video buffer. The energy consumption and LTE data usage queues are desired to be empty to satisfy the time-average constraints while the video chunk queue should not be empty to avoid rebuffering. However, filling the video buffer to Q_{max} is not desirable due to the possibility of queue overflow and the waste of resources, and hence, we use $\frac{Q_{max}}{2}$ instead of Q_{max} when defining $L[r]$.

Then, the Lyapunov drift $\Delta[r]$, which is desired to be minimized, is defined as the difference between the Lyapunov function of the rth *request event* and that of the $(r+1)$th *request event*.

$$\Delta[r] = L[r+1] - L[r]. \tag{10}$$

Conventional Lyapunov optimization framework minimizes penalty. Since our work maximizes video utility, we define penalty as $-\log(b)T$ to follow the Lyapunov framework's concept of penalty minimization. Finally, the drift-plus-penalty ratio (DPP_r) is defined as:

$$DPP_r[r] = \frac{\Delta[r] + V\cdot penalty[r]}{T[r]} = \frac{\Delta[r]}{T[r]} - V\log(b[r]), \tag{11}$$

where V is a positive constant parameter to allow a trade-off between virtual queues' backlogs and the gap from a theoretical optimal video utility. By taking an action to greedily minimize the drift-plus-penalty ratio every *request event*, we can achieve time-average video utility maximization deviating by at most $O(1/V)$ from optimality while satisfying time-average resource constraint bound of $O(V)$.

6 REQUEST Algorithm

In this section, we introduce REQUEST, an online algorithm to request video chunks with near-optimal video quality while satisfying resource constraints and preventing video rebuffering. The REQUEST architecture is illustrated as Fig. 2. At the start of every *request event*, REQUEST observes current virtual queue backlogs ($Q_e[r]$ and $Q_d[r]$), video queue backlog ($Q_r[r]$), and estimated LTE and Wi-Fi link throughput ($\tilde{r}_l[r]$ and $\tilde{r}_w[r]$). The detailed expressions for $Q_r[r]$ and $\tilde{r}_l[r]$ are provided later in (13) and (16), respectively. $\tilde{r}_w[r]$ is calculated in a fashion similar to $\tilde{r}_l[r]$. By using these values, REQUEST determines the four parameters, i.e., video bitrate ($b[r]$), *request interval* ($T[r]$), and the number of chunks to request through LTE and Wi-Fi ($N_l[r]$ and $N_w[r]$), which minimize the drift-plus-penalty ratio, $DPP_r[r]$, in (11). Let the four-tuple ($N_l[r]$, $N_w[r]$, $b[r]$, $T[r]$) be the *request policy* $\pi[r]$ for the rth *request event*.

The main algorithms of REQUEST include *request interval adaptation* and *chunk request adaptation*. The *request interval*

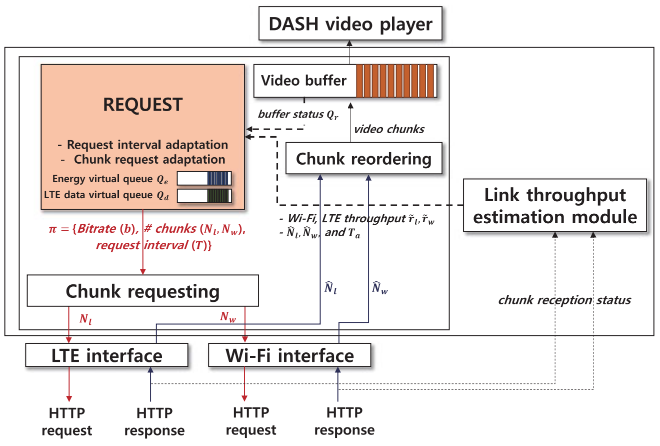

DASH video player
REQUEST
- Request interval adaptation
- Chunk request adaptation
Energy virtual queue Q_e
LTE data virtual queue Q_d
Video buffer
buffer status Q_r
video chunks
Chunk reordering
Link throughput estimation module
- Wi-Fi, LTE throughput $\bar{r}_l, \bar{r}_w$
- $\hat{N}_l, \hat{N}_w$, and T_a
π = {Bitrate (b), # chunks (N_l, N_w), request interval (T)}
Chunk requesting
$\hat{N}_l$
$\hat{N}_w$
chunk reception status
N_l
N_w
LTE interface
Wi-Fi interface
HTTP request
HTTP response
HTTP request
HTTP response

Fig. 2. REQUEST software architecture.

adaptation controls *request interval* according to chunk reception status, and the *chunk request adaptation* determines π based on DPP_r minimization algorithm.

6.1 Request Interval Adaptation

Delayed Start Time of Request Event. If all the chunks requested in the rth *request event* are successfully downloaded within $T[r]$, $t[r+1]$ is equal to $t[r]+T[r]$. However, in practice, some chunks requested via either LTE or Wi-Fi may not be successfully received on time, i.e., within $T[r]$, due to severe throughput fluctuation. In such a case, we should delay the start time of the next *request event* for successful chunk receptions. If we request $N_l[r]$ and $N_w[r]$ chunks by LTE and Wi-Fi at the rth *request event*, respectively, the delay is allowed until the following conditions are satisfied.

i) We wait until all the chunks requested over LTE ($N_l[r]$) are successfully received.
ii) If a chunk is requested over Wi-Fi before $t[r]+T[r]$, we wait until $t[r]+T[r]$ for full reception of the chunk.

Let the number of actually received chunks during the rth *request event* by LTE and Wi-Fi be $\hat{N}_l[r]$ and $\hat{N}_w[r]$. With the above two conditions, $\hat{N}_l[r]=N_l[r]$ and $\hat{N}_w[r]\le N_w[r]$. The reason why we do not guarantee the reception of all the chunks requested over Wi-Fi is that we quickly start a new *request event*, and adapt the number of chunks and bitrate for the next *request event* to resolve link throughput degradation.

If the chunk download of the rth *request event* is not completed within $T[r]$ and it takes additional $T_a[r]$ time to finish the remaining chunk reception, the start time of the $(r+1)$th *request event* $t[r+1]$ becomes $t[r]+T[r]+T_a[r]$. Fig. 3 shows the examples for the start time of the next *request event* when chunks are requested with $N_l=4$ and $N_w=5$ for the rth *request event*.

Fig. 3a shows the case that all the requested chunks are received within $T[r]$. In Fig. 3b, download of the last chunk (the fourth chunk) requested over LTE does not finish by $t[r]+T[r]$, and $t[r+1]$ is delayed by $T_a[r]$. In Fig. 3c, both of the last chunks requested over LTE and Wi-Fi have not been received until $t[r]+T[r]$, and the download over LTE finishes earlier than the one over Wi-Fi so that $T_a[r]$ is

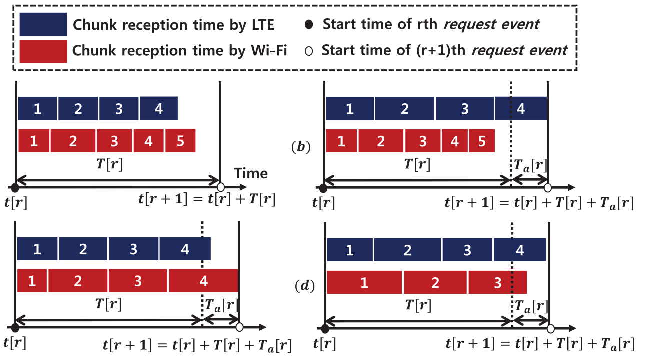

Chunk reception time by LTE
Chunk reception time by Wi-Fi
Start time of rth request event
Start time of (r+1)th request event
(a)
(b)
(c)
(d)
T[r]
Time
$T_a[r]$
t[r]
t[r + 1] = t[r] + T[r]
t[r + 1] = t[r] + T[r] + T_a[r]
1
2
3
4
5

Fig. 3. Examples for request interval and start time of the next request event according to the chunk download results.

determined by the download completion time of the fourth Wi-Fi chunk. As shown in Fig. 3d, even though there are two chunks left to request over Wi-Fi, those chunks will not be requested over Wi-Fi after $t[r]+T[r]$.

*Request Interval Adaptation.*When a delay of T_a occurs, the delay is compensated by reducing the next *request interval* by T_a. If $\hat{N}_w[r-1]$ chunks are received over Wi-Fi and the delay of $T_a[r-1]$ occurs, $T[r]$ is reduced by $T_a[r-1]$ as follows:

$$T[r]=t_p(\hat{N}_w[r-1]+N_l[r])-T_a[r-1]. \tag{12}$$

By doing so, we can compensate for the amount of buffer that has been reduced due to the delay. For example, even though the rth *request event* is delayed by $T_a[r-1]$, we can ensure that there are $\hat{N}_w[r]+T_b/t_p$ chunks in the buffer at $t[r+1]$ by reducing $T[r]$ by $T_a[r-1]$. Without this adaptation, the delays accumulate and initial buffer-time may not be guaranteed at the start time of a *request event*, thus causing rebuffering. According to our request interval adaptation, the video buffer backlog evolves as:

$$Q_r[r+1]=Q_r[r]+(\hat{N}_w[r]+N_l[r])-(T[r]+T_a[r])/t_p. \tag{13}$$

Lemma 2. *If $Q_r[1]=N_b$, $T[r]$ is determined by (12), and $Q_r[r]$ evolves as (13) for all $r\in\{1,2,3,\cdots\}$, then for all integers $R>0$, there is no video stall if $T_a[R]<N_b\cdot t_p$.*

Proof. From (12) and (13) we have:

$$\begin{aligned}Q_r[r+1]-Q_r[r]=&(\hat{N}_w[r]+\hat{N}_l[r])-(T[r]+T_a[r])/t_p,\\ Q_r[r+1]-Q_r[r]=&(\hat{N}_w[r]+\hat{N}_l[r])\\ &-(\hat{N}_w[r-1]+N_l[r])\\ &+(T_a[r]-T_a[r-1])/t_p,\\ Q_r[r+1]-Q_r[r]=&\hat{N}_w[r]-\hat{N}_w[r-1]\\ &+(T_a[r]-T_a[r-1])/t_p.\end{aligned} \tag{14}$$

Summing over $r\in\{0,1,\ldots,R-1\}$ provides:

$$\begin{aligned}Q_r[R]-Q_r[1]=&\hat{N}_w[R-1]-(T_a[R-1]-T_a[0])/t_p,\\ Q_r[R]=&N_b+\hat{N}_w[R-1]-(T_a[R-1])/t_p,\\ Q_r[R]\ge&N_b-T_a[R-1]/t_p>0.\end{aligned} \tag{15}$$

□

6.2 Chunk Request Adaptation

Estimated LTE and Wi-Fi Throughput. For the rth request event, the estimated LTE/Wi-Fi throughput, $\tilde{r}_l[r]$ and $\tilde{r}_w[r]$, are calculated by using the exponential weighted moving average (EWMA) with the coefficient β.

$$\tilde{r}_l[r] = \beta\tilde{r}_l[r-1] + (1-\beta)r_l[r-1], \tag{16}$$

where $r_l[r-1] = \frac{N_l[r-1]t_p b[r-1]}{\tau_l[r-1]}$, i.e., the average chunk download speed by LTE during the $(r-1)$th *request event*, and $\tau_l[r-1]$ is the total download time for $N_l[r-1]$ chunks of bitrate $b[r-1]$ by the LTE interface. $\tilde{r}_w[r]$ is calculated in a similar fashion.

Arrival Rate of Energy/LTE Data Virtual Queues. In order to reduce energy and LTE data waste when a user stops watching video in the early stage of playback, we put the expected energy/LTE data waste as well as the expected energy/LTE data consumption in the arrival process of the energy/LTE data virtual queue in the stage of determining the optimal $\pi[r]$. We define $\bar{E}_w[r]$ and $\bar{D}_w[r]$ as the expected energy and LTE data waste during the rth *request event*, respectively.[6]

The arrival rate $e_e[r]$ of energy virtual queue is given as

$$e_e[r] = e_v[r] + e_{e,w}[r] + e_{e,l}[r] + \bar{E}_w[r]\cdot T[r]/(t[r]+T[r]), \tag{17}$$

where $e_v[r]$ is the energy consumption for video playback including CPU and display power, $e_{e,w}[r]$ and $e_{e,l}[r]$ are the energy consumption for Wi-Fi and LTE interfaces, respectively. We multiply the expected waste by $\frac{T[r]}{t[r]+T[r]}$ to reflect the effect of waste amount on the average power from the beginning of video playback until the end of the current *request event*. The expected energy consumption for network interface n_i, where $n_1 = l$ and $n_2 = w$ for LTE and Wi-Fi, respectively, is calculated as follows:

$$\begin{aligned} e_{e,n_i}[r] = {} & \tilde{P}_{n_i}[r]t_{n_i,d}[r] \\ & + P_{n_i,tail}\cdot\min\big((T[r]-t_{n_i,d}[r]), t_{n_i,tail}\big), \end{aligned} \tag{18}$$

where $\tilde{P}_{n_i}[r]$ is the average power for network n_i with the expected throughput $\tilde{r}_{n_i}[r]$, and $t_{n_i,d}[r]$ $\left(=\frac{t_p b[r] N_{n_i}[r]}{\tilde{r}_{n_i}[r]}\right)$ is the expected download time for $N_{n_i}[r]$ chunks of bitrate $b[r]$ by network n_i with the expected throughput $\tilde{r}_{n_i}[r]$. $\tilde{P}_{n_i}[r]t_{n_i,d}[r]$ is the energy consumption of network interface n_i for chunk download and the remaining term denotes the energy consumption for tail time [22].

Likewise, the arrival rate $e_d[r]$ of LTE data virtual queue is

$$e_d[r] = t_p b[r] N_l[r] + \bar{D}_w[r]\cdot T[r]/(t[r]+T[r]). \tag{19}$$

For $\bar{E}_w[r]$ and $\bar{D}_w[r]$, we assume that the probability that a user stops watching video during $T[r]$ takes a uniform distribution. During $T[r]$, the video chunk remaining in the buffer, which are received via Wi-Fi in the previous *request event*, will be played back for $t_pQ_r[r]$, and the chunk received by LTE in the current *request event* will be played back for the remaining time, i.e., $T[r]-t_pQ_r[r]$. We adopt the approach in [13] to derive $\bar{E}_w[r]$ and $\bar{D}_w[r]$ as closed forms (20) and (21), respectively.

$$\begin{aligned} \bar{E}_w[r] = {} & \frac{1}{T[r]}\Bigg\{\int_0^{t_p\cdot Q_r[r]}\big\{\big(t_p\cdot Q_r[r]-t\big)\cdot e_{e,w}[r-1]\big\}dt + \int_0^{t_p\cdot Q_r[r]}\left(\frac{\min(t_{d,l},t)}{t_{d,l}}\cdot e_{e,l}[r]\right)dt \\ & + \int_0^{t_p\cdot Q_r[r]}\left(\frac{\min(t_{d,w},t)}{t_{d,w}}\cdot e_{e,w}[r]\right)dt + \int_{t_p\cdot Q_r[r]}^{T[r]}\left\{\left(\frac{\min(t_{d,l},t)}{t_{d,l}} - \frac{t-t_p\cdot Q_r[r]}{t_p\cdot N_l[r]}\right\}\cdot e_{e,l}[r]\right)dt \\ & + \int_{t_p\cdot Q_r[r]}^{T[r]}\left(\frac{\min(t_{d,w},t)}{t_{d,w}}\cdot e_{e,w}[r]\right)dt\Bigg\} \\ = {} & \frac{1}{T[r]}\left\{\frac{1}{2}(t_pQ_r[r])^2 e_{e,w}[r-1] + \left(T[r]-\frac{1}{2}t_{l,d}\right)e_{e,l}[r] + \left(T[r]-\frac{1}{2}t_{w,d}\right)e_{e,w}[r] + \frac{(T[r]-t_pQ_r[r])^2}{2t_pN_l[r]}e_{e,l}[r]\right\}. \end{aligned} \tag{20}$$

$$\begin{aligned} \bar{D}_w[r] = {} & \frac{1}{T[r]}\cdot\left\{\frac{1}{2}\big(t_p\cdot Q_r[r]\big)^2\cdot e_{e,w}[r-1] + \left(T[r]-\frac{1}{2}t_{d,l}\right)\cdot e_{e,l}[r]\right. \\ & \left. + \left(T[r]-\frac{1}{2}t_{d,w}\right)\cdot e_{e,w}[r] + \frac{\big(T[r]-t_p\cdot Q_r[r]\big)^2}{2t_p\cdot N_l[r]}\cdot e_{e,l}[r]\right\} \\ = {} & \left(1-\frac{t_{l,d}+t_pN_l[r]}{2T[r]}\right)e_d[r]. \end{aligned} \tag{21}$$

*Chunk Request Policy Determination Algorithm.*Now, we find the optimal $\pi[r]$ which minimizes $DPP_r[r]$ as follows:

$$\begin{aligned} \min_{\pi[r]} \quad & DPP_r[r], \\ \text{subject to} \quad & T[r] = t_p\Big(\hat{N}_w[r-1]+N_l[r]\Big) - T_a[r-1], \\ & \frac{t_pb[r]N_l[r]}{T[r]} \le \tilde{r}_l[r], \frac{t_pb[r]N_w[r]}{T[r]} \le \tilde{r}_w[r], \\ & N_l[r], N_w[r] \in \{0,1,2\ldots,Q_{max}\}, \\ & 1 \le N_l[r]+N_w[r] \le Q_{max}. \end{aligned} \tag{22}$$

The first constraint of (22) is to follow the chunk request policy in Section 4. The second constraint is to prevent requesting more chunks than what the estimated link throughput permits. The third and fourth constraints represent that the video should be requested in a unit of chunk. The run time algorithm to obtain the optimal $\pi[r]$ is summarized in Algorithm 1. At the start of a new *request event*, update Q_e, Q_d, and Q_r, and estimate $\tilde{r}_l$ and $\tilde{r}_w$ (line 5). For all the possible $\pi[r]$ (lines 6–9), the algorithm checks whether π satisfies

6. When updating $Q_e[r]$ and $Q_d[r]$ at $t[r]$, since the video has been played back continuously, we set $\bar{E}_w[r-1]$ and $\bar{D}_w[r-1]$ to 0, and then update $Q_e[r]$ and $Q_d[r]$.

the second constraint, and if π is feasible, it calculates DPP_r (lines 10–11). The π minimizing DPP_r is selected as the optimal request policy (line 13).

Algorithm 1. Optimal Chunk Request Policy Determination

```
Initialize :
 1: $Q_e \leftarrow 0$, $Q_d \leftarrow 0$, $Q_r \leftarrow T_b/t_p$, and $N_{w,prev} \leftarrow 0$
During video playback :
 2: while Video playsbf do
 3:     if New request event starts then
 4:         $N_{w,prev} \leftarrow \hat{N}_w$, $DPP_{min} \leftarrow \infty$
 5:         Update $Q_e$, $Q_d$, $Q_r$, Estimate l $\tilde{r}_l$ and $\tilde{r}_w$
 6:         for $N_t \in \{1, 2, \ldots, Q_{max}\}$ do
 7:             for $N_l \in \{1, 2, \ldots, N_t\}$ do
 8:                 $N_w \leftarrow N_t - N_l$, $T \leftarrow t_p(N_{w,prev} + N_l) - T_a$
 9:                 for $b \in \{b_1, b_2, \ldots, b_m\}$ do
10:                     if $\pi = \{N_l, N_w, T, b\}$ is feasiblebf then
11:                         Calculate $e_e$, $e_d$, $L$, $\Delta$, and $DPP_r$
12:                         if $DPP_r \leq DPP_{min}$ then
13:                             $DPP_{min} \leftarrow DPP_r$, $\pi^{opt} \leftarrow \pi$
14:         Request video chunks according to $\pi^{opt}$
15:     else
16:         Wait for the next start time of request event
```

7 Performance Evaluation

Comparison Scheme. Although many schemes for video streaming have been proposed to improve QoE or energy efficiency, there have been few studies to consider DASH-based video streaming by using both LTE and Wi-Fi simultaneously. We select the energy-efficient HTTP adaptive streaming algorithm (EE-HAS) proposed in [5] as a comparison scheme. To the best of our knowledge, EE-HAS is the only scheme that can be implemented as an application without modifying other protocol stacks. EE-HAS considers energy consumption and LTE data usage constraints to determine the optimal bitrate of video chunks. It emphasizes a trade-off relationship between energy consumption and video quality by using α, which is a real number between zero and one. EE-HAS with smaller α works with more weight on the energy consumption minimization. The major difference between REQUEST and EE-HAS is that EE-HAS divides each video chunk into two segments, which are requested over LTE and Wi-Fi in parallel.

Performance Metrics. the following metrics are used in this work.

[i) *Average video quality:* We measure the average video bitrate during the entire playback time. For REQUEST, the average bitrate is calculated as $\frac{\sum_{r=1}^{R} b[r](\hat{N}_l[r]+\hat{N}_w[r])}{\sum_{r=1}^{R}(\hat{N}_l[r]+\hat{N}_w[r])}$, where R is the index of the last *request event*.

ii) *Rebuffering:* Rebuffering occurs if a video buffer becomes empty or some chunks have not been successfully received within the time limit due to abrupt throughput degradation or Wi-Fi disconnection in the middle of a chunk download. Frequent rebuffering events severely degrade QoE of user [2], [30], [43].

iii) *Amount of energy and LTE data waste:* If a user stops watching a video at time instant t, let the waste of energy and LTE data at t be $E_w(t)$ and $D_w(t)$, respectively. We analyze the effect of energy and LTE data waste on overall power and LTE data usage rate, respectively, by dividing $E_w(t)$ and $D_w(t)$ by t. We refer to $E_w(t)/t$ and $D_w(t)/t$ as the time-normalized energy waste and LTE data waste, respectively.

Parameters. In both simulation and measurement, we use $\alpha = 0.5$ for (16), $Q_{max} = 10$, $t_p = 4$ (s), and $V = 1$. Video starts after downloading two initial chunks, i.e., $T_b = 8$ s. For the measurement, we use 596-second Big-Buck-Bunny video clip.[7] encoded at 20 bitrates, i.e., {0.045, 0.089, 0.128, 0.177, 0.218, 0.256, 0.323, 0.378, 0.509, 0.578, 0.783, 1.009, 1.207, 1.474, 2.087, 2.410, 2.944, 3.341, 3.614, 3.936} (Mbps). For the simulation, each trial runs for 600 s, and uses the same video encoded at 10 bitrates, i.e., {0.23, 0.33, 0.48, 0.69, 0.99, 1.43, 2.06, 2.96, 5.03, 6.00} (Mbps).[8]

7.1 Prototype-Based Evaluation

We implemented both REQUEST and EE-HAS by modifying the open-source DASH Android application, ExoPlayer. We use Samsung Galaxy S5 smartphone (SM-G900) which runs on Android 5.0. For both schemes, the same power model of SHV-E120 smartphone [22] is used to estimate the energy consumption for chunk download and video playback.[9] We conduct an experiment in a lab environment, where average LTE throughput ranges from 10 to 15 Mbps. We set the maximum throughput of Wi-Fi to 10 Mbps by controlling the QoS option in the setting window of Wi-Fi AP. In the middle of video playback, we degrade Wi-Fi throughput to less than 2 Mbps by adding a contending node with saturated UDP uplink traffic for 180 s.

7.1.1 REQUEST with Various Resource Constraints

We first evaluate REQUEST with various power, p_{av} (W), and LTE data, d_{av} (Mbps), constraints, i.e., $\{(p_{av}, d_{av}) | p_{av} \in \{1.5, 2, 3\}$, $d_{av} \in \{1, 2, 3, 4\}\}$. Fig. 4 shows the average bitrate, power, and LTE data usage rate during the entire playback. During the period when there is no Wi-Fi background traffic from other connected devices, REQUEST fully utilizes Wi-Fi link to save LTE data usage while satisfying the power constraint, and hence, the average LTE data usage is generally less than the LTE data constraint. Obviously, REQUEST can achieve almost the highest bitrate with the highest power and LTE data constraints. Besides, with tight resource constraints, REQUEST tends to select lower average bitrate to satisfy the resource constraints. For all the cases, REQUEST satisfies the constraints.

7.1.2 REQUEST without Resource Waste Consideration

REQUEST considers the expected energy and LTE data waste when calculating the arrival rate of energy and LTE data virtual queues using (3) and (8). To evaluate the resource saving

7. http://www-itec.uni-klu.ac.at/ftp/datasets/DASHDataset2014/BigBuckBunny/4sec/.

8. For simulation, we used video encoded with higher bitrate than the measurement case, since LTE and Wi-Fi throughput measured in real environments was much higher than the highest video bitrate used in the measurement.

9. Our work is independent of power models, i.e., any power model can be used for this REQUEST algorithm.

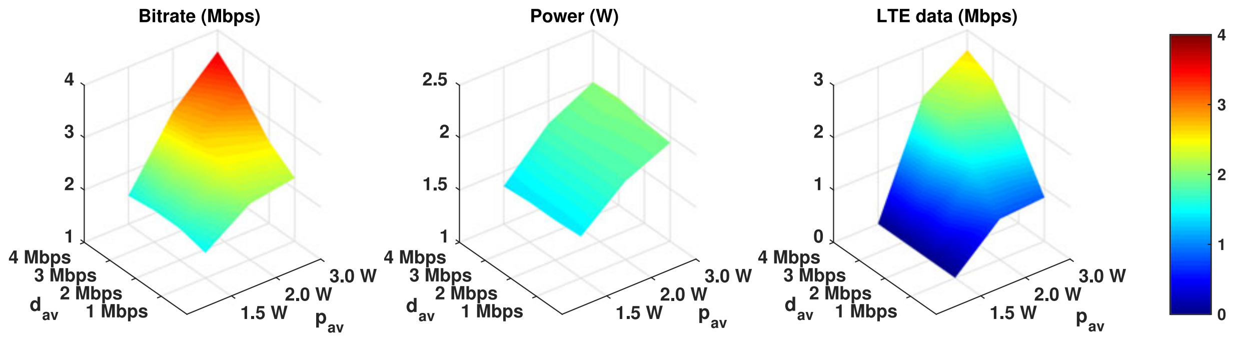
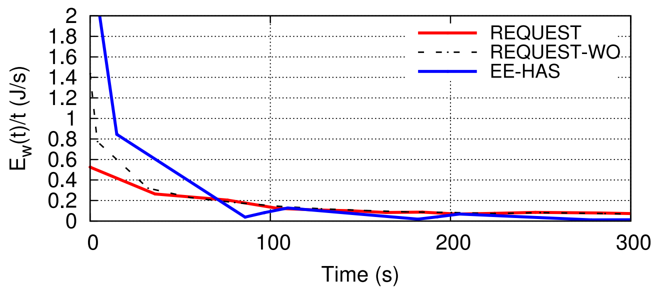
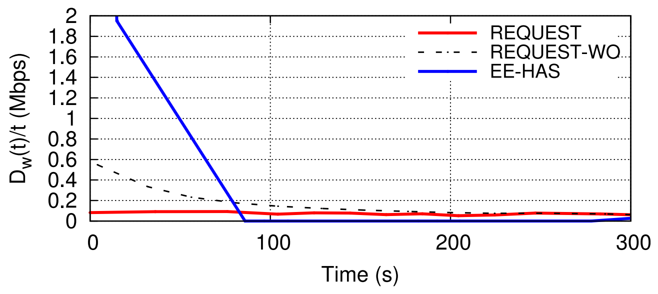


Fig. 4. Average bitrate, power, and LTE data usage rate of REQUEST. The x-label and y-label of each colormap are power and LTE data constraints, respectively.



(a) Time-normalized energy waste.



(b) Time-normalized LTE data waste.

Fig. 5. Time-normalized energy and LTE data waste comparison for REQUEST, REQUEST-WO, and EE-HAS.

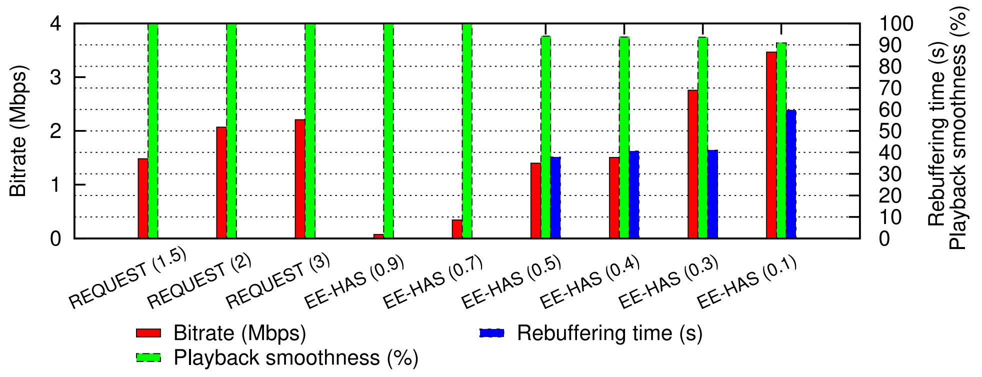


Fig. 6. Average bitrate, playback smoothness, and rebuffering time of REQUEST (p_{av}) and EE-HAS (α) with various p_{av} and α. For all the cases, $d_{av} = 1\ Mbps$.

by considering the expected waste, we implemented another version of REQUEST, namely REQUEST-WO, which does not consider the expected waste. Fig. 5 shows the expected waste, i.e., $E_w(t)$ and $D_w(t)$, and normalized waste, i.e., $E_w(t)/t$ and $D_w(t)/t$, when p_{av}=2 W and d_{av}=1 Mbps. For EE-HAS, $\alpha = 0.3$. Since EE-HAS is prefetching aggressively, it is more wasteful than REQUEST in the early stage of video playback. Before the video plays for 50 s, both $E_w(t)/t$ and $D_w(t)/t$ of REQUEST are smaller than half of those of REQUEST-WO. REQUEST reduces the waste by selecting lower bitrate and a smaller number of chunks to request compared to REQUEST-WO at the beginning of the video streaming. Since the resource waste is relatively high compared to the total resource usage if a user stops watching video in a few seconds, REQUEST can save more resources when the user quits video early. REQUEST can be a more resource-saving option, and enabling REQUEST-WO can be also an alternative if the user simply wants to watch higher video quality with more potential waste.

7.1.3 Comparison Study

For comparative evaluation of REQUEST, we run REQUEST and EE-HAS with various p_{av} and α. Fig. 6 presents the average bitrate and total rebuffering time during the entire playback time of the 596-second video. Playback smoothness (%) is defined as $\frac{100 \cdot \text{Total playback time}}{\text{Total playback time} + \text{rebuffering time}}$. EE-HAS with small α increases bitrate but suffers from long total rebuffering time, i.e., 41.0 s for $\alpha = 0.3$, and total 59.6 s for $\alpha = 0.1$ when the background traffic starts to degrade Wi-Fi throughput abruptly. Even though EE-HAS with large α, which operates more energy efficiently, does not suffer rebuffering, the average bitrate is quite low, i.e., below 100 kbps with $\alpha = 0.9$. However, REQUEST retains 100 percent playback smoothness for all the power constraints with over 1 Mbps average bitrate. In addition, the average bitrates of EE-HAS with $\alpha = 0.4$ and REQUEST with $p_{av} = 1.5$ are similar, i.e., over 1.5 Mbps, but EE-HAS shows only 93.6 percent playback smoothness (40.5 s

TABLE 2
Bitrate and Rebuffering of EE-HAS ($\alpha = 0.1$) and REQUEST ($p_{av} = 2$) for Five Traces

Trace	EE-HAS						REQUEST	
	RT (s)	# ER	ER time (s)	# OR	OR time (s)	Bitrate (Mbps)	RT (s)	Bitrate (Mbps)
1 (office)	4	1	4	0	0	5.27	0	5.17
2 (station)	16.39	3	5.33 ± 1.89	1	0.4	3.03	0.30	5.14
3 (station2)	43.83	3	8 ± 3.27	3	6.61 ± 3.81	2.00	0	2.22
4 (cafe)	12	3	4	0	0	4.06	0	5.17
5 (cafe2)	32.65	3	6.67 ± 1.89	16	0.79 ± 0.16	2.59	0	5.23

rebuffering time) while REQUEST provides seamless playback. In summary, in contrast to EE-HAS which provides either too low bitrate without rebuffering or high bitrate with long rebuffering time, REQUEST achieves enhanced quality without rebuffering.

7.2 Trace-Driven Simulation

To comparatively evaluate the performance of REQUEST, we implement both REQUEST and EE-HAS using Matlab. For a fair comparison, measured LTE and Wi-Fi TCP throughput traces are used. We first measured the LTE and Wi-Fi TCP throughput using *Iperf* with SM-G900 every second in various places, e.g., office, subway station, and cafeteria. A desktop in a lab is used as a server for *Iperf*.

We use $\alpha = 0.1$ for EE-HAS and $p_{av} = 2\ W$ for REQUEST. d_{av} is 1.5 Mbps. We classify rebuffering events into two types, i.e., i) rebuffering due to empty video buffer (empty buffer rebuffering) and ii) rebuffering due to absence of video chunks to be played now even though video buffer is filled with video chunks to be played later (out-of-order rebuffering). The out-of-order rebuffering may occur especially with EE-HAS, where a chunk may be divided into two segments that are requested over LTE and Wi-Fi, separately in parallel. If these segments are not received by the player on time, a partially received chunk cannot be decoded and played back, thus causing out-of-order rebuffering. In Table 2, # ER, and ER time denote the number of empty buffer rebuffering events and the average buffering time per event. # OR and OR time are those of the out-of-order rebuffering. RT denotes the total rebuffering time of REQUEST, and bitrate (Mbps) is the average bitrate during playback. For all the cases, rebuffering time of REQUEST is almost zero while EE-HAS suffers from frequent rebuffering events, e.g., 19 rebuffering events and total 32.65 s rebuffering time for the third trace. The bitrate of REQUEST is similar to or higher than that of EE-HAS while REQUEST avoids rebuffering even in mobile and dense environments where LTE and Wi-Fi throughput are often unstable. Therefore, REQUEST may consume more data than EE-HAS in some cases. This is expected since REQUEST's purpose is to ensure better quality while satisfying the data and energy constraints.

8 Discussion

In this section, we discuss how well REQUEST controls the trade-off between resource usage and video quality. In Section 5, we adopt Lyapunov optimization framework to maximize time-average video utility and satisfy time-average energy/LTE data resource constraints. Time-average resource constraints are translated into virtual queues with (3) and (8) which are based on traditional virtual queue concept of Lyapunov optimization. We point out that drift-plus-penalty algorithm with the current version of virtual queues tends to utilize resources conservatively. This behavior allows the time-average resource constraints to be easily satisfied, but time-average utility is farther from the optimum.

Furthermore, in Section 7, V is fixed without run time adaptation where V is a positive constant to provide a trade-off between virtual queues' backlogs and the gap from a theoretical optimal video utility. We expect that run time adaptation of V controls the trade-off between resource usage and video utility better, thus enhancing video utility more under resource constraints.

8.1 Modification of Resource Virtual Queues

We revisit the Equations (3) and (9) for queue backlog evolution $Q_e[r]$ and Lyapunov function $L[r]$. $L[r]$ is related to the squared backlogs, which is good when they are close to zero. Since backlog cannot be negative, we use the max function as in (3). However, this max function makes resource utilization more conservative. This makes it difficult to enhance video utility even though there are enough resources left. To allow more aggressive resource utilization, we call $\tilde{Q}_e[r]$ the modified virtual queue with max function removed from $Q_e[r]$, which is defined as follows:

$$\tilde{Q}_e[r+1] = \tilde{Q}_e[r] + e_e[r] - p_{av}T[r]. \tag{23}$$

$\tilde{Q}_d$ is similarly defined to $\tilde{Q}_e$ as follows:

$$\tilde{Q}_d[r+1] = \tilde{Q}_d[r] + e_d[r] - d_{av}T[r]. \tag{24}$$

Then, the modified $L[r]$ is defined as follows:

$$\begin{aligned} L[r] =& \frac{1}{2}\Big(\big(\max(\tilde{Q}_e[r],0)\big)^2 + \big(\max(\tilde{Q}_d[r],0)\big)^2 \\ &+ \Big(\frac{Q_{max}}{2} - Q_r[r]\Big)^2\Big) \end{aligned} \tag{25}$$

Table 3 is a simple example which shows a comparison of the values of Q_e and $\tilde{Q}_e$ for the same arrival e_e and departure $p_{av}T$. When $r = 4$, since $\sum_{r=1}^{4} e_e[r] = 62$, and $\sum_{r=1}^{4} p_{av}T[r] = 70$, $\sum_{r=1}^{4} e_e[r] < \sum_{r=1}^{4} p_{av}T[r]$ which satisfies time-average energy constraint until $r = 4$. At $r = 5$, $e_e[5](= 20)$ is greater than $p_{av}T[5](= 15)$ by 5, and $Q_e[5] = Q_e[4] + 5 = 5$ which makes $L[5]$ $(\leq \frac{25}{2})$ a relatively large positive number. Therefore, $DPP_r[5]$ becomes large, which causes to act as a

TABLE 3
Simple Example of $\tilde{Q}_e$

r	1	2	3	4	5	6
$e_e[r]$	8	22	12	20	20	21
$p_{av}T[r]$	10	20	10	30	15	20
$Q_e[r+1]$	0	2	4	0	5	6
$\tilde{Q}_e[r+1]$	-2	0	2	-8	-3	-2
$\max(\tilde{Q}_e[r+1], 0)$	0	0	2	0	0	0

drift reduction, i.e., towards less resource usage to reduce energy queue backlogs at the next request event. This behavior can be one reason for unnecessarily dropping the bitrate at the next request event. However, since the equation $\sum_{r=1}^{5} e_e[r] < \sum_{r=1}^{5} p_{av}T[r]$ is still satisfied, it does not need to operate conservatively at all.

On the other hand, $\tilde{Q}_e$ is smaller than zero, and hence, it does not affect the increase of drift. This tendency enhances the video utility by aggressively utilizing resources.

8.2 Adaptation of V

V provides a trade-off between video quality and resource utilization. With large V, REQUEST decides to use more resources to enhance video utility, On the other hand, with small V, REQUEST tends to lower video utility to prevent increasing backlogs of virtual queues. Since it is difficult to theoretically obtain the optimal V, we propose a run time adaptation algorithm which heuristically controls V based on the variation of virtual queues' backlogs. A heuristic V adaptation algorithm is described in Algorithm 2.

Algorithm 2. V Adaptation Algorithm

```
Initialize :
 1: $V \leftarrow 1, \tilde{Q}_e \leftarrow 0, \tilde{Q}_{e_{prev}} \leftarrow 0, \tilde{Q}_d \leftarrow 0$, and $\tilde{Q}_{d_{prev}} \leftarrow 0$
During video playback :
 2: whileVideo playsbf do
 3:     if New request event starts then
 4:         Update $\tilde{Q}_e$ and $\tilde{Q}_d$ based on $\pi^{opt}$
 5:         if $\tilde{Q}_e > \tilde{Q}_{e_{prev}} > 0$ or $\tilde{Q}_d > \tilde{Q}_{d_{prev}} > 0$ then
 6:             $w \leftarrow \min\big(2, \frac{1}{2}\big(\tilde{Q}_e/\tilde{Q}_{e_{prev}} + \tilde{Q}_d/\tilde{Q}_{d_{prev}}\big)\big)$
 7:             $V \leftarrow V/w$
 8:         if $\tilde{Q}_e < \tilde{Q}_{e_{prev}} < 0$ and $\tilde{Q}_d < \tilde{Q}_{d_{prev}} < 0$ then
 9:             $w \leftarrow \min\big(2, \frac{1}{2}\big(\tilde{Q}_e/\tilde{Q}_{e_{prev}} + \tilde{Q}_d/\tilde{Q}_{d_{prev}}\big)\big)$
10:             $V \leftarrow V \cdot w$
11:         $\tilde{Q}_{e_{prev}} \leftarrow \tilde{Q}_e$ and $\tilde{Q}_{d_{prev}} \leftarrow \tilde{Q}_d$
12:         Return V to Algorithm 1
13:     else
14:         Wait for the next start time of request event
```

The main concept of Algorithm 2 is that it reduces V when virtual queue's backlog continues to increase, and conversely, increases V when backlog keeps decreasing in negative numbers. Since it is difficult to satisfy the time-average resource constraints with the current value of V when either $\tilde{Q}_e$ or $\tilde{Q}_d$ continues to increase, it is desirable to reduce V so that Algorithm 1 selects a policy which reduces the backlog as an optimal policy. If both $\tilde{Q}_e$ and $\tilde{Q}_d$ are less than zero and continue to decrease, it means that it would be possible to satisfy time-average constraints even if Algorithm 1 selected a policy with higher video utility. In this case, it is better to increase V to make Algorithm 1 select a policy with higher video utility by using more resources.

In order to determine how much to increase and decrease V, we define a weighting factor w. w is determined in relation to the amount of variation of $\tilde{Q}_e$ and $\tilde{Q}_d$, derived as lines 2 and 2 in Algorithm 2. The weighting factor w is limited to less than two to prevent performance degradation due to abrupt change of V. As shown in line 8, the algorithm increases V only if both $\tilde{Q}_e$ and $\tilde{Q}_d$ are less than zero and continue decreasing, i.e., both energy and LTE data resources are used less than time-average constraints. On the contrary, when either $\tilde{Q}_e$ or $\tilde{Q}_d$ is larger than zero and continues to increase, the algorithm decreases V to impose more weight for reducing virtual queue's backlog to satisfy both time-average resource constraints.

TABLE 4
Simulation Results for Performance Comparison Among REQUEST, REQUEST-Q, and REQUEST-QV

	REQUEST	REQUEST-Q	REQUEST-QV
Bitrate (Mbps)	5.10 ± 0.47	5.24 ± 0.57	5.29 ± 0.66
$\delta_p (= \frac{p - p_{av}}{p_{av}})$	-0.33 ± 0.11	-0.21 ± 0.14	-0.19 ± 0.14
$\delta_d (= \frac{d - d_{av}}{d_{av}})$	-0.98 ± 0.03	-0.50 ± 0.37	-0.42 ± 0.38

8.3 Trace-Driven Simulation

Through trace-driven Matlab simulation, we evaluate the performance of REQUEST with the modified virtual queues and V adaptation algorithm. We use LTE and Wi-Fi throughput traces for 600 s in five difference places. Time-average constraints, p_{av} and d_{av} are configured to $p_{av} \in 2, 2.5, 3(W)$, and $d_{av} \in 1, 2, 3, 4, 6$ (Mbps), respectively. We run Matlab simulation for 75 cases which are the combinations of five throughput traces, three energy constraints, and five LTE data constraints. Table 4 shows the performance of three versions of REQUEST in terms of three evaluation metrics, i.e, bitrate, power usage deviation δ_p, and LTE data usage deviation δ_d. δ_p and δ_d indicate how well the resources are used during video playback. $\delta_p (= \frac{p - p_{av}}{p_{av}})$ is defined as the ratio of the difference between the average power p and target power p_{av} to the target power. If $\delta_e < 0$, the average power during video playback p is less than the target power p_{av}, i.e., a smartphone saves the energy compared to the target energy usage. On the other hand, $\delta_p > 0$, a smartphone consumes power more than p_{av}, i.e., it uses excessive battery energy. δ_d is defined similar to δ_p, and has a similar characteristic.

V is fixed to one in REQUEST while REQUEST-Q modifies the virtual queues from Q_e and Q_d to $\tilde{Q}_e$ and $\tilde{Q}_d$, respectively. REQUEST-QV adapts V with Algorithm 2 and the modified virtual queues. As shown in Table 4, REQUEST-QV outperforms REQUEST and REQUEST-Q in terms of the bitrate. With these throughput traces, the δ_p and δ_d of REQUEST, REQUEST-Q, and REQUEST-QV are less than zero, which means that actual resource usages are less than target resource usages. However, REQUEST-QV utilizes resources more aggressively while not exceeding the target resource usages by adapting V in run time. This results in the enhanced video bitrate of REQUEST-QV.

9 Conclusion

In this paper, we propose REQUEST which utilizes both LTE and Wi-Fi networks to provide seamless video streaming under resource constraints. The proposed chunk request policy of REQUEST ensures that all the video chunks are received even in unstable network environments. REQUEST achieves near-optimal time-average video quality while satisfying time-average resource constraints by adopting Lyapunov optimization framework. Our prototype-based evaluation and trace-driven simulation demonstrate that REQUEST provides seamless video streaming by using both LTE and Wi-Fi links in real-world environments.

In addition, enhancements in terms of virtual queue remodeling and tradeoff coefficient V adaptation are finally discussed, and the corresponding performance improvements are verified.

Acknowledgments

This work was supported by the Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2015-0-00300, Multiple Access Technique with Ultra-Low Latency and High Efficiency for Tactile Internet Services in IoT Environments). An earlier version of this paper appeared in the *Proceedings of the 25th ACM Multimedia* (*MM*), October 2017 [1].

References

[1] J. Koo, J. Yi, J. Kim, M. A. Hoque, and S. Choi, "REQUEST: Seamless dynamic adaptive streaming over HTTP for multi-homed smartphone under resource constraints," in *Proc. ACM Multimedia Conf.*, 2017, pp. 934–942.

[2] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for Internet video," in *Proc. ACM SIGCOMM Conf. SIGCOMM*, 2013, pp. 339–350.

[3] Cisco, "Cisco visual networking Index: Forecast and methodology, 2016–2021," Cisco White Paper, Document ID:1465272001663118, Sept. 2017.

[4] D. H. Bui, K. Lee, S. Oh, I. Shin, H. Shin, H. Woo, and D. Ban, "GreenBag: Energy-efficient bandwidth aggregation for real-time streaming in heterogeneous mobile wireless networks," in *Proc. IEEE 34th Real-Time Syst. Symp.*, 2013, pp. 57–67.

[5] Y. Go, O. C. Kwon, and H. Song, "An energy-efficient HTTP adaptive video streaming with networking cost constraint over heterogeneous wireless networks," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1646–1657, Sep. 2015.

[6] S. Bae, D. Ban, D. Han, J. Kim, K.-h. Lee,Lee, S. Lim, W. Park, and C.-k. Kim, "StreetSense: Effect of bus Wi-Fi APs on pedestrian smartphone," in *Proc. ACM Internet Meas. Conf.*, 2015, pp. 347–353.

[7] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, pp. 536–550, Apr. 2013.

[8] M. A. Hoque, M. Siekkinen, and J. K. Nurminen, "Using crowdsourced viewing statistics to save energy in wireless video streaming," in *Proc. ACM 19th Annua. Int. Conf. Mobile Comput. Netw.*, 2013, pp. 377–388.

[9] "ExoPlayer," 2016. [Online]. Available: http://google.github.io/ExoPlayer/

[10] J. Chen, A. Ghosh, J. Magutt, and M. Chiang, "QAVA: Quota aware video adaptation," in *Proc. ACM 8th Int. Conf. Emerging Netw. Experiments Technol.*, 2012, pp. 121–132.

[11] "Galaxy S5 Download Booster," 2016. [Online]. Available: http://galaxys5guide.com/samsung-galaxy-s5-features-explained/galaxy-s5-download-booster, Accessed on: Oct. 18, 2016.

[12] "Airplug," 2016. [Online]. Available: http://www.airplug.com/solution.php, Accessed on: Oct. 18, 2016

[13] W. Lee, J. Koo, S. Jin, and S. Choi, "EQ-Video: Energy and quota-aware video playback time maximization for smartphones," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 1045–1048, Jun. 2015.

[14] X. Zhu, P. Agrawal, J. P. Singh, T. Alpcan, and B. Girod, "Distributed rate allocation policies for multihomed video streaming over heterogeneous access networks," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 752–764, Jun. 2009.

[15] J. Wu, B. Cheng, C. Yuen, Y. Shang, and J. Chen, "Distortion-aware concurrent multipath transfer for mobile video streaming in heterogeneous wireless networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 4, pp. 688–701, Apr. 2015.

[16] J. Wu, J. Yang, Y. Shang, B. Cheng, and J. Chen, "SPMLD: Subpacket based multipath load distribution for real-time multimedia traffic," *J. Commun. Netw.*, vol. 16, no. 5, pp. 548–558, Oct. 2014.

[17] C. Xu, T. Liu, J. Guan, H. Zhang, and G.-M. Muntean, "CMT-QA: Quality-aware adaptive concurrent multipath data transfer in heterogeneous wireless networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 11, pp. 2193–2205, Nov. 2013.

[18] J. Wu, C. Yuen, B. Cheng, M. Wang, and J. Chen, "Streaming high-quality mobile video with multipath TCP in heterogeneous wireless networks," *IEEE Transactiosn Mobile Comput.*, vol. 15, no. 9, pp. 2345–2361, Sep. 2016.

[19] J. Wu, B. Cheng, M. Wang, and J. Chen, "Energy-efficient bandwidth aggregation for delay-constrained video over heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 30–49, Jan. 2017.

[20] J. Wu, B. Cheng, M. Wang, J. Chen, J. Wu, B. Cheng, M. Wang, and J. Chen, "Quality-aware energy optimization in wireless video communication with multipath TCP," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2701–2718, Oct. 2017. [Online]. Available: https://doi.org/10.1109/TNET.2017.2701153

[21] S. Dimatteo, P. Hui, B. Han, and V. O. Li, "Cellular traffic offloading through WiFi networks," in *Proc. IEEE 8th Int. Conf. Mobile Ad-Hoc Sensor Syst.*, 2011, pp. 192–201.

[22] J. Koo, W. Lee, Y. Park, and S. Choi, "PIMM: Packet interval-based power modeling of multiple network interface-activated smartphones," in *Proc. ACM 6th Int. Conf. Future Energy Syst.*, 2015, pp. 111–120.

[23] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in *Proc. ACM 10th Int. Conf. Mobile Syst. Appl. Services*, 2012, pp. 225–238.

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

[25] X. Li, M. Dong, Z. Ma, and F. Fernandes, "GreenTube: Power optimization for mobile video streaming via dynamic cache management," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 279–288.

[26] W. Hu and G. Cao, "Energy-aware video streaming on smartphones," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 1185–1193.

[27] "Gpac," 2016. [Online]. Available: https://gpac.wp.mines-telecom.fr

[28] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the internet," *IEEE Multimedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011.

[29] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," in *Proc. ACM SIGCOMM Comput. Commun.*, 2014, pp. 187–198.

[30] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, "BOLA: Near-optimal bitrate adaptation for online videos," in *Proc. IEEE 35th Annu. Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[31] J. Kim, G. Caire, and A. F. Molisch, "Quality-aware streaming and scheduling for device-to-device video delivery," *IEEE/ACM Trans. Netw.*, vol. 24, no. 4, pp. 2319–2331, Aug. 2016.

[32] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.

[33] V. Mhatre and K. Papagiannaki, "Using smart triggers for improved user performance in 802.11 wireless networks," in *Proc. ACM Annu. Int. Conf. Mobile Comput. Netw.*, 2006, pp. 246–259.

[34] I. Ramani and S. Savage, "SyncScan: Practical fast handoff for 802.11 infrastructure networks," in *Proc. IEEE 24th Annu. Joint Conf. Comput. Commun. Societies*, 2005, pp. 675–684.

[35] J. Shi, L. Meng, A. Striegel, C. Qiao, D. Koutsonikolas, and G. Challen, "A walk on the client side: Monitoring enterprise WiFi networks using smartphone channel scans," in *Proc. IEEE 35th Annu. Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[36] K.-H. Kim, H. Nam, and H. Schulzrinne, "WiSlow: A Wi-Fi network performance troubleshooting tool for end users," in *Proc. IEEE Conf. Comput. Commun.*, 2014, pp. 862–870.

[37] S. Rayanchu, A. Patro, and S. Banerjee, "Airshark: Detecting non-WiFi RF devices using commodity WiFi hardware," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, 2011, pp. 137–154.

[38] Q. Song and A. Jamalipour, "A network selection mechanism for next generation networks," in *Proc. IEEE Int. Conf. Commun.*, 2005, pp. 1418–1422.

[39] Q.-T. Nguyen-Vuong, N. Agoulmine, and Y. Ghamri-Doudane, "A user-centric and context-aware solution to interface management and access network selection in heterogeneous wireless environments," *Comput. Netw.*, vol. 52, no. 18, pp. 3358–3372, 2008.

[40] "RFC 7233 (HTTP/1.1 range requests)," 2014. [Online]. Available: https://tools.ietf.org/html/rfc7233

[41] Z. Yan and C. W. Chen, "RnB: Rate and brightness adaptation for rate-distortion-energy tradeoff in HTTP adaptive streaming over mobile devices," in *Proc. ACM 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 308–319.

[42] M. J. Neely, "Dynamic optimization and learning for renewal systems," *IEEE Trans. Autom. Control*, vol. 58, no. 1, pp. 32–46, Jan. 2013.

[43] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *Proc. ACM SIGCOMM Conf.*, 2011, pp. 362–373.

Jonghoe Koo received the BS and PhD degrees from the Department of Electrical and Computer Engineering, Seoul National University, Korea, in 2011 and 2017, respectively. He has been a research engineer with Samsung Research, Samsung Electronics, Seoul, Korea, since 2017. His current research interests include energy-aware mobile computing and reliable video streaming.

Juheon Yi received the BS degree in electrical engineering from Seoul National University (SNU), Seoul, Korea, in 2016. He is currently working towards the MS degree at SNU. His research interests include wireless networking and mobile computing, in particular, mobile deep learning framework.

Joongheon Kim (M'06–SM'18) received the BS and MS degrees in computer science and engineering from Korea University, Seoul, Korea, in 2004 and 2006, and the PhD degree in computer science from the University of Southern California (USC), Los Angeles, CA, in 2014, with two additional MS degrees in electrical engineering and computer science (specialization in high performance computing and simulations). He has been an assistant professor with Chung-Ang University, Seoul, Korea, since 2016. In industry, he was with LG Electronics Seocho R&D Campus (Seoul, Korea, 2006–2009), InterDigital (San Diego, CA, 2012), and Intel Corporation (Santa Clara, CA, 2013–2016). He is a senior member of the IEEE; and a member of the ACM and IEEE Communications Society. He was awarded the Annenberg graduate fellowship with his PhD admission from USC (2009).

Mohammad Hoque received the MSc degree in computer science, in 2010, and the PhD degree from Aalto University, Finland, in 2013. He is a postdoctoral researcher with the University of Helsinki. His research interest includes energy-aware computing, distributed systems, and data analytics.

Sunghyun Choi (S'96–M'00–SM'05–F'14) received the BS (summa cum laude) and MS degrees from the Korea Advanced Institute of Science and Technology, in 1992 and 1994, respectively, and the PhD degree from the University of Michigan, Ann Arbor, in 1999. He is a professor with the Department of Electrical and Computer Engineering, Seoul National University (SNU), Korea. Before joining SNU in 2002, he was with Philips Research USA. He was also a visiting associate professor at Stanford University, from June 2009 to June 2010. His current research interests include area of wireless/mobile networks. He authored/coauthored more than 170 technical papers and book chapters in the areas of wireless/mobile networks and communications. He has co-authored (with B. G. Lee) a book *Broadband Wireless Access and Local Networks: Mobile WiMAX and WiFi*, Artech House, 2008. He holds more than 100 patents, and has tens of patents pending. He has served as a general co-chair of COMSWARE 2008, and a technical program committee co-chair of ACM Multimedia 2007, IEEE WoWMoM 2007, and COMSWARE 2007. He has also served on program and organization committees of numerous leading wireless and networking conferences including ACM MobiCom, IEEE INFOCOM, IEEE SECON, and IEEE WoWMoM. He is also currently serving on the editorial boards of the *IEEE Transactions on Mobile Computing* and the *IEEE Wireless Communications*. He has served as a guest editor of the *IEEE Journal on Selected Areas in Communications* (*JSAC*) and *ACM Wireless Networks* (*WINET*). From 2000 to 2007, he was an active voting member of the IEEE 802.11 WLAN Working Group. He has received a number of awards including the Presidential Young Scientist Award (2008), IEEK/IEEE Joint Award for Young IT Engineer (2007), KICS Dr. Irwin Jacobs Award (2013), Shinyang Scholarship Award (2011), the Outstanding Research Award (2008), and the Best Teaching Award (2006) both from the College of Engineering, SNU; and the Best Paper Award from IEEE WoWMoM 2008. He is a fellow of the IEEE.