Video Capacity and QoE Enhancements over LTE

Sarabjot Singh, Ozgur Oyman, Apostolos Papathanassiou, Debdeep Chatterjee, Jeffrey G. Andrews

Department of Electrical and Computer Engineering, University of Texas at Austin

Intel Corporation, Santa Clara, California

Stored Video is an Increasingly Dominant Source of Wireless Traffic

- Youtube, Netflix, ESPN, other clips shared via servers
- Rebuffering is the key QoE impairment for stored video
  - Frame losses, delay, retransmissions, etc. all boil down to rebuffering
  - A 1% increase in buffering leads to an average decrease of 3 minutes in user engagement.
  - Viewers watch 32% more video rebuffering is eliminated.
  - Viewers experiencing a single start-up failure return 54% less.
- Metrics like *rebuffering percentage* – the average % of time spent rebuffering rather than viewing video – should play a key role in determining "outage" in a video delivery system

The Need for Adaptive Streaming

- The rate available to users in a wireless network is highly unpredictable and time/space-variant
  - SINR varies dramatically over time and space due to channel and interference fluctuations (interference usually the more important)
  - Network congestion is also a major factor: at peak times even users with very high SINR may get a low rate
- Clearly, there is a need to dynamically adapt video streaming to such conditions if one wishes to avoid rebuffering
  - Send high resolution video to users in "good" conditions
  - Send low resolution video to users in congested or low SINR conditions
- Youtube currently allows viewers to pick from a few resolution levels manually, for example. Other apps do it automatically.

Related Work

- Video capacity for LTE in the context of real-time video was evaluated (similar to here) and reported in

  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.

- A cross-layer sum utility optimization, where the QoE was abstracted for various services as a function of the allocated rate in the context of HSPA in

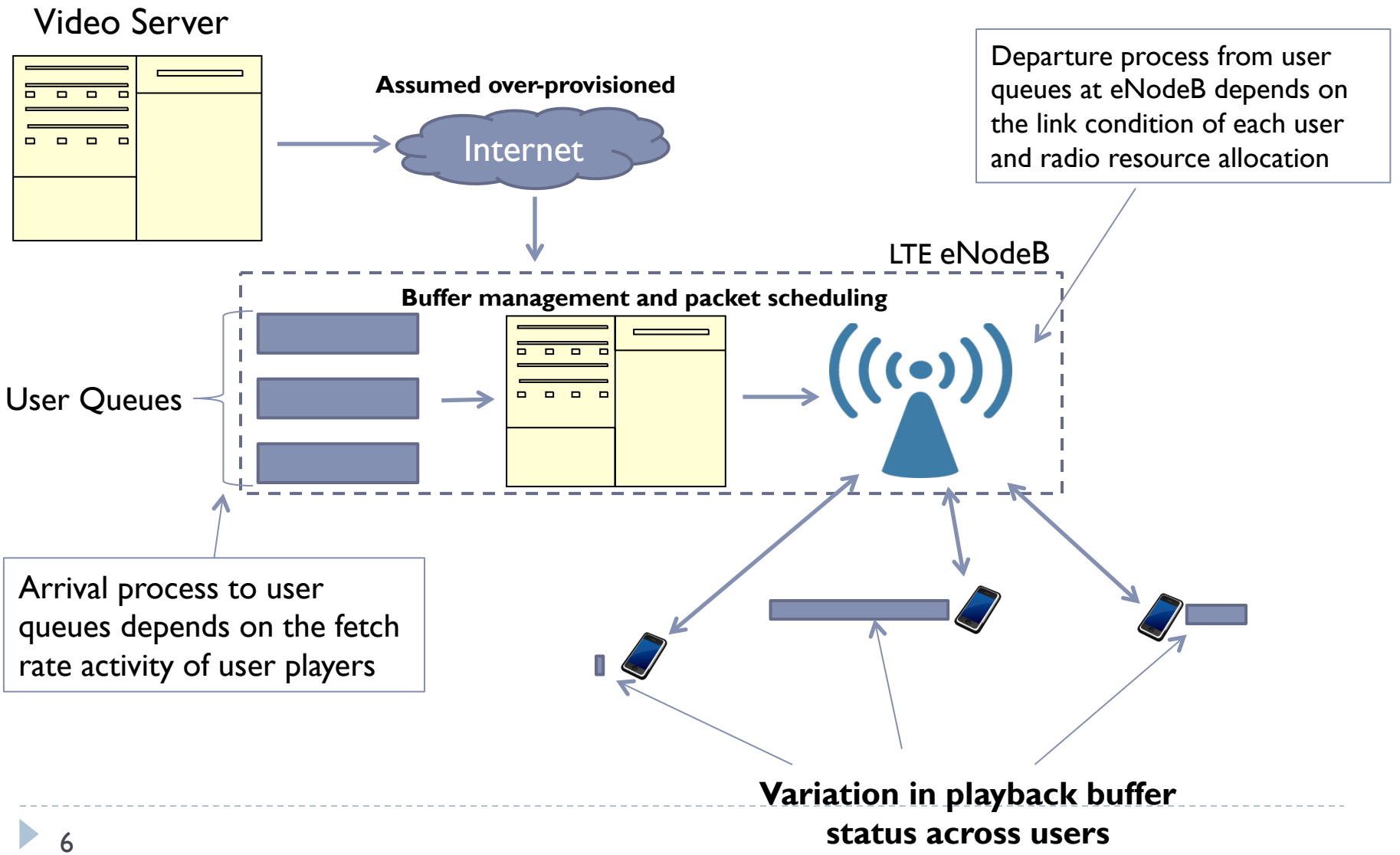
  S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, "QoE-driven crosslayer optimization for High Speed Downlink Packet Access," Journal of Communications, vol. 4, no. 9, 2009.

- Resource Management based on users' playback buffer status and rebuffering percentage have not been investigated
- Minimal work on characterizing and optimizing video streaming over LTE, esp. stored video

Our Contributions

- Define a QoE-aware outage criteria for buffered video streaming services
  - Users experiencing rebuffering percentage greater than a threshold are declared as in outage.
- Introduce the concept of rebuffering outage capacity
  - Only count the users in coverage towards system capacity
- Proposed a QoE-aware RRM framework that works in conjunction with adaptive streaming to optimize capacity
  - A standard QoE-agonistic resource manager cannot take full advantage of adaptive streaming capabilities
- Evaluate all this in the context of LTE downlink (3GPP Rel. 8/9)

System Architecture



Key Metric: Rebuffering Outage Capacity

Rebuffering Outage Capacity: The number of users in a cell that can simultaneously stream video subject to two constraints:

1. Each user rebuffers a fraction of time less than A^{out}
2. A fraction A^{cov} of these users meet the first constraint (on average)

Formally:

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

The expectation is over multiple user geometry realizations, and p_{rebuf} is fraction of time spent rebuffering.

Optimization based on playback buffer status

- A combined "barrier function" utility function is proposed:

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

  - $\bar{r}$ is the average rate, *f* is the number of video frames in playback buffer.
  - α, β and $f_{\min}$ are tunable parameters. Higher values give more emphasis on the rebuffering portion.

- ***Lemma 1***: Maximizing the sum log utility across all uses results in the optimum user *j** (in each LTE sub-frame) given by

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

Usual PF solution

playback buffer aware term

- $S_{frame,j}$ is the size of the video frame in transmission, d_j is the instantaneous data rate and $R_{\text{smththrpt,j}}$ is smoothed average of delivered throughput to user *j*

QoE-aware Radio Resource Management

- If users additionally feedback information about how often they rebuffer (i.e. p_{rebuf}), this can be further exploited to improve QoE
- QoE-aware prioritization results in a novel PFBF (proportional fair with barrier for frames) scheduler with the choice of the user to be scheduled (in each LTE subframe) given by:

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

where

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

  - V_j accounts for the long term affect of rebuffering percentage, with k being the total number of users
  - A user who has encountered more rebuffering over time would be given higher priority

Simulation Model (LTE Release 8/9)

- Video traffic transmission is simulated focusing on a center cell in a 19 cell hexagonal grid
  - Half of resources used for video
  - Other half is reserved for voice/data
- Downlink users are chosen randomly from a larger population dropped uniformly in the center cell.
  - 100,000 LTE sub-frames were simulated for each of these users
  - Base stations in all other cells generate interference to the selected users corresponding to full buffer operation
  - Statistics obtained after averaging over 30 distinct random drops

Parameters	Assumption
Channel model	3GPP Case 1 with 3D antenna pattern SCM-UMa (15 degrees angular spread)
Downlink transmit power	46 dBm
MIMO Mode	4x2 SU-MIMO for the downlink
Cellular Layout	Hexagonal grid, 19 cell sites, 3 sectors per site
Distance-dependent path loss (dB)	$L = I + 37.6\log_{10}(R)$, R in kilometers, I=128.1
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 (0.5λ spacing)
Outer-loop for target FER control	10% FER for 1st HARQ transmission
HARQ scheme	Chase combining
HARQ delay	8 ms
Max HARQ Retx	4
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
CQI Delay	5 ms
Intersite Distance	500 m

Adaptive Streaming

- Video library at the server contains 5 videos each at different quality levels (perhaps 7 or 8 such levels)
  - Example characteristics of the video traces* of a certain target quality level (32-34 dB)

Video Source	Quantization Parameter (PSNR)	Average Bitrate (including overhead) (Kbps)
Sony_1080	34 (33.5dB)	225.1
Citizen Kane	38 (32.7dB)	97.1
Die Hard	42 (32.5dB)	49.4
NBC News	34 (33.6 dB)	259.9
Matrix-Part 1	42 (33.6 dB)	45.8

- Each user assigned a video randomly.
- Each user adapts video quality and rate based on its perceived end-to-end throughput.
  - The throughput estimate is obtained by averaging over multiple packets
  - Player then chooses the video stream level of bit rate less than the estimated throughput

*Traces taken from
http : //trace.eas.asu.edu/videotraces2/svc _single

Quality-Capacity tradeoff use cases

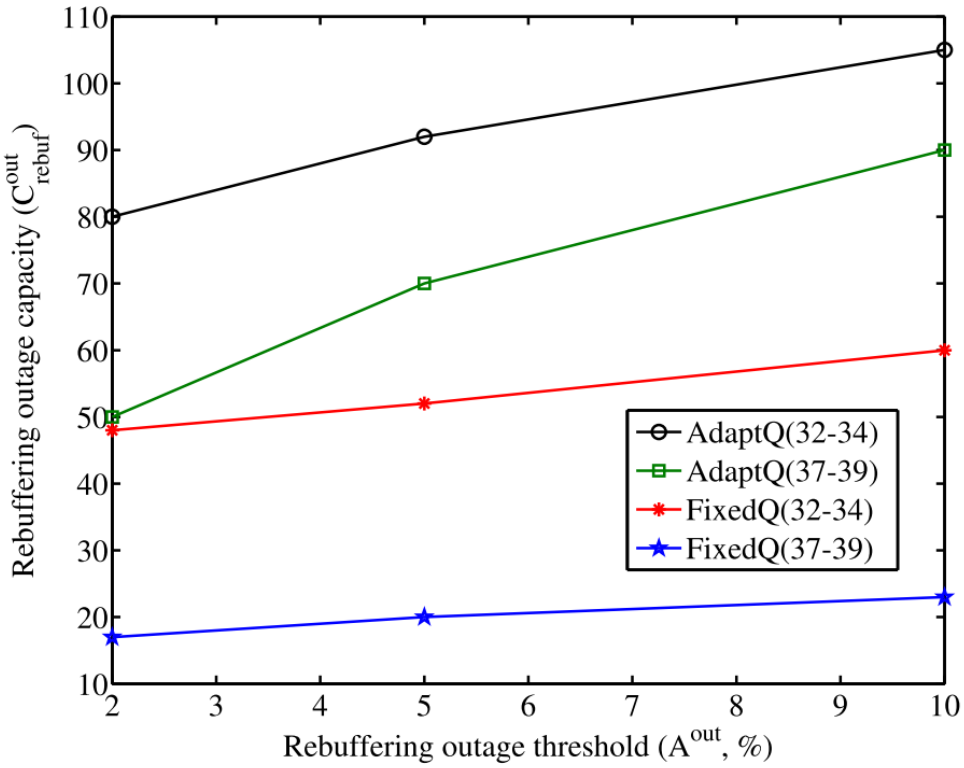
Four use cases are chosen to evaluate this tradeoff:

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

3. AdaptQ(32-34): users adapt according to link conditions.
   - Minimum quality level of 24-26 dB
   - Maximum of 32-34 dB PSNR.
4. AdaptQ(37-39): same as above except with maximum quality of 37-39 dB PSNR for very high resolution.

Quality-Capacity Tradeoff

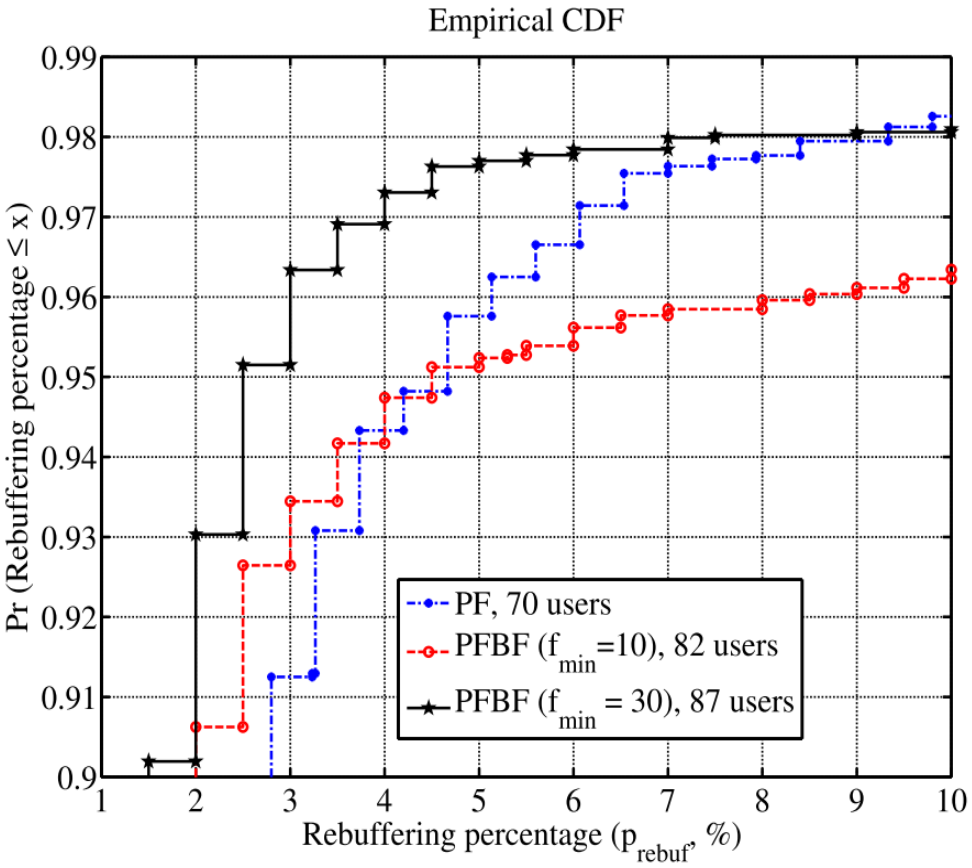
- Adaptive streaming provides significant capacity gains, in the range of **100% to 300%.**
  - Compare AdaptQ(37-39) with FixedQ(37-39)
- Availability of higher quality levels reduces video capacity of the system in the absence of a QoE aware RRM.
  - Compare AdaptQ(37-39) with AdaptQ(32-34)
- Relaxing the rebuffering outage threshold allows packing more users in the system.



A^{cov} of 95% used.
Proportional fair scheduling used in these results

QoE enhancements with improved RRM

- Rebuffering outage capacity at A^{cov}=95%
- $\alpha = \beta = 1$
- Higher value of f_{min} gives more emphasis on the playback buffer occupancy while allocating resources.
- Capacity gain in the range of 20-25% **in addition** to adaptive streaming gains.



AdaptQ(37-39) use case

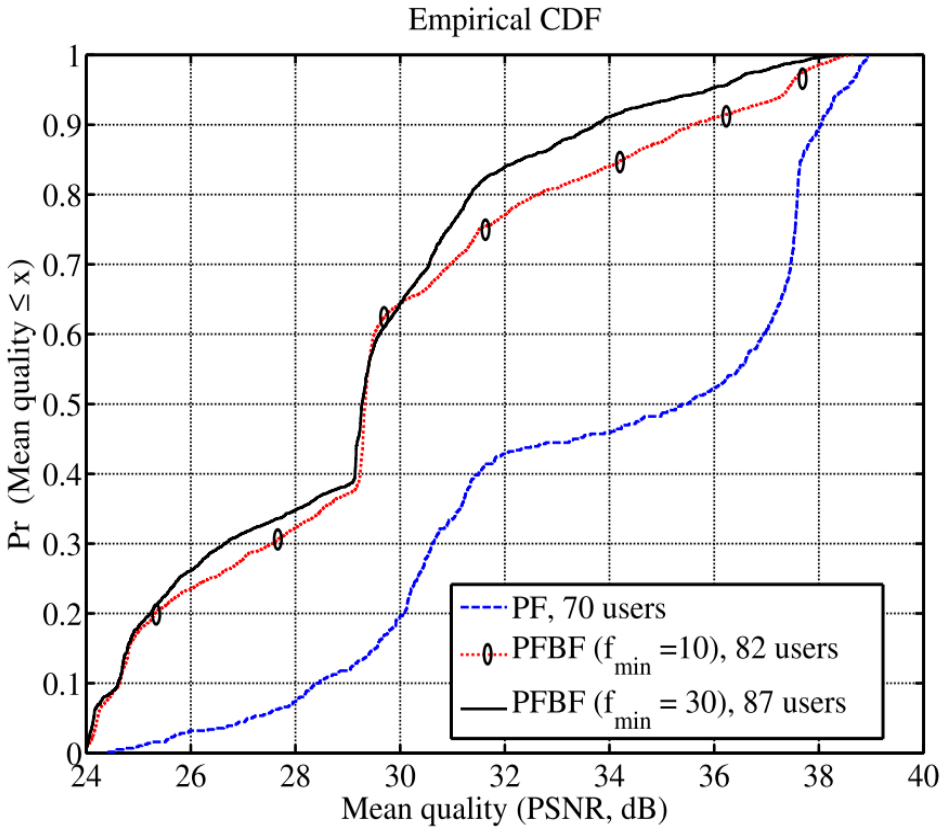
Summary & Future Directions

- Rebuffering outage capacity a logical metric for stored video capacity
- Adaptive streaming alone is not enough
  - Key to decreasing the rebuffering percentage, and hence increasing the rebuffering outage capacity.
  - However, "disadvantaged" users suffer since "privileged" user devices request and can successfully receive high resolution videos
- Thus, proposed a QoE-aware RRM that introduces further tunable parameters to increase fairness and capacity further
- Future Directions
  - Centralized video rate adaptation (i.e. at eNodeB), may be able to better balance competing requests and provide more fairness
  - Need of a composite QoE metric that can account for rebuffering percentage and video quality at the same time.
  - QoE-aware Characterization of Live Streaming Video in LTE

Backup Slides

QoE enhancements with RRM

- Higher value of f_{min} looses out in video quality while giving more more emphasis on the playback buffer occupancy.
- The gains in terms of rebuffering percent can be further increased by sacrificing the video quality.
- Proposed PFBF provides the flexibility to tune resource allocation as per user preferences (quality v.s. rebuffering)



Backup Slides

Rate adaptation for Adaptive Streaming

The throughput estimate R_{thrpt} is the throughput value averaged over the last P IP packets downloaded by the client, or,

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

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.

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

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.

Backup Slides

Proof of Lemma 1: The system can not be instantly moved to the optimal solution. 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. 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 $(x)_0 = \max(x, 0)$.

Backup Slides

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,

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