Cache-At-Relay: Energy-Efficient Content Placement for Next-Generation Wireless Relays

Melike Erol-Kantarci
Electrical and Computer Engineering Department
Clarkson University
mail: merolkkan@clarkson.edu

Abstract—Uploading and downloading content have recently become one of the major reasons for the growth of Internet traffic volume. With the increasing popularity of social networking tools and their video upload/download applications, as well as the connectivity enhancements in wireless networks, it has become a second nature for mobile users to access on-demand content on-the-go. Urban hot spots, usually implemented via wireless relays, answer the bandwidth need of those users. On the other hand, same popular contents are usually acquired by a large number of users at different times, and fetching those from the initial content source each and every time, make inefficient use of network resources. In-network caching provides a solution to this problem by bringing contents closer to the users. Although in-network caching has been previously studied from latency and transport energy minimization perspectives, energy-efficient schemes to prolong User Equipment (UE) lifetime have not been considered. To address this problem, we propose the Cache-At-Relay (CAR) scheme which utilizes wireless relays for in-network caching of popular contents with content access and caching energy minimization objectives. CAR consists of three Integer Linear Programming (ILP) models, namely Select Relay (SER), Place Content (PCONT) and Place Relay (PREL) which respectively solve content access energy minimization, joint minimization of content access and caching energy and joint minimization of content access energy and relay deployment cost problems. We have shown that PREL significantly minimizes the content access energy consumption of UEs while PCONT provides a compromise between the content access and the caching energy budgets of the network.

Index Terms—G, content caching, content placement, in-network caching, LTE, relay, content access power, wireless networksG, content caching, content placement, in-network caching, LTE, relay, content access power, wireless networks4

I. INTRODUCTION

Content production and consumption have become the primary use of today’s Internet. Data intensive applications such as on-demand video streaming have drastically increased the Internet traffic volume. In proportion to those applications, with the advances in mobile devices and growing network bandwidth, users have become eager for content than ever before. Traditional Content Delivery Networks (CDNs) or Peer-to-Peer (P2P) networks aim to improve the performance of the core network, while for mobile wireless users, the access network forms the performance bottleneck. The under-utilization of network resources by CDNs and P2P networks has been recently addressed in Content Centric Networking (CCN) where optimization of transport and storage energy as well as latency minimization have been the core issues studied [1]. CCN has introduced two important concepts; that are content naming and in-network caching. Although in-network caching has been known in the form of caching content in distributed locations such as servers or routers, the addressed problems have been limited to optimal cache management, cache placement at the core network hierarchy and forward caching. In CCN, the content travels towards the end-user depending on user demands [2].

In the evolution of wireless networks, the traditional macro cell architecture is expected to be replaced by an heterogeneous architecture with a mix of macro and smaller cells, that are also known as femtocells and picocells in LTE-Advanced (LTE-A) and 4G networks. In addition, wireless relays are primarily intended for extending cell edge coverage and improving spectrum [3] where layer-3 relays are standardized and defined as low power Base Stations (BS) in the recent LTE releases. Despite those primary functionalities, relays can be also used as urban hot spots [4]. Particularly with the increasing number of social network subscribers and their content intensive demands, urban hot spot deployments are gaining significant importance. Users upload and download video files which cause a huge bandwidth demand and a bottleneck when each and every user accesses the macro cell BS. Thus, there is an emerging need to cache contents closer to the hot spots rather than the content server or the BS that are spatially remote. Given the fact that storage costs are dropping [5], employing caches at wireless relays has become practically feasible.

In the literature, caching at wireless front end devices have been considered with a focus on minimizing latency and transport energy [6], [7]. Furthermore, periodic broadcasting or multicasting have been proposed to replace caching [8]. Broadcasting and multicasting are costly and they consume large amount of bandwidth. Although latency and transport energy are important, in wireless networks, one of the handicaps is the limited battery of hand-held devices. In [9], [10], we studied content placement considering the energy-efficiency of end user devices and presented our initial findings.

In this paper, we propose the Cache-At-Relay (CAR) scheme which utilizes wireless relays for in-network caching of popular contents with an objective of enhancing the energy-efficiency of UEs and relays. CAR consists of three Integer Linear Programming (ILP) models each of which addresses a significant and distinct case of caching at wireless relays. The first model, Select Relay (SER) minimizes the uplink power of the UEs when the relays are pre-deployed in known locations and the cached contents are known a priori. Note that, we aim minimizing content access energy to make UE more energy efficient and we use uplink energy minimization term to account for the energy spent by the UE to access the base station. The second ILP model, namely Place Content (PCONT) aims to minimize the uplink power of the UEs and the caching power of the relays when the relay locations are known. Thus, PCONT solves the content placement problem in pre-deployed relays with uplink and caching power minimization objective. The third model, Place Relay (PREL) aims to jointly minimize the number of relays and the uplink power of the UEs. PREL addresses the trade off between relay deployment cost and uplink energy budget.
The preliminary performance evaluations of SER and PREL has been conducted in [9]. In this paper, we have shown that PREL significantly minimizes the uplink energy consumed by UEs. However, PREL is not limited in caching power budget. PCONT provides a compromise between the uplink energy budget of the UEs and the caching energy budget of the relays. Moreover, cache utilization of PCONT outperforms SER. Meanwhile under run time constraints the performance of SER is favorable to PCONT and PREL.

The rest of the paper is organized as follows. Section II provides a brief summary of related works and outlines the fundamental differences of our work from the existing works in the literature. The system model and problem definition are provided in Section III and the proposed solutions are introduced in Section IV. Section V gives the performance evaluations of the proposed models, and finally Section VI concludes our paper.

II. RELATED WORK

Content caching has been widely studied in the context of Content Delivery Networks (CDN). The fundamental idea behind CDN is to cache popular contents in distributed locations that are likely to be closer to the users. CDN is generally deployed by the content publisher and the major goal of caching in CDN is to reduce latency, increase reliability and avoid hot spots. Caching contents in the distributed network elements is a more recent idea that has been studied in the course of Content-Centric Network (CCN) and it is known as in-network caching. In a CCN, data delivery is initiated by an interest packet generated by the user which is followed by a data packet response from a content provider, in this case this can be any node in the network which has the demanded content [11]. In [12], the authors have shown that caching at the Radio Access Network (RAN) rather than the core network improves the latency performance.

In recent years, energy-efficient content delivery problem in a CCN. They have proposed an ILP model and an online distributed algorithm to make caching decisions that minimize the caching and the downlink transport energy in a multicast scenario. In our paper, we are addressing content placement in a unicast scenario with an objective of uplink power minimization, and we utilize wireless relays for in-network caching rather than peer nodes.

In [7], the authors have analyzed the optimum ways of assigning files to small cell access points. The main objective of the study is to minimize the download time of videos. We tackle a similar problem from an energy-efficiency perspective. We consider caching at relays and solve the content placement problem with an objective of minimizing the uplink power that is spent for the interest packets and control messages. We additionally minimize relay deployment cost by selecting optimum locations for relays. Although relay placement can be studied from coverage perspective [13], [14], we consider that the relays are deployed to serve urban hot spots rather than extending cell coverage. Thus we assume the number of relays are adequate to provide coverage.

In [5], the authors have proposed an optimized caching mechanism where en-route routers cache the contents. Efficient content retrieval for wireless nodes has been explored through an optimization scheme whose objective is to minimize latency. The authors have also addressed content discovery which is done prior to content retrieval. Similar en-route caching approaches have been explored in a few more previous studies. In [15] the authors have utilized the idea that caching contents closer to the users would reduce the delay, hence they investigated caching at the gateway nodes, router nodes and intermediate nodes. A specific application of in-network caching in multicast on-demand video delivery has been presented in [16]. The authors have exploited relay stations to store the initial parts of a video to provide those to the nodes that join the multicast session after it has started. Users download content from the multicast server and from the relay at the same time where the latter is called as “patching stream”. However recent studies have claimed that multimedia contents are demanded repeatedly rather than being demanded simultaneously [6], making broadcast and multicast techniques inefficient. In this respect, a Markov decision process has been defined to determine when and what to cache at relays and how to transmit those in an energy efficient manner has been studied in [6]. The authors have proposed using adjustable transmit power at the relays. However, in practice, when a relay is serving as a hot spot, adjustable transmit power may become impractical. On the other hand, partially caching a video content in the BS and the relay and delivering it via Cooperative Multi Point (CoMP) technique has been studied in [17]. Furthermore, in [18] the authors have considered collaboration among wireless service providers for content delivery. An auction mechanism has been designed to allow caching in the presence of non-cooperative service providers.

III. SYSTEM MODEL AND PROBLEM DEFINITION

A. Network Model

Consider a connected network \( G = (V, E) \) where \( T \) is the set of UEs and \( T \subset V \), and \( A \) is the set of relays where \( A \subset V \). Here \( |T| = N_{UE} \) and \( |A| = N_{R} \). Each \( i \in A \) is a wireless relay node equipped with a cache of identical storage capacity of \( S_{R} \) and identical communication range of \( D_{R} \). Relays use \( M \) set of antennas; for Line of Sight (LOS) communication with the BS and for communicating with the UEs [19] and they are deployed at random grid points in the cell coverage area, serving as hot spots. The communication range of the BS is denoted by \( D_{BS} \). Each \( n \in T \) user is randomly deployed in the cell.

B. Content Request Model

In [20], the authors propose a popularity-based content delivery scheme in mobile CNDs. They use an average value of 10 views and a weighted moving average to compute popularity of a video. More recent research reveals that social networks impact the content requests in a different way than it is perceived in traditional networks. In [21], the authors findings indicate that social networks create super popular subsets of contents which include less pieces to be accessed more and some pieces to be completely ignored. The purpose of our study is to focus on energy-efficiency of caching at relays. We assume caches at relays have been accumulated with popular contents. Users demand contents randomly among the popular content set. The number of popular contents is limited to \( C \) where popularity of a video and cache management is based on traditional CDN approaches. Each video content is partitioned in chunks of size \( S_{E} \) bits, and cached either at the BS or the relays. UEs can download chunk \( k \) of a content either from the BS or a relay. The source for chunk \( k+1 \) is not necessarily the same with the source of chunk \( k \).

C. Energy Model

In an LTE network, UE transmit power for the physical uplink shared channel (PUSCH) transmissions is defined for a subframe as [22], [23]:

\[
P_{u}[dBm] = \min\{P_{max}, P_{D} + 10\log_{10}M_{u} + \gamma PL + \Delta_{mc3} + f(j)\}
\]

where \( P_{max} \) is the maximum allowed power configured by higher layers with an upper limit of 23 dBm, \( P_{D} \) is a cell specific parameter with 1dBm resolution, \( M_{u} \) is the number of assigned resource blocks as indicated in the uplink (UL) scheduling grant, \( PL \) is the downlink path loss estimate calculated by the UE in dB, \( \gamma \) is a cell specific path loss compensation factor that takes values from 0.4 to 1 in steps of 0.1. Finally, \( \Delta_{mc3} \) is signaled by the Radio Resource Control (RRC) unit and \( f(i) \) is the closed loop power correction function for subframe \( j \). Equation (1) can
be simplified for open loop PUSCH where $\Delta_m$, $\Delta_h$, and $f$ are omitted and $\gamma$ is assumed to be 1 to compensate for full path loss. Then for open loop PUSCH the uplink transmit power calculation becomes [24]:

$$P_u[dBm] = \min\{P_{max}, P_O + 10\log_{10} M_d + PL\}$$

(2)

Here, pathloss is the only distance related term where the free space PL can be written as:

$$PL[dBm] = 20\log_{10}(d) + 20\log_{10}(f) - 147.55[m]$$

(3)

where $d$ is the distance between the UE and the relay or the BS. For an LTE network, we can set $f=2.6GHz$ [25].

We have utilized the simplified model presented above while [26] covers detailed issues in energy consumption of UEs. It has been demonstrated in [27] that power consumption of a UE increases as a piece-wise linear function of the uplink transmission power. Measurements from various smart phones have been formulized in:

$$P_{UE}[mW] = \begin{cases} \sigma_L P_u + \omega_L, & \text{if } P_u \leq \theta \\ \sigma_H P_u + \omega_H, & \text{if } P_u > \theta \end{cases}$$

(4)

Here $\sigma_L$ and $\omega_L$ are the parameters of high power consumption regime with steep slope while $\sigma_H$ and $\omega_H$ are the parameters for low power consumption regime. We use $P_{UE}$ for uplink energy calculation in:

$$E_{UE}[mJ] = \frac{P_{UE} \cdot N_c \cdot S_k}{T}$$

(5)

where $N_c$ is the number of contents demanded from a relay, $S_k$ is the size of the content and $T$ is the throughput.

D. Problem Definition

Let $n \in T$ be a UE demanding a popular video $V_i$ that has been previously demanded and contents $\{k_1, k_2, k_3, \ldots, k_c\}$ has been cached in the BS and some of the relays where the set of relays containing the desired pieces ($R_c$) is $R_c \subseteq A$. Each $r_k \in R_c$ is equipped with M-antenna and is able to communicate with $n > 0$ UEs. Each $n$ generates an interest packet to demand content $k_c$ from the BS or a relay and sends it through the uplink channel. Each $n$ has a battery capacity of $B_n$ that is consumed upon uplink transmission, downlink reception, computation, display, etc. Considering the limited uplink energy budget, selecting relays that incur minimum uplink energy, placing content for incurring least uplink energy and caching energy, in addition to joint minimizing relay deployment cost with uplink energy consumption are the problems tackled in this paper. We formulate these in three subproblems. The list of frequently used notations is given in Table 1.

**Definition 1.** Given fixed relay deployment and a priori knowledge of content locations, determine the best relay to minimize the total uplink power of UEs.

**Definition 2.** Given fixed relay deployment, place contents in the cache of the relays to minimize the uplink power of the UEs and the caching power of the relays.

**Definition 3.** Given infinite caching power at relays, jointly minimize the number of relays and the uplink power of UEs.

We aim to solve these three problems under the proposed CAR suite of optimization models. Our assumptions are:

**Assumption 1.** Relay and content placement are independent from the coverage problem where the coverage is assumed to be dealt at the planning phase. We assume an urban scenario with densely deployed stationary relays. Thus, $N_c$ relays are assumed to provide adequate coverage and spectrum for the UEs and possible interference problems are omitted. Furthermore, nomadic relays are avoided due to handover problems.

**Assumption 2.** Popular contents are assumed to be acquired at least once before they are demanded by a UE. Transmissions are assumed to be unicast. Broadcast and multicast options are avoided due to the asynchronous access patterns of users.

**Assumption 3.** Each UE is allowed to download one chuck of a content from a single relay or BS. Multiple or parallel downloads are restricted. The rest of the content may be downloaded from another cache.

IV. CAR

We propose the Cache-At-Relay (CAR) scheme which utilizes wireless hot spot relays for in-network caching of popular contents for the purpose of improving the energy-efficiency of UEs and relays. Cache-At-Relay (CAR) scheme is composed of three ILP models each of which addresses a significant and distinct case. The first model, Select Relay (SER) is the simplest model and it aims to minimize the uplink power of the UEs when the relays are deployed in known locations and the cached contents are known a priori. The problem is reduced to selecting the closest relay or the BS depending on the uplink power budget. The second ILP model, namely Place Content (PCONT) aims to minimize uplink power of the UEs and caching power of the relays when the relay locations are known. Thus, PCONT solves the content placement problem in pre-deployed relays. The third model which is the Place Relay (PREL) aims to jointly minimize the relay deployment cost and the uplink power of the UEs. The details of the models are given in the following sections.

A. SER

SER is simply aiming to choose the closest relay to the UE. To be consistent with other models, it is formulated with an ILP model whose objective is to select the source of a content to minimize the uplink power of UE. In SER, location of relays and the contents that they cache is known a priori. SER selects whether to download the content from a relay or the BS. The objective function of the ILP model is given in eq. (6) where $\alpha_{xy,i}$ is a binary variable that is 1 if the $i^{th}$ UE is downloading $k^{th}$ content from the relay located at $(x, y)$ and 0 otherwise, and $\beta_{i,BS}$ is a binary variable that is 1 if the $i^{th}$ UE is downloading the $k^{th}$ content from the BS. We assume that the BS keeps a copy of all the contents that has been previously accessed and the demands are assumed to be within this set of most popular contents. $\rho_{xy,i}$ is a binary variable that is 1 if the $k^{th}$ content is cached at the relay at $(x, y)$. The uplink power consumed by UE $i$ to access the relay at $(x, y)$ for $k^{th}$ content is denoted by $P_{x,y,relay}^{r,relay}$ and the uplink power consumed to access the BS for $k^{th}$ content is denoted by $P_{x,y,BS}^{r,BS}$. $d_{xy,i}$ is the distance between UE $i$ and the relay at $(x, y)$ and $d_{i,BS}$ is the distance between UE $i$ and the BS.

$$\min \sum_i \sum_k \sum_x \sum_y P_{x,y,relay}^{r,relay} \alpha_{xy,i} + \sum_k \sum_x \sum_y P_{x,y,BS}^{r,BS} \beta_{i,BS}$$

(6)

The objective function in eq. (6) is solved subject to:

$$\alpha_{xy,i} d_{xy,i} \leq D_R, \quad \forall x, y, i, k$$

(7)

$$\alpha_{xy,i} = 0, \quad \forall d_{xy,i} > D_R$$

(8)

$$\alpha_{xy,i} \leq R_{xy} h^{xy,k}_i, \quad \forall x, y, i, k$$

(9)

$$\alpha_{xy,i} d_{xy,i} \leq d_{i,BS}, \quad \forall x, y, i, k$$

(10)

$$\beta_{i,BS} + \sum_x \sum_y \alpha_{xy,i} = H_t, \quad \forall x, y, i, k$$

(11)
\( \beta_{x,y,i}^{k} \) is a binary variable that is 1 if UE \( i \) is downloading content \( k \) from the relay at \( (x, y) \)

\( \delta_{x,y,i}^{k} \) is the binary value indicating whether UE \( i \) is receiving content \( k \) from the BS

\[ \varphi_{x,y,i}^{k} = \begin{cases} 1 & \text{if } \delta_{x,y,i}^{k} = 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ \eta_{x,y}^{k} \] is a binary variable that is 1 if there is a relay at \( (x, y) \) and UE \( i \) requests \( k \) from the relay at \( (x, y) \)

\[ \eta_{x,y} = \begin{cases} 1 & \text{if } \varphi_{x,y,i}^{k} = 1 \\ 0 & \text{otherwise} \end{cases} \]

Equations (7) and (8) ensure that UEs can download contents from the relays if they are in the cell coverage of those relays. Equation (9) guarantees that a UE can download a demanded content from a relay at \( (x, y) \) only if there is a relay at that location and that relay has the demanded content. A UE does not select to download content from the relay if the BS is closer than the relay which is formulated with the constraint in eq. (10). If a content is demanded by a UE than it is either downloaded from a relay or a BS. This is formulated in eq. (11). According to eq. (12), a content can be downloaded from the BS if the BS is closer to the UE than the relays that cover the UE and that have the content in their cache.

### B. PCONT

PCONT is an ILP model to optimally place contents in relays whose locations are known ahead. The objective of PCONT is formulated in eq. (13) where the uplink power to the relay and the BS as well as the caching power at the relay and at the BS is aimed to be minimized. Here, \( P_{cache,k} \) and \( P_{cache,k}^{BS} \) denote the caching power budget of the relay and the caching power consumed at the BS for content \( k \), respectively.

\[
\min \sum_{i} \sum_{k} \sum_{x} \sum_{y} P_{tr,relay}^{x,y,i,k} R_{x,y} d_{x,y,i}^{x,y,i} + \sum_{i} \sum_{k} \sum_{x} \sum_{y} P_{tr,BS}^{x,y,i,k} d_{x,y,i}^{x,y,i}
\]

Here \( \Delta_{x,y,i}^{k} \) is a binary variable that is 1 if the \( k \)th content is cached at the relay \( (x, y) \) and UE \( i \) downloads it from that particular relay.

\[ \Delta_{x,y,i}^{k} = \delta_{x,y,i}^{k} \eta_{x,y}^{k} \]

\[ \Delta_{x,y,i}^{k} \] is a binary variable that is 1 if the \( k \)th content is cached at the relay \( (x, y) \) and UE \( i \) requests it from the BS. We assume that if a content is not requested by any of the UE’s BS still keeps a copy of it. Thus \( \eta_{BS} = 1, \forall k \).

\[ \Delta_{x,y,i}^{k} = \eta_{BS} \beta_{x,y,i}^{k} \]

\( P_{cache,k} \) and \( P_{cache,k}^{BS} \) are calculated as [11]:

\[ P_{cache,k} = S_k P_{eff}^{k} \eta_{x,y}^{k} \]

and

\[ P_{cache,k}^{BS} = S_k P_{eff}^{k} \eta_{BS}^{k} \]

where \( P_{eff}^{k} \) and \( P_{eff}^{BS} \) are the power efficiency values of the storage employed at the relays and the BS, respectively, and \( S_k \) is the size of the chunk of a content. The objective function of PCONT is solved subject to:

\[
S_k \sum_{k} \eta_{x,y}^{k} \leq S_{tr}, \quad \forall x, y
\]

\[
\beta_{x,y,i}^{k} \leq \eta_{BS}^{k} H_{x,y}^{i}, \quad \forall x, y, k
\]

\[
\alpha_{x,y,i}^{k} \leq \eta_{x,y}^{k} R_{x,y}^{i}, \quad \forall x, y, i, k
\]

\[
\eta_{x,y}^{k} \leq R_{x,y}^{i}, \quad \forall x, y, k
\]

\[

\sum_{x} \sum_{y} \eta_{x,y}^{k} + \eta_{BS}^{k} \geq 1, \quad \forall x, y, k
\]

\[
 \sum_{x} \sum_{k} \sum_{y} P_{cache,k}^{x,y,i,k} + \sum_{x} \sum_{k} \sum_{y} P_{cache,k}^{BS}^{x,y,i,k}
\]

\[
 \Delta_{x,y,i}^{k} \leq \Delta_{x,y,i}^{k} \quad \forall x, y, i, k
\]

\[
 \Delta_{x,y,i}^{k} - \eta_{x,y}^{k} - \Delta_{x,y,i}^{k} \geq 1, \quad \forall x, y, i, k
\]

\[
 \Delta_{x,y,i}^{k} \leq \beta_{x,y,i}^{k}, \quad \forall x, y, i, k
\]

\[
 \Delta_{x,y,i}^{k} - \eta_{BS}^{k} - \beta_{x,y,i}^{k} \geq 1, \quad \forall x, y, i, k
\]
Equation (18) ensures that the cached contents do not exceed the reserved storage for caching at the relays. BS is assumed to be relaxed in storage. A content can be downloaded from a relay at \((x, y)\) only if there is a relay at \((x, y)\) and it has the content in its cache as given in equations (19) and (20), respectively. A content can be cached at the relay only if there is a relay at \((x, y)\) as denoted in eq. (21). If a content is cached at the relay \((x, y)\) then at least one UE should be downloading that content according to eq. (22). A content should be available at least in one relay as depicted in eq. (23). Furthermore, a requested content is ensured to be cached at the relay within communication range of UE \(i\) or the BS which is formulated in eq. (24). Eq. (25-30) are linearization constraints for the binary variables \(\Delta_{x,y}^i\) and \(\Delta_{x,y}^{i,BS}\). The rest of the ILP model is inherited from SER with the equations (7-11).

### C. PREL

PREL jointly minimizes the relay deployment cost, i.e. number of relays, and uplink power. Uplink power reduces when relays are closer to UEs therefore intuitively placing more relays with cached contents will reduce uplink power. However, deploying relays is costly therefore minimizing the number of relays is desired because it reduces the Capital Expenditures (CAPEX) of operators. PREL aims to address this trade off by adopting a normalized weight factor, \(\rho\). PREL assumes that each relay keeps all of the contents, hence content placement is not an issue and BS does not hold copies of contents. The objective function of PREL is given in:

\[
\min \quad \rho \sum_{x,y} R_{xy} p_{\text{off}} + (1 - \rho) \sum_i \sum_x \sum_y P_{tr,relay} \psi_{xy}^{ik}
\]

\[
(31)
\]

where \(\psi_{xy}^{ik}\) is a binary variable that is 1 if there is a relay at \((x, y)\) and UE \(i\) is receiving \(k_{th}\) content from the relay at \((x, y)\). \(p_{\text{off}}\) is a constant to scale the second term.

\[
\psi_{xy}^{ik} = R_{xy} \alpha_{xy,i}^k
\]

\[
(32)
\]

The objective function of PREL is solved subject to the following constraint set.

\[
\alpha_{xy,i}^k \leq R_{xy} H_i^k, \quad \forall x, y, i, k
\]

\[
(33)
\]

\[
\sum_x \sum_y \alpha_{xy,i}^k = H_i^k, \quad \forall i, k
\]

\[
(34)
\]

\[
\sum_i \sum_k \alpha_{xy,i}^k \geq R_{xy}, \quad \forall x, y
\]

\[
(35)
\]

\[
\psi_{xy}^{ik} \leq R_{xy}, \quad \forall x, y, i, k
\]

\[
(36)
\]

\[
\psi_{xy}^{ik} \leq \alpha_{xy,i}^k, \quad \forall x, y, i, k
\]

\[
(37)
\]

\[
\psi_{xy}^{ik} - R_{xy} - \alpha_{xy,i}^k \geq -1, \quad \forall x, y, i, k
\]

\[
(38)
\]

Equation (33) ensures that a UE can download content from \((x, y)\) if there is a relay at \((x, y)\). Each UE downloads content from one and only one relay and if there is a relay at \((x, y)\) at least one UE downloads at least one content from that relay. These are formulated in eq. (34) and eq. (35), respectively. The latter ensures there are no redundant relays. Equations (36-38) handles the linearization of the binary variable \(\psi_{xy}^{ik}\). Finally, PREL inherits the rest of the constraints from the equations (7-11) of SER.

### V. PERFORMANCE EVALUATION

We consider a single cell area with 1km radius. Relays are randomly deployed at grid points in the terrain. The range of a relay is set to 200m, based on the experiments conducted in [28, 29] where the maximum achievable range for an outdoor LTE-A relay is found to be 300m for similar number of UEs considered in [9] and in this paper. We assume the number of active users who are interested in content downloading varies from 20 to 100. Based on the statistical data from [30] and [31], an average of 300 active users per cell in urban scenarios is reasonable where we assume only 20-100 UEs are interested in content downloading. UEs are randomly deployed within the cell boundaries. For SER and PCONT the number of relays vary from 10 to 30. For PREL, the number of relays are determined by the optimization. We assume \(ST_k = 400MB\) for each relay. We assume the cache employed at the relays is Dynamic Random Access Memory (DRAM) whose power efficiency is given by \(p_{\text{off}} = 2.50 \times 10^{-3}[mWatt/bit]\). On the other hand, BS is assumed to employ a Reduced latency DRAM whose power efficiency is \(p_{\text{off}} = 3.75 \times 10^{-3}[mWatt/bit]\) [11]. Each content is assumed to be 8MB, based on the measurements from Youtube revealing an average video size of 8.4MB [32]. The UEs are assumed to be stationary during the optimization epoch as well as the video download. This assumption is valid for users connected to a small cell base station within a house or building. For mobile users, that potentially handover from one cell to another, we assume they at least stay in the same cell during the optimization epoch. For mobile users, the optimization schemes need to run periodically. We assume the maximum number of contents demanded by each user varies between 10 to 40. Contents correspond to the popular contents cached at the relay. We use CPLEX to solve the ILP formulations of the CAR scheme. The results are averaged over 10 runs.

In our results, we calculate the uplink energy assuming an uplink data rate of 8Mbps and considering contents of size 8MB being downloaded either from the relay or from the BS. In this case, uplink power is translated into energy spent using eq. (4). We set the power threshold \(\theta = 10dBm\), \(\sigma_i = 4, \sigma_{UL} = 1.2, \sigma_H = 61, \omega_H = 0.52\) which are the measured values for a Samsung GT-B 3730 2600MHz band UE and we set \(P_0 = -75dBm\) [27].

Our performance metrics are uplink energy, caching energy and cache utilization. For PREL, varying values of the weight factor is also evaluated. For those parameters, we show the number of relays and the uplink energy per relay. We also provide a run time analysis to demonstrate the complexity of our schemes.

In Fig. 1, we compare the total uplink energy of SER, PCONT and PREL under varying number of UEs. We assume the number of relays is set to 30 for SER and PCONT and \(\rho = 0.5\) for PREL and the number of contents is set to 10. As the number of UEs increase the uplink energy increases as expected since aggregated demand of users increases proportionally. However the rate of increase as well as the actual value of uplink energy consumption is the least for PREL. Thus, PREL improves the uplink energy overhead for the user terminals significantly. On the other hand, caching energy spent at the relay nodes is given in Figure 2 where the performance of the models is reversed. Note that, we compute caching energy only for the duration of the download however contents might be cached for a longer time increasing caching energy. PREL incurs the highest caching energy per each second the content is stored while SER incurs the least caching energy. The reason for this behavior is SER requires less number of contents to be cached at relays. UEs download content from the BS because the locations of relays are not optimized. For PREL, content placement is not included in the model, thus each relay is assumed to keep all of the contents, resulting in high caching energy expenditure. Note that the absolute values of uplink energy values are higher than caching energy values, in addition uplink power is on UEs energy budget which is expected to be far more restricted than the relay’s energy budget. In Table
II, we present the results for the baseline case where there are no relays. The uplink energy is significantly higher that any of the three CAR schemes. For better visualization we do not show the baseline results on Fig. 1. Caching energy is computed as $2.04 \times 10^{-4}$ mJ for each second contents are stored. This is computed considering all contents are stored at the BS since there are no relays in baseline scenario. For most cases, this is higher than the CAR schemes as well since storage efficiency of the DRAM at the BS incurs more energy consumption.

In Fig. 3, we present the total uplink energy for varying number of relays for SER and PCONT. We assume the number of UEs is set to 60 and the number of contents is 10. PREL is not included because the number of relays is the outcome of the ILP model. As seen in Figure 3, total uplink power of SER and PCONT drops for increasing number of relays. As the number of relays increase, they are placed at locations closer to UEs, hence the uplink power is reduced. As a consequence, the total caching energy increases due to more contents being acquired from relays instead of BS. The total caching energy is given in Figure 4. Again comparing these two plots, PCONT favors uplink energy while SER favors caching energy. However, the contribution of uplink energy is more significant which is in the order of tens of Joules, therefore in total PCONT incurs less energy expenditure than SER. Meanwhile, the same discussion of UEs’ energy budget versus relays’ energy budget as above holds in these set of results.

Figure 5 presents the cache utilization for SER and PCONT. Storage of relays is set to 300MB to avoid storage imposed limits. For 60 UEs and 10 contents, the utilization of caches in PCONT is significantly higher. With the increasing number of relays cache utilization of PCONT drops because contents are optimally distributed to relays. On the other hand, SER does not experience a significant change in cache utilization as the number of relays increase because contents are distributed uniformly among relays.

In Figure 6 and 7, we present the total uplink energy and total caching energy for SER, PCONT and PREL for varying number of contents. We assume the number of UEs is 60 and the number of relays is 20 for SER and PCONT while $\rho = 0.5$ for PREL. In terms of uplink energy, PREL outperforms SER and PCONT, it incurs less than one fourth of energy consumption of SER and less than half of the uplink energy required for PCONT. Meanwhile, the caching power requirements for PREL and PCONT is higher than SER. This is because as the content and relay placement is optimized, more contents are retrieved from the relays instead of the BS. Note that, increase in the number of contents, drastically increases the uplink and caching energy in comparison to the previous results. This is natural since
TABLE II

<table>
<thead>
<tr>
<th>UPLINK ENERGY TO BS WHEN THERE ARE NO RELAYS</th>
<th>N=20</th>
<th>N=40</th>
<th>N=60</th>
<th>N=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total uplink energy to BS (10^{-4} mJ)</td>
<td>17.23</td>
<td>32.86</td>
<td>44.84</td>
<td>74.47</td>
</tr>
</tbody>
</table>

more contents will be cached and downloaded when the number of contents is high.

For PREL, we evaluate the impact of $\rho$ on results. We present the uplink energy per relay for two representative weight factors, i.e. 0.5 and 0.8 in Figure 8. Uplink energy per relay is observed to be higher for $\rho = 0.8$ in comparison to the case when $\rho = 0.5$. Higher $\rho$ values imply that PREL aggressively reduces the number of relays and uplink energy is compromised. Thus, the tradeoff between uplink energy and relays gets biased toward relay minimization. This is also verified by the results presented in Figure 9 which shows the number of relays for $\rho = 0.5$ and $\rho = 0.8$. When $\rho = 0.8$, less relays can be employed to serve the same number of UEs. Caching energy per relay is not impacted by $\rho$, since all the relays cache all of the contents.

In Table 3, we present the total uplink energy, total caching energy and the number of relays for $\rho = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ for PREL. We set the number of UEs to 60 and the number of contents to 30. As seen in the table, as the weight increases, the model tries to reduce the number of relays more aggressively. As a result, the total uplink energy increases and the number of relays reduces. Meanwhile total caching energy also reduces since this is proportional to number of relays.

In Figure 10, we plot the runtime of SER, PCONT and PREL executed on a Core2 2.2GHz Intel processor and 4GB of memory. SER is the simplest model and its runtime is significantly lower than the others. PCONT and PREL have similar run times, however even for high number of UEs their execution time is less than a minute which shows that our models have reasonable complexity.

In summary, PREL significantly minimizes the uplink energy incurred by UEs while it has the highest caching energy consumed by relays. The weight factor impacts the aggressiveness of PREL where higher weight factor values indicate minimization of relays are given more emphasis than minimization of uplink energy. PCONT provides a compromise between the uplink energy budget of the UEs and the caching energy budget of the relays.
TABLE III
PERFORMANCE OF PREL UNDER VARYING $\rho$ VALUES.

<table>
<thead>
<tr>
<th>Weight Factor ($\omega$)</th>
<th>Total Uplink Energy [J]</th>
<th>Total Caching Energy [J]</th>
<th>Number of Relays</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>20.87</td>
<td>82.52</td>
<td>20.80</td>
</tr>
<tr>
<td>0.7</td>
<td>12.13</td>
<td>79.68</td>
<td>19.30</td>
</tr>
<tr>
<td>0.8</td>
<td>18.53</td>
<td>75.42</td>
<td>18.20</td>
</tr>
<tr>
<td>0.8</td>
<td>33.77</td>
<td>75.23</td>
<td>18.00</td>
</tr>
<tr>
<td>0.9</td>
<td>46.66</td>
<td>73.18</td>
<td>17.40</td>
</tr>
</tbody>
</table>

We have proposed the Cache-At-Relay (CAR) scheme which jointly optimizes content access energy, caching energy and relay deployment costs in three Integer Linear Programming (ILP) models. The ILP models are Select Relay (SER), Place Content (PCONT) and Place Relay (PREL) which respectively address uplink energy minimization, joint minimization of uplink and caching energy, and joint minimization of uplink energy and relay deployment cost. We have shown that PREL significantly minimizes the uplink energy consumption of UEs while PCONT provides a compromise between the content access energy budget of the UEs and the caching energy budget of the relays.

For future work, we aim to utilize nomadic relays to overcome the relay deployment cost issue. Emergent problems are handover, determining the travelling path of the nomadic relay and location-dependent cache management. In addition, we are planning to study joint optimization of cache management and energy-efficiency. Selecting which contents to store, popular content placement in relay caches is still an open problem. In particular, response time becomes a significant dimension of the investigated problem. In our future work, we are planning to pursue joint optimization of the response time and uplink energy expenditure. Furthermore, in this paper, we assume popular videos are limited with a small number and leave the discussion on maintaining those videos in the caches as a future work.

VI. CONCLUSION

In this paper, we have addressed the content placement problem in wireless relays with an objective of energy efficiency.

Moreover, the cache utilization of PCONT outperforms SER while under run time constraints the performance of SER is favorable to PCONT and PREL. Our results mostly focus on maintaining those videos in the caches as a future work. We aim to utilize nomadic relays to overcome the relay deployment cost issue. Emergent problems are handover, determining the travelling path of the nomadic relay and location-dependent cache management. In addition, we are planning to study joint optimization of cache management and energy-efficiency. Selecting which contents to store, popular content placement in relay caches is still an open problem. In particular, response time becomes a significant dimension of the investigated problem. In our future work, we are planning to pursue joint optimization of the response time and uplink energy expenditure. Furthermore, in this paper, we assume popular videos are limited with a small number and leave the discussion on maintaining those videos in the caches as a future work.

REFERENCES


