Delay-Conscious Caching Scheme for Energy-Efficient Video Streaming

Árpád Huszák
Department of Networked Systems and Services
Budapest University of Technology and Economics, Budapest, Hungary
huszak@hit.bme.hu

***Abstract*—Wireless multimedia streaming became a popular service of today's mobile networks, however due to the limited battery capacity energy efficient solutions are required to extend the availability of the mobile device. Although energy efficiency is getting more and more critical factor, the experienced quality of the customer can not be deteriorated. In this paper an energy efficient caching scheme is proposed that keeps the balance between the power consumption and the acceptable playout delay caused by caching that has direct influence on the perceived quality. According to the obtained results the proposed delay-conscious caching model can serve as a guideline for the offline playout delay determination in order to keep the delay as low as possible, but reduce the consumed energy as well.**

Keywords-component; video streaming; energy efficiency; caching; quality of service;

I. Introduction

By the escalation of mobile network connection speeds, multimedia services became the most popular services over the Internet. Recent studies have shown that video streaming is responsible for 25-40% of the Internet traffic [1], but according to the forecasts [2], two-thirds of the world's mobile data traffic will be video by 2017. Of course, energy-efficiency will play a significant role in this improvement, because mobile devices typically have a limited energy budget. While the energy capacity of mobile phone batteries has improved slowly, [3], the phone complexity trend still follows the Moore's Law. So energy efficient methods must be investigated to control the power usage of mobile devices. It must be always kept in mind that the quality of the service is the most important criteria of the customer. The energy efficient solutions will be acceptable for the users, if the usage of these methods is imperceptible from the quality point of view. Multimedia services are very sensitive to quality degradation, because strict conditions (e.g. delay, jitter, packet loss rate, bandwidth, etc.) must be fulfilled to provide high quality service for the customers.

Video caching is an admitted method for media streaming, because it smooths the delay and bandwidth variation during the streaming. However it can be also used for energy reduction purposes, when the network interface is periodically switched on and off in order to reduce power consumption. The deployed cache will hide the effect of altering *ON*/*OFF* states by introducing an extra delay. Of course, the tolerable caching delay depends on the type of video streaming. We can divide streaming media applications into three subtypes: one–way pre–recorded, one–way live and two–way interactive media. Caching solution is not acceptable for interactive media due to very low tolerable delay. For one-way live streams (e.g. internet TV) few seconds delay is still allowed. Many smoothing algorithms have been proposed [4], but the analyzed performance metrics are often different and not optimized all together

In this paper an energy efficient caching scheme is introduced that keeps the balance between the power consumption and the acceptable delay in case of *ON*/*OFF* energy reduction scheme. Utilizing the relation between the adopted playout delay and the power consumption, the aim of this work was to give guidelines for the playout delay determination in order to keep the delay as low as possible, but also reduce the consumed energy. Analytical considerations were used to introduce the proposed model, but in order to analyze the efficiency of the scheme, a simulated WLAN environment was deployed. The model is applicable for 3G and 4G networks, too. According to the obtained simulation results, the proposed delay-conscious caching method can help to find the appropriate parameters (playout delay, video bitrate) of the energy efficient video streaming scheme.

The rest of the paper is organized as follows. The related works are presented in Section II. In Section III the analytical model of the proposed delay-conscious caching scheme is introduced, while the evaluation of the method follows in Section IV. Finally, Section V concludes the paper.

II. Background and Related Works

Energy issues in mobile environment have been studied in many papers investigating the consumption characteristics of wireless interfaces [5][6]. According to the published results, one of the most energy consuming components in mobile devices is the wireless interface since networking activities cause up to 50% of the total energy consumption in hand-held devices and up to 10% in laptops.

The interfaces are designed to handle different states in order to decrease the power consumption [7][8]. It can be described with three states: active (*ON*), idle and sleeping (*OFF*). In active state the user is either receiving (Rx) or transmitting (Tx) packets. The idle state means that the user is participating actively but is not transmitting or receiving. In sleeping or *OFF* state the user device powers off to save energy. Typically, the transition to the active state is longer than switching the interface off. During the state changes no traffic is forwarded, however energy is consumed while

978-1-4673-5828-6/13/$31.00 ©2013 IEEE

switching from *OFF* to *ON* state, or contrary. The required time for changing states was analyzed in [10]. The authors found that about 1s is needed to switch the interface to active state and 0.1 s to 0.2 s to switch it off.

Numerous power consumption models were published [11][12] in the last decade. In this paper a bitrate dependent WLAN interface power consumption model was used based on [11], where higher transmission rates cause a slight increase in power consumption in the active state. If energy efficiency is not an important issue, active and idle states are used, however the consumption can be significantly decreased if the device switches to sleep (*OFF*) mode instead of idle. Due to long transition times, switching to *OFF* state is acceptable only if the interface can stay offline for a longer time period.

Different power reduction methodologies have been presented in the last decade ranging from physical to the application layer. For instance, µPM [13] enables WLAN interface to enter power saving modes even between MAC frames, without noticeable impact on the traffic flow. Energy efficient solutions for video streaming applications were also investigated by authors of [14] and [15]. They proposed to use additional buffer to temporally cache video stream data that can be transmitted to the client using burst rate. After the burst, the WLAN, 3G or 4G interface can be switched into a sleep state to improve the energy consumption efficiency.

Bertozzi et al. [16] also proposed a playout buffer aware mechanism, which switches off the WLAN interface until the playout buffer is reduced to a threshold level and then turns it on again, however the buffer delay setup was analyzed slightly. During the offline period, the access point buffers the incoming video stream packets for the mobile device and thus creates bursty traffic.

III. Delay-Conscious Caching Scheme

Video streaming typically generates asymmetric network traffic requiring high downlink bandwidth, while the uplink traffic is negligible in most of the cases. If UDP is used without any control protocol extension, there are no packets delivered from the client to the media server at all. Assuming that no other application is running in the background that requires network connection, we propose to use a caching scheme that optimizes power consumption of the client during the video streaming

The proposed scheme of this paper focuses on one-way live video allowing few seconds delay, however the goal is to keep the delay caused by caching as low as possible. In traditional streaming solutions the caching delay is used to provide continuous playout even if the jitter is high, however caching can be used to heap up a short part of the streamed video and switch off the network interface. While the network interface is down, the video stream is played from the cache. Due to the continuous playout, the caching scheme will be effective only if the downlink bandwidth (Bw_d) is higher than the video bitrate (λ_v), because during the active (*ON*) period (t_{ON}) the cache can be filled only if $\lambda_v < Bw_d$ (Figure 1.).

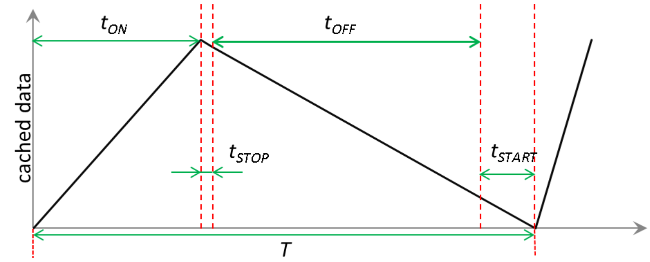

Figure 1. Heap up and offline playout cycle (*T*)

The t_{ON} time depends on the video bitrate (λ_v), downlink bandwidth (Bw_d) and the caching delay (d) as the equation shows.

$$t_{ON} = \frac{d \cdot \lambda_v}{Bw_d - \lambda_v} , \qquad (1)$$

During the interface state changes (t_{STOP}, t_{START}) and the *OFF* period, the reception of stream packets is not possible. Due to the unavailability of the client while its interface is down two different approaches can be considered.

In the first one the media server and the client uses synchronized clocks to determine when is the client ready to accept stream packets. In this case the length of the heap up and offline playout cycle (*T*) must be always the same. Unfortunately it can not be guaranteed that the download rate is constant and during a t_{ON} period exactly d length video is heaped up.

The more realistic solution is to use control messages to inform the media server that the client's interface is up and ready to receive the next portion of the video stream. Similarly, the server must inform the client that no more packets will be sent within the actual cycle and it can switch off its interface. The proposed communication scheme between the server and client is shown in Figure 2.

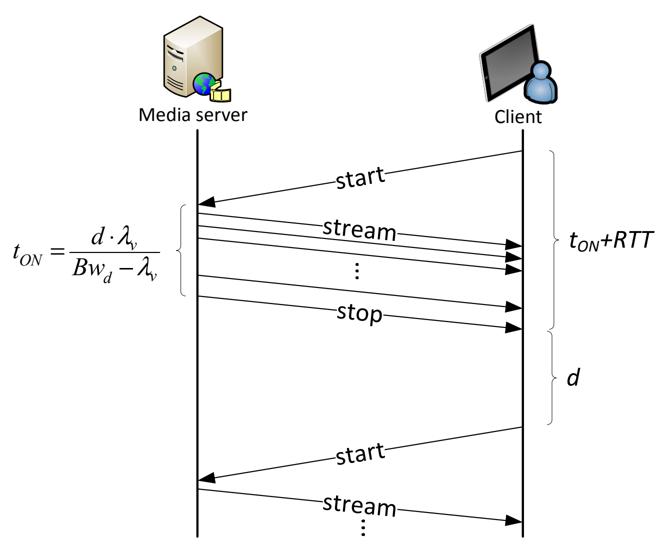

Figure 2. Relation of caching delay and consumtion

When the client interface is up, a *start* message is sent to the server. Afterwards, the server starts or continues to forward the video stream packets with higher bitrate than the video rate (λ_v) in order to heap up longer video that will emptied from the cache during the *OFF* period. The *start* messages must contain the caching delay (d) parameter, because it is required by the media server to determinate the t_{ON} time. The other way is that instead of t_{ON} time, the media server calculates the amount of streamed data (S_v) that must be forwarded during the *ON* period.

$$S_v = \lambda_v (d + t_{ON}) = \lambda_v T \qquad (2)$$

The client must be also informed that the server forwarded all packets belonging to d+t_{ON} length video stream and the interface can be switched off. Due to continuous playout the cache will be starved after d time. Before the heaped up video stream is played out, the interface must be switched on to ask for the next portion of the stream by sanding the *start* message.

The cached video length must be long enough to provide seamless streaming for the time of a cycle (Figure 1.), therefore the caching delay must be determined as follows:

$$d = t_{STOP} + t_{OFF} + t_{START} \qquad (3)$$

Of course each state has its own energy consumption characteristic as already discussed in the previous section. It was already shown in [15][17] that the energy consumption can be reduced by increasing the cache size. Longer caching delay makes it possible to keep the interface longer in offline state. Unfortunately, the prolonged playout delay is not acceptable for live video streaming.

In this paper we investigate the relation between the consumption and playout cache size in order to find the optimum. The consumed energy is the sum of required energy in *ON* state, interface *start* and *stop* state.

$$W = P_{ON} \cdot t_{ON} + P_{stop} \cdot t_{stop} + P_{start} \cdot t_{start} \qquad (4)$$

By decreasing the *ON* time and the number of state changes, the overall consumption is reduced because the consumption is higher when the interface is up (P_{ON}>P_{start} and P_{ON}>P_{stop}). The result is the same if the *OFF* state time is increased, leading to longer caching delay that is illustrated in Figure 3. based on analytical calculations. When the caching delay is set to lower values, frequent state changes are generated causing higher consumption, because the overall t_{start} and t_{stop} period will increase. In these periods the energy is consumed, but no traffic is forwarded on the link.

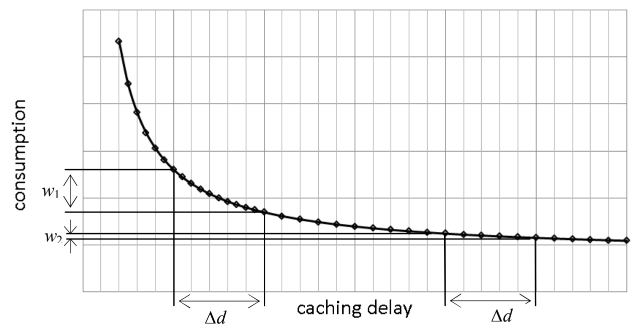

Figure 3. Relation of caching delay and consumtion

The goal is to reduce to consumption, but keep the caching delay as low as possible. Due to the exponential characteristic of the consumption and delay relation, the difference in consumption becomes minimal if the caching delay is high. As it can be seen in Figure 3. the difference of consumption values is significantly higher (w_1> w_2) if lower caching delays are used, however the difference of delay (Δd) is the same in both cases. It is not worth to set too high caching delays, because the improvement in consumption will be minimal, but the user's quality of experience will be significantly depreciated.

Assuming that the consumption and the playout delay is equally important, the optimal setup is when ε unit of change in the caching delay leads to same ε change in consumption. The theoretical optimum can be expressed using the consumption function $W(d)$:

$$W(d_0 + \varepsilon) = w_0 - \varepsilon \qquad (5)$$

We can formalize the optimum caching delay in other way, too. In the proposed delay-conscious caching scheme, the theoretical caching delay optimum (d_0) is the point where the derivative of $W(d)$ is equal to -1.

$$\frac{\Delta W(d)}{\Delta d} = -1 \qquad (6)$$

If the caching delay is higher than the theoretical optimum (d>d_0) the obtained consumption improvement can be quite negligible, but the delay, which is an important parameter of the QoS and QoE, can become even unacceptable for the users. On the other hand, if lower caching delay is set up (d<d_0), the consumption will significantly increase, but the experienced quality will not be proportionately better.

The playout delay sensitivity depends on the multimedia service. In some cases longer caching delay can be enabled (e.g. video on demand), while in other cases the low delay can be more important than the consumed energy (e.g. live TV stream). The proposed theoretical caching delay determination model can be easily extended with weights. In this case the gradient of the tangent is -β_c/ β_d at the $W(d_0)$

point, where β_c is the weight of the consumption and β_d is the weight of the playout delay.

$$\frac{\Delta W(d)}{\Delta d} = -\frac{\beta_c}{\beta_d} \tag{7}$$

In order to formalize the consumption in function of caching delay based on equation (1) and (4) while assuming that $P_{start} \cdot t_{start} = P_{stop} \cdot t_{stop} = W_s$, the following statement can be considered

$$W(d) = \left(\frac{d\lambda_v}{Bw_d - \lambda_v} \cdot P_{ON} + 2W_x \right) \cdot \frac{l_v}{d + t_{OFF}}, \tag{8}$$

where l_v stands for the length of the streamed video and $l_v/(d+t_{OFF})$ determines the required number of cycles (*ON*, *OFF*, *start* and *stop*). Rearranging the equation (8) and simplifying the relation:

$$d = \frac{2W_x l_v (Bw_d - \lambda_v)}{W(Bw_d - \lambda_v) - (l_v \lambda_v P_{ON})} \equiv \frac{c}{W - \beta(\lambda_v / Bw_d)} \tag{9}$$

We can conclude the caching delay is inversely proportional to consumption and the ratio of video/download bitrate.

In real networks the $W(d)$ function can not be exactly formulated, because it depends on number of other continuously changing parameters, like available bandwidth depending on channel conditions, video stream bitrate (e.g. VBR video), distance from access point or mobile base station, consumption values regarding to different states, etc. Due to the complexity of the consumption and caching delay relation, experimental relations were used to determinate the acceptable caching delay to find the balance between the consumption and caching delay.

IV. SIMULATION EVAULATION

To analyze the consumption of an energy-efficient streaming service with a delay-conscious caching scheme, the ns-2 [18] simulation environment was used. The simulation tool made it possible to change the characteristics of the video stream features as well the network conditions.

The importance of energy efficiency is more conspicuous due to limited battery capacity of mobile equipment. Therefore, a mobile client in a WLAN environment was simulated. The WLAN interface power consumptions was modeled with a bitrate dependent consumption model based on [11], where higher transmission rates cause a slight increase in power consumption. The reference download rate (Bw_{d0}) was 1 Mbps and the reference consumption (P_{ONREF}) that belongs to idle mode of the interface was set to 1.3 W. The power consumptions of different states within the heap up and offline playout cycle are shown in Table I.

TABLE I. CONSUMPTION VALUES

state	state length	power
ON	t_{ON}	$P_{ON} = P_{ONREF}\left[1 + \left(0.01\frac{Bw_d}{Bw_{d0}}\right)\right]$
OFF	t_{OFF}	0 W
start	1 s	0.7 W
stop	0.2 s	0.15 W

In the analyzed wireless network the downlink bandwidth was set to 2 Mbp. Due to the delay sensitiveness of multimedia services, UDP transport protocol was used to deliver the video stream packets and the control messages.

In the simulations the 45 seconds long *deadline* reference test sequence was used with 1374 CIF frames [19]. The video was coded with FFmpeg [20] MPEG-4 codec to produce different bitrate streams.

In the first measurement the impact of video bitrate on the consumption was analyzed. The video stream bitrate determines the amount of data that must be delivered to store a *d* length video for the offline period playout. The available download rate was constant, thus in the case of higher quality video streams (requiring higher video bitrate), more time will needed to deliver the data from the server to the client. The increased *ON* time will lead to higher consumption. In the measurement the coded video bitrate was 100 bps to 1.6 Mbps, while the download bandwidth was constantly 2 Mbps. The length of cached video stream was set to 2 s, 10 s, 18 s that also affects the length of the *ON* period and the cached data amount. The measured consumption values are presented in Figure 4.

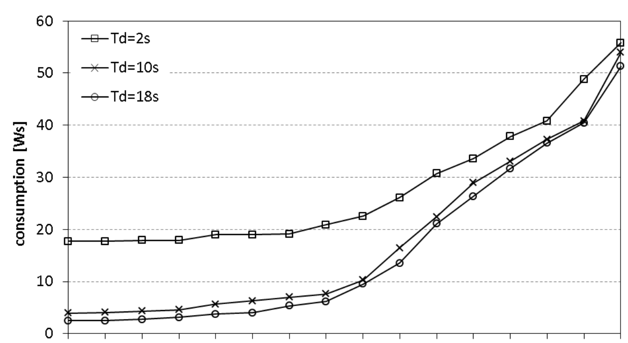

Figure 4. Consumption values of different bitrate video streams using 2s, 10s, 18s playout caching delay

The required energy to deliver a 45 s video was measured in Ws (Watt-seconds, 1 Joule=1 Ws). Note that the video bitrate scale in Figure 4. is linear from 20 kbps to 100 kbps and 200 kbps to 1.6 Mbps. According to (8), the video bitrate and consumption relation is linear. In this figure the caching delay impact is also introduced. The difference of consumptions between applying 10 s and 18 s caching delay was very low, only 1.62 Ws in average. Comparing 2 s and 10 s caching delays, the average difference was 9.9 Ws, however in case of higher video rates the difference becomes lower. According to the results, 10 s caching delay will reduce the consumption, but it is not worth to extend the length of the delay. because the improvement in consumption will be negligible.

In the following scenario, the duration of the offline period and the corresponding caching delay was analyzed by changing the delay (d) from 2 s to 20 s. The video length was 45 s, hence the number of heap up and offline playout cycles was 3 to 23. Every cycle contains a *start* and *stop* state with a constant $t_{start}+t_{stop}$=1.2 s long time period (Table I.), thus the caching delay (d) can not be less than the sum of *start* and *stop* times. In the *start* and *stop* states the consumption is not zero, however no traffic is delivered. For example, if the caching delay is d=2 s, the *OFF* state time will be only 0.8 s long. The efficiency of consumption reduction depends on the ratio of *start*, *stop* and *OFF* state time periods.

$$\eta = \frac{t_{ON}}{t_{start} + t_{stop} + t_{OFF}} \tag{10}$$

In Figure 5. the measured consumption values are shown for video bitrate and downlink bandwidth ratios (V/D).

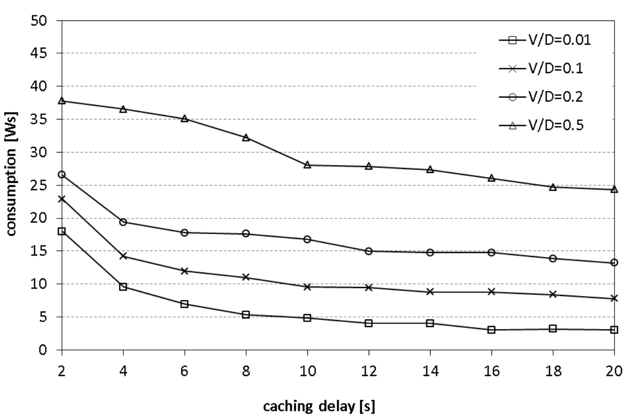

Figure 5. Consumption in function of caching delay for different video bitrate and download bandwidth ratios (V/D)

As it was derived in equation (9), the consumption is inversely proportional to caching delay. The simulation results confirmed the analytical calculations. The aim of this paper is to give guidelines for the offline playout delay determination in order to keep the delay as low as possible, but also reduce the consumed energy. Figure 5. shows that using very low delay, the consumption is high, but for higher caching delay the difference of consumption values are becoming lower. According to this characteristic, it is proposed the find the balance between the caching delay and the consumption. If we want to find the appropriate d parameter based on the proposed delay-conscious caching scheme, we have to find the point, where the derivative of the curve is -1. For the set of obtained simulation values we can analyze the difference of consumption values of close delay parameter values. For example in case of V/D=0.1 we can found that $W(d=4)-W(d=6)=2.22\,Ws$, so the caching delay should be set within the range of 4s and 6s. Based on the simulation results, for energy efficient video services, where the video bitrate and downlink bandwidth ratio is lower (e.g. V/D=0.01 to 0.2) a 4–6 s caching delay is recommended. For higher V/D values longer delay (e.g. d=10 s) is suggested.

The energy consumption efficiency depends on the video bitrate that is in strong correlation with the streamed video quality. In the following experiments we analyzed the achievable video quality for different offline playout delay parameter sets. First, the encoding of the 45 s long test sequence was performed with different bitrates. In the measurement we used the widely accepted peak signal–to–noise ratio (PSNR) video quality measurement algorithm. The objective quality values of the encoded MPEG-4 test sequences are introduced in Figure 6.

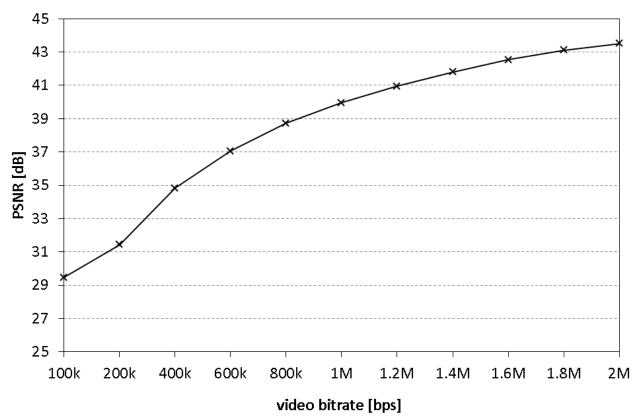

Figure 6. PSNR quality of different birate video streams

Using the coded video sequences, the achievable quality was analyzed where the consumption of the streaming was the same. By reducing the video bitrate, so the quality, the amount of stream data will be lower, therefore the *ON* state of the heap up and offline playout cycle can be decreased. If the consumption can remain at the same value, shorter offline periods can be allowed, hence lower caching delay can be set up. The obtained PSNR quality will be lower, but the delay will be reduced that is also a very important QoE parameter. Some applications are not enjoyable if the delay is too long.

The following Figure 7. shows the PSNR qualities that can be achieved in case of near same energy consumption. The achievable video quality was analyzed by changing the caching delay, but consuming the same amount of energy (10 W, 20 W or 30 W). According to the simulation results, by increasing the caching delay the video quality can be improved. Unfortunately, the video bitrate can not be decreased under a certain level, so if the allowed energy consumption is too low, the client can not be served width low caching delay. In the simulations 11 different video bitrate was used. It was shown that the same quality was achieved with different caching delay setups (e.g. 20 Ws energy consumption when the caching delay is 12 s to 20 s). Of course, it is worth to set the lowest one, because the quality remains the same by consuming the same energy.

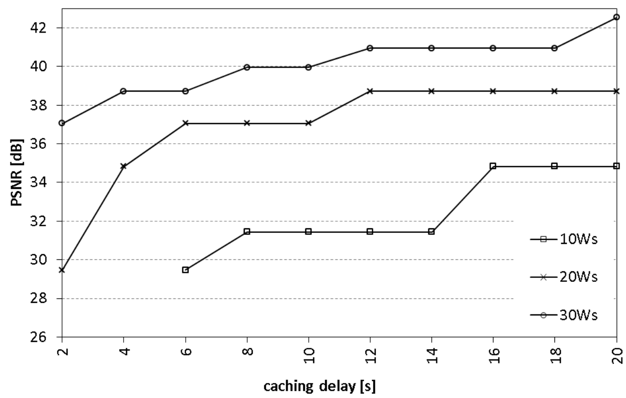

Figure 7. Achievable video quality

The simulation results show that the proposed delay-conscious caching method can help to find the appropriate parameter set of the energy efficient video streaming scheme. The experimented quality of the user is influenced by the coding parameters (e.g video bitrate) and the playout delay that can be minimized using the proposed solutions.

V. CONCLUSIONS

In this paper a mobile wireless scenario for multimedia streaming was considered and an energy efficient delay-conscious caching scheme was introduced. Buffering the video stream data to hide the link variations is a widely used solution, even for power consumption reduction purposes. The allowed cache size, thus the playout delay determines the offline state of the interface and the overall power consumption. The aim of this work was to find the balance between the power consumption and the acceptable playout delay caused by caching that has direct influence on the perceived quality. Based on analytical and simulation results the proposed delay-conscious caching model could serve as guideline for the offline playout delay determination in order to keep the delay as low as possible, but reduce the consumed energy as well.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n° 288502 (CONCERTO project).

REFERENCES

[1] Global internet phenomena report. Technical report, Sandvine, 2011.

[2] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017, white paper, February 6, 2013

[3] Powers, R.A., “Batteries for low power electronics,” Proceedings of the IEEE , vol.83, no.4, pp.687-693, Apr 1995

[4] Feng and J. Rexford, “A comparison of bandwidth smoothing techniques for the transmission of prerecorded compressed video”, Proc. of the IEEE INFOCOM, Apr.1997, pp. 58-66.

[5] G. Anastasi, M. Conti, E. Gregori, A. Passarella, and L. Pelusi, “An energy-efficient protocol for multimedia streaming in a mobile environment”, Journal of Pervasive Computing and Communications, 1(4):301--312, Dec 2005.

[6] M. Anand, E.B. Nightingale, and J. Flinn, “Self-Tuning Wireless Network Power Management”, Wireless Networks, 11(4), 2005.

[7] Phillips, C.; Singh, S., “An Empirical Activity Model for WLAN Users,” INFOCOM 2008. The 27th Conference on Computer Communications. IEEE , vol., no., pp.2065-2073, 13-18 April 2008

[8] Zeadally, S., Khan, S.U., Chilamkurti, N., “Energy Efficient Networking: past, present, and future”, 2012.Journal of Supercomputing, vol. 62, no. 3, pp. 1093-1118, 2012

[9] J-P. Ebert, S. Aier, G. Kofahl, A. Becker, B. Burns and A. Wolisz. Measurement and Simulation of the Energy Consumption of an WLAN Interface. In TKN Technical Reports Series, 2002.

[10] Sukyoung Lee, “Power Saving Scheme for Integrated WLAN and Cellular Networks”, ICCS 2006, Part II, LNCS 3992, pp. 985–991, 2006

[11] Ebert, J.P., Aier, S., Kofahl, G., Becker, A., Burns, B., and Wolisz, A., “Measurement and Simulation of the Energy Consumption of an WLAN Interface”. TKN Technical Reports Series, 2002

[12] Enrico Rantala, Arto Karppanen, Seppo Granlund, Pasi Sarolahti, “Modeling energy efficiency in wireless internet communication”, ACM SIGCOMM Workshop on Networking, Systems, and Applications for Mobile Handhelds, MobiHeld 2009, Barcelona, Spain, August 17, 2009

[13] Liu, J., and Zhong, L. “Micro power management of active 802.11 interfaces”, MobiSys (2008), pp. 146–159.

[14] Shivajit Mohapatra, Radu Cornea, Nikil Dutt, Alex Nicolau, and Nalini Venkatasubramanian, “Integrated power management for video streaming to mobile devices,” ACM Multimedia, Nov. 2003

[15] Xin Li, Mian Dong, Zhan Ma, and Felix C Fernandes , “GreenTube: Power Optimization for Mobile Video Streaming via Dynamic Cache Management “, Proceedings of the 20th ACM Multimedia (MM'12), Nara, Japan, October 2012.

[16] D. Bertozzi, L. Benini, and B. Ricco., “Power Aware Network Interface Management for Streaming Multimedia”, Wireless Communications and Networking Conference, 2002, volume 2, pages 926–930. IEEE, 2002.

[17] Mohapatra, S., Cornea, R., Dutt, N., Nicolau, A., and Venkatasubramanian, N., ”Integrated power management for video streaming to mobile handheld devices”, In Proc. of the 11th ACM international conference on Multimedia (2003), pp. 582–59.

[18] Ns-2 – Network Simulator, http:///www.isi.edu/nsnam/ns

[19] ITU (International Telecommunication Union), ftp://ftp3.itu.int/video-site/sequences/Deadline/

[20] FFmpeg, http://www.ffmpeg.org/