Video Capacity and QoE Enhancements over LTE

Sarabjot Singh*, Ozgur Oyman†, Apostolos Papathanassiou†, Debdeep Chatterjee†, and Jeffrey G. Andrews*
*Department of Electrical and Computer Engineering
University of Texas at Austin, Texas, USA
†Intel Corporation, Santa Clara, California, USA

***Abstract*—Quality of Experience (QoE) has taken a center stage in the performance evaluation of multimedia delivery technologies. The Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) system is the latest generation of wireless cellular technology expected to deliver higher data rates and meet the burgeoning data demand. With the projected dominant share of video services in mobile traffic, providing satisfactory QoE to video users is a key objective for LTE system design. In this paper, we present a QoE-based evaluation methodology to assess the LTE system video capacity in terms of the number of unicast video consumers that can be simultaneously supported for a given target QoE. We define and use the notion of rebuffering outage capacity to quantify the video service capacity. Our evaluation further incorporates adaptive streaming, a promising technology for video delivery over wireless, and presents its consequent QoE-capacity tradeoff. The impact of QoE-based outage criteria is also investigated on the downlink video capacity. Finally, we propose a QoE-aware radio resource management (RRM) framework which allows the network operator to further enhance the video capacity. Our results demonstrate that there is a significant potential to optimize video capacity through QoE awareness both at the application level and radio access network (RAN) level.**

I. Introduction

Today, video communication over mobile broadband is challenging due to limitations in bandwidth and difficulties in maintaining high reliability, quality, and latency demands imposed by rich multimedia applications. In the meantime, mobile video traffic is growing at an immense rate due to significant consumer demand, with the projected share of video constituting more than two-thirds of the total mobile traffic by 2015 [1]. The Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) is considered to be the latest wireless cellular technology whose roll-out has already begun in certain parts for the world. LTE is expected to cater to the high bandwidth and low latency demands of video applications. However, it is still possible that these enhancements will not be sufficient to meet the anticipated future demand for video with satisfactory QoE levels. Consequently, network operators are constantly in search of solutions which allow them to provide improved capacity and QoE for their video users with their limited resources [2].

One of the key emerging video QoE enhancing solutions is adaptive streaming, which aims to optimize and adapt the video configurations over time in order to deliver the best possible user experience [3]. QoE enhancements from adaptive streaming could be in the form of enhanced video quality, reduced startup time or fewer rebuffering events. The adaptive streaming portion of Internet video is anticipated to grow at an average of 77% a year toward supporting 51% of Internet video by 2015 [4].

Prior work [5] reported capacity of a LTE system for a fixed rate (non-adaptive) video service called near-real time video service subject to specified QoS constraints, without considering metrics suited for buffered streaming. A heuristic scheduling algorithm that allocated resources to users based on their corresponding mean opinion score (MOS) requirements was proposed in [6]. MOS was derived from metrics like average rate and loss rate while the mapping was generated based on a trained random neural network. A cross-layer design, where the MOS was abstracted for various services as a function of the rate allocated to the same, aimed at maximizing the thus defined sum utility functions in a HSDPA network was proposed in [7]. However, RRM based on metrics like user's playback buffer status and rebuffering percentage, which form the crux of the buffered video services has never been investigated. In addition, with adaptive streaming solutions prior works generally left the radio-level RRM completely agnostic of the streaming service mechanisms leading to suboptimal performance. To the best of our knowledge, the work presented in this paper is the first to 1) define a QoE-aware outage criteria for buffered video streaming services, 2) introduce the concept of rebuffering outage capacity and assess LTE capacity for stored video services as per the defined outage criteria and in terms of rebuffering outage capacity, 3) evaluate the impact of adaptive streaming on the rebuffering outage capacity and user QoE, and 4) propose a QoE-aware RRM framework that works in conjunction with adaptive streaming to optimize capacity under certain target QoE constraints.

II. Adaptive Streaming Framework over LTE

Quality of experience expectations of users differ depending on the type of video application being consumed. Buffered or stored video services like YouTube and Netflix are very popular over Internet. In this scenario the content (video) is pre-encoded and available at the server beforehand. The playback starts after an initial startup delay and continues while the content is being downloaded.

Definition 1. *Rebuffering* is the state of streaming invoked when the playback buffer is emptied in which the video playback is stalled while the buffer is being filled up.

The work by the first author was done during an internship with Intel Corp.
Corresponding author : sarabjot@utexas.edu.

978-1-4577-2053-6/12/$31.00 ©2012 IEEE

Since the video content is reliably downloaded and played in this service, the QoE is primarily affected by rebuffering events [8]. Another type of video services is real-time streaming, where content is live generated, so the client player does not have the chance to rebuffer leading to a scenario where video frame delay is equivalent to loss of frames. We focus on buffered video streaming in this paper and real-time streaming results would be presented in a future publication.

In a cellular setting, for many clients it may not be feasible to support high data rate video streams due to spatio-temporal variations of link conditions or/and because of the geometry itself. Hence, it would be wise to send a lower quality (and consequently lower bit rate) version of the corresponding video to ensure uninterrupted play out. In this paper we deal with the pull-based paradigm where the client (rather than server) plays the central role by carrying the intelligence that drives the video adaptation. This paradigm has been gaining more traction recently with the increasing popularity of HTTP-based adaptive services (HAS) [9].

The client player is modeled as operating in two modes: buffering (transient) and steady-state. The two modes differ in terms of video chunk request rates from the video server and hence the arrival rates at the LTE eNB. In steady-state mode, chunks are requested at a periodic rate, whereas in transient mode the next chunk is fetched from the server as soon as the last requested chunk is downloaded. Detailed operation of adaptive streaming clients is described in [10]. The client player can be in transient mode in two scenarios: 1) The client is rebuffering and playback is stalled, or 2) The client buffer is under a specified threshold B_{thresh}. The rate adaptation algorithm tries to match the source rate to the link throughput. Hence, user throughput estimation plays a crucial role. The throughput estimate R_{thrpt} is the throughput value averaged over the last P IP packets downloaded by the client, or,

$$R_{\text{thrpt}} = \frac{1}{P} \sum_{i=L_P-P}^{L_P} \frac{S_{\text{packet}}(i)}{T_{\text{download}}(i) - T_{\text{fetch}}(i)}, \qquad (1)$$

where L_P is the index of the last packet downloaded by the client, T_{download} is the time the packet enters into client queue, T_{fetch} is the time when it enters the eNB queue and S_{packet} is the packet size. Averaging helps avoid the effect of short-term throughput variation on rate adaptation. The effect of TCP retransmissions would be reflected in R_{thrpt} as the download time would increase drastically for the corresponding TCP packets. The client starts by fetching the video at the lowest bit rate available. In steady-state phase, the maximum bit rate video that is supported by the client's throughput estimate, R_{thrpt}, is fetched with representation level of

$$Q_{\text{rep}}^{\text{sup}} = \arg\max_i b_i; b_i \leq R_{\text{thrpt}} \text{ over } i = 1, 2 \ldots N, \qquad (2)$$

if $b_1 > R_{\text{thrpt}}$ then $Q_{\text{rep}}^{\text{sup}} = 1$, where b_i denotes the bitrate of encoded video of representation level i and N denotes the highest quality or representation level. In steady-state mode, $Q_{\text{rep}}^{\text{fetch}} = Q_{\text{rep}}^{\text{sup}}$, i.e. the fetched quality is the maximum supported quality at the current throughput. While in buffering mode, in our implementation, the client would keep reducing the fetched bit rate by one quality level stepwise, or, otherwise, the one governed by supported quality, subject to the minimum bit rate video available as,

$$Q_{\text{rep}}^{\text{fetch}} = \max(1, \min(Q_{\text{rep}}^{\text{prev}} - 1, Q_{\text{rep}}^{\text{sup}}))$$

where the $Q_{\text{rep}}^{\text{prev}}$ is representation level of the latest received chunk. Thus, in the transient mode quality level is never scaled up.

III. QoE-aware Radio Resource Management

In the following sections, we propose a RRM algorithm which works in conjunction with adaptive streaming framework. According to our proposed solution, the management of the radio resource in LTE can be seen as a combination of a QoE aware prioritization engine and a playback buffer aware downlink scheduler.

A. Scheduler Design

Conventional scheduler designs (like proportional fair [11]) are based on average rate based utility functions which may correlate poorly with playback buffer state of the users.

Definition 2. A combined utility function $U(\bar{r}, f)$ is proposed:

$$U(\bar{r}, f) = \log(\bar{r}) - \alpha \exp(-\beta(f - f_{\text{min}})), \qquad (3)$$

where $\bar{r}$ is the average user rate [1], f is number of video frames in the playback buffer.

Remark 1. As long as the video frames in the playback buffer, f, exceed the minimum value of f_{min} then (3) is approximately the utility corresponding to proportional fair (PF). When f drops below the minimum value then the utility function rapidly decreases forcing the scheduler to serve users with video frames below the minimum (since the gradient rapidly increases). The parameters α and β determine the rate at which the penalty for violating the constraint increases. The function as in the second term is called barrier functions since they serve to perform a flexible "barrier" around the feasible region. A similar barrier function based utility for average rate guarantee was introduced in [12].

Lemma 1. *If the resource allocation problem is defined as maximizing the sum-utility across all k (say) users or,*

$$\max\ S_U(\vec{r}, \vec{f}) = \sum_{i=1}^{k} U_i, \qquad (4)$$

where each U_i is given by (3), then the resultant scheduling decision in every scheduling opportunity is the choice of the user given by

$$j^* = \arg\max_j \left\{ \frac{\alpha d_j}{S_{\text{frame},j}} \exp(\beta(f_{\text{min}} - f_j)) + \frac{d_j}{R_{\text{smththrpt},j}} \right\}, \qquad (5)$$

where $S_{\text{frame},j}$ is the size of the video frame in transmission, d_j is the instantaneous data rate, f_j is the number of frames

[1]The actual units of metrics are irrelevant as the scheduling decision depends only on maximum gradient direction (refer to appendix A).

in the playback buffer and $R_{\mathrm{smththrpt},j}$ is smoothed average of delivered throughput of user j.

Proof: See Appendix A. ∎

Note that the proposed metric requires the feedback of each user's playback status.

B. QoE aware prioritization

The priorities among users for resource allocation are adjusted based on the dynamic feedback of the QoE metric of rebuffering percentage (p_{rebuf}).

Definition 3. *Rebuffering percentage* is the percentage of the total streaming time spent rebuffering.

To include fairness in terms of rebuffering percentage, the scheduling metric is modified by scaling with the following defined fairness parameters

$$V_j = \begin{cases} 1 + \frac{k \times p_{\text{rebuf},j}}{\sum_{i=1}^{k} p_{\text{rebuf},i}} & \text{if } \sum_{i=1}^{k} p_{\text{rebuf},i} > 0, \\ 1 & \text{otherwise.} \end{cases}$$

With this incorporation the scheduling decision becomes,

$$j^* = \arg\max_j \left\{ V_j \left(\frac{\alpha d_j}{S_{\text{frame},j}} \exp(\beta(f_{\min} - f_j)) + \frac{d_j}{R_{\mathrm{smththrpt},j}} \right) \right\} \tag{6}$$

Hereafter, the above metric is referred to as the PFBF (proportional fair with barrier for frames).

IV. Simulation Model

A. LTE System Model

A dynamic system-level simulation of LTE air-interface based on MATLAB platform with detailed abstractions of application, transport, MAC and physical layers is used for performance evaluation. The capacity was evaluated for an air-interface configuration with 10MHz carrier bandwidth along with Frequency-division duplex (FDD). The maximum number of hybrid automatic repeat request (HARQ) retransmissions was four. Video traffic transmission was simulated from the center cell only to reduce the simulation time and complexity. The base stations in all other cells generated interference patterns corresponding to a full buffer mode of operation. Users were picked randomly from a user population of 684 dropped uniformly in the cell. For each configuration, statistics were collected from thirty different random drops of users in cell area. Wide-band channel quality information (CQI) feedback over physical uplink control channel (PUCCH) is sent every 5 milliseconds from the mobile terminal. HARQ ACK/NACK messages are sent every 6 milliseconds from the mobile terminal and received at the base-station with a delay of 2 milliseconds. The CQI feedback and the HARQ ACK/NACK feedback processes are assumed errorless. For the link to system mapping, the Mutual Information Effective SINR Metric (MIESM) [13] method is used. Packet error rate (PER) is then obtained by using the computed effective signal-to-interference and noise ratio (SINR) in the additive white Gaussian noise (AWGN) PER versus SINR curve corresponding to the used modulation and code rate. Ideal channel estimation over the demodulation reference signals (DM-RS) was assumed. An overhead of 28% is assumed for downlink control channels and reference signals. A frequency selective scheduling scheme is used. Half of the available downlink resources are assumed to be allocated to best effort data services and the other half are available for video users. Table I summarizes the system simulation parameters.

TABLE I: LTE simulation parameters

Parameters	Assumptions
Channel model	3GPP Case 1 with 3D antenna pattern, SCM-UMa (15 degrees angular spread)
Downlink transmit power	46 dBm
MIMO Mode	4×2 SU-MIMO for the downlink
Cellular Layout	Hexagonal grid, 19 cell sites, 3 sectors per site
Distance-dependent path	loss L=I + 37.6log10(.R), R in kilometers, I=128.1
Lognormal Shadowing	Similar to UMTS 30.03, B 1.141
Shadowing standard deviation	8 dB
Number of antennas at UE	2
Number of antennas at cell	4
Antenna configuration at UE	Co-polarized antennas
Antenna configuration at eNB	Co-polarized (0.5 spacing)
Outer-loop for target FER control	10% FER for 1st HARQ transmission
Link adaptation	MCSs based on LTE transport formats according to [14]
HARQ scheme	Chase combining
DL overhead	3 for PDCCH
UE speed	3km/h
Scheduling granularity	5 RB subband
Receiver type	MMSE-IRC
Feedback mode	Wideband PMI based on LTE 4-bit CB, subband CQI
Intersite Distance	500 m
User distribution	Users dropped uniformly in the entire cell

B. Video Traffic and Quality Modeling

An analytical traffic model for near-real time streaming of video (NRTSV) has been proposed in [13]. This model attempts to capture the variability in the traffic by using a truncated Pareto distribution for packet sizes and inter-arrival time between packets. However, such methods may not generate realistic video traffic because it has been observed that compressed video traffic is highly dependent on the characteristics of the video source. Also our main objective, application level performance analysis (like quality evaluation) is virtually impossible using the specified model which is agnostic of encoding dependency among frames. Owing on these reasons we resort to publicly available video traces for traffic generation [15], [16]. These traces list the frame in encoding order and provide the corresponding sizes, quality and other characteristics. In our simulations each frame is first preceded by a Network Abstraction Layer Units (NALU) prefix which can be approximated as a 10 byte header. Overhead of 20 bytes each for the transport and network layers leading to a total encapsulation overhead of 40 bytes per packet is used as described in [15]. We consider a network with a maximum transfer unit (MTU) of 1500 bytes. During the

TABLE II: Video trace details

Video source	Quantization Parameter/ PSNR	Average bitrate (Kbps)	Quantization Parameter/ PSNR	Average bitrate (Kbps)
Sony_1080	34/ 33.5dB	225.1	28/ 37.7dB	498.8
Citizen Kane	38/ 32.7dB	97.1	28/ 39.4dB	333.4
Die Hard	42/ 32.5dB	49.4	43/ 37.8dB	102.7
NBC News	34/ 33.5dB	259.9	28/ 37.2dB	570.7
Matrix-Part1	42/ 33.6dB	45.8	34/ 38.6dB	98.4

transient phase, fetch rate is equal to one group of pictures (GOP) per frame period. While in steady-state, the client switches to periodic fetching with the fetch rate equal to frame rate. Quality adaptation was done by switching among traces of the corresponding average bit rate on a GOP by GOP basis and the throughput estimation (R_{thrpt}) was done by averaging over $L_P = 100$ IP packets. A B_{thresh} of 1 second is used in simulations. An initial startup delay of 1 second and a rebuffering period (the time for which the client rebuffers after playback buffer starvation with playback stalled) of 0.5 second, with a session time of 100 seconds, is simulated. A larger startup delay would imply a more relaxed latency and hence would enable higher video capacity. The investigation regarding this impact is left for future study.

All videos used were H.264 SVC single layer with encoding type: Main (Level 2.1). The resolution was CIF 352× 288 [2]. The videos are variable bit rate and encoded to target fixed quality (quantization parameter). The video details are shown in Table II, with the first two column corresponding to a set of traces with quality level in the range of 32-34 dB PSNR and the last two columns to that of set of traces with quality level in the range of 37-39 dB PSNR. In each Monte-Carlo trial, a user is randomly assigned one out of the five available video sources at the video server. For each simulation run, the quality of the received video is also evaluated based on the offset distortion files for the videos used in the simulation. These offset distortion files specify (for each frame) the quality degradation of subsequent frames if they were replaced by the current frame. This captures the previous-decode-frame-replacement error concealment methodology.

V. Video Capacity Evaluation

For evaluating the number of users that can be satisfactorily supported by the network, two criteria need to be defined: 1) outage threshold (A^{out}) (which defines a satisfied user) and 2) coverage threshold (A^{cov}) (which defines the fraction of satisfied users of the whole population). As an example [13] defined A^{out} as 2% frame losses and A^{cov} as 98% for near-real-time services. For our purpose, the outage threshold is defined in terms of maximum allowable rebuffering percentage.

Definition 4. *Rebuffering outage capacity* ($C^{\text{out}}_{\text{rebuf}}$) is the number of active users that can simultaneously stream video where users are satisfied A^{cov} percentile of the time, with a user being counted as satisfied if and only if the rebuffering percentage in its video streaming session is less than or equal to A^{out}, or

$$C^{\text{out}}_{\text{rebuf}} = \mathbb{E}\left[\arg\max_{K}\left\{\frac{\sum_{i=1}^{K}\mathbb{1}(p_{\text{rebuf,i}} \leq \mathrm{A}^{\text{out}})}{K} \geq \mathrm{A}^{\text{cov}}\right\}\right], \tag{7}$$

where $\mathbb{E}[.]$ denotes the expectation and $\mathbb{1}$ is the indicator function. The expectation is over multiple user geometry realizations.

A. Quality-Capacity tradeoff

Two scenarios are chosen to evaluate this tradeoff:

- FixedQ(32-34): users fetch a video stream with fixed quality in the range of 32-34 dB PSNR (Table II, column 2).
- FixedQ(37-39): users fetch a video stream with fixed quality in the range of 37-39 dB PSNR (Table II, column 4).

Note that a narrow (in terms of PSNR) slab of quality values is used to generalize the nomenclature as all the videos are not available at exactly the same quality value. From Table II, the average load to the system for FixedQ(32-34) is 135.5 Kbps and that for FixedQ(37-39) is 320 Kbps. The rebuffering outage capacity for the mentioned two cases is shown in Fig. 1 for varying value of rebuffering outage threshold A^{out} and a fixed coverage threshold $\mathrm{A}^{\text{cov}} = 95\%$. As expected, the maximum number of simultaneous users that can be supported for FixedQ(32-34) are higher than that for FixedQ(37-39) which demonstrates the evident capacity-quality trade off. Also, with the increase in the A^{out}, $C^{\text{out}}_{\text{rebuf}}$ increases monotonically. Further if the clients are allowed to do video adaptation depending on link condition (as described in Section II), it would be interesting to assess the rebuffering outage capacity improvement. Thus, two further cases are incorporated:

- AdaptQ(32-34): users adapt according to link conditions. The representation levels available from the server range from the quality level of 24-26 dB up to the maximum representation level having the corresponding quality in the range of 32-34 dB PSNR.
- AdaptQ(37-39): users adapt according to link conditions. Configuration is same as AdaptQ(32-34) with the exception of the maximum available quality being in the range of 37-39 dB PSNR.

The respective performance gains are shown in Fig. 1. As expected, with respect to the defined outage criteria, adaptive streaming proves to be very effective in increasing the rebuffering outage capacity. Similar trends of monotonic increase in $C^{\text{out}}_{\text{rebuf}}$ with increase in A^{out} are observed. The relative gain from FixedQ(37-39) to AdaptQ(37-39) is much higher than that from FixedQ(32-34) to AdaptQ(32-34). This is because the clients have more video representation levels to switch to in the former. Note that allowing more representation levels for adaptive streaming leads to decrease in rebuffering outage capacity (compare AdaptQ(37-39) and AdaptQ(32-34)) because of the content agnostic RAN (proportional fair

[2]More details can be found at http : //trace.eas.asu.edu/videotraces2/svc_single

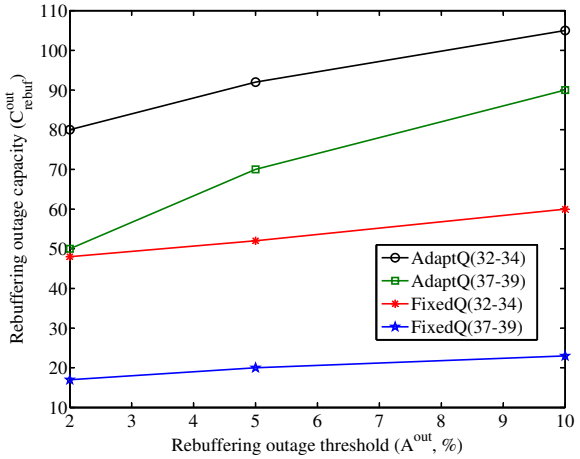

Fig. 1: Variation in the rebuffering outage capacity with the rebuffering outage threshold across various configurations

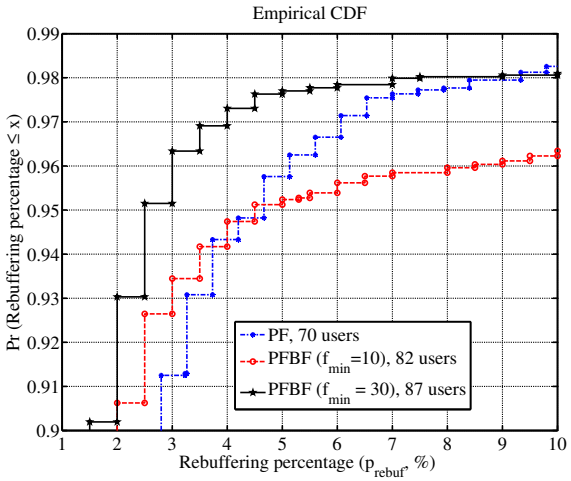

Fig. 2: CDF of rebuffering percent across users

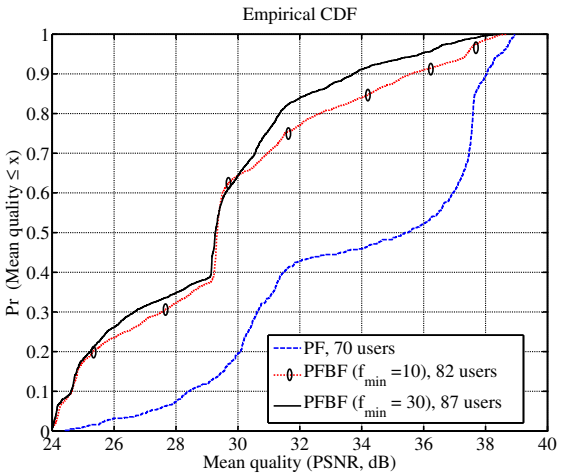

Fig. 3: CDF of mean session quality across users

resource allocation is used for computing the results) and greedy client based implementation of HAS services. This motivates the use of a QoE-aware resource management to work in conjunction with adaptive streaming.

B. QoE enhancements with RRM

For the performance analysis of the proposed algorithm of Section III, a form of the PFBF metric, (6), with $\alpha = \beta = 1$ is used. A system with AdaptQ(37-39) configuration is simulated. Fig. 2 shows the distribution of rebuffering percent across users for different values of $f_{\min}$ and users. Higher value of $f_{\min}$ implies higher fairness being guaranteed by the network to users in terms of their playback buffer. As compared with the baseline of PF based scheduling, PFBF allows the network to accommodate more users keeping the 95^{th} value of rebuffering percent almost same, with the exact gains depending on $f_{\min}$. Fig. 3 shows the distribution of video quality across users. As can be seen from Fig. 3, in order to gain from the rebuffering aspect, some quality[3] needs to be sacrificed. Under the assumption that user experience is more sensitive to playback stall than to nominal quality degradation, the proposed technique would lead to increase in the number of satisfied users in the system and hence translate into capacity gains. For example, fixing $\mathrm{A}^{\mathrm{cov}} = 95\%$ and $\mathrm{A}^{\mathrm{out}} = 5\%$, the rebuffering outage capacity for PF, PFBF with $f_{\min} = 10$ and PFBF with $f_{\min} = 30$ is 70, 82 and 87 respectively. Thus, the proposed resource allocation provides a capacity gain in the range of $\sim 20\%$, while providing the operator the flexibility to dynamically tune the parameters based on user preferences.

VI. Conclusions

In this paper, we introduce the notion of rebuffering outage capacity for evaluating the capacity of downlink LTE air-interface for buffered video services. This is the first work to evaluate the performance based on rebuffering percentage as the outage criteria. A clear quality-capacity trade off was demonstrated through the results. The impact of adaptive streaming on the rebuffering outage capacity is also assessed and it is shown that adaptive streaming can be instrumental in decreasing the rebuffering percentage, and hence, increasing the rebuffering outage capacity. Even with adaptive streaming, some users located in bad channel conditions could suffer because of the greedy client based implementation of rate adaptation. A key takeaway was that the availability of more video representation levels corresponding to higher bit rate video could lead to a decrease in the number of active users that can be supported by the network. A QoE-aware radio resource management with tunable parameters is thus proposed to work in conjunction with adaptive streaming which allows the network operator to maximize the number of satisfied customers. Our results show that with QoE aware-

[3]Although quality trend is shown with PSNR here due to constraints of video traces, other quality metrics like MS-SSIM and VQM can also be used

ness in the RAN there is a significant potential to optimize rebuffering outage capacity and thus profits while provisioning a guaranteed QoE.

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015."

[2] O. Oyman, J. Foerster, Y.-J. Tcha, and S.-C. Lee, "Toward enhanced mobile video services over WiMAX and LTE," *IEEE Commun. Mag.*, vol. 48, pp. 68–76, Aug. 2010.

[3] O. Oyman and S. Singh, "Quality of experience for HTTP adaptive streaming service," *IEEE Commun. Mag.*, Apr. 2012.

[4] TDG Research. [Online]. Available : http://www.tdgresearch.com.

[5] A. Talukdar, M. Cudak, and A. Ghosh, "Streaming video capacities of LTE air-interface," in *IEEE International Conference on Communications (ICC)*, pp. 1–5, May 2010.

[6] K. Piamrat, K. D. Singh, A. Ksentini, C. Viho, and J.-M. Bonnin, "QoE-Aware scheduling for video-streaming in High Speed Downlink Packet Access," in *Wireless Communications and Networking Conference (WCNC), 2010 IEEE*, pp. 1–6, Apr. 2010.

[7] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, "QoE-driven cross-layer optimization for High Speed Downlink Packet Access," *Journal of Communications*, vol. 4, no. 9, 2009.

[8] R. Mok, E. Chan, and R. Chang, "Measuring the quality of experience of HTTP video streaming," in *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pp. 485–492, May 2011.

[9] 3GPP, "Transparent end-to-end packet-switched streaming service (PSS); progressive download and dynamic adaptive streaming over HTTP (3GP-DASH) (Release 10)," TS 26.247, 2011.

[10] S. Akhshabi, A. C. Begen, and C. Dovrolis, "An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP," in *Proceedings of the second annual ACM conference on Multimedia systems*, MMSys, (New York, NY, USA), pp. 157–168, ACM, Feb. 2011.

[11] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *IEEE Vehicular Technology Conference*, vol. 3, pp. 1854–1858, May 2000.

[12] P. Hosein, "QoS control for WCDMA high speed packet data," in *4th International Workshop on Mobile and Wireless Communications Network*, pp. 169 – 173, Sept. 2002.

[13] R. Srinivasan, J. Zhuang, L. Jalloul, R. Novak, and J. Park, "IEEE 802.16m evaluation methodology document (EMD)," tech. rep., IEEE 802.16 Broadband Wireless Access Working Group, Jan. 2009.

[14] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 10)," TR 36.213, 2011.

[15] P. Seeling and M. Reisslein, "Video transport evaluation with H.264 video traces," *IEEE Communications Surveys Tutorials*, vol. PP, pp. 1–24, Sept. 2011.

[16] G. Van der Auwera, P. David, and M. Reisslein, "Traffic and quality characterization of single-layer video streams encoded with the H.264/MPEG-4 advanced video coding standard and scalable video coding extension," *IEEE Trans. Broadcast.*, vol. 54, pp. 698–718, Sept. 2008.

VII. ACKNOWLEDGMENT

The authors would like to acknowledge the help of Chunming Han with the system level simulations.

APPENDIX A
PROOF OF LEMMA 1

Proof: The system can not be instantly moved to the optimal solution of (4). Also due to changing the channel conditions the optimal solution also changes. By taking scheduling decisions governed by the greatest ascent direction the system can move to the "present" optimal solution. Thus, the user, which when scheduled, results in movement along the maximum utility function gradient direction is chosen. The proposed utility function can be subdivided into two utility functions as

$$U(\bar{r}, f) = \log(\bar{r}) - \alpha \exp(-\beta(f - f_{\min})) = U^1(\bar{r}) + U^2(f).$$

The solution to the first part of the utility function (U^1) is the well known proportional fair scheduler[12]. The gradient of the second utility function at n^{th} LTE subframe is denoted as $U'_j(f_j(n))$ (superscript 2, is omitted from the following analysis). The update equation for f_i after every LTE subframe can be written as

$$f_i(n+1) = \left(f_i(n) - \frac{1}{N}\right)_0 + \frac{d_i(n)}{S_{\text{frame},i}(n)},$$

where N is the video frame period in terms of LTE subframes, S_{frame} is the size of the current frame in transmission and $(x)_0 = \max(x, 0)$. Parameterizing the movement along the ray corresponding to serving user j by ϵ, the objective function can be written as,

$$S^{\epsilon}_{U,j}(\vec{f}) = \sum_{i=1}^{k} U_i(f_i(n) + \epsilon(f_i(n+1) - f_i(n)))$$

$$= \sum_{i=1, i\neq j}^{k} U_i\left(f_i(n) - \frac{\epsilon}{N}\right) + U_j\left(f_j(n) + \epsilon\left(\frac{d_j(n)}{S_{\text{frame},j}(n)} - \frac{1}{N}\right)\right).$$

Taking the derivative with respect to ϵ at $\epsilon = 0$ we get,

$$S'_{U,j} = -\frac{1}{N}\sum_{i=1, i\neq j}^{k} U'_i(f_i(n)) + U'_j(f_j(n))\left(\frac{d_j(n)}{S_{\text{framesize},j}(n)} - \frac{1}{N}\right).$$

Gradient in the direction corresponding to serving user j is,

$$S'_{U,j} = U'_j(f_j(n))\frac{d_j(n)}{S_{\text{framesize},j}(n)} - \sum_{i=1}^{k} U'_i(f_i(n))\frac{1}{N}.$$

Since the second term is common to all the directions, the maximum gradient direction is given by,

$$j^* = \arg\max_j S'_{U,j} = \arg\max_j \left\{U'_j(f_j(n))\frac{d_j(n)}{S_{\text{framesize},j}(n)}\right\},$$

which for $U(f) = -\alpha \exp(-\beta(f - f_{\min}))$ becomes the choice of the user in each scheduling interval given by,

$$j^* = \arg\max_j \left\{\frac{\alpha d_j}{S_{\text{framesize},j}} \exp(\beta(f_{\min} - f))\right\}.$$

■