Panayiotis Kolios, Katerina Papadaki, Vasilis Friderikos

Energy efficient mobile video streaming using mobility

Article (Accepted version) (Refereed)

Original citation:
Kolios, Panayiotis, Papadaki, Katerina and Friderikos, Vasilis (2016) *Energy efficient mobile video streaming using mobility.* Computer Networks, 94. pp. 189-204. ISSN 1389-1286

DOI: 10.1016/j.comnet.2015.10.026

Energy Efficient Mobile Video Streaming Using Mobility

Panayiotis Kolios*, Katerina Papadaki†, Vasilis Friderikos‡

***Abstract*—Undeniably the support of data services over the wireless Internet is becoming increasingly challenging with the plethora of different characteristic requirements of each service type. Evidently, about half of the data traffic shifted across the Internet to date consists of multimedia content such as video clips or music files that necessitate stringent real-time constraints in playback and for which increasing volumes of data should be shifted with the introduction of higher quality content.**

This work recasts the problem of multimedia content delivery in the mobile Internet. We propose an optimization framework with the major tenet being that real-time playback constraints can be satisfied while at the same time enabling controlled delay tolerance in packet transmission by capitalizing on pre-fetching and data buffering. More specifically two strategies are proposed amenable for real time implementation that utilize the inherent delay tolerance of popular applications based on different flavours of HTTP streaming. The proposed mechanisms have the potential of achieving many-fold energy efficiency gains at no cost on the perceived user experience.

***Index Terms*—Mobile video, Energy efficiency, Cellular Networks, Store-carry and Forward**

I. Introduction

Cellular networking solutions are expected to provide a significant boost in meeting the requirements for higher datarates and increased traffic demand while maintaining and even reducing the communication energy expenditure [1]. Evidently, the interest has shifted towards the deployment of heterogeneous access technologies that will allow not only an increase in aggregated throughput but further alleviate the burden of heavy loaded cells by offloading traffic over Wireless Local Area Networks (WLANs), either through Wi-Fi access points or femtocells or co-existence of both. However it remains to be seen how the multiple WLANs will affect the network performance and user experience and how they will eventually contribute to the total energy consumption. Clearly, short range communication achieved by such small cell deployments would result in increased energy efficiency gains on the transmission link; however both the embodied energy (i.e., the manufacturing energy cost) and operational power consumption of the multiple WLAN circuitry will most certainly contribute substantially to the total energy budget. It is important to note also that such heterogenous topologies are only applicable in urban scenarios where a wired infrastructure is in place for WLAN installment. For the suburban and rural areas (with approximately 80% of the total BS nodes being deployed [2]) the absence of the wired network precludes the realization of such heterogenous deployments.

In this work we consider a complementary to previous proposals networking solution for cellular networks that is applicable for urban, suburban and rural scenarios. Broadly speaking these cellular network deployments are similar in the sense that provide ubiquitous connectivity with the difference being on the actual density of the nodes (i.e., BSs) in the network; which can be translated to available network capacity. For example as the density of nodes decrease, i.e., as we move from urban to suburban and rural network deployments, the delay required to achieve substantial savings in terms of energy consumption changes. Therefore, the set of mechanisms detailed hereafter can be equally applied to all different network settings which will translate to different benefits/gains depending on the specific network topology/deployment.

An interesting observation to be made at this point is that the dramatic increase in data usage demand is accredited to the full integration and support of Internet services over wireless networks. Together with the increasing processing capabilities of mobile terminals, customers have come to expect the same broadband experience in both fixed and mobile Internet access. Mobile network operators already report many-fold increase in data traffic over their networks (5000% in the case of AT&T [3]) while Cisco forecasts that this increase in data traffic demand will continue to grow exponentially over the several coming years [4]. While voice will still retain its popularity, more bandwidth hungry applications including file sharing, News feed updates and, especially, multimedia content consumption diminish the impact of voice services on the system performance. Also, while file sharing is currently contributing the dominating percentage in data traffic demand, it is anticipated that video content will surpass within the current year and will contribute to the excess of 60%-70% of the total traffic in the near future according to Cisco [4] and Qualcomm [5], respectively.

Interestingly, this shift from voice-like communication to mobile Internet widens significantly the elasticity in message delivery delays solely due to the broad range of user expectations for each individual service offered. While interactive applications such as voice communication and Web browsing set a stringent delay deadline constraint on the end-to-end path, file downloading, email access, News feeds updates and video and audio services to name a few, allow for a significant flexibility in delivery delay. For example, the downloading

*Department of Informatics, Centre for Telecommunications Research, King's College London, Strand, WC2R 2LS, London, England, e-mail:*panayiotis.kolios@kcl.ac.uk, ‡vasilis.friderikos@kcl.ac.uk

†Department of Management, London School of Economics, Houghton Street, London WC2A 2AE, London, England, e-mail: k.p.papadaki@lse.ac.uk, tel: +44 20 79556538 (corresponding author)

experience from requesting a movie file will not be affected severely or at all if total download time takes instead of 5 minutes, 6 minutes or more; especially if significant energy efficiency gains can be achieved as we will discuss in the sequel. Furthermore, while streaming a movie file (from, lets say, Youtube) it does not affect the user experience if the playback buffer contains 30 seconds of extra playback time or the total file size (which can be in the order of few minutes) at any one time. This is to say that as long as playback is not interrupted, the perceived user experience remains the same irrespective of the buffer size and thus the delay in fragment transmissions. Obviously, the same conditions hold for audio downloading/ streaming as well.

The advantage of the increase in delay tolerance is that the BS (and equally, the users) can defer transmission of packets for time instances in which the underlying channel conditions of the wireless link are anticipated to be in a more favourable state. Doing so can result to a number of improvements in communication as we have previously discussed in [6] with energy efficiency being the prominent example since reduced power would be needed to successfully transmit the buffered data packets. Figure 1 serves as an illustration of the multiplicity of transmit opportunities that arise from an increase in delay tolerance during packet delivery.

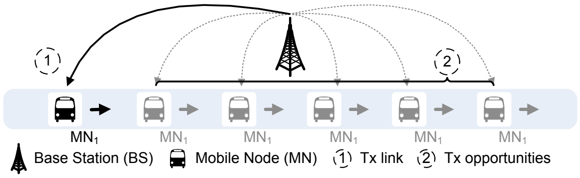


Fig. 1. Illustration of the proposed mechanical forwarding scheme in cellular networks.

The figure illustrates a single mobile terminal that moves along a stretch of road with replicas of the terminal showing its future positions. The particular road segment is covered by a cellular basestation that provides communication to the terminal in motion. For the sake of this example, it is assumed that the terminal requests to download a file (equivalently upload a file) as soon as it enters the cell coverage. Traditionally, the file exchange will commence immediately and will last for as long as it takes for the file to be delivered in a best effort manner. Evidently, this strategy is followed for simple file transfers or multimedia delivery services that have real-time constraints (as discussed in Sec. II below). Clearly, this strategy fails to take into consideration the inherent delay tolerance in content delivery. Instead, the terminal can postpone the request for a later time (and always under the delay deadline constraints) until it gets closer to the basestation as shown by the set of future transmit opportunities (set of links in 2) in the figure. By doing so, localized transmissions can be achieved that can greatly reduce the energy consumption in transmission. More importantly, it should be emphasized that this postponement can be achieved even in the case when real-time playback constraints are imposed as in the case of mobile video delivery. In this case, data exchange can commence immediately so as to allow for playback to begin while additional data transfers for buffering of content ahead of playback can be done only during time instances with favourable link gains. This is to say, that only few media fragments are delivered during bad channel conditions while larger volumes exchanged only during good conditions.

It should be emphasized here that the proposed scheme differs from drive-through Internet [7], [8]. The latter concept deals with content delivery when there is intermittent connectivity while here we assume that connectivity is perpetual. Instead what is proposed here is that informed decisions are made on content delivery based the channel gains anticipated in the future that arise due to terminal mobility and delivery delay. The proposed approach is also different from delay tolerant networking where end-to-end communication is achieved over time due to network topology restrictions [9]. The network topology considered here follows the cellular structure where mobile terminals are directly connected to the BS in a single hop. As before, in the proposed solution communication is deliberately delayed to improve on the system energy efficiency.

In our previous work [10], we termed this method of delivering delay tolerant traffic in cellular networks as Mechanical Forwarding (MF) due to the fact that the actual mobility of nodes is used within the cell. Even though the method of delivering delay tolerant traffic is based on the same principles as in paper [10], this paper deals explicitly with video whereas [10] considers generic delay tolerant traffic. As a result, we develop a model that optimizes traffic specifically for video. In our previous studies we considered the delivery of chunks of messages as one file with a single message delivery delay. Here on the other hand, we extend that concept and show how it can be applied to the delivery of multimedia content with real-time delivery constraints. More specifically we concentrate on video/ audio content delivery, where one of the constraints is for the user experience to remain uninterrupted. An initial idea of this concept was described in the magazine article [11], without the mathematical formulation, implementation of the model and numerical simulations.

In the following section (Section II) we review the current trends in video content delivery over wireless networks and introduce the terminology that is going to be used in the rest of the paper. The system model is detailed in Section III. MF schemes are derived in Section IV where the mathematical framework is detailed and the message forwarding decisions policies are derived. Further, the performance of the proposed MF paradigm is evaluated in Section V and the potential energy efficiency gains achieved over current implementations are illustrated. Section VI comments on related work found in the literature and finally the outcomes of this work are summarized in Section VII.

II. Video Content Delivery Methods

In this section we review the video content delivery methods that have evolved over the years and introduce the terminology to be used in the forthcoming sections. A more detailed description of the different technologies can be found in [12], [13] and [14].

To date, there exist two distinct content delivery methods; either through managed or unmanaged networks. Video delivered to subscribers over managed networks uses dedicated infrastructure (for example, cable TV) that supports multicast transmissions with stringent quality of service requirements for high quality viewing experience. Nevertheless, the inherently high monetary cost of deploying and maintaining such infrastructure makes it very unattractive to many content creators. Hence this work does not investigated managed networks but instead focuses on unmanaged networks. Unmanaged networks (also known as over-the-top (OTT) solutions) capitalize on the traditional Internet infrastructure to shift video files similar to any other datagram service. As such, OTT solutions have become increasingly popular to both amateur and professional content creators due to the flexibility and ease of developing multimedia services for the Internet. For example, Warner Bros in the USA offers online viewing options for its movies not only through its dedicated website but also through its Facebook page. Moreover, and due to the flexibility of OTT solutions, several different delivery methods have been made possible.

In the sequel all existing delivery methods are identified and described. These methods can be classified as follows:

1) Live content delivery
2) Video-on-Demand (VoD) (time shifted content delivery)
    a) Traditional Streaming (TS)
    b) Progressive download (PD)
    c) Adaptive streaming (AS)
3) Electronic sell-through (EST)

Live content delivery is supported through the many real-time protocols available on the Internet. In addition, time-shifted content delivery can be easily accessed through the myriad of available repositories such as the iPlayer and Radioplayer of BBC in the UK. Clearly, the advantage of this approach is that personalized schedules can be achieved to serve the individual needs. As a result video-on-demand (VOD) services have aroused as one of the most popular alternatives to live viewing for time-shifted content delivery. At the same time, electronic sell-through (EST) services have been developed to enable the purchase of content which can in turn be legally downloaded and owned by users for future playback. Once content is purchased, EST services employ either customary file transfer protocols or peer-to-peer links to deliver the multimedia to the user in a best-effort manner. On the contrary, live viewing and VOD posse real-time constraints on the delivery of multimedia files and as a result special considerations need to be made when delivering such OTT services. A large body of work has been looking at different ways of dealing with this real-time constraints and as a result three distinct methods have emerged, namely traditional streaming, progressive download and adaptive streaming; with an overview of each method provided in the sequel. Furthermore, table I summarises the characteristics and application of the different delivery methods.

TABLE I
VIDEO DELIVERY METHOD CHARACTERISTICS

Method	Delay Constraints	Example Service
Live	Real-time	Silverlight, VLC Player, Adobe Flash Player
VoD	Real-time playback and varying buffering	YouTube, Vimeo, Spotify
EST	Elastic	Apple iTunes, Amazon, Google Play

A. Traditional Streaming (TS)

Similar to other best-effort protocols available on the Internet, traditional streaming is a client-server model in which upon request, the server transmits portions of the multimedia content sequentially and playback commencing right after packet reception. The transmission rate always meets the media playback rate and thus minimal buffering of content takes place at the client. Hence, a thin terminal with marginal resource requirements is needed at the client side. However, since minimal buffering takes place, the connection between the client and the server is kept alive through the session. Evidently, with TS a relatively stable data rate is required to avoid gaps in the playback of multimedia content. In addition the media stream is discarded as soon as the content has been consumed at the client and thus no time-shifting is possible. To date, the approach is used mostly for live content delivery either from TV or radio broadcasters through one of the available OTT solutions such as Microsoft's Windows Media Services and Apple's QuickTime player.

With TS, either the client requests portions of the multimedia content based on its achieved data rate and buffer status or alternatively report to the server its achievable data rate and buffer status which then decides on the transmission intervals.

B. Progressive Download (PD)

Contrary to TS, progressive download heavily capitalizes on buffering to minimize playback stalls. Using PD, a media file is requested from the server which then delivers the file in a best-effort manner using in most cases HTTP over the TCP protocol. PD capitalizes on broadband connections to deliver enough portions of the video/audio file which are buffered on the client ahead of playback and downloading continuous until all the portions of the file have been delivered. Playback is decoupled from downloading with the former commencing as soon as enough content has been received. PD has become quite a popular multimedia delivery method employed by many content providers including YouTube (for entertainment), Vimeo (for better quality content) and Instagram (for short clips or slideshow).

Unlike TS, the received media file is stored at the client (in most cases in the cache memory of the browser) while the session lasts so that is can be replayed; avoiding in this way the need of future retransmissions. Also, the server maintains multiple copies of the media file in various qualities to support devices of different capabilities. Upon initial request of the media file, the server chooses the copy of the media file that matches the device capabilities and uses that file throughout the rest of the session. As with TS, this approach works well for the fixed Internet where the data rate does not change abruptly. For the mobile Internet however, the approach can result in playback stalls and re-buffering in locations with low coverage and especially for users on the movie for whom the wireless channel conditions change continuously. Moreover, since the downloading is decoupled from the playback, considerable resource can go to waste if the user decides not to watch part or the whole media file by the time the whole download completes.

C. Adaptive Streaming (AS)

Adaptive streaming is the most recent take-up to multimedia content delivery that aims to address the shortfalls of PD while encompassing all the good elements of both TS and PD. For example, AS communicates media files through HTTP over TCP similar to PD. It also employs buffering but limits the buffer size to a small portion of the video that is application controlled. As with PD, the client requests media files from the server (i.e., it is a pull-based protocol) based on the network conditions and the buffer status. On top, AS is advertised to provide a highly flexible, scalable and robust solutions by introducing two main novelties.

Firstly, instead of communicating large media portions (as is the case with PD), AS chops media files into fragments (meta-data about each media file fragment is contained within manifest files). Each fragment is time synchronized with every other fragment that exists of the same media file across different playback qualities (i.e., the different encoding stream). In turn, similar to TS, the client and the server communicate regularly to request/forward media fragments of the same or different encoding stream depending on the available data rate and the buffer status. In this way, AS ensures that the quality of experience is maximized while avoiding any playback freezing or re-buffering.

Evidently, smaller media fragments enable better request/forwarding interaction and improves the switching decisions to match viewing experience to the underlying network capabilities. For this reason, AS is considered by many as the best alternative to meet the growing user demands for video/ audio delivery and all major vendors have produced variants of this technology: Apple for example, has created the HTTP Live Streaming technology which is described within the IETF draft [15], Microsoft developed the Smooth Streaming architecture while Adobe's HTTP Dynamic Streaming and Cisco's Videoscape platforms are among the other alternatives.

With AS, a finite buffer is set to avoid waste of resources when playback is abandoned. Most vendors have set a buffer limit of 30 seconds to be maintained ahead of playback at all times in order to ensure that re-buffering is reduced.

D. Energy efficiency aspects of the aforementioned technologies

In this work we consider the energy efficiency aspects of all three variants in media streaming as described above, in addition to EST. More specifically, we consider all three major sources of energy consumption in media content delivery and storage as described below:

1) the transmit power consumption to successfully communicate fragments over the wireless interface
2) the circuit power consumption to receive and process the bits at the client site
3) the power consumption of the memory components to store media fragments ahead of playback time and for future replaying

We devise simple and practical message (i.e., fragment) forwarding strategies from adaptive streaming technologies, that take into account not only the playback experience but further the energy efficiency merits of content delivery strategies. It should be emphasized that the proposed forwarding strategies do not suggest the change of either AS or EST approaches. However, both AS and EST allow for dynamic variation in the delivered and buffered data volume and this flexibility is capitalized here to optimize both the delivery volume and the buffer size in order to achieve energy efficiency gains while ensuring uninterrupted content playback. Furthermore, we show that instead of simply fixing the buffered content duration to a fixed value[1], an optimized buffer size can result in substantial energy efficiency gains. Optimized buffering is also implemented for EST. Overall, two energy efficient (EE) variants of AS and EST are proposed in this work that optimize the buffering that needs to take place prior to multimedia playback, namely EE-AS and EE-EST. Finally we show that the proposed algorithms are computationally light to implement and have very small execution times.

III. System Model

Without loss of generality we consider the downlink of a media file to a single user within an LTE cell of radius D meters. Clearly the same method can be applied effortlessly to the uplink and for multiple users. Also, the model is agnostic of the cellular technology being used for Internet access. However we make explicit references to documentation of LTE (and beyond) systems and thus we simply assume that an LTE cell is considered. As mentioned above, a user could potentially download a file for a later playback (i.e., the EST case) or stream a video from a content provider (i.e., the case of TS, PD and AS).

Let v be the instantaneous velocity of the user. Note that location and direction of travel information of the users might be obtained via triangulation or from a GPS sensor. Also it could be the case that location information is already available to the users either because it is required for control of information or it has been acquired for location based services. The work in [16] and [17] detail the location information logging and reporting procedures in LTE. Based on the location of the user, the distance depended channel gains can be calculated; for an LTE system the recommendations in [18] are used. Additionally, shadowing could be used (especially if an urban

[1] As explained above, most vendors recommend a 30 sec buffer availability ahead of playback.

environment is considered) using the shadowing maps of the cell coverage area.

Let $h_t(d_t)$ represent the channel gain at a distance d_t of the user from the BS at time t. For a bandwidth B allocated to the user by the system and with channel codes approximating the Shannon capacity (as illustrated in [18]), the achievable datarate R can be calculated as follows:

$$R(d_t) = B\log_2\left(1+\frac{P_{Tx}h_t(d_t)}{N_0 B}\right) \tag{1}$$

where in the above expressions $h_t(d_t)$ is the channel gain, P_{Tx} is the transmit power and N_0 is the noise power spectral density.

Since AS technologies are shown to be the preferred method of multimedia content delivery, the work hereafter focuses on the derivation of AS variants that in addition to maintaining an uninterrupted playback experience, take a more vigilant approach to energy efficient transmissions. As such, our focus is on pull-based approaches where the terminal is allowed to control the rate by which fragment requests are made to the media server, without the BS interaction. The only prerequisite is that the terminal has an estimate of its downlink datarate at each location within its path. The authors in [19] conducted empirical measurement studies to illustrate that such location based estimates of the downlink datarate are not only valid but also have high correlation across different times of the day. Similarly we assume here that within the cell coverage area there are distinct regions D_m of downlink datarate $R(m)$, where $m \in \mathcal{M} = \{1,\dots,M\}$, which are defined depending on the distance from the terminal node (figure 2a).

We consider a media file of N fragments of total size F bits. Each fragment has a fixed duration so that all fragments across the different encoded media files are time synchronized. Therefore each fragment contributes to ς seconds of playback time. Let $k \in \mathcal{K} = \{1,\dots,K\}$ be the k^{th} media encoding stream with nominal encoding rate $Z(k)$. Clearly then, the fragment size with encoding rate $Z(k)$ is $F(k) = Z(k)\varsigma$. It follows that the transmission time of a single fragment of media encoding stream k in region m within the cell is $\psi(k,m) = \frac{F(k)}{R(m)}, \forall\ k \in \mathcal{K},\ m \in \mathcal{M}$ and the minimum transmission time in each region is $\tau(m) = \min_k(\psi(k,m))$. These encoding rates are usually of multiplicative nature [23], with $Z(k) = 2Z(k-1) = 4Z(k-2) = \dots$ and therefore the transmit duration for a fragment of the k^{th} media stream is an integer multiplicative $p(k) = 2^{k-1}$ to the minimum transmission time $\tau(m)$.

IV. Mathematical programming formulation

In the following a network flow formulation is introduced for the problem of finding the best time periods for a client to request fragments of a media file from all the available encoding rates. The objective is to find the minimum energy decision policies for communicating the media file to the user while achieving an acceptable viewing experience and eliminating freezing/ re-buffering of playback under the varying downlink rate conditions.

A. *Space-time-stream Network*

A mobile user is considered within the cell that travels with an instantaneous velocity v. Figure 2a shows an illustrative example with the mobile user shown at four consecutive locations within a single cell of $m = 4$ downlink datarate regions.

First, we construct a space-time network whereby the location of the user is captured every $\tau(m)$ consecutive units of time. The current position of the mobile node at time epoch t is represented by (MN,t), for $t = 0,1,\dots,T$. The space-time network is shown in figure 2b for one mobile user MN.

Further, we consider the existence of K possible encoding streams. Recall from section III that $\tau(m)$ is the time required to transmit a single fragment of the media file with the lowest encoding quality in region $m \in \mathcal{M}$. Therefore, if a fragment of double the encoding rate (when $k = 2$) needs to be transmitted, communication will occur over two consecutive time periods, and so on (we elaborate more on this in the sequel). In order to distinguish between different data encoding streams, K node replicas are created for every position of the mobile node at every time period. In this way the 3-dimensional *space-time-stream network* is created as shown in figure 2.

The purpose of this space-time-stream network is to capture the dynamics of the network topology as it evolves over time. On this network we create a network flow problem as follows: Specifically, we define a network $G = (V,L)$ of nodes V and links L with costs and capacities on each link and with demands and supplies at each node, where satisfying demands/supplies in G at minimum cost represents finding the minimum energy decision policies for communicating the media file to the user while achieving the acceptable viewing experience (i.e. eliminating freezing/re-buffering).

We initially consider the simple case of one mobile node MN (the model can be easily extended to many mobile nodes and we do that in the simulations in section V). The nodes of the network consists of the set V_p:

$$V_p = \{(MN,t,k) : t = 0,1,\dots,T; k = 1,\dots,K\}, \tag{2}$$

where (MN,t,k) represents the stage of mobile user MN being in time epoch t when it streams quality k. Thus the set of nodes of the space-time network V consists of nodes in V_p, the BS node and a sink node S: $V = \{BS,S\} \cup V_p$. Figure 2c shows an example of network G, where $K = 3$ encoding streams are considered.

Further, we define the links of G as $L = L_1 \cup L_2 \cup L_3$, which consists of three distinct subsets defined as follows:

L_1 The set of transmission links from the BS to the mobile node: $(BS,j) \in L_1$, where $j = (MN,t,k) \in V_p$ for $t = 1,\dots,T$ and $k = 1,\dots,K$.

L_2 Set of links that represent buffering of data at consecutive time instances. In this case no transmission occurs: $(i,j) \in L_2$, where $i = (MN,t,k) \in V_p$ and $j = (MN,t+1,k) \in V_p$ for all $t = 0,\dots,T-1$ and $k = 1,\dots,K$. These are the links that physically transfer data fragments using the mobility of the mobile node.

L_3 Set of links that connect the last set of replicated

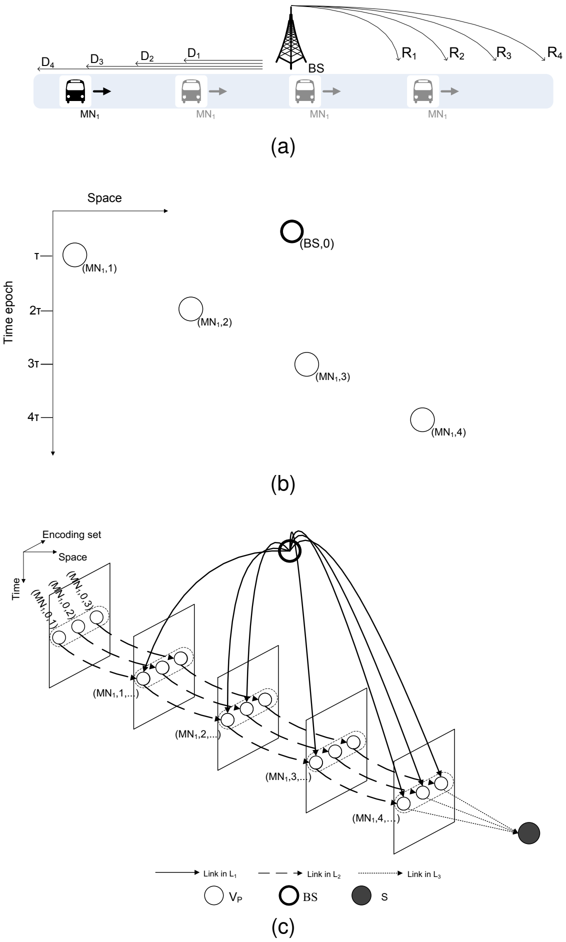


Fig. 2. The illustrations show the graphical representation of the terminal-based solution where the original cellular topology (fig. 2a) is converted into a static space-time network (fig. 2b) that is used to build the directed network graph in figure 2c.

terminal nodes at time T to the super sink node S: $(i, S) \in L_3$, where $i = (MN, T, k) \in V_p$ for all $k = 1, \dots, K$.

To aid our discussion we introduce the following notation: for $i = (MN, t', k') \in V_p$, we denote by $t(i)$ the time epoch t' and we denote by $k(i)$ the encoding stream k', i.e. $t(i) = t'$, $k(i) = k'$.

The example in figure 2c contains all such links to aid visualization. Links L_1 represent transmissions from the BS to the mobile node MN, and if the link is (BS, i), where $i = (MN, t, k)$, this means that the node MN will receive the data fragment sent from the BS via data encoding quality k at time t. L_2 links represent movement of data fragments via the mobility of MN, where these data fragments that have already been received at previous time periods. L_3 are virtual links.

As node replication is done every $\tau(m)$ units of time, links in L_1 can only forward a single media fragment while L_2 links connect the mobile node at consecutive intervals and thus conceptually unlimited data can be buffered in the mobile node's memory from one time period to the next. Therefore links in L_2 can carry ideally the total number of all media fragments.

Our graph G has links L_1 from the BS to all nodes $j \in V_p$ for all time periods except time period 0. However, not all of these L_1 links can have non-zero flow at the same time. We represent with each link $(BS, j) \in L_1$ a transmission that started at time one time period ahead at time $t(j) - 1$ in the case of $k = 1$; in the case of $k = 2$, it started $t(j) - 2$ time periods ahead and for general k, $t(j) - 2^{k-1}$ periods ahead. Suppose that there is flow on the link $(BS, j) \in L_1$ with $j = (MN, t, 3)$, since this transmission will take 4 time periods, it needs to start at time $t(j) - 4$ to end at time $t(j)$. The time that the transmission ends is time $t(j)$ and the time that the transmission begins depends on $k(j)$ and in fact it is $t(j) - 2^{k(j)-1}$. Then there cannot be any flow on the L_1 links (BS, i) with $i = (MN, t', k')$ for $t(j) - 4 \leq t' \leq t(j)$ and for all k'. This is because there is physically only one L_1 link and the rest are just virtual replicas, so only one can be used. The illustration in figure 2c shows an example of a set of three copies of the media stream ($k = 3$) that take 1, 2 and 4 (2^{k-1}) time periods to complete respectively. Therefore while a transmission takes place over $p(k)$ time epochs, no other L_1 link within that time interval should be active.

When $k = 1$ for a media fragment of size $F(1) = \tau(m)R(m)$ the transmission time is simply $\tau(m)$ units of time since we are in the m^{th} region. For a downlink rate $R(m)$ in location m within the cell that matches the fragment size $F(k)$, there is a valid link at every time period, i.e., $(BS, j) \in L_1$, $\forall j = (MN, t, 1) \in V_p$. On the other hand, if the downlink rate cannot match the fragment size (the case where a higher encoding media stream exists) the transmission events take longer than a single time period. For example, when $k = 2$ we have $F(2) = 2\tau(m)R(m)$ then transmissions of a single fragment takes $p(2) = 2$ time periods to complete, and for $k = 3$ it takes $p(k) = 2^{k-1} = 2^2 = 4$ periods to complete.

For $j = (MN, t, k) \in V_p$ we let the set $O(j)$ be the set of L_1 links before time $t(j)$ that cannot co-transmit with link $(BS, j) \in L_1$. Formally,

$$O(j) = O(MN, t, k) = \begin{cases} (BS, i) \in L_1, i = (MN, t', k') \\ \text{for } t - 2^{k-1} \leq t' \leq t, \ \text{ for all } k' \end{cases} \tag{3}$$

We will introduce constraints that prohibit the co-transmission of (BS, j) with any link in $O(j)$.

Given the above constraints the feasible directed paths from the BS to the sink S represent all feasible ways of serving the media file to the mobile node. Therefore we define $x_{ij} \in \mathbb{Z}$ to be the flow variable on link $(i, j) \in L$ which represents the number of data fragments that go through the link (i, j). Let N be the total number of data fragments that we want to sent to MN from the BS, then the capacity of link (i, j) is given by

$$u_{ij} = \begin{cases} 1 & for \ \ (i, j) \in L_1 \\ N & for \ \ (i, j) \in L_2 \cup L_3, \end{cases} \tag{4}$$

where note that L_1 links are designed to carry only one data fragment, whereas the other links can carry the whole load.

The aim of our problem is to carry N data fragments from the BS to the MN during the T time epoch via some encoding stream and this is represented as a path from BS to the sink node S on G. If we let $b(i)$ represent the supply of data fragments at node i (negative $b(i)$ indicates demand) that are waiting to be transmitted or physically transferred to other nodes that have a demand for data fragments. We must have $\sum_{i\in V} b(i) = 0$, which means supply must equal demand. Therefore, on G we put a supply of N fragments at node BS and a demand of N data fragments at node S. Any feasible flow x on G should satisfy these supply/demands along with the capacity and co-transmission constraints discussed above.

$$b(i) = \begin{cases} +N & \text{i = BS} \\ -N & \text{i = S} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Thus a feasible flow x on G is a path from BS to S satisfying some constraints and we want to pick the one that finds a good tradeoff between energy consumption and quality of media content playback. To do this we define costs on the arcs as follows:

For energy consumption, if it is an L_1 link we use the energy consumed during transmission, and if it is an L_2 link the storage power, and there is no cost for L_3 links. The energy consumed in transmission of a fragment of size $F(k)$ for link $(BS, j) \in L_1$ calculated as follows,

$$E(BS,j) = P_{\text{Cx}}\psi(k(j),m) + P_{\text{Tx}}(BS,j)\psi(k(j),m), \tag{6}$$

where P_{Cx} is the circuit power consumption and P_{Tx} is the transmit power over link $(BS, j) \in L_1$ defined as,

$$P_{\text{Tx}}(BS,j) = \frac{N_0 B}{h(BS,j)} 2^{R(m)/B} - 1 \tag{7}$$

with $h(BS, j)$ being the channel gain of link $(i, j, k) \in L_1$.

For $(i, j) \in L_2$, we must have by definition of L_2 links $t(j) = t(i) + 1$ and $k(i) = k)j)$, and the energy consumed to keep data in the terminal's memory for future playback is simply,

$$E(i,j) = P_{\text{Rn}}(k(i))\tau(m), \; (i,j) \in L_2. \tag{8}$$

where $P_{\text{Rn}}(k)$ is the retention power consumption of the memory units per fragment of encoding rate $Z(k)$.

Finally, given the existence of multiple media streams, there is a trade-off between minimizing the energy consumption and maximizing the quality of media content playback. This trade-off is expressed here as a weighted cost function between these two conflicting objectives,

$$c_{ij} = \begin{cases} E(i,j) - \lambda Z(k) & (i,j) \in L_1 \cup L_2 \\ 0 & (i,j) \in L_3 \end{cases} \tag{9}$$

with the weighting factor λ adjusted accordingly to give more emphasis to the media quality for higher values and more weight is given to the energy expenditure for lower λ values.

B. Linear Programming Formulation

Given the network $G = (V, L)$ with capacities u_{ij}, demand/supplies b_i and costs c_{ij} we would like to find the least costly path from BS to S that satisfies the capacity, demand/supply constraints along with the co-transmission constraints of L_1 links. This network flow problem can be formulated as a linear program (LP) as follows:

$$(\text{P1}) \min_{x} \sum_{(i,j)\in L} c_{ij} x_{ij} \;\; \text{s.t.} \tag{10}$$

$$\sum_{(BS,j)\in O(j)} x_{BS,j} \leq 1 \;\; j \in V_p \tag{11}$$

$$\sum_{j:t(j)=t'} \; \sum_{i:(i,j)\in L_1 \cup L_2} x_{ij} \geq \left\lceil \sum_{l=1}^{t'} \frac{\tau(m(l))}{\varsigma} \right\rceil \quad t' = 1, \ldots, T \tag{12}$$

$$\sum_{j:(i,j)\in L} x_{ij} - \sum_{r:(r,i)\in L} x_{ri} = b(i) \;\; i \in V \tag{13}$$

$$0 \leq x_{ij} \leq u_{ij}, \;\; x_{ij} \in \mathbb{Z} \tag{14}$$

The objective function in problem (P1) minimizes the weighted sum cost of communicating N fragments of data from the BS to the MN (or in G it finds the least costly feasible path from node BS to node S). Equations (13) are the flow conservation constrains that ensure that supply or demand at each node is satisfied and (14) are capacity constraints and non-negativity constraints that ensure that the flow is within capacity. The objective with constraints (13) and (14) constitute the typical minimum cost flow problem that can be found in [20]. Constraints (11) and (12) are specific to our optimization problem.

Constraint inequalities (11) ensure that at each time instance, no concurrent transmissions take place from the BS to the MN using different media encoding streams. Specifically, they ensure that at most one link from the set $O(j)$ (see equation (3)) of L_1 links can transmit for each $j \in V_p$.

Constraints (12) ensure that at each time instance there is enough downlink data for uninterrupted playback. The right hand side of the constraints is the lower bound needed for uninterrupted playback. The left hand side is the total flow of data fragments that go into the group of nodes (MN, t', k) in V_p of time epoch t'. The constraint ensures that at each time t' the number of data fragments that arrive at MN via transmission or via physical transfer is not less than the lower bound.

Since we require the downlink of fragments of media files, the integrality constraint on the flow variables is a natural assumption. Note that the constraints in (12) enforce the real-time viewing requirements for smooth playback. With this constraints, a minimum buffer is maintained for uninterrupted playback but no upper bound is placed on the playback buffer (the maximum upper bound is of course either the memory availability or the maximum fragment count). However, this property is the fundamental change we propose in this paper as opposed to the static buffer size that is required to be maintained in current implementations of AS technologies. The key benefit of dynamic playback buffer as explained before is to allow for the best choice of fragment requests to be made depending on current and future networking conditions observed/anticipated.

In the case of progressive download, the constraints in (12) are replaced by,

$$\sum_{j=1}^{p(k)N} x_{ijk} = N, \ (i \mapsto j)^k \in L_1 \tag{15}$$

where the decision on the selected stream is done at the session setup and the same stream is maintained to the session termination. On the other hand, for the traditional streaming service the maximum available media stream is selected to match the downlink datarate. Mathematically, the rate of fragment requests for a selected media stream k can be described as follows,

$$x_{ijk} = 1 \ \forall\, j \,\text{such that} \sum_{t=1}^{j} \tau(m) \geq \beta\varsigma, \ \beta = 1, \ldots, N \tag{16}$$

Finally in the case of EST the constraints in (12) are redundant and thus the optimization problem is identical to problem (P1) excluding those constraints in equation (12).

C. Algorithmic Implementation

Recall that in the derivation of the space-time network above, node replication was done under the assumption that a constant velocity is maintained over the time horizon T; that constant velocity was the instantaneous velocity of the mobile node (or the average velocity over the recent past). Even though such an assumption would be valid for highway scenarios or scenarios with predicable mobility patters, in some case would be highly unlikely that a mobile node would maintain its velocity over long time durations. To remedy this problem and to allow for corrective actions to be taken on out-dated decisions, the forwarding paths can be recomputed at frequent time intervals. By doing so, the robustness of the forwarding decisions is increased and the maximum energy gains are obtained.

As such, algorithm 1 presents the iterative re-computation of the forwarding paths. We let $\bar{T}$ be the end of the horizon of running the algorithm.

Algorithm 1 Iterative-MF scheme.

```
1: Compute estimates of location and velocity information.
2: Update supply/demand parameters.
3: Re-construct space time network using steps 1, 2.
4: Solve problem (P1) for time horizon $t \longrightarrow t + T$ where $t$
   is the current time.
5: Execute decisions for the first $\theta$ time periods.
6: Go to step 1 after $t + \theta$ units of time, unless $t + T > \bar{T}$
   and in this case stop.
```

After updating the location and velocity of the mobile node in step 1 of algorithm 1, a new space-time network is build based on this transit information and the remaining media content to be delivered (steps 2 and 3). The forwarding decisions are being re-computed from the optimal solution to the LP on this space-time-stream network (step 4). From those end-to-end forwarding paths, the decisions for the first θ time units are followed (step 5) before a new iteration takes place (step 6 in algorithm 1).

Each iteration of the algorithm solves the LP (P1). We know that the complexity of the LP is polynomial in the input size, which in itself depends on the size of the space time network, which is $|V_p| = T \times K$. If we let this polynomial be f then $f\left(|V_P|\right)$ is the complexity of solving the LP, and thus the complexity of algorithm 1 is $(\bar{T}T) f\left(|V_P|\right)$. Thus, when the cell is small (large), for a constant terminal velocity, the residence time is smaller and thus T (and $|V_p|$) will be smaller (bigger). When the terminal velocity is higher, for a constant cell size, T (and $|V_p|$) will be smaller. Thus, the computational complexity increases when the cell size increases or when the terminal velocity decreases.

The proposed algorithm is applicable to run at the client side with network support or at the network side assisted by the mobile terminal. This by no means excludes other possibilities such as for example that the algorithm runs on a cloud.

V. Numerical Investigations

We evaluate in this section the performance gains of the proposed MF enhancements to media streaming technologies (AS and EST) as introduced in the previous sections. We compare the proposed energy efficient adaptive streaming (EE-AS) and energy efficient electronic sell-through (EE-EST) variants of the respective media delivery methods (i.e., AS and EST) to both TS and PD.

A. Communication Model

A 10MHz LTE system is considered with the distance depended signal attenuation expressed as follows [21],

$$L(d) = 105.5 + 21\log_{10}(d) \tag{17}$$

where d is the transmission distance measured in kilometers and a noise figure of 5 dB is included. Other location specific parameters such as shadowing and antenna gain pattern can be considered for the channel gain h, however we neglect such parameters as we would like to show here the potential gains for an arbitrary network layout. Also note that this is a valid system scenario in highway deployments where shadowing effects are minor and BS antennas are directed on the roads. The cell radius is $D = 800$m and four rate regions are assumed with achieved downlink datarates $R(1) = 200$Kbps, $R(2) = 400$Kbps, $R(3) = 800$Kbps and $R(4) = 1600$Kbps within regions $D_1 = D$, $D_2 = D/2$, $D_3 = D/4$ and $D_4 = D/8$, respectively. The noise power spectral density is N_0 =-174dBm/Hz. The transmit power is P_{Tx} can then be calculated from (7). The terminal circuit power consumption is $P_{Cx} = 150$mW and the power consumed by the storage memory per chunk of density 512Mb is approximately 18.12mW for a Mobile DRAM unit [22].

B. Media Streaming requirements

We consider the download of a media file from a server which maintains three copies of the video at different encoding streams. The nominal encoding rates are $Z = 100, 200, 400$Kbps [23]. This video is fragmented into $N = 55$ chunks of $\varsigma = 2$sec duration.

C. Mobility Model

The mobile terminal is simply assumed to travel at an average velocity of $v = 15$m/s. While we make the assumption that the user initial location at call setup is uniformly distributed along the cell radius, we explicitly show the solutions of the different schemes at several locations within the cell to illustrate the preference to the different media delivery technologies in relation to the users location and velocity.

D. Numerical Results

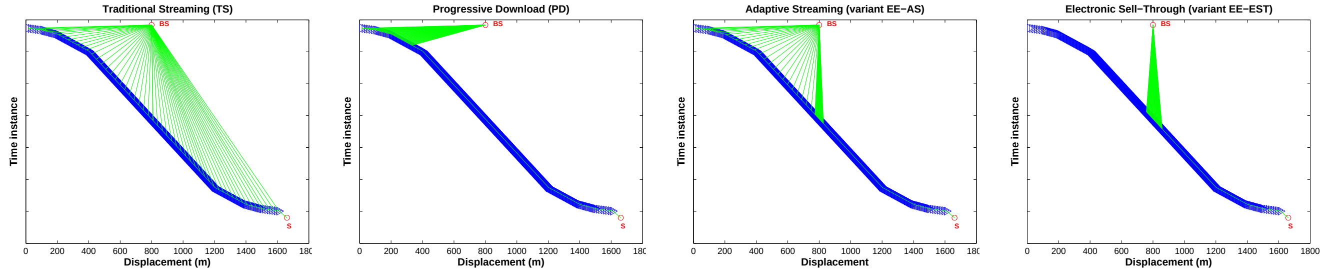
Figure 3 illustrates the solutions to TS, PD, EE-AS and EE-EST for a mobile node crossing a single cell. The transmission instances (indicated as green links) on the space-time network, provide an understanding on the operation of the different delivery schemes (results are shown for λ=0). With the selected encoding rate matching the maximum downlink datarate, the TS scheme forwards fragments at constant time intervals to maintain uninterrupted playback, as shown in figure 3a. PD on the other hand selects a copy of the media stream that satisfies the original downlink datarate achieved at call setup and maintains the connection until the complete file is downloaded, figure 3b. EE-AS requests adequate fragments to maintain smooth playback while approaching the cell center and steps up the fragment request rate at the area close to the BS to achieve minimum energy communication. While moving away from the BS, the mobile node already has the complete file in buffer for future playback and thus no communication occurs at a later stage as shown in figure 3c. Moreover, with no real-time constraints, EE-EST utilizes only the best locations within the cell for communication (i.e., in the locality of the BS in this case) and thus communication occurs at the closest points to the BS until the complete file is downloaded, figure 3d.

Figure 4 further shows the Energy Consumption Gains (ECG) of the proposed EE-AS and EE-EST schemes to TS and PD i.e., the ratio of energy consumption of TS and PD to the proposed schemes for different video quality performance (i.e., for varying λ values). It is evident that the proposed schemes provide substantial energy efficiency gains to both TS and PD. More specifically, in the case of EE-AS when energy consumption is the primary objective, the energy gains account for a factor greater than 2× and 3× to TS and PD, respectively. Moreover, for the EE-EST case, the gains are more than 70 and 100 times the achieved performance of TS and PD respectively. Note that the ECG values of EE-EST change abruptly from $\lambda = 0$ to $\lambda = 0.0001$. This simply indicates that the costs are more sensitive to small values of λ between that range. Clearly this is the case due to the fact that the transmit power consumption is significantly lower around the BS and contributes to smaller percentage to the total energy cost. However we keep the varying range of λ the same as for the other schemes for consistency.

Evidently these gains can be sacrificed for higher quality video delivery (increasing values of λ). In figure 5 the distribution of delivered fragments (of different quality) are shown. For the three nominal encoding rates considered in this work, the fragment quality is simply classified as 'Low Quality', 'Standard Quality' and 'High Quality'. As opposed to PD that downloads low quality fragments due to the initial bad conditions over the wireless link, TS can improve the performance as the channel gains increase by simply requesting higher order encoding streams. That is to say, whenever the channel quality is better (and hence the achievable data rate is higher) a better media quality fragment can be send to the client for playback (as shown in Fig. 5a). However the superior performance of EE-AS and EE-EST is apparent with significant portions of the highest quality fragments being delivered. These results also serve as a sensitivity analysis of the λ parameter on the forwarding strategies to be followed. As shown in the figure, properly selecting the λ parameter results to the delivery of different quality video fragments; with higher λ delivering substantially better video fragments especially for EE-AS and EE-EST.

Further, figure 6 illustrates the variations in the energy consumption when the initial location of the user at call setup is varied from the cell edge to the cell center. Fig. 6a depicts the absolute energy cost of transmitting the video content by all approaches, demonstrating the significant improvements in energy efficiency that can be achieved by deferring transmissions to locations closer to the BS (i.e., the solutions of EE-AS and EE-EST as shown in the figure). In that respect it also demonstrates the sensitivity of the proposed forwarding approach for video delivery requests conducted across the whole cell. Noticeably, the energy consumption drops as delivery requests are made closer to the cell center (i.e., at displacements close to 800m that approach the cell radius) while that energy increases when requests are made by users moving away from the cell center as shown in the figure. As expected, the gains achieved by EE-AS to TS (in figure 6b) increase when the user's initial location is closest to the BS. This is clearly the case as the TS will have to transmit over longer distances as the user moves away from the BS at future time instances. Interestingly, the gains also increase when compared to PD (5-fold reductions in the area around 600 to 400 meters away from the BS as shown in figure 6b). This is due to the fact that EE-AS needs to transmit less data over the bad channel gains improving its performance. The performance of EE-AS and PD is the same in the area near the BS. This is clearly the case because transmitting as soon as possible at the shortest distances away from the BS achieves the maximum gains.

For the case of EE-EST, figure 6c shows the energy efficiency gains compared to both TS and PD. Clearly the gains of EE-EST drop to meet the performance of PD as the initial user position is shifted towards the cell center. This is the case because PD improves its performance as the initial location of the user is within a closer vicinity to the BS which experiences better channel gains. Interestingly, the performance gains of EE-EST also drop when compared to TS. While the BS needs to transmit over longer distances as the user crosses the cell center and moves away from the cell in the case of TS, for the case of EST the BS sends over the best channel gains but the terminal has to store the complete file over longer time periods and ahead of playback time; reducing in that way the overall energy efficiency gains of EE-EST. Nevertheless, it is evident



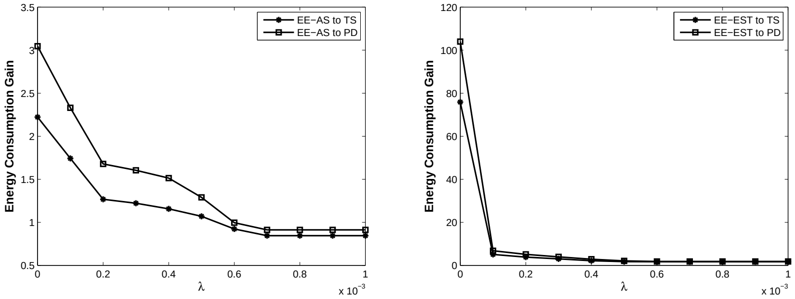
(a)

(b)

(c)

(d)

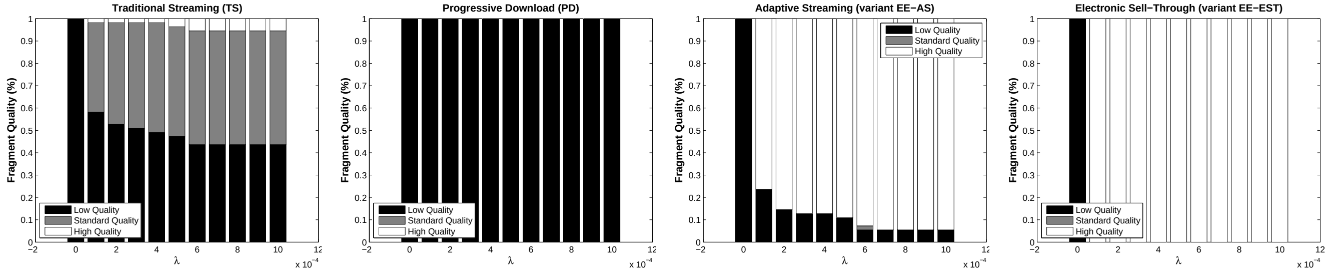
Fig. 3. Solutions to TS, PD, EE-AS and EE-EST are illustrated in figures 3a, 3b, 3c and 3d respectively.



(a)

(b)

Fig. 4. Energy efficiency gains of the proposed MF strategies (i.e., EE-AS and EE-EST) as compared to both TS and PD for a user crossing a cell.



(a)

(b)

(c)

(d)

Fig. 5. Distribution of received fragments from the different achieved quality levels for varying values of weighting parameter λ.

that the proposed MF schemes (either EE-AS or EE-EST) take into account all these conditions and sources of energy waste to achieve the maximum possible energy efficiencies for each case. As shown, the proposed MF schemes can in many cases provide many-fold energy efficiency gains in communicating multimedia streams over the wireless channel.

As alluded from the above, the proposed optimization problem can be executed online (as shown by Algorithm 1) to ensure that the best possible gains are achieved. As shown in Fig. 6b a $5\times$ increase in battery life (i.e., the tipping point in the curve that evaluates EE-AS against PD in the latter figure) should be achieved by EE-AS when compared to the same video content delivery that employs basic progressive download (currently employed by YouTube, for example). Much higher saving in battery life can be expected compared to Traditional Streaming as shown by the increasing performance gains in the same figure. In absolute terms, for a transmission starting at the cell edge and for λ set to 0, traditional streaming consumes 12 joules, progressive download consumes 7.8 joules and the proposed EE-AS consumes a mere 4 joules. The same rule applies for EE-EST and the current state-of-the-art PD and TS techniques.

It is worth pointing out that in this paper power consumption losses due to BS cooling, various intermediate-frequency and base-band operations (including for example analogue-to-digital conversion) as well as losses due to the inherent inefficiencies of the Power Amplifier (PA) have not been taken into account. As vividly pointed in [24] the difference between the actual power transmission (as considered in this paper) and the power supply from the grid (i.e., the overall power consumption of the BS) differ by approximately one order of magnitude. For example a BS with a power transmission of 20W has a total power consumption of approximately 1200 Watts.

Finally, a sensitivity analysis has been conducted on the forwarding strategies for EE-AS followed under varying traveling

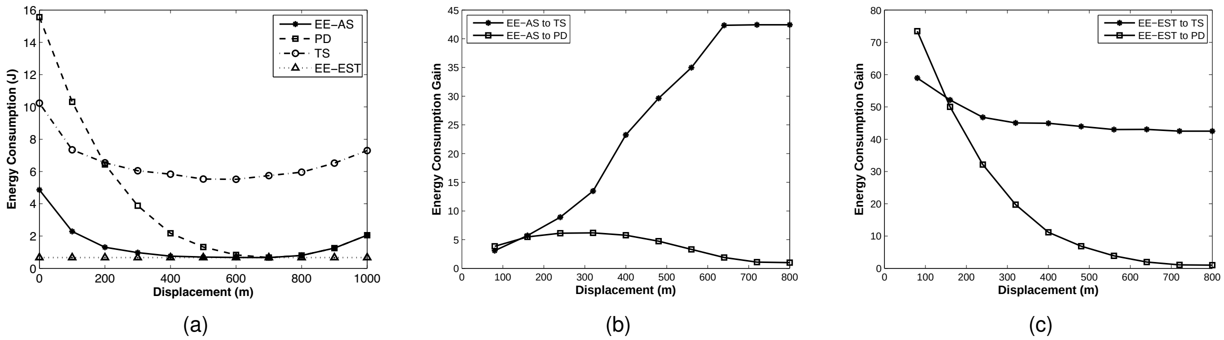


Fig. 6. Fig. 6a plots the absolute energy costs of EE-AS while Figs. 6b and 6c plot the energy efficiency gains of EE-AS and EE-EST when the initial location of the user being changed. At a displacement of 0 meters the user is at the cell edge while at a displacement of 800m, the user is at the cell center.

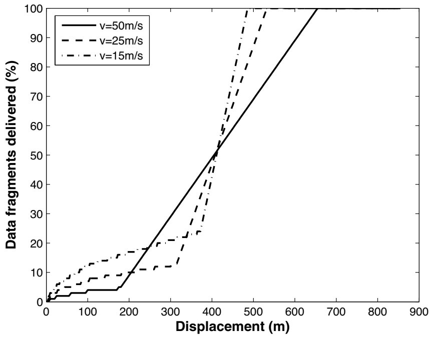


Fig. 7. Percentage of data fragments delivered for varying traveling speeds.

speeds. To do so, the weighting parameter was set to $\lambda = 0$ while all other parameters have remained as defined in Section V-A. Figure 7 illustrates the fragment delivery decisions made while a mobile node travels from the cell edge to the cell center at different travelling speeds. Clearly, depending on the travelling speed forwarding opportunities change and as a result the delivery decisions change accordingly. At lower speeds, more fragment deliveries are made while the terminal approaches the BS, as compared to higher speeds, to ensure that playback does not freeze. However there are ample opportunities at the cell center for bulk data transfer (since the node spends more time near the BS) and thus the gradient of the curve is bigger in that area. Evidently the opposite happens for larger speeds where less transmissions need to take place at the edge of the cell since the mobile node approaches the BS faster but then bulk fragment delivery commences sooner to make up for the sparse delivery opportunities closer to the cell center.

VI. Related Work

A. State-of-the-art solutions

Of the 66.88 million adults living in the UK [25], 93% personally use a mobile phone and currently approximately 58% use their mobile device to access the Internet up from 15% three years ago [26]. Even so, the additional burden of the data traffic generated by this percentage of subscribers was adequate in increasing system utilization high enough for all customers to experience degraded performance and frequent service outages. In response, mobile network operators were forced to place caps on data bundles to limit the impact of system overload. The same pattern is repeated in other European countries and around the world. Special focus has been placed in minimizing the energy consumption from the grid rather than the RF (transmitted) energy consumption [27] as well understanding the fundamental trade-offs between energy consumption and delivered quality of service to the end (mobile) user [28].

Moreover, this increase in system utilization has further exacerbated the power usage deficiencies of current cellular deployments and has stigmatized the telecommunication sector with a significant contribution to the total CO_2 emissions on the planet [29][30][31]. Furthermore, with the increase in data usage, mobile operators have seen a dramatic increase in the operational expenditure cost while users have suffered from limited mobile usability due to the short battery recharging periods. In the former case, significant research effort has been placed in manufacturing hardware components with reduced power requirements. More specifically, the Base Station (BS) components at the radio access network have received significant attention as they are responsible for approximately 50% of the total energy profile of a typical mobile operator (with another 30% attributed to the core and switching network) [29]. Further, different BS architectures are also being discussed by the research community; including distributed antenna systems with potentially remote radio heads [32]. Alternatively, software solutions are being considered with dynamic spectrum sharing [33][34] and sleep modes being the prominent two solutions to save energy under low utilization periods [35][36][37][38]. All these techniques are especially important if the trend of shrinking cell sizes is to continue in order to support the increased datarate requirements of today's and tomorrow's mobile communication needs. Clearly the same two paths (in terms of hardware and software advances) are also being considered in the case of mobile phones [39]. However significant research and development has already been conducted in the past for reducing the energy waste of the inherently battery constraint mobile devices and thus little additional gains are expected from such mechanisms.

B. Progress in mobile video delivery

While mobile video content delivery continuous to attract the attention of even more customers and content providers around the globe, it is only with the recent advances in media content delivery methods that such demanding services seem viable and cost efficient. A big step towards this path has been the introduction of full IP based architectures in cellular networks and the merge of the Internet and the mobile domains. Even so, there are still significant challenges to overcome in reducing costs to provide such multimedia services over the Internet and into mobile devices, ranging from providing a standard audio/video encoding format at the application layer, optimizing the message flows at the networking layer and providing flexibility at the physical layer in coping with the upcoming step increase in system utilization, [40][41][42]. Further, while HTTP Adaptive Streaming shows to be the preferred technology to be adopted for OTT media services, techniques that allow HTTP Progressive Download with rate adaptation are also being proposed for those legacy software on current mobile terminals [43]. Also significant work is done to improve efficiency of broadcasting protocols over wireless cellular systems, with the work in [44] detailing joint application-layer information, MAC layer priority handling and Reed-Solomon coding to handle bursts errors and increase the user perceived video quality.

Moreover, and even though significant effort is being place in realizing scalable and efficient content delivery techniques, only minor research has addressed the energy efficiency implications of the different delivery methods in use today. The work in [45] details empirical results from measurements on the energy consumption of video download/playback from different access technologies. The authors indicate that switching between wireless wide area networks to WLANs can possible achieve savings in energy in the case of PD. Similar to the latter work, the study in [46] details mobile phone measurements from different radio access technologies and propose a delay flexible scheduling algorithm to minimize operational energy consumption on mobile phones. Further, [48] details a testbed implementation for evaluating the energy efficiency of different access technologies for mobile content delivery, arguing that switching to the fastest available access network, provides the best energy efficiency gains. However, to our knowledge there has been no study until now to address the energy consumption performance of the various media delivery techniques currently employed and no effort has been made in optimizing the media delivery algorithms for energy efficiency. In this work we address the latter issues and show that intelligently selecting the media delivery fragment request rate can achieve significant energy efficiency gains. A number of key challenges for mobile video are detailed in [47].

The actual mobility of nodes has been considered in the past by several authors in providing performance improvements in wireless networks under different scenarios. In infrastructureless networks, mobility has been used extensively as the means to provide connectivity in sparse networks [9][49][50][51], increase capacity in ad-hoc networks [52] and reduce the energy consumption in communication [53][54][55]. In cellular networks, the work in [56] details how location and mobility information of moving terminals can be used to assist network protocols increase system performance. In a similar fashion to our work, the authors use mobility estimates from sensory inputs for bit rate adaptation over the wireless channel, access point association, neighborhood maintainable in mesh deployments and path selection in vehicular networks. In contrast to the aforemention studies however, we detail here the energy efficiency aspects of media content delivery (with and without real-time viewing requirements) in cellular networks. Moreover, the proposed MF schemes leverage the benefits of the new Adaptive Streaming technologies to provide high quality energy efficient content delivery over the air.

C. The proposed Mechanical Forwarding applicability

A brief word on the applicability of this scheme. Clearly the proposed MF strategy provides benefits when nodes are mobile. Some scenarios that this is the case include motorways, railway systems, urban commute either via automotives or pedestrian mobility. We argue that for the majority of the cases that the users are static they could be serviced by a WLAN that could provide energy savings as shown by previous studies. However, for the case of mobile users that are traditionally serviced by macrocells (for example to reduce handover signalling) the proposed scheme promises to provide significant benefits at no expense on the perceivable user experience.

Notably the proposed video streaming techniques can be integrated into a BS packet scheduling engine under the widely used assumption that some form of per-service differentiation is allowed. This is the case for example in LTE networks where in addition to QoS for different flows a large number of additional channel quality related information can be used. Having said that, within LTE standardization the scheduling algorithm used is not specified in the standard and it is BS (or Evolved Node B as it called in LTE nomenclature) vendor specific.

In parallel to the developments of video streaming techniques, the emergence of Information Content Networking (ICN) architectures will allow content to be cached topologically closer to the mobile user. As a result real time fast feedback based video player buffer management at the application layer would be possible. Extension of the proposed schemes to such emerging architectures would be an interesting future avenue of research.

In addition to the above, the proposed scheme could be considered as a natural fit to the emerging split-architectures for 5G wireless networks [57], [58], [59]. In split architectures signaling nodes are responsible for the coverage and are usually assumed to deliver low rate services, over long ranges; whereas pico-cells data nodes can be activated and deactivated depending on the traffic demand and it is designed for high rate and small ranges. This decoupling between control plane and data rate can be considered as a physical one in the general case and therefore depending on the mobility of the user and location of targeted pico-cells the macro controller can utilize the proposed technique in order to stream video content to the users.

VII. Conclusions

In this paper, energy efficient strategies for media content delivery over mobile cellular networks are being investigated. We consider simple to implement, practical solutions for the problem of delivering audio/video files to mobile terminals with and without real-time viewing constraints. By capitalizing on the actual mobility of nodes we devise energy efficient forwarding strategies to maximize energy gains at no expense on the perceivable user experience. These schemes, termed Mechanical Forwarding, can achieve many-fold savings in energy consumption and are of very low computational complexity. A mathematical programming problem has been derived which can be solved optimally at low computational cost. Translating the achieved energy savings into battery life extension of the mobile terminal is left as an future research. Finally, the proposed MF schemes can be considered as an energy saving feature of adaptive streaming systems and can be implemented today with minor modification to the current protocol implementations as discussed above.

Acknowledgments

The work reported in this paper has formed part of the Green Radio Core 5 Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, `www.mobilevce.com`. This research has been funded by EPSRC and by the Industrial Companies who are Members of Mobile VCE.

References

[1] L.M. Correia, et al, Challenges and enabling technologies for energy aware mobile radio networks, *IEEE Communications Magazine*, Vol. 48, No. 11, Nov. 2010, Page(s):66 - 72.

[2] Information supplied by a European Operator.

[3] Federacl Communication Commision, Public Notice No. 6, DA 09-2100, Sept. 2009, http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-09-2100A1.pdf.

[4] Cisco, Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20102015, Feb. 2011.

[5] Qualcomm, LTE Broadcast: Evolving and going beyond mobile, Aug. 2014, https://www.qualcomm.com/documents/lte-broadcast-evolving-and-going-beyond-mobile.

[6] P. Kolios, V. Friderikos and K. Papadaki, Future Wireless Mobile Networks, *IEEE Vehicular Technology Magazine*, Vol:6, No: 1, Mar. 2011, Page(s): 24 - 30.

[7] W.L. Tan, W.C. Lau, O. Yue and T.H. Hui, Analytical Models and Performance Evaluation of Drive-thru Internet Systems, *IEEE Journal on Selected Areas in Communications*, Vol: 29, No:1, Jan. 2011, Page(s): 207 - 222.

[8] I. Psaras and L. Mamatas, On Demand Connectivity Sharing: Queuing management and load balancing for User-Provided Networks, *Computer Networks*, Vol: 55, No: 2, Feb. 2011, Page(s):399 - 414.

[9] K. Fall and S. Farrell, DTN: An Architectural Retrospective, *IEEE Journal on Selected Areas in Communications*, Vol. 26, No. 5, Jun. 2008, Page(s):828 - 836.

[10] P. Kolios, V. Friderikos and K. Papadaki, Energy Efficient Relaying via Store-Carry and Forward within the Cell, *IEEE Transactions on Mobile Computing*, Nov. 2012.

[11] P. Kolios, V. Friderikos and K. Papadaki, Energy-aware mobile video transmission utilizing mobility, *IEEE Network*, Vol. 27, No. 2, Mar. 2013.

[12] A. Begen, T. Akgul and M. Baugher, Watching Video over the Web: Part 1: Streaming Protocols, *IEEE Internet Computing*, Vol: 15 , No: 2, Mar. 2011, Page(s): 54.

[13] A. Begen, T. Akgul and M. Baugher, Watching Video over the Web, Part II: Applications, Standardization and Open Issues, *IEEE Internet Computing*, Dec. 2010.

[14] A. Zambelli, IIS Smooth Streaming Technical Overview, *Microsoft Corporation*, Mar. 2009.

[15] R. Pantos and W. May, HTTP Live Streaming, *IETF*, Internet Draft draft-pantos-http-live-streaming-05, http://tools.ietf.org/html/draft-pantos-http-live-streaming-05.

[16] ETSI 3rd Generation Partnership Project, LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Stage 2 functional specification of user equipment positioning in E-UTRAN, *ETSI Technical Specification 136 305 V9.0.0 Release 9*, Oct. 2009.

[17] 3GPP, Technical Specification Group Radio Access Network Study on Minimization of drive-tests in Next Generation Networks, *3GPP*, TR 36.805 v9.0.0, Dec. 2009.

[18] 3GPP, LTE, Evolved Universal Terrestrial Radio Access (E-UTRA), Radio Frequency (RF) system scenarios, *3GPP TR 36.942 version 12.0.0 Release 12*, Oct. 2014, Page(s):17.

[19] J. Yao, S. Kanhere, M. Hassan, Improving QoS in High-speed Mobility Using Bandwidth Maps, *IEEE Transaction of Mobile Computing*, May 2011.

[20] R.K.Ahuja, T.L.Magnanti and J.B.Orlin, Network Flows: Theory, algorithms, and applications, *Prentice Hall*, 1993, Page(s):318.

[21] 3GPP, LTE, Evolved Universal Terrestrial Radio Access (E-UTRA), *Radio Frequency (RF) system scenarios, 3GPP TR 36.942 version 9.0.1 Release 9, Apr. 2010.*

[22] Micron, Mobile DRAM Power-Saving Features and Power Calculations, *Micron Technology, Inc.*, Technical Note TN-46-12.

[23] Apple, Best Practices for Creating and Deploying HTTP Live Streaming Media for the iPhone and iPad, Technical Note TN2224, http://developer.apple.com/library/ios/ #technotes/tn2224/_index.html.

[24] Gunther Auer, et al., How Much Energy is Needed to Run a Wireless Network?, *IEEE Wireless Communications*, Vol: 18, No: 5, Oct. 2011, Page(s):40 - 49.

[25] Office for National Statistics, Internet Access, http://www.statistics.gov.uk/cci/nugget.asp?id=8

[26] Ofcom, Facts & Figures, http://media.ofcom.org.uk/facts/

[27] Hauke Holtkamp, Gunther Auer, Samer Bazzi, Harald Haas, Minimizing Base Station Power Consumption, *IEEE Journal on Selected Areas in Communications, vol. 32, no. 2, February 2014.*

[28] M. Aykut Yigitela, Ozlem Durmaz Incelb, Cem Ersoy, QoS vs. energy: A traffic-aware topology management scheme for green heterogeneous networks, *Computer Networks, vol. 78, no. 26, pp. 130-139, 2014.*

[29] S. Vadgama, Trends in Green Wireless Access, *Fujitsu Sci. Tech. J.*, Vol: 45, No: 4, Oct. 2009, Page(s):404 - 408.

[30] P. Grant and S. Fletcher, Mobile basestations: Reducing energy, *E&T Engineering and Technology Magazine*, Vol: 6, No: 2, Feb. 2011.

[31] G. Cook and D. Pomerantz, Clicking Clean: A Guide to Building the Green Internet, *GreenPeace*, May 2015.

[32] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe and T. Thomas, LTE-advanced: Next-generation wireless broadband technology *IEEE Wireless Communications*, Vol: 17, No: 3, Jun. 2010, Page(s):10 - 22.

[33] G. Salami and R. Tafazolli, Inter-operator Dynamic Spectrum Sharing (Analysis, Cost and Implications), *International Journal of Computer Networks*, Vol: 2, No: 1, Mar. 2010, Pages:47 - 61.

[34] T.W. Ban, W. Choi and D.K. Sung, Capacity and energy efficiency of multi-user spectrum sharing systems with opportunistic scheduling, *IEEE Transactions on Wireless Communications*, Vol: 8, No: 6, Jun. 2009, Page(s):2836 - 2841.

[35] C.S. Bontu and E. Illidge, DRX Mechanism for Power Saving in LTE, *IEEE Communications Magazine*, Vol. 47, No. 6, Jun. 2009, Page(s):48 - 55.

[36] S. Zhouy, J. Gongy, Z. Yangy, Z. Niuy and P. Yang, Green Mobile Access Network with Dynamic Base Station Energy Saving, *ACM MobiCom'09*, Sept. 2009.

[37] M.A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, Optimal Energy Savings in Cellular Access Networks, *IEEE International Conference on Communications Workshops, 2009*, Jun. 2009, Page(s):1 - 5.

[38] M.A. Marsan, and M. Meo, Energy efficient wireless Internet access with cooperative cellular networks, *Computer Networks*, Vol: 55, No: 2, Feb. 2011, Page(s):386 - 398.

[39] K. Pentikousis, In Search of Energy-Efficient Mobile Networking, *IEEE Communications Magazine*, Vol. 48, No. 1, Jan. 2010, Page(s):95 - 103.

[40] K.J. Ma, R. Bartos, S. Bhatia and R. Nair, Mobile video delivery with HTTP, *IEEE Communications Magazine*, Vol: 49, No:4, Apr. 2011, Page(s):166 - 175.

[41] T. Stockhammer, Dynamic Adaptive Streaming over HTTP Design Principles and Standards, *Second W3C Web and TV Workshop*, Feb. 2011.

[42] S. Akhshabi, A.C. Begen and C. Dovrolis, An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP, *Second ACM conference on Multimedia systems*, Feb. 2011.

[43] K.J. Ma, M. Li, A. Huang and R. Bartos, Video Rate Adaptation in Mobile Devices via HTTP Progressive Download of Stitched Media Files, *IEEE Communications Letters*, Vol: 15, No:3, Mar. 2011, Page(s):320 - 322.

[44] K. Kang and W.J. Jeon, Differentiated Protection of Video Layers to Improve Perceived Quality, *IEEE Transactions on Mobile Computing*, to appear.

[45] Y. Xiao, R.S. Kalyanaraman and A. Yla-Jaaski, Energy Consumption of Mobile YouTube: Quantitative Measurement and Analysis, *International Conference on Next Generation Mobile Applications, Services, and Technologies*, Sept. 2008.

[46] N. Balasubramanian, A. Balasubramanian, A. Venkataramani, Energy consumption in mobile phones: a measurement study and implications for network applications, *ACM SIGCOMM Internet measurement conference*, Nov. 2009.

[47] Moo-Ryong Ra, Mobile Videos: Where are We Headed?, *IEEE Internet Computing, vol.19, no. 1, pp. 86-89, Jan.-Feb. 2015*, doi:10.1109/MIC.2015.9

[48] A.D. Zayas and P.M. Gmez, A testbed for energy profile characterization of IP services in smartphones over live networks, *Mobile Networks and Applications*, Vol: 15, No: 3, June 2010, Page(s):330 343.

[49] C. Chen and Z. Chen, Exploiting Contact Spatial Dependency for Opportunistic Message Forwarding, *IEEE Transactions on Mobile Computing*, Vol. 8, No. 10, Oct. 2009, Page(s):1397 - 1411.

[50] J. Zhao and G. Cao, VADD: Vehicle-Assisted Data Delivery in Vehicular Ad Hoc Networks, *IEEE International Conference on Computer Communications*, Apr. 2006, Page(s):1 - 12.

[51] W. Zhao, et. al, Capacity Enhancement using Throwboxes in DTNs, *IEEE International Conference on Mobile Ad hoc and Sensor Systems*, Oct. 2006, Page(s):31 - 40.

[52] M. Grossglauser and D.N.C. Tse, Mobility increases the capacity of ad hoc wireless networks, *IEEE/ACM Transactions on Networking*, Vol: 10, No: 4, Aug. 2002, Page(s):477 - 486.

[53] A. Venkateswaran, V. Sarangan, T. La Porta and R. Acharya, A Mobility-Prediction-Based Relay Deployment Framework for Conserving Power in MANETs, *IEEE Transactions on Mobile Computing*, Vol. 8, No. 6, Jun. 2009, Page(s):750 - 765.

[54] Y. Dong, WK. Hon, D. Yau and JC. Chin, Distance Reduction in Mobile Wireless Communication: Lower Bound Analysis and Practical Attainment, *IEEE Transactions on Mobile Computing*, Vol. 8, No. 2, Feb. 2009, Page(s):276 - 287.

[55] S. Chakraborty, Y. Dong, D. Yau and J. Lui, On the Effectiveness of Movement Prediction to Reduce Energy Consumption in Wireless Communication, *IEEE Transactions on Mobile Computing*, Vol. 5, No. 2, Feb. 2006, Page(s):157 - 169.

[56] L. Ravindranath, C. Newport, H. Balakrishnan and S. Madden, Improving Wireless Network Performance Using Sensor Hints, *NSDI*, Mar. 2011.

[57] A. Capone, et al., Rethinking cellular system architecture for breaking current energy efficiency limits, *Sustainable Internet and ICT for Sustainability*, Oct. 2012.

[58] H. Ishii, Y. Kishiyama and H. Takahashi, A novel architecture for LTE-B :C-plane/U-plane split and Phantom Cell concept,*IEEE Globecom Workshop*, Dec. 2012.

[59] P.K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, A. Benjebbour, Design considerations for a 5G network architecture, *IEEE Communications Magazine*, Vol. 52, No. 11, Nov. 2014, Page(s):65 - 75.