Advanced Downlink LTE Radio Resource Management for HTTP-Streaming

Thomas Wirth⋆, Yago Sánchez⋄, Bernd Holfeld⋆, Thomas Schierl⋄
Wireless Communication and Networks⋆ and Multimedia Communications⋄
Fraunhofer Heinrich Hertz Institute
Einsteinufer 37, 10587 Berlin
Berlin, Germany
{thomas.wirth, yago.sanchez, bernd.holfeld, thomas.schierl}@hhi.fraunhofer.de

ABSTRACT

Video traffic contributes to the majority of data packets transported over cellular wireless. Future broadband wireless access networks based on 3GPP's Long Term Evolution offer mechanisms for optimized transmission with high data rates and low delay. However, especially when packets are transmitted in the LTE downlink and if services are run over-the-top (OTT), optimization of radio resources in a multi-user environment for video services becomes infeasible. The current market trend is moving to OTT solutions, also for video transmission, where an emerging standard based on HTTP streaming - DASH - is expected to have a huge success in the upcoming years. The solution presented in this paper consists of a novel technique, which combines LTE features with knowledge on DASH sessions for optimization of the wireless resources. The combined optimization yields an improved transmission of videos over cellular wireless systems which are based on LTE and LTE-Advanced.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Performance

Keywords

LTE/LTE-A, Radio Resource Management, DASH, HTTP-Streaming, Cross-layer Design

1. INTRODUCTION

Video content delivery over wireless is expected to grow exponentially in the coming years. Already in today's mobile wireless networks, every second bit contains video data and this is expected to grow to approximately two-thirds by 2016 [1]. The major driver behind this evolution is the availability of new smart phone and tablet generations with higher processing power, long battery lifetime and high definition (HD) display technology.

Furthermore, future LTE [2] and LTE-A [3] radio access networks can provide sum rates ranging from 100 Mbps up to 1 Gbps, as well as short latencies of approximately 10-20 ms and thus enable services with a new quality-of-experience (QoE). This allows deployment of better services, which will enforce the success of video applications in mobile environments. Since the wireless channel is a shared medium, all users will experience a data rate of several Mbps depending on the cell load and the users current channel condition. Therefore, in order to provide a video service of high quality, the radio resources have to be used efficiently, especially when the system load is very high and when many users compete for medium access on the shared medium.

Recently, there has been a shift in paradigm regarding video delivery techniques where the mobile network operator (MNO) and service or content provider do not necessarily coincide. Many service providers have adopted over-the-top (OTT) alternatives in the last years to deliver video content directly to end users and many others are on the process of doing so. This is in many cases disadvantageous for the MNO who is skipped in the value chain. Proof of such a support are proprietary solutions which are already deployed, e.g. [4] and [5]. OTT services generally use generic HTTP data connections, which bypass the MNO and transmit data from the Internet directly to the user. Due to this fact, data is treated as simple web objects, i.e. HTTP resources, which simplifies the implementations of content delivery networks (CDNs) and servers, which are simple web servers. However at the same time, this leads to the network components to be in principle not aware of the characteristics of the transported data and makes an optimization of the wireless resources more difficult.

In an effort to unify existing streaming formats based on HTTP and enable interoperability between servers and clients, many standardization bodies have worked on the development of a standard for adaptive streaming over HTTP as in [6], [7] or [8]. Among this, MPEG's dynamic adaptive streaming over HTTP, as explained in Section 2, is expected to be very successful.

The migration from RTP/UDP based media delivery to HTTP has shown to be very beneficial and as aforementioned has become much support from the market. Although media streaming has been associated with RTP/UDP due to its lower latency, relying on the HTTP-transmission control

MM'12, October 29–November 2, 2012, Nara, Japan.

protocol (TCP) for media delivery has shown to be a very valuable solution for scenarios, where stringent delay constraints are not considered. TCP allows reliable end-to-end transmission by utilizing error detection, flow control, and congestion control. Furthermore, such applications facilitate the deployment of servers and CDNs, since these work as simple web servers and/or web caches, respectively.

Considering the forthcoming broadband access technology based on LTE, its building blocks of the physical layer (PHY layer) like diversity transmission and reception, turbo-encoding and retransmission protocols, e.g. hybrid automatic repeat request (HARQ), have shown to be very robust also for higher-layer protocols like TCP. However, in this new paradigm of video streaming, the clients are the ones responsible for managing the video sessions with no centralized entity dealing with fairness among users, see also Section 2.

In LTE, the scheduling entity is placed at the base station (BS) which a.o. is responsible for the multi-user radio resource management (RRM) of time, frequency and antenna resources. A lot of research has been done designing scheduling algorithms with the aim to optimize resource assignment, e.g. the maximum throughput (MAX) scheduler only serves the best users, while the proportional fair (PF) [9] scheduler gives a more fair resource assignment by taking the rate history of each user into account. However, these schedulers are typically agnostic to video characteristics and characteristics of the alternative videos in an adaptive service, which would allow for another dimension of optimization and fair resource sharing among the users. In this work, we propose a new cross-layer technique, which takes dynamic adaptive streaming over HTTP (DASH) related information at the BS scheduler into account, in order to optimize multi-user resource assignment for video transmission over the LTE downlink.

2. DASH

Dynamic adaptive streaming over HTTP [10] is an emerging standard for adaptive HTTP streaming. In a DASH service, the content is typically offered at the server at different versions, which may include different resolutions or qualities and therefore different bit-rates. Thereby, it allows users to select the most adequate video version for them, i.e. the version that matches at best their equipment capabilities and/or current connectivity characteristics. To allow for dynamic adaptation at each time instance, each of the mentioned video versions is divided in small video parts corresponding to relatively small time intervals, usually referred to as segments. The MPEG DASH standard basically defines the formats of these segments and the format of the media description representation (MPD), which is an XML document describing the characteristics of the media available at the server.

The segment formats are defined for two encapsulation methods MPEG-2 TS [11] and ISO base Media File Format [12], putting some constraints on how to create the content with focus on easing the adaptation of media, i.e. switching from segments from one version of the content to segments from another version of the same content.

The MPD contains important information for the users that allows them to make the decisions on which content to download, as well as some description of the content. There are many attributes in the MPD describing important information about the content such as *@codecs* which refers to the codec used for encoding the content and lets the user know if it is able to process the content. The *@bandwidth* and *@minBufferTime* tags allow clients to derive the bit-rate requirements of the media and the necessary pre-buffer size in order to avoid playout interruptions. For a more comprehensive and detailed description of DASH, the reader is referred to the standard or to [10].

2.1 Client Behavior

The behavior of the client is not specified in the DASH standard. Thus, the implementation of a DASH client is left free. There may be many different algorithms for applying adaptation based on statistics of previously downloaded segments, buffer fullness, or limitation on buffer storage capacity, etc.

For the study carried out within this paper, a simple client implementation is considered which takes into account the buffer fullness at each time instant and based on some thresholds, quality switching is performed. Furthermore, no limitation in buffer storage capacity is considered, which means that clients will continue downloading future media data without any restriction as long as the channel characteristics allows for it. The pseudo-code of the client behavior is presented below:

```
quality=highest;
prebuffer(initial_prebuffer_length);
while(more_data_in_server)
  download_next_segment();
  if(current_buffer_length < threshold_down *
   initial_prebuffer_lenght && quality > lowest)
    quality--;
  elseif(current_buffer_length > threshold_up *
   initial_prebuffer_lenght && quality < highest)
    quality++;
  elseif(current_buffer_length == 0)
    prebuffer(initial_prebuffer_length);
```

3. RADIO RESOURCE MANAGEMENT

As pointed out in Section 1, the main goal of this work is to present a novel cross-layer scheduling technique, which is part of the RRM of the LTE downlink. The task of the scheduler is to assign radio resources to all user equipments (UEs) in a multi-user scenario. Radio resources are elements from the time, frequency, and spatial domain of the LTE system. The proposed scheduler takes into account information saved within the MPD regarding the available content stored on the DASH server. The information can consist of but is not limited to the number of available video versions for each content and/or the corresponding video characteristics, which could be the required bit-rate for the requested video-stream. Goal of the RRM is to optimize the allocation of the available resources to the users, such that the quality-of-service (QoS) for all users with respect to the video quality and interruption-free video transmission is optimized.

As aforementioned, the proposed scheduling policy takes into account information provided in the MPD, namely the bitrate of the available versions for performing the optimization of resource allocation. This requires scheduling techniques which can take rate constraints as input parameters. Related work is the minimum rate constraint scheduling investigated in [13] and [14]. The authors modify MAX and PF policies to support rate constraints by introducing a token counter for each data queue. This token counter is used to penalize resource assignment for users when a certain target is met or favor users that do not meet a minimum rate constraint. Similar to this approach, we define the radio resource assignment (RRA) problem for broadcast channels under long-term policy *pol* at time instance i to be given by

$$\mathbf{R}_{pol}(i) = \arg\max_{\mathbf{r}(i)\in\mathcal{C}(i)} \sum_{k=1}^{K} \gamma_k \, w_{pol,k}(i) \, r_k(i) \quad (1)$$

where $\mathbf{r}(i) = [r_1(i), r_2(i), ..., r_K(i)]^T$ denotes the vector of achievable data rates inside the achievable rate region $\mathcal{C}(i)$ and $\mathbf{R}_{pol}(i)$ is the vector that contains the scheduled data rates of each user by employing the intended policy $w_{pol,k}(i)$. A user k is associated with $w_{pol,k}(i)$ as its policy-dependent weighting factor, $r_k(i)$ as its achievable rate and γ_k as priority indicator which signals delay- or throughput-awareness of the user. We have $\gamma_k = 1$ if all users are equal-prioritized.

3.1 Proposed Solution

The proposed solution consists of a cross-layer approach, where LTE's scheduler can make use of information defined in DASH, namely defined in the MPD, to optimize radio resource assignment among the user terminals. In order to get this information, different approaches can be followed. One possibility is that the LTE base station (eNB) is aware of such services and MPD sniffing or inspection techniques are implemented. Mechanisms that implement deep packet inspection (DPI) are already widely used in wireless networks, e.g. for service flow detection. This is a very simple approach but has limitations, e.g. if the MPD is encrypted. One more sophisticated approach requires active collaboration of the mobile terminals. In LTE, there are some techniques in order to set up dedicated radio bearers between the users and the BS by exchanging evolved packet system (EPS) messages. Currently, the defined EPS messages only allow providing information about a single bitrate point. However, a simple extension of these messages allows to map user specific information into these EPS messages and transmit this information to the BS. This is a more robust solution, since when the MPD is encrypted or other security mechanisms are applied to the MPD for avoiding external non-allowed entities to read the MPD's information, it allows the BS to know information about the DASH service.

The scheduling algorithm can now be defined by modifying γ_k in Eq. (1) such that each user k experiences a different priority indicator, depending on its current channel quality and/or particular service flow. γ_k is then updated in each time instant i by an exponential weighted moving average filter according to α_k, refer to Eq. (2) and Eq. (3). The value of α_k depends on the minimum data rate requirement R_k^{min} from the MPD and will be set to a small $\epsilon > 0$, if the minimum data rate requirement is met.

$$\gamma_k(i) = e^{\alpha_k(i-1)\frac{\alpha_k(i)}{\|\alpha(i)\|}} \quad (2)$$

$$\alpha_k = max\{R_k^{min} - \bar{R}_k, \epsilon\} \quad (3)$$

Here, $\bar{R}_k$ is the average data rate over the past time interval targeting to fulfill R_k^{min}. In contrast, if the maximum data rate requirement R_k^{max} is achieved, the priority indicator can directly be set to a small value ϵ, such that the user experiences a very low weight and is deferred from the scheduling list within the next scheduling time slot.

For the scheduling policy $w_{pol,k}$, see Eq. (1), the conventional scheduling approaches MAX and PF scheduling will be used as in [13].

4. SIMULATION DESCRIPTION

The proposed cross-layer algorithms for advanced RRM are evaluated with a multi-cell system-level simulator. Investigations within a multi-cell environment yield a realistic evaluation because this scenario is interference limited in contrast to noise limited investigations. This is also the scenario to be expected in commercially deployed LTE networks with full frequency reuse. In addition, the simulator implements all the key features of LTE and LTE-Advanced which will be described in the next section.

4.1 System-Level Simulator

The cellular system is modeled using a hexagonal cell-structure, where one Macro BS serves three sectors. All UEs are placed in a center cell, which is surrounded by two rings of Macro-cell interference. This is a common approach for system-level simulations for LTE/LTE-A systems. The channel model used is based on 3GPP SCME [15]. The simulator models LTE compliant subcarrier grouping to subchannels, multi-user MIMO techniques to support up to 4 spatial layers, frequency selective link adaptation based on exponential effective SINR mapping (EESM), LTE compliant adaptive modulation and coding (AMC), and channel quality identifier (CQI) reporting. The CQI reports are used as inputs for the scheduling algorithm. The main simulation parameters can be found in Table 1.

Parameter	**Value**
Number of cells	19 sites / 3 sectors per site
Inter-site distance	500 m
Carrier frequency	2.6 GHz
System bandwidth	10 MHz
Channel model	3GPP SCME [15]
BS parameters:	
Transmit antennas	4
Transmit power	46 dBm
Antenna gain	17 dBi
Antenna downtilt	10°
FWHM	58° azimuth, 6.2° elevation
UE parameters:	
Receive antennas	2
CQI reporting interval	10 ms

Table 1: Simulation assumptions

4.2 Video Traffic Model

The video traffic model used in the simulation is based on an encoded *football* video stream with three operating point, see Table 2. The lower target rate R^{min} of the modified schedulers is set to the lowest quality **Q1** while R^{max} is set to **Q3**. It is assumed that each UE within the simulation scenario targets to receive the top video quality **Q3**.

5. RESULTS

One of the main factors that have a big impact on the QoE experienced by the users of a video service is the appearance of interruptions. It is widely accepted that video

Quality	Q1 33.1-33.6 dB	Q2 35.2-35.8 dB	Q3 38.1-38.8 dB
Rate [Kbps]	685.39	1083.82	1680.85

Table 2: Different operating points of the downloaded video stream

interruptions are really annoying for the users and a service of quality can be measured by the lack of interruptions. Therefore, in order to assess the quality of the proposed techniques, Figure 1 shows the probability that no interruptions occur during a DASH session, where the initial pre-buffer at the UE is set to 20 seconds and the number of UEs in the simulation environment is increased from 10 to 60 user terminals. When the MAX policy is applied, the interruption probability increases already for 30 UEs to more than 10 % and in the scenario with 60 UEs, less than 20 % of the users receive the video content without suffering any interruption for the 5 minutes long DASH session. The standard PF scheduler performs a bit better, but also yields a high interruption probability for the scenario with 60 UEs. The according to Eq. (2) and Eq. (3) modified MAX and PF schedulers outperform the state-of-the-art schedulers. For the modified PF policy, the probability that a user suffers an interruption is much lower than for the other schedulers and even for 60 UEs only reaches approximately 30 %.

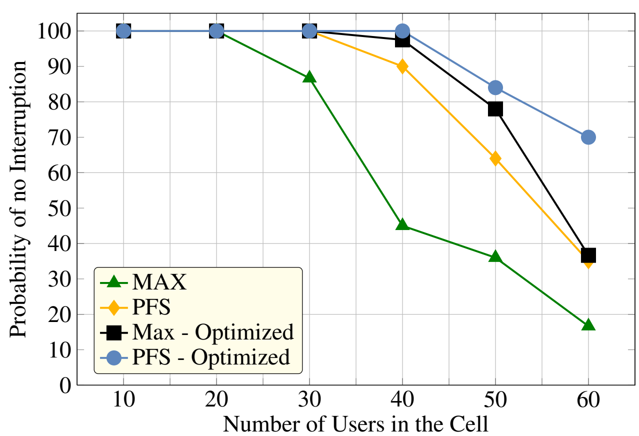

Figure 1: Probabilities of no interruptions for different LTE cross-layer scheduling techniques for video over DASH transmission

6. CONCLUSIONS

Most of the traffic in future mobile networks has been reported to be video traffic. It is foreseen, that this will exponentially increase in the upcoming years. Therefore, efficient allocation of the wireless resources is a crucial issue to provide QoS to the users. Due to the recent changes in the video transmission paradigm to over-the-top (OTT) services, i.e. HTTP streaming, optimization of mobile resources has become a more difficult task. In this work, we have presented a novel cross-layer technique that optimizes the allocation of resources at the base station, taking into account information of the video sessions. The proposed approach shows a huge improvement over the state-of-the-art techniques, considerably reducing the probability that users suffer playout interruptions and therefore increasing sharply the QoE of the users.

7. REFERENCES

[1] Cisco Networks, “Cisco visual networking index: Global mobile data traffic forecast update, 2011-2016,” Feb. 2012, White Paper.

[2] T. Haustein and T. Wirth et. al., “Measurements of multi-antenna gains using a 3GPP-LTE air interface in typical indoor and outdoor scenarios,” *14th European Wireless Conference*, Jun. 2008.

[3] 3GPP, “http://www.3gpp.org/lte-advanced.”

[4] A. Zambelli, “IIS smooth streaming technical overview,” Microsoft Cooperation, 2009. [Online]. Available: http://download.microsoft.com/download/4/2/4/4247C3AA-7105-4764-A8F9-321CB6C765EB/IIS_Smooth_Streaming_Technical_Overview.pdf

[5] A. Pantos and W. May, “HTTP live streaming, version 6. IETF internet-draft [updated 2011, march 31; cited 2011, april 6]. expires 2011, october 2.” [Online]. Available: http:///tools.ietf.org/html/draft-pantos-http-live-streaming-06

[6] “Open IPTV forum - release 2 specification, HTTP adaptive streaming, draft v0.06 - june 7,” 2010.

[7] “3rd generation partnership project; technical specification group services and system aspects; transparent end-to-end packet-switched streaming service (PSS); protocols and codecs (release 9); 3GPP TS 26.234 v9.3.0 (2010-06), section 12: Adaptive HTTP streaming.”

[8] “Information technology - dynamic adaptive streaming over HTTP (DASH) - part 1, media presentation description and segment formats, ISO/IEC DIS 23009-1,” 2011.

[9] F. Kelly, “Charging and rate control for elastic traffic,” *European Transactions on Telecommunications*, 1997.

[10] I. Sodagar, “The mpeg-dash standard for multimedia streaming over the internet,” *Multimedia, IEEE*, vol. 18, no. 4, pp. 62 –67, april 2011.

[11] “Information technology–generic coding of moving pictures and associated audio information: Systems,” ISO/IEC 13818-1, 2007.

[12] “Information technology–coding of audio-visual objects –part 12: ISO base media file format,” ISO/IEC 14496-12, 2008.

[13] M. Andrews, L. Qian, and A. Stolyar, “Optimal utility based multi-user throughput allocation subject to throughput constraints,” in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 4, march 2005, pp. 2415 – 2424 vol. 4.

[14] L. Qian, N. Song, D. Vaman, X. Li, and Z. Gajic, “Power control and proportional fair scheduling with minimum rate constraints in clustered multihop td/cdma wireless ad hoc networks,” in *Wireless Communications and Networking Conference, 2006. WCNC 2006. IEEE*, vol. 2, april 2006, pp. 763 –769.

[15] 3GPP-TR25996, “Spatial channel model for multiple input multiple output (MIMO) simulations.”