Exploiting Scalable Video Coding for Content Aware Downlink Video Delivery over LTE

Ahmed Ahmedin[1], Kartik Pandit[1], Dipak Ghosal[1], and Amitabha Ghosh[2,*]

[1] Department of Computer Science, University of California, Davis, CA
{kdpandit,ahmedin,dghosal}@ucdavis.edu
[2] UtopiaCompression Corporation, Los Angeles, CA
amitabhg@utopiacompression.com

Abstract. We propose a content aware scheduler to allocate resources for video delivery on the downlink of a Long Term Evolution (LTE) network. We consider multiple users subscribe to a video streaming service, and request videos encoded in H.264 Scalable Video Coding format. The scheduler maximizes the average video quality across all users by assigning resource blocks based on their device capabilities, link qualities, and available resources. We measure video quality using two full reference metrics: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index. We formulate the video delivery problem first as an integer linear program (ILP), and then reduce it to the multiple choice knapsack problem (MCKP). To solve the MCKP, we propose two fast heuristics with reduced processing overhead at the eNodeB, and a fully polynomial-time approximate scheme (FPTAS) using dynamic programming and profit-scaling. Our evaluation results indicate that the heuristics are within a factor of $\frac{1}{2}$, and the FPTAS is very close to the optimal obtained from an ILP solver. We also propose a signaling mechanism to implement the content aware scheduler in existing LTE systems, and evaluate the impact of signaling delay on video distortion using both indoor and outdoor measurements collected from AT&T and T-Mobile networks.

Keywords: LTE, Scalable Video Coding, content aware optimization, scheduler, network optimization, FPTAS, water-filling.

1 Introduction

The continuous growth in cellular data traffic is encouraging service providers to introduce new services and compete with each other to deliver the highest quality at the lowest price. Multimedia delivery is one of the most rapidly evolving services, as smart handheld devices (e.g., iPhone, iPad, tablet) and high-speed 4G technologies (e.g., LTE, WiMAX) are fast getting adopted [1]. It is projected that 70% of the cellular data traffic will be from video by 2016 [2].

The user equipments (UEs) in a cellular network can be very diverse, ranging from battery and hardware constrained cell phones, to more powerful tablets

* A. Ghosh did this work as a postdoctoral research associate at Princeton University.

M. Chatterjee et al. (Eds.): ICDCN 2014, LNCS 8314, pp. 423–437, 2014.

with sophisticated transcoding features. Different users are also susceptible to different video qualities due to limited bandwidth and random channel variations resulting from shadowing, multipath fading, etc. These factors can cause the UE buffer to underflow during video playback. The eNodeB (term used for LTE base transceiver station) can also run out of resources without satisfying all the requests. In particular, when a large number users demand high quality videos at the same time, severe buffer underflows may occur for multiple users.

The H.264 Scalable Video Coding (SVC) [8] has emerged as a suitable coding standard for compressing high-quality video bitstreams. It supports a variety of devices using three different scalability options: (1) temporal scalability, where complete frames can be dropped from a video using motion dependencies; (2) spatial scalability, where videos are encoded at multiple resolutions; and (3) quality scalability, where decoded samples of lower qualities can be used to predict samples of higher qualities to reduce the bit rate required to encode the higher qualities. A UE can use any of these scalability options, or combine them based on the type of the video and user requirements. By leveraging multiple profiles supported by SVC that differ in compression, bit rate, and size, the video quality can be adapted based on link quality, device capability, and available resource blocks (referred to as physical resource blocks or PRBs in LTE).

There has been a lot of work in content aware networking for wireless video delivery, including choosing the best network code for video transmission over mesh networks [3], cross-layer solution with more protection for packets carrying important parts (e.g., I-frames) [4], and streaming SVC videos over WiMAX [7]. A similar method to [4] for content aware video delivery on the uplink of a wideband code division multiple access (WCDMA) network is proposed in [6]. Video frame scheduling under deadline constraints in the downlink is discussed in [5], while SVC tools for wireless are introduced in [8]. The performance of SVC over LTE is characterized in [9].

In this paper, we present a content aware PRB scheduler to deliver SVC encoded videos to multiple users on the downlink of an LTE network. Our goal is to maximize the average video quality across all users for a fixed number of PRBs. The PRB scheduler in the eNodeB decides the profile levels of the videos, and the number of PRBs to assign to each user depending on its decoding capability and link quality between the eNodeB and the UE. We assume that these link qualities can be estimated from feedback signals, such as channel quality indicator (CQI) and hybrid automatic repeat request (HARQ).

Our key contributions are the following:

- We formulate the PRB scheduling problem as an integer linear problem (ILP), and reduce it to the multiple choice knapsack problem (MCKP) [15].
- We propose a greedy heuristic and a water-filling heuristic to solve the MCKP with reducing processing complexity at the eNodeB.
- We also propose a fully polynomial-time approximation scheme (FPTAS) using dynamic programming and profit-scaling to solve the MCKP.
- We compare the performance of the heuristics and the FPTAS with the optimal by solving the ILP using CPLEX [18], a state-of-the-art ILP solver

developed by IBM. Our results indicate that the heuristics perform within a factor of $\frac{1}{2}$, and the FPTAS is very close to the optimal.
– We propose a signaling mechanism to implement the content aware PRB scheduler in an existing LTE system, and evaluate the impact of signaling delay on video distortion using both indoor and outdoor (urban and suburban) measurements collected from AT&T and T-Mobile networks.

The rest of the paper is organized as follows. In Section 2, we describe our system model and formulate the PRB scheduling problem. In Section 3, we first map the PRB scheduling problem to the MCKP, and present two heuristics and an FPTAS to solve the MCKP. Section 4 presents our evaluation results of the proposed heuristics and the FPTAS. In Section 5, we describe a signaling mechanism to implement the content aware PRB scheduler in an existing LTE system, and also present our evaluation results of this modified architecture based on measurement data. Finally, we conclude in Section 6.

2 LTE System Model

In this section, we first describe a high-level architecture of the content aware PRB scheduler in an LTE downlink, and define two video quality metrics. We then present the LTE video model and formulate the PRB scheduling problem.

2.1 Content Aware LTE Downlink Architecture

We consider the downlink of a single eNodeB in an LTE network where multiple users request SVC-encoded videos from a video server (e.g., YouTube). The Core Network (CN) establishes a non-guaranteed bit rate Evolved Packet System (EPS) bearer that provides Internet Protocol (IP) services to the UEs. The scheduler at the eNodeB allocates a certain number of PRBs to send the video as a unicast to each UE. A schematic diagram of this architecture is shown in Figure 1. The solid lines indicate different interfaces that already exist between different nodes in the EPS bearer. The dotted lines are the new conceptual interfaces we propose, the implementation of which is described in Section 5.

We envision that the content aware PRB scheduler is conceptually associated with the eNodeB. When a UE requests a video, the video server responds with the quality and transcoding information of that video. The eNodeB obtains this information from the UE, and sends it along with the set of available PRBs to the PRB scheduler. The PRB scheduler also obtains the channel quality from the UE, and then computes the number of PRBs and a video rate to be assigned to the UE corresponding to an SVC profile level. The profile level is sent to the UE, and the PRB assignment is sent to the scheduler at the eNodeB. The scheduler then allocates the assigned number of PRBs to the video flow.

In the downlink physical layer, LTE uses orthogonal frequency-division multiple access (OFDMA), and allocates radio resources in both time and frequency domains. The time domain is divided into LTE downlink frames, which are split

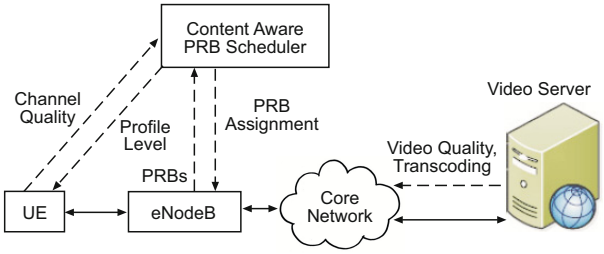


Fig. 1. A content aware architecture for video delivery over LTE downlink. The solid lines indicate interfaces that already exist in an LTE system; the dotted lines are the new conceptual interfaces proposed to implement the PRB scheduler.

into Transmission Time Intervals (TTIs), each of duration 1 millisecond (ms). The LTE downlink frame has a duration of 10 ms corresponding to 10 TTIs. Each TTI is further subdivided into two time slots, each of duration 0.5 ms, and each 0.5 ms time slot corresponds to 7 OFDM symbols. In the frequency domain, the available bandwidth is divided into subchannels of 180 kHz each, and each subchannel comprises 12 adjacent OFDM subcarriers. As the basic time-frequency unit in the scheduler, a PRB consists of one 0.5 ms time slot and one subchannel. The minimum unit of assignment for a UE is one PRB, and each one can be assigned to only a single UE. Additionally, the LTE downlink makes use of adaptive modulation and coding.

It is important to note that the content aware PRB scheduler only determines the number of PRBs needed for each UE, but not the specific PRBs that will finally be allocated. This job is left for a TTI level scheduler, which is a key component of the existing eNodeB design. Several TTI level schedulers that map PRBs to UEs have been studied in literature [27]. We propose to integrate the content aware PRB scheduler with any given TTI level scheduler using a two-level approach, similar to the one proposed in [28]. The PRB scheduler behaves like an upper-level scheduler, assigning the PRBs on a frame-by-frame basis. Within a frame, any TTI level scheduler that maximizes throughput or is proportionally fair can be used to map the PRBs to the UEs.

2.2 Video Quality Metrics

The content aware PRB scheduler requires the video quality and transcoding information to compute a PRB assignment. In this paper, we use two full-reference metrics that use the distortion-free version of a video as the reference. The first one is peak signal-to-noise ratio (PSNR) [24], and the second one is structural similarity (SSIM) index. For a video stream, these metrics are computed by averaging their values over all the video frames. For a frame of size $u \times v$ (in pixels), the PSNR of the i^{th} frame can be computed as [24]:

$$\text{PSNR}(i) = 10 \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}(i)} \right), \tag{1}$$

where MAX is the maximum possible pixel value (typically, 255), and MSE is the mean square error, defined as:

$$\text{MSE}(i) = \frac{1}{uv} \sum_{k=0}^{u-1} \sum_{l=0}^{v-1} [I_i(k,l) - R_i(k,l)]^2, \tag{2}$$

where I_i and R_i represent the i^{th} frames of the received video and reference video, respectively. Thus, the video PSNR is given by: $\text{VPSNR} = \frac{1}{m} \sum_{i=0}^{m} \text{PSNR}(i)$, where m is the total number of frames in the video.

The second metric SSIM takes into account the inter-dependency between different pixels, and, therefore, more consistent with the perception of the human eye [10]. The SSIM of the i^{th} frame can be computed on two windows x and y as [24]:

$$\text{SSIM}_{x,y}(i) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{3}$$

where μ_x and σ_x^2 are the mean and variance, respectively, for window x; likewise, μ_y and σ_y^2 are the mean and variance, respectively, for window y. The covariance of x and y is σ_{xy}. The two variables c_1 and c_2 are to stabilize the division with weak denominator. Thus, the video SSIM is given by: $\text{VSSIM} = \frac{1}{m} \sum_{i=0}^{m} \text{SSIM}(i)$.

The SVC standard [8] defines 21 profiles that differ in capabilities and target specific classes of applications. The term "level" specifies a set of constraints indicating the required decoder performance for a certain profile, such as maximum picture resolution, frame rate, bit rate, etc. Table 1 shows the VPSNR and VSSIM values for the movie trailer MIB3 encoded at different SVC levels. The reference video is encoded at Baseline Level 4.

Table 1. MIB3 trailer attributes for different SVC levels

Levels/Attributes	**VPSNR**	**VSSIM**	**Rate (Kbps)**
L1.3 (96 × 72)	36.7617	0.72761	146
L2.2 (192 × 144)	37.684451	0.8625723	304
L3.0 (320 × 240)	38.36902	0.9254554	452
L4.0 (640 × 480)	Reference	Reference	1162

2.3 Video Model

We consider a total of N UEs and M available PRBs in the LTE system, with each PRB having a fixed bandwidth, denoted by B. Suppose each UE i can decode up to a set $L_i = \{l_{ij}\}$ of video profile levels. Each profile level $l_{ij} \in L_i$ requires a certain number α_{ij} of PRBs depending on channel conditions for smooth video playback without incurring buffer underflow. We assume that all the M PRBs are available to adapt the video quality only, and are not used

for any other purpose, such as reliability or other application requirements. We assume that each UE i uses a forward error correction (FEC) code for protection, with coding rate T_i and modulation scheme m_i. Suppose $R_i(l_{ij})$ denotes the total downlink rate required for UE i to receive the video at level l_{ij} including all levels below it. This rate can be computed as [23]:

$$R_i(l_{ij}) = \alpha_{ij} m_i T_i B \log_2 \left(1 + \frac{P g_i}{N_0}\right), \tag{4}$$

where P denotes the transmission power of the eNodeB; g_i is the channel gain from the eNodeB to UE i; and N_0 is the noise power. We assume that the channel gain g_i can be estimated using CQI measurements.

Suppose $Q_i(l_{ij})$ denotes the average quality observed while receiving the video at level l_{ij}. Since we measure video quality using VPSNR or VSSIM, $Q_i(l_{ij})$ accordingly refers to these quantities when UE i receives the video at level l_{ij}. We assume that there exists a monotonic, one-to-one relationship between the observed video quality and the corresponding rate.

2.4 PRB Scheduling Problem Formulation

We assume that the eNodeB is capable of sending videos at the basic profile level. To reduce distortion, however, a higher level is required, but at the expense of more number of PRBs. Depending on the link quality and available number of PRBs, the scheduler at the eNodeB chooses a certain level l_{ij}, and assigns the corresponding number α_{ij} of PRBs to each UE i. Suppose x_{ij} is a decision variable that is 1 if level l_{ij} is assigned to UE i, and 0 otherwise. We consider that these levels are chosen in such a way that it maximizes the average video quality over all UEs. We formulate this PRB assignment problem as:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} \sum_{l_{ij} \in L_i} x_{ij} Q_i(l_{ij}) \\ \text{subject to} \quad & \sum_{i=1}^{N} \sum_{l_{ij} \in L_i} x_{ij} \alpha_{ij} \leq M \\ & \sum_{l_{ij} \in L_i} x_{ij} = 1, \ \ \forall i \\ \text{variables} \quad & x_{ij} \in \{0, 1\}, \ \ \forall i, \ \forall l_{ij} \in L_i \end{aligned} \tag{5}$$

where the first constraint ensures that the total number of PRBs assigned to the UEs does not exceed the available number of PRBs, and the second constraint chooses exactly one profile level for each UE i. This is an ILP because of the integer variables x_{ij}, and, therefore, NP-hard.

3 Solutions to PRB Assignment Problem

In this Section, we first reduce the content aware PRB scheduling problem into the MCKP, and then present two fast heuristics and an FPTAS to solve it.

3.1 Reduction to Multiple-Choice Knapsack Problem

The PRB assignment problem (5) can be cast as the Multiple-Choice Knapsack Problem [15], which is a generalization of the classical 0-1 Knapsack Problem [13]. A similar reduction for video delivery over WiMAX is given in [7]. In MCKP, we are given a set of items subdivided into N mutually disjoint classes, $K_1, \ldots, K_N$, and a knapsack of total capacity c. Each item $j \in K_i$ has a profit p_{ij} and a weight w_{ij}. The goal is to choose exactly one item from each class so as to maximize the total profit without exceeding the capacity. The MCKP can be written as:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} \sum_{j \in K_i} p_{ij} y_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{N} \sum_{j \in K_i} w_{ij} y_{ij} \leq c \\ & \sum_{j \in K_i} y_{ij} = 1, \ \ \forall i \\ \text{variables} \quad & y_{ij} \in \{0, 1\}, \ \ \forall i, \ \forall j \in K_i \end{aligned} \tag{6}$$

where y_{ij} is the decision variable that takes the value 1 if item j is chosen from class K_i, and 0 otherwise.

It is easy to see the mapping between the PRB assignment problem and the MCKP. The number of classes in the MCKP corresponds to the number of UEs, and the knapsack capacity c corresponds to the number M of available PRBs. The items in each class are the videos encoded at different profile levels. The decision variable y_{ij} corresponds to the variable x_{ij} that decides whether or not to choose level l_{ij} for UE i. The weight w_{ij} corresponds to the number of PRBs α_{ij} assigned to UE i, and the profit p_{ij} is the video quality $Q_i(l_{ij})$ experienced by UE i when receiving the video at level l_{ij}.

An important thing to decide is how frequently to solve the MCKP optimization, which defines the optimization horizon for the PRB assignment problem. The reason to consider this is the following: As channel conditions change over time, the solutions returned by the optimization might become stale if updated channel parameters are not used. Therefore, it is necessary to rerun the optimization whenever this happens, and also when the UEs start or end a video session. We discuss this issue in Section 5.

3.2 Fast Heuristics for PRB Assignment

The existing work on MCKP [15] offers various approximation algorithms that are not easily implementable in practical LTE networks. We propose two fast and simple heuristics that are easy to implement and show good performance.

The first heuristic (Algorithm 1) is a greedy algorithm similar to [16] with asymptotic worst-case running time $O(\sum_i |L_i|)$. The second heuristic (Algorithm 2) follows a technique similar to Water-Filling [22] by first assigning PRBs

to the users with better channel conditions, and then distributing the rest of the PRBs to other users. Note that, typically the number of users served by a single eNodeB can be at most a few hundreds, and therefore, the sorting in both heuristics can be accomplished efficiently using any standard sorting algorithm.

Algorithm 1. Greedy heuristic for content aware PRB assignment.

1. For each UE i, sort the profile levels in increasing order of required PRBs. 2. Pick the UEs in a round robin fashion.
3. For each UE i, choose the highest level l_{ij^*} from the sorted sequence that does not exceed the remaining PRB budget out of M total.

Algorithm 2. Water-Filling heuristic for content aware PRB assignment.

1. Sort the UEs in descending order of channel gains.
2. Pick the UEs from this sorted sequence starting from the first.
3. Follow steps 1, 2, and 3 in the Greedy heuristic, i.e., for each UE i, assign the highest profile level l_{ij^*} that does not exceed the remaining PRB budget.

3.3 An FPTAS for MCKP Using Dynamic Programming

The classical 0-1 Knapsack Problem admits an FPTAS via dynamic programming and profit-scaling [16]. Using a similar approach, we present an FPTAS for the MCKP to solve the PRB assignment problem. We first formulate a dynamic program.

Let $y_i(q)$ denote the minimum weight of a solution to MCKP with total profit q, and classes $K_1, \ldots, K_i$. If no solution exists, we set $y_i(q) = c + 1$. We use an upper bound U to specify the termination point of this (finite horizon) dynamic program. We initialize $y_0(0) = 0$, and $y_0(q) = c + 1, \forall q = 1, \ldots, U$. Then, the recursion can be written as:

$$y_i(q) = \min \begin{cases} y_{i-1}(q - p_{i1}) + w_{i1}, & 0 \leq q - p_{i1} \\ y_{i-1}(q - p_{i2}) + w_{i2}, & 0 \leq q - p_{i2} \\ \vdots & \\ y_{i-1}(q - p_{in_i}) + w_{in_i}, & 0 \leq q - p_{in_i} \end{cases} \tag{7}$$

where n_i is the number of items in class K_i.

If the argument to the min function is empty, it returns $c + 1$. The optimal profit is $\max\{q | y_N(q) \leq c\}$, with a runtime complexity $O(U \sum_{i=1}^{N} n_i) = O(nU)$, where $n = \sum_{i=1}^{N} n_i$, is the total number of videos across all classes. This type of

recurrence admits an FPTAS [16]. The approach relies on appropriately scaling the profits in the above recursion. Accordingly, we define a new set of profits, $\tilde{p}_{ij} = \lfloor \frac{p_{ij}}{K} \rfloor$, with K appropriately chosen to satisfy the tight inequality $K \leq \frac{\epsilon z^*}{N}$, where z^* is the optimal value of the objective function in the MCKP, and ϵ is a positive quantity that decides the approximation factor. With this condition is satisfied, the DP has an approximation factor $(1-\epsilon)$ [16]. The following analysis shows how to choose the value of K.

Let $p_{\max}$ be the item with the highest profit across all classes. If we choose $K = \frac{\epsilon p_{\max}}{N}$, then the above condition is clearly satisfied. Let the optimal value of the scaled problem be z_s^*. Then, it is clear that $z_s^* \leq N\tilde{p}_{\max}$, where $\tilde{p}_{\max} = \lfloor \frac{p_{\max}}{K} \rfloor$. Since $\tilde{p}_{\max} \leq \frac{p_{\max}}{K} = \frac{N}{\epsilon}$, we obtain $z_s^* \leq \frac{N^2}{\epsilon}$. Consequently, we can replace the upper bound U in the recursion by $\frac{N^2}{\epsilon}$. Since U can be computed in linear time, we get an overall running time of $O(\frac{nN^2}{\epsilon})$. The dynamic program using this technique of profit scaling is described in Algorithm 3. The objective value of the MCKP with the original profits can be obtained by examining the items that are chosen from each class in the solution of the algorithm.

Algorithm 3. Dynamic Program Scaling of Profits

```
Compute an upper bound U.
Set y_0(0) = 0, and y_0(q) = c + 1, ∀q = 1, ..., U.
for i = 1, ..., N do
    for q = U, ..., 0 do
        y_i(q) = min_{j∈{K_i|q≥p̃_ij}} (y_{i-1}(q − p̃_ij) + w_ij).
z_s^* = max{q|y_N(q) ≤ x}.
```

4 Performance Evaluation

In this section, we compare the performance of the two heuristics and the FPTAS with the optimal obtained from CPLEX.

4.1 Experimental Setup

In our simulations, we uniformly distribute the UEs around the eNodeB, and randomly map each UE to a video. We use LTE system parameters defined in the 3GPP standard [21]. The focus of this study is primarily in measuring the performance at the physical and MAC layers. We acknowlege that different content distribution networks (CDNs) may employ different techniques at higher layers which might affect the metrics evaluated here. The transmission power P of the eNodeB is 46 dBm; the noise figure N_0 is 7 dB; the transmission frequency F is 925 MHz; the eNodeB antenna height h_b is 30 meters; and the UE antenna height h_m is 1.5 meters. We follow the path loss model described in [17], and use

the statistical tool R [19] to generate the channel model. The path loss G for a UE that is d meters away from the eNodeB is given by:

$$G = 69.55 + 26.16\log_{10}(F) - 13.82\log_{10}(h_b) \\ -ch + (44.9 - 6.55\log_{10}(h_b))\log_{10}(d), \quad (8)$$

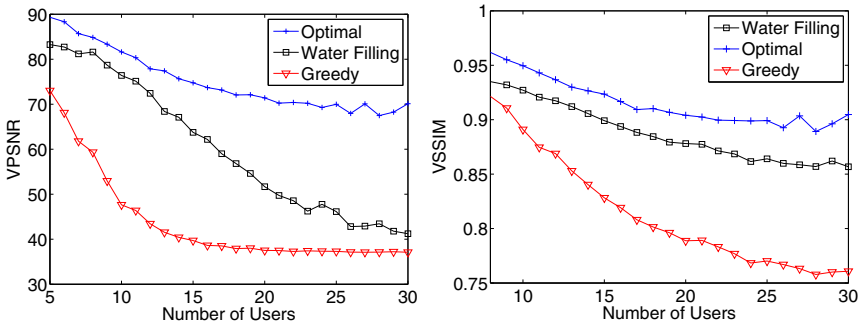
where the parameter ch depends on the city size. The number of available PRBs M in our simulation is set to 50, which is the same number of PRBs in an LTE frame when the channel bandwidth is 10 MHz. The spectral efficiency $m_i T_i$ for UE i depends on the CQI and is given in the LTE standard [21].

4.2 Simulation Results

In our simulation, we assume that a user experiences buffer underflow if it is not assigned the required number of PRBs to support a download data rate at least equal to the playback rate. We first compare the performance of the Greedy and the Water-Filling heuristics with the optimal. The results for VPSNR and VSSIM are averaged over 1000 iterations, where, at each iteration, we randomly map the UEs to the videos and generate channel conditions according to (8).

As shown in Figure 2(a) and 2(b), the three plots representing Greedy, Water-Filling, and Optimal follow a similar trend, i.e., the video quality decreases with increasing number of UEs. This is expected because the number of PRBs allocated per UE decreases with increasing number of UEs for a fixed PRB budget. We also note that the difference in VPSNR and VSSIM values obtained from the heuristics and those of the optimal increases with more number of users. However, the difference is less predominant for the Water-Filling algorithm than the Greedy one. This is because of the following: In the Greedy algorithm, the UEs are picked up at random and assigned PRBs for the highest profile level possible. In contrast, the Water-Filling algorithm first sorts the UEs in decreasing order of channel gains, and then assigns the PRBs corresponding to the highest levels. Thus, for the same rate requirement between two users, the user with good channel condition will need fewer PRBs in the Water-Filling algorithm, and, therefore, more PRBs will be left to satisfy the profile levels of other users. In the Greedy algorithm, the chance of picking up a user with good channel condition decreases as the number of users increases, and so it performs increasingly worse as compared to the Water-Filling algorithm for more number of users.

We now compare the performance of the FPTAS with the optimal obtained from CPLEX. As discussed before, the asymptotic running time of the FPTAS is $O(\frac{nN^2}{\epsilon})$, where n is the total number of videos, and ϵ decides the approximation factor, which is at least $(1 - \epsilon)$ in our implementation of the dynamic program. We applied the FPTAS for the same channel and video models for three different values of ϵ, namely, 0.25, 0.5, and 0.95. The results for VPSNR, shown only for $\epsilon = 0.5$ and $\epsilon = 0.95$ in Figure 3(a) and 3(b), respectively, indicate that the FPTAS performs very close to the optimal.



(a) (b)

Fig. 2. Comparison of (a) VPSNR and (b) VSSIM obtained from the Greedy and Water-Filling heuristics with that of the optimal from CPLEX

5 Content Aware LTE Architecture and Signaling

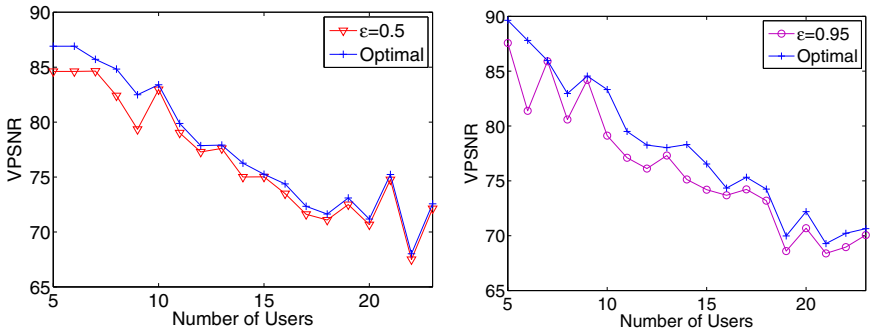
In this section, we first propose a new signaling mechanism and a modification to the LTE architecture to implement the PRB scheduler. We then evaluate the performance of this modified architecture using measurement data.

5.1 Signaling Mechanism and Architecture Modification

We reuse the IP services of the EPS bearer to implement the content aware PRB scheduler. The signaling mechanism, as shown in Figure 4, takes place as follows: Upon receiving a video request from the UE, the video server responds with the levels, rates, and VPSNR/VSSIM information of that video. The UE sends this information to the eNodeB, which, in turn, forwards it to the PRB scheduler. The UE also sends the CQI and the Reference Symbol Received Power (RSRP) to the PRB scheduler. The PRB scheduler also obtains the set of available PRBs from the eNodeB, and then runs the optimization to compute the PRB assignment and the profile level assignment for each UE. The profile level is sent to the UE, while the PRB assignment is sent to the scheduler in the eNodeB. Finally, the UE requests the video at the assigned profile level from the video server.

We note that there can be delays associated with signaling that may affect the performance of the algorithm. This may require re-running the optimization. We show the effect of this delay under various channel scenarios, and give a method to choose when to re-run the optimization.

We propose to implement the PRB scheduler at two different places, motivated by the emerging trend of software defined networking (SDN) toward an open architecture at the switches and routers. The first is to include the PRB scheduler in the Mobility Management Entity (MME), where it can handle communications and negotiations between the server and the network. The MME



(a) (b)

Fig. 3. Comparison of VPSNR obtained from the FPTAS and the optimal from CPLEX for different values of ϵ: (a) $\epsilon = 0.5$, and (b) $\epsilon = 0.95$

can keep track of the PRBs assigned to the UEs, and run the optimization with appropriate parameters during a handover. Although there is one instance of the PRB scheduler for each eNodeB, they are all located within a single MME. The PRB scheduler can also be placed at the eNodeB itself. However, this has some disadvantages, the biggest one being the difficulty of modifying every eNodeB to accommodate the PRB scheduler. We note that there is no security vulnerability of breaching user privacy in this modified architecture. The eNodeB treats each video simply as another flow, and it is the UE that requests a content aware profile level and PRB assignment.

5.2 Measurement Based Evaluation

We evaluate the performance of the modified architecture using real data sets collected from AT&T and T-Mobile networks by doing a drive-test and measuring delays using an Android device and Qualcomm eXtensible Diagnostic Monitor (QxDM) [20]. A sample plot for an outdoor suburban measurement data is shown in Figure 5. The plot captures four quantities: reference signal received power (RSRP), reference signal received quality (RSRQ), received signal strength indicator (RSSI), and CQI variation, as a time series for about 14 minutes. The data is then fit into a lognormal distribution, as shown in Figure 6(a), which is then used to obtain the urban data. The outdoor urban data is generated using the spatial channel model in [25].

The PRB scheduler depends on UE reports sent to the eNodeB. There is a network delay between the server and the UE, which can be tens to a few hundreds of milliseconds. Thus, depending on the environment, the channel conditions may change between the time the UEs request the video profile levels determined by the PRB scheduler, and the time the server starts sending the packets. As a result, the decisions taken by the scheduler may be obsolete.

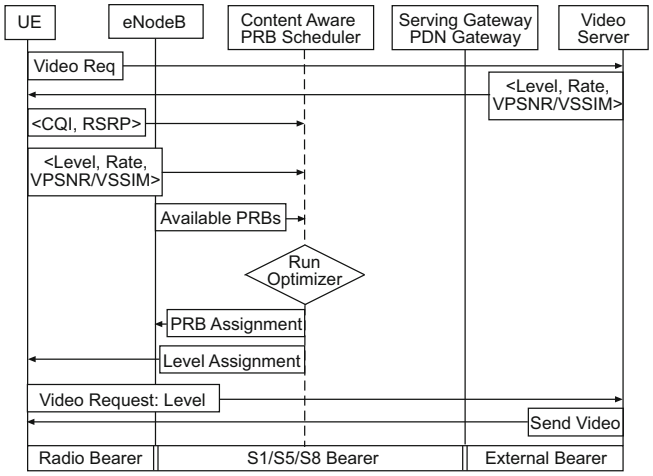


Fig. 4. Signaling to implement the content aware PRB scheduler in LTE

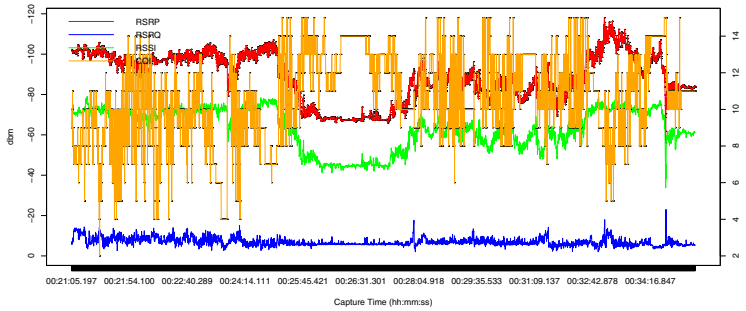


Fig. 5. RSRP, RSRQ, RSSI, and CQI data in an outdoor suburban environment

We measure this delay in AT&T and T-Mobile networks for different technologies. For an LTE network, the delay is 50-150 ms; for an HSDPA+ network it is 160-450 ms; and for an on-campus Wi-Fi network, the delay is 7-20 ms.

We evaluate the impact of this signaling delay on video distortion for both indoor and outdoor environments. Figure 6(b) shows the distortion per user in the outdoor for both urban and suburban areas. We observe that the impact of delay becomes more predominant with increasing number of users. We also see that the urban environment has more distortion than the suburban one. This is due to more severe variation in link quality in the urban environment than the suburban one, and can result from more multi-path fading, shadowing, and Doppler effect. The indoor environment has (plot not shown here) very little effect on distortion due to negligible variation in channel conditions.

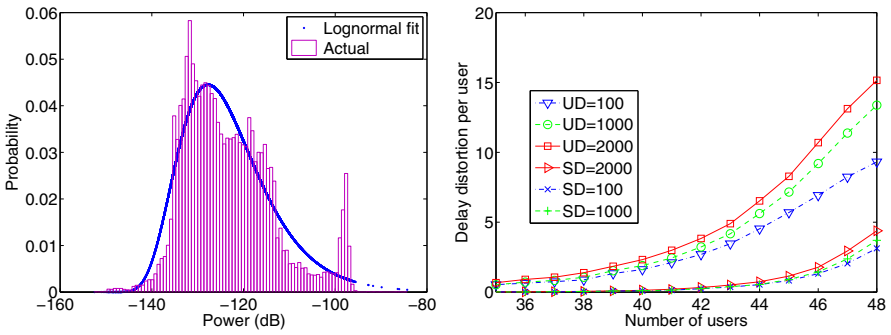


Fig. 6. (a) A Lognormal fit to the RSRP in an outdoor suburban environment; (b) Distortion as a function of the number of users in both urban and suburban outdoor environments for different delays; UD: urban delay; SD: suburban delay

6 Conclusion

We propose a content aware PRB scheduler for downlink video delivery in LTE based on SVC. The eNodeB in our scheme maximizes the average video quality across all users based on their link qualities, device capabilities, and available PRBs. We propose two fast heuristics and an FPTAS to solve this optimization problem, and compare their performance with the optimal. Our results show that the heuristics are a factor 1/2 away from the optimal, while the FPTAS is very close to the optimal. We also propose a signaling mechanism and a modification to the LTE architecture to implement the PRB scheduler. We evaluate the effect of signaling delay on this modified architecture using real measurement data. Our results show that, even after factoring in real channel variations and delays, the PRB scheduler still performs very well.

References

1. Driving, LTE.: Adoption: Mass-Market Pricing and a Large Device Ecosystem Emerge as Key Factors, Analysys Mason (2013)
2. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update (2013)
3. Seferoglu, H., Markopoulou, A.: Video-Aware Opportunistic Network Coding over Wireless Networks. IEEE J. Select. Areas Commun. 27(5), 713–728 (2009)
4. Haghani, E., Ansari, N., Parekh, S., Colin, D.: Traffic-Aware Video Streaming in Broadband Wireless Networks. In: IEEE WCNC, pp. 1–6 (2010)
5. Zhu, X., Girod, B.: Distributed Media-Aware Rate Allocation for Wireless Video Streaming. IEEE Trans. Circuits Syst. Video Technol. 20(11), 1462–1474 (2010)
6. Pandit, K., Ghosh, A., Ghosal, D., Chiang, M.: Content Aware Optimization for Video Delivery over WCDMA. EURASIP J. Wirel. Commun. Netw. (2012)
7. Sharangi, S., Krishnamurti, R., Hefeeda, M.: Energy-Efficient Multicasting of Scalable Video Streams Over WiMAX Networks. IEEE Trans. Multimedia 13(1), 102–115 (2011)

8. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. IEEE Trans. Circuits Syst. Video Technol. 17(9), 1103–1121 (2007)
9. McDonagh, P., Vallati, C., Pande, A., Mohapatra, P., Perry, P., Mingozzi, E.: Investigation of scalable video delivery using H.264 SVC on an LTE network. In: IEEE WPMC, pp. 1–5 (2011)
10. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. Image Processing 13(4), 600–612 (2004)
11. Nikolaos, T., Boulgouris, N.V., Strintzis, M.G.: Optimized Transmission of JPEG2000 Streams over Wireless Channels. IEEE Trans. Image Processing 15(1), 54–67 (2006)
12. Pisinger, D.: A Minimal Algorithm for the Multiple Choice Knapsack Problem. Eur. J. Oper. Res. 83, 394–410 (1994)
13. Kellerer, H., Pferschy, U., Pisinger, D.: Knapsack Problems. Springer (2004)
14. Gens, G., Levner, E.: An Approximate Binary Search Algorithm for 0-1 MCKP. Information Processing Letters 67, 261–265 (1998)
15. Sinha, P., Zoltners, A.A.: The Multiple Choice Knapsack Problem. Operations Research 27(3), 503–515 (1979)
16. Lawler, E.L.: Fast Approximation Algorithms for Knapsack Problems. Mathematics of Operations Research 4(4), 339–356 (1979)
17. Holma, H., Toskala, A.: LTE for UMTS - OFDMA and SC-FDMA Based Radio Access. John Wiley & Sons (2009)
18. IBM CPLEX Optimizer, `http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/`
19. The R Project for Statistical Computing, `http://www.r-project.org/`
20. Qualcomm eXtensible Diagnostic Monitor (QXDM Professional) (2007)
21. Requirements for further advancements for E-UTRA (LTE-Advanced), 3GPP TR 36.913, `http://www.3gpp.org/ftp/Specs/html-info/36913.htm`
22. Proakis, J.G.: Digital Communications. McGraw-Hill, New York (2001)
23. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons (2006)
24. Bovik, A.C.: The Essential Guide to Video Processing. Elsevier (2009)
25. Radio Frequency (RF) Requirements for LTE Pico eNodeB. 3GPP TR 36.931
26. Sesia, S., Touffik, I., Baker, M.: LTE - The UMTS Long Term Evolution: From Theory to Practice. Wiley (2011)
27. Capozzi, F., Piro, G., Grieco, L.A., Boggia, G., Camarda, P.: Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey. IEEE Communications Surveys and Tutorials 15(2), 678–700 (2013)
28. Piro, G., Grieco, L.A., Boggia, G., Camarda, P.: A Two-Level Scheduling Algorithm for QoS Support in the Downlink of LTE Cellular Networks. In: European Wireless Conference, pp. 246–253 (2010)