Energy Prioritized Caching for Cellular D2D Networks

S. Sinem Kafiloğlu, Gürkan Gür† and Fatih Alagöz
Department of Computer Engineering, Bogazici University
Istanbul, Turkey
†Zurich University of Applied Sciences (ZHAW)
Winterthur, Switzerland
Email: {sinem.kafiloglu, fatih.alagoz}@boun.edu.tr, gueu@zhaw.ch

Abstract—Multimedia content transmission heavily taxes network resources and puts a significant burden on wireless systems in terms of capacity and energy consumption. In this context, device-to-device (D2D) paradigm has a utilitarian value to alleviate the network burden by utilizing short-range transmissions with less energy cost. For the realization of proper D2D networking, caching in devices is essential. Additionally, caching needs to operate efficiently with the aim of realizing network-wide energy improvements. To this end, we study multimedia caching in cellular D2D networks and propose an energy prioritized D2D caching (EPDC) algorithm in this work. We also investigate the optimal caching policy in that system setting. The impact of device capacity and D2D transmission range on energy consumption is studied by focusing on different operation modes such as D2D and base station transmissions. According to the simulation results, EPDC algorithm has substantial energy-efficiency gains over the commonly utilized Least Recently Used (LRU) algorithm and content-based caching algorithms PDC and SXO. For larger D2D transmission ranges, this improvement becomes more evident.

I. INTRODUCTION

The multimedia traffic in data networks has surged with the exploding video consumption of users and substantially enriched services of multimedia providers. Based on the Cisco Annual Internet Report forecasts, multimedia continues to be of enormous demand today, and there will be escalated burden with the application requirements of the future with 8K and VR streaming, and cloud gaming even beyond the forecast period of 2023 in 5G networks [1]. In this context, the network operators have to meet certain QoS levels for provisioning consumer-driven content-based services. Nevertheless, the network resources are limited and burgeoning content consumption requests are observed leading to spiking energy consumption. In that regard, the energy efficiency (EE) problem emerges in these content-flooded networks. This problem needs to be resolved by practical networking paradigms in an efficient manner.

In-network caching is one of the techniques to alleviate the EE problem in future wireless networks. In this approach, the network load is reduced while transmission latency is lowered [2]. In [3], an energy-efficiency based in-network caching algorithm is proposed. It can perform better than the existing algorithms in term of both energy consumption and cache hit rate according to their simulation results. Fang et. al. also investigated in-network caching for achieving green content centric network paradigm [4]. Overall, in-network caching is a promising solution to tackle the EE challenges. In the same vein, Device-to-Device (D2D) communication technique is beneficial in terms of various aspects such as delay reduction, improved data rates, boosted spectral efficiency in addition to reduced energy consumption [5]. D2D mechanism can be used as a relaying factor to improve the power-efficiency of cellular networks [6]. It also improves the radio resource utilization to alleviate the communication coverage challenges in cellular systems [7]. Therefore, it is posed as a crucial enabler for boosting system capacity for heavy multimedia workloads and improving network-wide energy efficiency [8].

The simultaneous exploitation of edge caching and D2D networking is broadly studied, e.g. [9]. The major part of the literature focuses on the service improvement as a crucial element in cache-aided D2D networks [10], [11]. In that regard, some studied metrics are throughput [10], total service success probability [11], transmission delay [12] and offloading rate [12], [13]. Furthermore, the energy consumption and efficiency pose as vital metrics to be examined in cache-enabled D2D networks. In [8], the energy efficiency in a D2D caching HetNet is investigated in great detail without focusing on D2D caching optimization. On the contrary, in [14] a greedy caching algorithm is proposed and compared to the optimal case in terms of energy efficiency. That latter work resides not on D2D networks but Information-Centric Networks (ICNs). Lee et. al. propose caching policy and cooperation distance design in D2D caching networks for i) energy efficiency ii) throughput optimization separately in [15]. However, their energy model is limited to the device reception powers only.

In this work, we investigate the edge cache management techniques in D2D cellular networks owing to the fact that in-network caching is regarded as an efficient facilitator for improving system-wise energy consumption. We have a comprehensive service-mode based energy model and define content energy expenditures based on the prospective consumption across the network. We formulate the optimization for the cache replacement problem in D2D networks and solve it with dynamic programming. Besides, we propose a heuristic algorithm Energy Prioritized D2D Caching (EPDC) to im-
prove the time complexity of the content services. Finally, a rigorous performance analysis is performed while focusing on the service rate and energy consumption in different operation modes separately.

The key contributions of this work are as follows:

- We introduce a new energy expenditure model based on prospective energy consumption for content retrieval in cellular D2D networks.
- We optimally manage caches for the replacement problem in our D2D network architecture with the aim of keeping contents that would induce large energy cost.
- We propose an energy cost based heuristic, namely Energy Prioritized D2D Caching (EPDC) algorithm.
- We investigate the impact of cache capacity and D2D transmission range on the service capacity and energy consumption of several operation modes.

II. RELATED WORK

There is a plethora of work in the literature on content caching in D2D networks. We start our investigation by elaborating on papers for content-based D2D cache management. In [16], they have shown that Mandelbrot-Zipf (MZipf) distribution scale law in cache-aided D2D networks is the same with commonly utilized Zipf distribution and hence already existing studies utilizing the Zipf distribution become practically usable. Besides, they have shown that D2D cache network has improvement over the alternative BS unicast scenario. Congruently, we make use of the Zipf distribution in our content popularity model. However, instead of a comparison with no cache-aided D2D scenario, we specifically focus on the cache replacement for the energy aspect and make our comparison to conventional cache replacement techniques. Chen et. al. learn user preferences and manage caching accordingly, and they achieve a remarkable caching gain over the baseline that utilizes content popularity [17]. They exhaustively study the offloading probability while we focus on the energy consumption in great detail and inspect consumption in modes separately. Li et. al. propose caching in Fog Radio Access Network (f-RAN) enhanced with D2D technology [18]. Based on social preferences, AP communities are constructed and within a community, cooperative caching is utilized while communities are dynamically re-determined. For further improvement in transmission delay, D2D mode transmission is used and most appropriate device in every AP and most requested contents are selected for the caching scheme. They interrogate cache hit and delay metrics while in our study we deliberately focus on the network-wise energy consumption. Besides, we have a more general network architecture that can be tuned for custom network scenarios such as fog architecture if needed. Overall, these studies are on the content-based caching in D2D networks.

Now, we specifically discuss energy focused caching studies. The greening of computer networks becomes more challenging with the advent of 5G networking due to high multimedia traffic and low latency requirements [19]. To overcome the energy burden of 5G heterogeneous networks, in-network caching is broadly utilized in different segments. Yang et. al. propose a self-optimizing in-network caching algorithm for improving energy-efficiency in 5G networks [20]. However, the D2D communication mechanism is not utilized in their study. As D2D paradigm is a major component in our system, we particularly focus on the D2D network caching studies. Chen et. al. propose to place contents via mobility-awareness to optimize cache hit performance in [21]. Apart from that, they optimize content transmission powers of small base stations (SBSs) and devices to alleviate the energy cost. They demonstrate the improvement in terms of cache hit and energy efficiency in contrast to random, popular and mobility-aware caching with fixed data amount (MCF). Similar to our study, they interrogate the general 5G environment from the perspective of caching and EE. However, we restrain ourselves to a more specific 5G domain that is the “edge D2D network”. Above all, they analyze the cache placement problem while we focus on the replacement. In [22], they define an energy ratio metric and propose a sub-optimal caching scheme built upon this metric to improve EE. According to their study, their proposal performs better than the equal probability random, most popular random and cut-off random caching techniques in terms of EE. They have restricting assumptions such as no energy consumption in cache hits and the lack of universal source concept. Besides, we define EE as the total energy consumption over the successfully received data while they measure differently as the ratio of the energy in caching network over no-caching network.

Finally, we elaborate on some optimization studies in D2D caching networks. Chen et. al. optimize the caching and resource allocation for maximizing the offloading of services to local hits or from neighbouring devices in D2D operation mode [23]. However, this study is limited to offloading probability inspection only whereas we focus both on the service capacity and energy consumption of our entire D2D cellular network. In [24], they study cache partitioning, content placement and user association problems in D2D HetNets. First, they propose a game-based cache partitioning scheme to achieve the optimal cache space allocation. Building upon this, they formulate a service delay minimization problem for content placement and user association. Due to the mixed integer nonlinear nature of this problem, Lagrangian partial relaxation and partitioning into sub-problems is applied and solved accordingly. Even though this is a holistic caching optimization study in D2D networks, it lacks the energy efficiency research that is a very fundamental premise of our work. Besides, our content model is more realistic with chunk and layer dimensions that are not available in [24]. Choi et. al. propose a cache placement algorithm to maximize the overall content quality observed by users and also a resource allocation algorithm with the objective of maximizing the expected requested content quality of users constrained by tolerable delay [25]. They consider different content qualities in D2D caching networks and focus on the content quality while our main objective is to improve the energy efficiency of D2D network. In [26], Chen et. al. propose optimal proactive caching for improving offloading to D2D and also optimize D2D transmission power to maximize service percentage. Even though they consider overlaying in D2D communications
and investigate the D2D offloading gain and device energy consumption, their content model is limited. We utilize layering and chunking in our content model yielding a more comprehensive content structure.

In a nutshell, the core motivation of our study is to optimize the cache replacement problem of popularity, partition and quality based multi-attributed multimedia contents for minimizing the energy burden of D2D edge networks.

III. SYSTEM MODEL

In this section we present the network architecture, content model and content request management in our system. The network architecture is shown in Fig. 1. In our network architecture, there is one cell with one base station (BS) and D2D enabled user devices are scattered across this cell. As broadly utilized in the literature [27], [28], devices are spread across the cell according to Point Poisson Process (PPP) with mean density $\lambda_{Hu}$. There exists the universal source to serve the contents that are not available in any system unit (e.g. BS, device) at the edge network. In our network architecture, there are four main service operation modes: 1) local hit, 2) D2D mode, 3) BS mode (direct), 4) BS mode (from the universal source) as marked in Fig. 1.

![Network Architecture](image)

**Fig. 1: Network architecture.**

In our content model, request probabilities are determined by popularity that is based on the Zipf distribution. The probability of content requests are calculated via $Zipf(s, N_c)$ where $s$ determines the distribution skewness and $N_c$ is the total number of contents in the system. Content chunking is also used in communication systems with the aim of performance improvement [29], [30]. Accordingly, we utilize chunked contents in our model. Inter-chunk popularity differentiation is another approach used for improving transmission and caching performance in content-driven networks. In that regard, the request rate of content chunks are calculated by the commonly used $Weibull(\lambda, k)$ distribution with the scale and shape parameter $\lambda$ and $k$ respectively.

Our contents are scalable coded (SVC) videos. Each content is of high quality (HQ) consisting of one base and one enhancement layer. For standard quality content consumption, the base layer is sufficient whereas both layers are required for HQ visualization. The request rate for HQ contents in our network scenario is $p_c \in [0, 1]$. We define each layer of a given content’s particular chunk uniquely as a content unit tuple by its content, chunk and layer ids respectively as $\{i, j, k\}$. Each such tuple is mapped to a unique content unit identifier $u$ for tractability. The average content sizes are 322 and 474 Mbits for SQ and HQ SVC videos, respectively [9]. For the calculation of these values, QCIF-formatted and an hour long temporal scalable encoded multimedia are exploited [31]. In [32], the calculation formula for the mean video frame sizes of different layers is provided. Some utilized video samples are Citizen Kane, Jurassic Park I, Star Wars IV, Aladdin and Tonight Show. For the full video list, please refer to our previous work [9]. In Section IV, we preserve the content characteristics of our content model and only change the total number of contents to push the system to its limits for the feasibility inspection of the time complexity analysis. In the performance evaluation section, we also keep the total number of contents fixed and investigate the impact of some other system parameters as described in Section VI.

We extend the wireless D2D network in our previous work [9] with cellular support. In this context, the BS transmission power $P_{BS}$ and BS transmission path loss exponent $n_{BS}$ are introduced for the calculation of the BS power strength in the BS channel model. The general content request management scheme is shown in Fig. 2. First, the local cache is looked up for a requested content unit. If that unit is not found in the local cache, then D2D mode is chosen as long as that requested unit is available in at least one device in the reception range of (at most $R_{D2D}$ away from) the requester device. We assume the inter-device distances are known by user devices according to a signaling scheme to discover neighbour devices [33]. Thereby, devices in the reception range of each other can determine inter-node distances. The BS serves when that requested unit is not found in any device.
in the reception range of the requester. This case branches into two scenarios: \(i\) that unit is available in the BS cache, 
\(ii\) that unit is not available in the BS cache. In the scenario-\(i\), the BS cache directly serves that unit. In the scenario-\(ii\), that 
unit is first fetched from the universal source to the BS and then 
served to the requester to allow access to the exterior of the 
device network system. Note that the multi-mode service 
operations (local hit, D2D mode, BS mode) in cellular D2D 
networks are applicable as shown in [34]. All requests and 
network using operation modes (2-4) acquire the same 
network access priority. If a content unit cannot be served 
due to all channels being busy, then that unit and prospective 
units of that content are all dropped.

The BS and devices have storages for caching multimedia 
contents. In this paper, we mainly study on the cache replace-
ment techniques running both on user devices and the BS. 
In the cache replacement procedure, the system unit (device/ 
BS) evicts its cache until there is sufficient free space for the 
requested and retrieved video content unit. The decision for 
the evictions will be explained in the next section.

IV. OPTIMAL AND ENERGY PRIORITIZED D2D CACHING 
(EPDC)

5G systems is posed to be an enabler for connected multi-
media services. However, the tremendous demand for the 
media services raises the energy consumption of the wireless 
networks, leading to a fundamental cost and environmental 
issue that needs to be addressed. To this end, 5G communica-
tions need to improve energy efficiency performance [35]. The 
D2D edge network is one the 5G technologies that provides 
service capabilities closer to end users and thus reduces the 
energy burden on the network. Similarly, in-network caching 
is an important mechanism to reduce network transmissions 
and enhance the energy efficiency as well. In that regard, 
we formulate our cache replacement problem in D2D edge 
network for decreasing energy consumption in this section. 
Subsequently, optimal and heuristic approaches are employed 
to solve it for our system setup.

Initially, we solve the scalable video caching in D2D 
networks with the backhaul support optimally. For the ease 
of tractability, the model and system parameters are listed in 
Table I. Say the content units residing in the local cache of 
some user are denoted by \(C\). Some user requests a content 
unit \(u'\) and it is retrieved from some other system component 
via wireless transmission. If the size of content units in the set 
\(C\) and new arrived unit \(u'\) together exceed the cache capacity 
\(C_{Dev}\), then a subset of content units \(C\) residing in the local 
cache need to be replaced. The eviction decision is essentially 
a 0/1 Knapsack problem [14]. For each content unit you either 
decide to sustain that unit in the local cache or remove it. 
The content units that are decided to remain in the local 
cache are designated by \(\overline{C}\). In our network architecture at 
each caching decision phase, we want to minimize prospective 
environmental energy consumption. Thereof, we sustain content unit set \(\overline{C}\) that would induce the largest total energy expenditure. This 
way, content units with lower energy consumption will be 
evicted. In the prospective requests for them, they will be 
fetched from other system components with less cost. The 
optimization problem is stated as follows:

\[
\begin{align*}
\max \sum_{u \in C} E^{(u)}_{all} \ 1_{[u \in \overline{C}]} \\
\text{s.t.} \quad s(u') + \sum_{u \in C} s(u) \ 1_{[u \in \overline{C}]} \leq C_{Dev}
\end{align*}
\]

with the indicator function \(1_{[x]}\) set to one if \(x\) is true and zero 

| TABLE I: Notations. |
|----------------------|
| Parameter | Explanation |
|-----------|-------------|
| \(N_c\) | The total number of contents |
| \(s\) | The Zipf distribution skewness parameter |
| \(\Lambda\) | The Weibull distribution scale parameter |
| \(k\) | The Weibull distribution shape parameter |
| \(p_i\) | The probability of content \(i\) being requested |
| \(p_j\) | The probability of content chunk \(j\) being requested |
| \(p_k\) | The probability of high quality contents being requested |
| \(\{i, j, k\}\) | The unique content unit tuple (The \(i^{th}\) content’s chunk \(j\) 
of layer \(k\)) |
| \(u\) | The unique content unit identifier such that each 
\(\{i, j, k\} \rightarrow u\) |
| \(s(u)\) | The size of the content unit \(u\) |
| \(R_{D2D}\) | The maximum allowable distance between a requester and transmi-

The second scenario consumes \(E_{D2D}^{(u)}\) energy when the content unit will not be available in 
the local cache but retrievable from some other device with probability \((1 - p_{loc}^{(u)})P_{D2D}^{(u)}\). If no device in the transmission 
range (centered at the requester with radius \(R_{D2D}\)) can serve 
due to the lack of unit, the BS takes action and will serve the 
content unit from the BS cache with the energy expenditure 

\[
\text{otherwise. Our objective function } \sum_{u \in C} E^{(u)}_{all} \ 1_{[u \in \overline{C}]} \text{ in (1) gives the total energy consumption of units across the network that are not evicted from the cache set } \overline{C} \text{ and residing in } \overline{C}. 
\text{With constraint (2), we ensure that the aggregate size of the new unit } u' \text{ and the units not evicted from the cache does not exceed the cache capacity. We formulate the prospective energy consumption of any content unit } u \text{ as } E^{(u)}_{all} \text{ in (3). This energy is branched into the four scenarios that are shown in Fig. 1: (1) Local hit, (2) Transmission via D2D technique, (3) Transmission from the BS cache, (4) Transmission from the universal source across the BS.} 
\text{With the probability } p_{loc}^{(u)} \text{ the first scenario occurs and } E_{loc}^{(u)} \text{ will be the corresponding prospective energy expenditure for the local hit of content unit } u. \text{The second scenario consumes } E_{D2D}^{(u)} \text{ energy when the content unit will not be available in the local cache but retrievable from some other device with probability } (1 - p_{loc}^{(u)})P_{D2D}^{(u)}. \text{ If no device in the transmission range (centered at the requester with radius } R_{D2D} \text{) can serve due to the lack of unit, the BS takes action and will serve the content unit from the BS cache with the energy expenditure } E_{\overline{C}}^{(u)} \text{.}
level $E_{BS}^{(u)}$. Unless the relevant unit is in the BS cache, the transmission path from the universal source to the BS cache and then from there to the requester device will be utilized with 

$$(1 - p_{loc}^{(u)})(1 - p_{BS}^{(u)})$$

probability and $E_{BS}^{(u)}$ energy will be consumed. The expected energy consumption of unit $u$ based on listed four scenarios is given in (3). The functions corresponding to the probability of content unit availability for different system components are provided in the Appendix.

$$E_{all}^{(u)} := p_{loc} E_{loc}^{(u)} + (1 - p_{loc}) p_{BS}^{(u)} E_{BS}^{(u)}$$

The expected energy expenditure of the given four scenarios are listed as follows:

$$E_{loc}^{(u)} := p_{loc} \frac{s(u)}{C_{loc}}$$

$$E_{DD}^{(u)} := p_{DD} \frac{s(u)}{C_{DD}}$$

$$E_{BS}^{(u)} := p_{BS} \frac{s(u)}{C_{BS}}$$

$$E_{BS}^{(U)} := [p_{BS}(u) \frac{s(u)}{C_{BS}(u)}] + E_{BS}$$

The expected energy expenditures are calculated by the product of power in some service type and its expected service duration for a given content unit. During local hits, a device consumes the least energy compared to other system components. We define this power $P_{loc}$ as $\frac{\delta_{Dev}}{\delta_{loc}}$ while its expected service duration is $\frac{s(u)}{C_{loc}}$. The largest service rate among the system components is $C_{loc}$. The product of $P_{loc}$ and $\frac{s(u)}{C_{loc}}$ gives the expected local hit energy consumption $E_{loc}^{(u)}$ for some content unit $u$ as shown in (4). The product of the device transmission power $P_{DD}$ and its expected prospective service duration $\frac{\delta_{DD}}{\delta_{loc}}$ gives the expected D2D transmission energy $E_{DD}^{(u)}$ in (5) for unit $u$. $C_{DD}$ is calculated by taking the average distance $\frac{\delta_{Dev}}{\delta_{loc}}$ for future D2D transmissions. The devices do not know the cache status of others, and for future predictions they tune to average values for the calculation of prospective energy figures. In (6), the expected energy consumption for a transmission of some unit $u$ from the BS cache is calculated similar to the $E_{DD}^{(u)}$. When we consider the fourth scenario, the content unit retrieval basically branches into two parts: $i)$ from the universal source to the BS cache, $ii)$ from the BS cache to the requester device. The retrieval to BS consumes $E_{BS}^{(u)}$ expected energy while the rest has $E_{BS}^{(u)}$ as shown in (7). For the retrieval to the BS cache, its reception power is $P_{BS}(u) := \frac{\delta_{BS}}{\delta_{BS}}$ and its expected prospective reception period is $\frac{\delta_{loc}}{\delta_{BS}}$.

Dynamic programming algorithm is one of the techniques to solve the 0-1 knapsack problem optimally [36]. In our optimal solution, we utilize dynamic programming given in Algorithm 1. In this algorithm, if the cache capacity $C_{Dev}$ does not suffice for the residing content units set $C$ and the newly requested one $u'$, then a subset of $C$ will be evicted for the sake of newly arrived unit $u'$. Our main motivation is to keep content units in the local cache that would otherwise result in large prospective energy consumption. We consider the cache as the knapsack. Each content unit $u$ is an item that is weighted by its size $s(u)$ and valued by its energy consumption $E_{all}^{(u)}$. In dynamic programming, we break the optimization problem into subproblems, recursively solve these subproblems and store their results in $E_{cum}$ array as given in Algorithm 1. By backtracking $E_{cum}$, we select the units with the largest energy sum to continue residing in the cache. These units are designated by the set $C$. However, the time complexity of this algorithm is $O(|C| \cdot W)$ where $W$ is the residual cache capacity after the new unit is inserted and $|C|$ is the number of content units in a given device cache [14]. This complexity does not provide a practical calculation time. Thereof, we propose a new energy prioritized D2D caching algorithm given in Algorithm 2. The core idea of this greedy algorithm is to utilize our analytically calculated expected energy consumption of units $E_{all}^{(i)}$ to alleviate the energy burden on the network system. In this heuristic, at the eviction decision phase, all content units in the set $C$ are sorted based on their $E_{all}^{(i)}$ from the largest to the smallest one. As shown in Algorithm 2 starting from the least energy consuming unit $c_k$, units are eliminated from the cache until the free space is suitable for storing new unit $u'$. Thus, unit(s) with low energy burden across the network are eliminated first. Large energy consuming units over the network are preserved locally for less energy burden in future requests. Therefore, this scheme is named as energy prioritized D2D caching (EPDC). The time complexity of this procedure is bounded by the sorting procedure of elements in the set $C$ based on their $E_{all}^{(i)}$ values.

$$O(|C| \cdot \log(|C|))$$ with $|C|$ denoting the number of content

Algorithm 1 Dynamic Programming Caching Algorithm

```
Algorithm 1 Dynamic Programming Caching Algorithm

OPT(C, u', C_{Dev})

Capacity ← TotalSize(C);
if (Capacity + s(u') ≤ C_{Dev}) then
    return C ∪ {u'}; % the set kept in the cache
else
    % traverse all content units in the local device cache
    for all (i = 1 : 1 : |C|)
        % traverse weights with δ_{Dev} (= 0.01 Mbs) incremental units
        for all j = 1 : 1 : \[
            \left\lceil \frac{Capacity - s(u')}{{δ_{Dev}}} \right\rceil
        \]
            if ((i == 1) ∨ (j == 1)) then
                E_{cum}(i, j) ← 0; % first row and column are set to zero
            else if (s(i) < j) then
                E_{cum}(i, j) ← \max\{E_{cum}(i-1, j), E_{cum}(i-1, j - s(i))\};
            else
                E_{cum}(i, j) ← E_{cum}(i-1, j);
            end if
        end for
    end for
    % start from the last row and column of E_{cum} backtrack and mark units that will continue to reside in the local device cache as C
    C ← backtrack(E_{cum});
    return C ∪ {u'};
end if
```

units in a given device cache.

Algorithm 2 Energy Prioritized D2D Caching Algorithm (EPDC)

\[
\text{EPDC}(C, u', C_{\text{Dev}})\{
\]
\[
\text{Capacity} \leftarrow \text{TotalSize}(C);
\]
\[
\text{if} \ (\text{Capacity} + s(u') \leq C_{\text{Dev}}) \text{ then return } C \cup \{u'\}; \% \text{the set kept in the cache}
\]
\[
\text{else ssort units in } C \text{ based on their } E_{a\ell}(i) \text{ values in (3) in descending order}
\]
\[
\% C_{\text{sort}} = \{c_1, c_2, ..., c_k\} \text{ with } c_1 \text{ having largest } E_{a\ell}, c_k \text{ lowest}
\]
\[
C_{\text{sort}} \leftarrow \text{sort}(C, E_{a\ell}, \text{DESCEND});
\]
\[
\text{while } (j \geq 1) \text{ do return } C' \text{; }
\]
\[
\% \text{the set decided to be kept in the cache}
\]
\[
\text{if } (\text{Capacity} + s(u') - \sum_{k=1}^{j} s(c_k) \leq C_{\text{Dev}}) \text{ then }
\]
\[
\text{end if}
\]
\[
\text{if } j = 1; \%
\]
\[
\text{end while}
\]
\[
\}
\]

A. Time complexity analysis

Based on our complexity analysis, OPT algorithm has infeasible operation time. As shown in Algorithm 1, for all cached units OPT has an iteration range by the cache capacity with \( \delta \) incremental units. For devices we take \( \delta_{\text{Dev}} = 0.01 \) Mbs and for the BS cache \( \delta_{\text{BS}} = 0.1 \) Mbs respectively. The average time consumed for devices and the BS cache by all techniques is provided in Table II. Note that we test the time complexity in a laptop device with Intel core i7-6500 CPU @2.5GHz and 8 GB RAM in MATLAB R2019 environment. We apply any chosen algorithm not only in devices but also in the BS cache for the sake of completeness and hence due to the large BS cache capacity even in a setup with moderate number of contents, we present the impracticality of applying OPT in real-time or near real-time scenarios. With \( C_{\text{Dev}} = 150 \) Mbs and \( C_{\text{BS}} = 2.8 \) Gbs even with 150 contents, OPT performs poorer than EPDC in terms of service time especially for the BS caching. This is because EPDC algorithm is only bounded by the sorting of content units but not cache capacity while OPT algorithm acts an iteration over the cache capacity for each in-cache content unit.

| Technique | Total Time (s) | Local Caching | Total BS Caching Time (s) |
|-----------|---------------|---------------|-------------------------|
| LRU       | 0.57          | 1.11          |                         |
| PDC       | 0.53          | 2.65          |                         |
| SXO       | 0.82          | 9.69          |                         |
| OPT       | 18.94         | 7319.11       |                         |
| EPDC      | 6.01          | 5.80          |                         |

V. PERFORMANCE METRICS

We analyze optimal scalable video caching algorithm (OPT) in D2D cellular networks with backhaul support and energy prioritized D2D caching (EPDC) algorithm in terms of service rate and energy expenditure. The system parameters are listed in Table III. We compare OPT and EPDC algorithms to the commonly used baseline technique Least Recently Used (LRU). LRU technique does not consider content features. Therefore, it used to compare the improvement of our algorithms over a classical non-content aware cache replacement technique. In the literature, popularity based caching policies are also broadly utilized [37], [38], [39]. As another comparison technique, we utilize the Popularity-driven caching (PDC) algorithm [34], that manages the cache replacement with the aim of keeping popular contents in cache with a greater probability. This technique is aware of content attribute popularity and decides on the evictions accordingly. As multimedia content caching is the main focus of our algorithms, we compare our algorithms to this content-aware cache replacement technique PDC. We also compare our algorithms to the SXO technique [40]. In this technique, content features size and order (access rate) are considered for the cache replacement decision where large contents with low access rate are evicted first. Note that we will investigate the service rate and energy expenditures in different operation modes. In that regard, the energy consumption of our system consists of the following parts:

- \( E_{\text{loc}} \) for local hits
- \( E_{\text{D2D}} \) for D2D transmission
- \( E_{\text{BS}} \) for direct services of the BS
- \( E_{\text{BS(U)}} \) for indirect services of the BS
- \( E_{\text{block}} \) due to blocked services

\( E_{\text{loc}}, E_{\text{D2D}} \) and \( E_{\text{block}} \) are defined as in our previous work [9]. One of our new contributions is the BS integration to our D2D network architecture. Therefore, we define the BS energy expenditure branched into i) direct services and ii) indirect services as \( E_{\text{BS}} \) and \( E_{\text{BS(U)}} \) in (8) and (9), respectively. \( E_{\text{BS}} \) is the overall transmission energy consumption of the BS serving requests directly from the BS cache whereas \( E_{\text{BS(U)}} \) is the energy consumption for services from the universal source

| Parameter | Explanation |
|-----------|-------------|
| \( P_{\text{BS}} \) | The transmission power consumption of the BS |
| \( P_{\text{BS(U)}} \) | The reception power consumption of the BS |
| \( C_{\text{BS}}(n) \) | The channel capacity between the BS and the \( n^{th} \) device |
| \( C_{\text{BS(U)}} \) | The capacity between the universal content repository and the BS |
| \( N \) | The total number of devices |
| \( U \) | The set of content units uniquely identifiable by content, chunk, layer id |
| \( r_{u} \) | The request for the content unit \( u \) |
| \( s_{(u)} \) | The size of the content unit \( u \) |
| \( S_{\text{BS}}(n) \) | The set of services from the BS to the \( n^{th} \) device |
| \( S_{\text{BS(U)}}(n) \) | The set of services from the universal content repository to the \( n^{th} \) device across the BS |
across the BS to requesters.

\[
E_{BS} := \sum_{u \in U} \sum_{n \in N} \sum_{r_u \in S(n)} P_{BS} \cdot \frac{|s(u)|}{C_{BS}(n)}
\]  

(8)

\[
E_{BS(U)} := \sum_{u \in U} \sum_{n \in N} \sum_{r_u \in S(n)} P_{BS} \cdot \frac{|s(u)|}{C_{BS}(n)} + \frac{P_{BS(U)} \cdot |s(u)|}{C_{BS(U)}}
\]  

(9)

By the summation of all the expenditures, we obtain 
\[E_{total} := E_{loc} + E_{D2D} + E_{BS} + E_{BS(U)} + E_{block}\] the total energy consumption of our system.

VI. PERFORMANCE EVALUATION

We perform simulations for varying system parameters cache capacity and D2D transmission radius, and investigate how these parameters impact the network for the service rate and energy expenditure. We compare EPDC and OPT algorithms to the LRU, PDC [34] and SXO [40]. The simulation parameters are listed in Table IV. In this list, the content unit popularity and request parameters (s, \(\lambda\), k, \(p_k\)), the power of system components (local, BS, device), local hit and BS reception power tune parameters (\(\theta_{loc}\), \(\theta_{BS}\)), cache capacities (\(C_{Dev}\), \(C_{BS}\)), channel parameters, device density \(\lambda_{NHU}\) and the maximum transmission radius of D2D mode \(R_{D2D}\) are defined. In Table IV, the local and D2D power consumptions are in mW scale while the BS transmission/reception powers are at the order of Watts as observed in practical systems. The channel parameters B, \(d_0\) and \(n_{D2D}\) are adopted from our previous study [9]. We take the path loss exponent of the BS transmissions \(n_{BS}\) greater than that of the D2D mode \(n_{D2D}\) as broadly employed in the literature [41], [42]. We have an event-based simulator that processes content unit request arrivals and service completions. Since we elaborate on the caching of multimedia contents, we simulate content requests according to a popularity scheme that utilizes Zipf distribution with parameter \(s\). Each content is partitioned into chunks and the requests for these chunks follow the Weibull distribution with parameters \(\lambda\) and k. Due to the scalable content structure, our contents are layered into base and enhancement. The high quality (HQ) content consumption with both layers is realized by the HQ request probability \(p_k\). For the event-based simulator that acts upon each content unit request arrival/completion, network operations require power and channel parameters to actualize the arrivals. In that regard, these parameters are provided in the Table IV. In the following subsections, we vary the device cache capacity \(C_{Dev}\) and D2D transmission radius \(R_{D2D}\) to elaborate on how the D2D cache and service capability impact the system.

A. Impact of device cache capacity \(C_{Dev}\)

In this section, we investigate a key component in our system, the device cache capacity \(C_{Dev}\). We enlarge/shrink the cache capacity \(C_{Dev}\) to observe how it impacts the caching techniques.

First, we compare our caching technique EPDC with the OPT and the alternative strategies LRU, PDC and SXO in terms of the local hit service rate and energy consumption. For any fixed \(C_{Dev}\) value, the energy optimizing OPT attains the smallest local service rate and in return the lowest local energy expenditure given in Fig. 3a and 3b, respectively. Compared to the optimal case, our heuristic EPDC is slightly worse in terms of energy consumption for local hits. However, the energy improvement over LRU, PDC and SXO manifests the utilitarian nature of our proposal for local hits. This improvement of EPDC over the alternatives becomes greater for decreased cache capacity. In the large cache capacity regime, the cache capacity has a greater impact on the energy expenditure as requested contents can be found locally with a greater probability. With decreasing cache capacity, the impact of cache replacement procedure can be observed more clearly. For \(C_{Dev} = 300\) Mbits the local hit energy consumption improvement rate of EPDC over LRU (PDC / SXO) is 4.0% from 0.67 to 0.64 J (3.5% from 0.66 to 0.64 J / 3.2% from 0.66 to 0.64 J). With decreasing \(C_{Dev}\) to 100 Mbits, this improvement over LRU (PDC / SXO) is amplified to 14.4% from 0.29 to 0.24 J (17.0% from 0.29 to 0.24 J / 15.4% from 0.29 to 0.24 J).

In all caching techniques, for increasing \(C_{Dev}\), the direct BS service usage and energy expense decline since large device caches entail increased local and D2D service utilization. For an arbitrary \(C_{Dev}\) value, our EPDC technique achieves less direct BS mode energy consumption than LRU, PDC and SXO techniques. The direct BS energy improvement rate of EPDC over LRU (PDC) is at its best at \(C_{Dev} = 200\) Mbits with 7.3% from 1.65 to 1.53 kJ (12.5% from 1.75 to 1.53 kJ). This improvement of EPDC over SXO reaches to its maximum at \(C_{Dev} = 300\) Mbits with 12.4% 0.99 to 0.87 kJ. EPDC falls back from the OPT for the BS energy

### Table IV: Simulation parameters.

| Symbol | Explanation | Value |
|--------|-------------|-------|
| \(T_{sim}\) | The simulation duration | 400 s |
| \(s\) | The Zipf distribution skewness parameter | 1 |
| \(\lambda\) | The Weibull distribution scale parameter | 5 |
| \(k\) | The Weibull distribution shape parameter | 0.8 |
| \(p_k\) | The probability of high quality contents being requested | 1 |
| \(P_{loc}\) | The power consumption of local content unit retrieval | 40 mW |
| \(P_{D2D}\) | The transmission power consumption of a device | 80 mW |
| \(P_{BS}\) | The transmission power consumption of the BS | 6 W |
| \(P_{BS(U)}\) | The reception power consumption of the BS | 1.2 W |
| \(\theta_{loc}\) | The local hit service power parameter | 2 |
| \(\theta_{BS}\) | The base station reception power parameter | 5 |
| \(C_{Dev}\) | The device cache capacity | 150 Mbits |
| \(C_{BS}\) | The base station cache capacity | 2.8 Gbits |
| \(B\) | The channel bandwidth | 2 MHz |
| \(N_0\) | The noise power density | -158 dBm |
| \(d_0\) | The reference distance of device antenna | 1 m |
| \(n_{D2D}\) | The path loss exponent of D2D transmission | 3 |
| \(n_{BS}\) | The path loss exponent of BS transmission | 4.2 |
| \(\lambda_{NHU}\) | The mean device density according to Point Poisson Process (PPP) | 0.0015 \(\frac{\text{m}^{-2}}{\text{m}^{2}}\) |
| \(R_{D2D}\) | The maximum D2D transmission operation radius | 200 m |
(a) Locally served content units (B: of base layer, E: of enhancement layer, S: successful).

(b) The energy consumed for the content units (B: of base layer, E: of enhancement layer, S: successful, F: fail, Remark: the energy consumption of failed services is very small (green and yellow components) compared to the energy of successful services).

Fig. 3: Results for the local caches.

Fig. 4: The energy consumed for retrievals in BS(U) mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

Fig. 5: The energy consumed for retrievals in D2D mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

caching policies. In correlation, the energy consumption of D2D usages are not significantly impacted by the varying device cache capacity $C_{Dev}$ for any given caching technique. The reason is that the varying cache capacity particularly impacts the local hit results. When we compare EPDC to the alternatives LRU, PDC and SXO for any $C_{Dev}$, we observe a slight increase in D2D service rate. In that regard, the energy consumption due to D2D services for EPDC is increased very slightly compared to LRU, PDC and SXO strategies. The greatest increase in D2D energy consumption of EPDC over LRU (PDC / SXO) is 3.6% from 92.43 to 95.78 J (5.6% from 90.49 to 95.78 J / 4.0% from 92.12 to 95.78 J) for $C_{Dev} = 100$ Mbits as shown in Fig. 5. The alternatives do not consider prospective energy consumption of the local, D2D, BS and BS(U) modes. However, our technique makes use of them and favors the D2D paradigm in contrast to the BS transmissions since D2D technique is more beneficial than cellular (BS) content transmissions in terms of energy cost. Thereof, we observe slight increase in D2D service rate and corresponding energy consumption in EPDC technique compared to the alternative counterparts.

Consumption. However, this degradation is at most 6.0% for $C_{Dev} = 100$ Mbits. Nevertheless, the evident improvement over LRU, PDC and SXO demonstrates the advantage of our heuristic EPDC for the direct BS energy consumption.

In our construct for indirect services of the BS, a similar reasoning with direct BS services and corresponding energy expenditures is valid. EPDC consumes less energy than LRU, PDC and SXO shown in Fig. 4 for all $C_{Dev}$ values.

The indirect BS energy expenditure improvement rate of EPDC over LRU (PDC / SXO) is at most 11.1% from 5.63 to 5.00 J (at most 10.8% from 4.38 to 3.91 J / 8.1% from 3.40 to 3.12 J) for $C_{Dev} = 100$ (150 / 200) Mbits. EPDC achieves indirect BS energy results close to the optimal case with at most 1.6% difference. Hence, our heuristic is practical for the energy reduction of the BS regarding the indirect BS services.

For the D2D analysis, the corresponding service rate is not affected by the varying $C_{Dev}$ values evidently for different
We expect to observe a major increase in D2D usage due to its low power consumption. However, the limited device capacities compared to the BS and the high power of the BS than devices are dominating factors against reducing the overall network energy consumption of EPDC. Thereof, we do not observe a significant change in D2D preference as in BS preference between different policies.

Finally, we focus on the total service rate and energy consumption of the network. As already mentioned, the D2D service rate is not impacted by the varying $C_{Dev}$ for different caching policies. However, for increasing $C_{Dev}$, the local service capacity rises (in Fig. 3a) while the BS service capacity reduces in all caching policies. The increase in local and decrease in BS service capacities balance each other and the total service capacity is not changed evidently for varying $C_{Dev}$ in each caching policy. In all policies, with increasing $C_{Dev}$ the energy consumption trends for local service/D2D/BS/BS(U) service types are consistent with the corresponding service capacity behaviour. Although for all caching strategies the total service capacity of the network is not affected by different $C_{Dev}$ values, with increasing $C_{Dev}$ the total network energy consumption declines for all caching strategies and hence the network-wide energy efficiency is improved.

As shown in Fig. 6, our proposed algorithm EPDC falls back from the optimal with at most 3.5% (for $C_{Dev} = 100$ Mbits) in our investigated $C_{Dev}$ range and EPDC has very close results to the optimal policy especially for large $C_{Dev}$’s. Although for fixed $C_{Dev}$ values the measured total network service capacity does not change significantly between different techniques, the observed total network energy expenditure of our technique EPDC is less and thus energy-wise more rewarding compared to LRU, PDC and SXO as depicted in Fig. 6. The greatest total network energy improvement of EPDC compared to LRU (SXO) is with 6.7% from 5.08 to 4.74 kJ (8.4% from 5.18 to 4.74 kJ) for $C_{Dev} = 200$ Mbits. EPDC attains the maximum improvement over PDC at $C_{Dev} = 150$ Mbits with 9.6% from 7.15 to 6.46 kJ. Thus, we observe the beneficial nature of our proposal for energy-sensitive network scenarios.

### B. Impact of D2D transmission radius $R_{D2D}$

We investigate how different settings of D2D transmission radius impact the system. In that regard, we vary the D2D transmission radius $R_{D2D}$. Initially, we study the local hit service capacity and corresponding energy consumption. As depicted in Fig. 7a, the local hit service capacity changes negligibly with varying $R_{D2D}$ for all caching strategies. When a content is available in the local cache, the service is locally provided and other network resources are not used. Hence, the D2D features do not impact the local hit service capacity. Besides, for all caching policies the local hit energy consumption is not affected by $R_{D2D}$ (Fig. 7b).

When we compare our proposal EPDC to LRU, PDC, SXO and OPT, EPDC achieves lower local service capacity and, in correspondence with this, local energy consumption improvement compared to LRU, PDC and SXO caching policies for any fixed $R_{D2D}$ value in the inspection range. Given in Fig. 7b, the greatest improvement for local hit energy...
consumption of our proposal over the alternative LRU (PDC / SXO) occurs at $R_{D2D} = 240$ m attaining 9.8% from 0.39 to 0.35 J (11.6% from 0.40 to 0.35 J / 11.1% from 0.39 to 0.35 J). Even the least local energy consumption improvement of EPDC over LRU (PDC / SXO) is 5.6% (6.2% / 6.3%) at $R_{D2D} = 80$ m. As expected, EPDC is not as good as the optimal caching in terms of local energy expenditure. Nevertheless, the largest rise observed in terms of local energy consumption of our proposal over the optimal policy is 3.3% only (from 0.35 to 0.34 J) at $R_{D2D} = 240$ m. Hence, the local energy expenditure improvement of EPDC over the baseline LRU and other comparison strategies is an indicator of the energy-wise utility of EPDC for local services.

Next, we perform an analysis on the direct BS services. The increasing D2D transmission radius $R_{D2D}$ amplifies the chance of finding requested content units in neighbours and subsequently the D2D service allocations are boosted. The BS services are less required and the direct BS service rate declines for all caching policies. In accordance with this change, in all policies the direct BS mode energy consumption reduces with increasing $R_{D2D}$. In our analysis, similarly a decrease is observed for indirect BS service rates with respect to increasing $R_{D2D}$. Also, decrease in indirect BS energy consumption with respect to increasing $R_{D2D}$’s is observed at all caching techniques as shown in Fig. 8.

For EPDC, the services supplied from the universal source across the BS (indirect BS services) hits lower service rate than the other algorithms for all $R_{D2D}$’s. Accordingly, as shown in Fig. 8, EPDC achieves improvement in indirect BS energy consumption over the comparison algorithms for all $R_{D2D}$’s. The optimal policy and our proposed algorithm EPDC are designed for minimizing the total network energy expenditures. Hence, omitting such long routes (from the universal source across the BS to requesters) with large power levels is instrumental for reducing overall system energy consumption. Considering policy comparisons for direct BS services, the corresponding energy expenditure improvement over LRU (PDC / SXO) is 12.2% (19.4% / 10.2%) for the largest $R_{D2D} = 240$ m. EPDC falls back to the LRU, PDC and SXO with decreasing $R_{D2D}$ in terms of direct BS energy consumption because the D2D service rate declines with decreasing $R_{D2D}$ and this enforces the system to get direct BS services for units available in the BS cache.

Fig. 9: The energy consumed for retrievals in D2D mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

For all caching policies, the D2D service rate and corresponding energy consumption increases with increasing $R_{D2D}$ as illustrated in Fig. 9. The reason is that the number of devices available in reception range of requesters increases in that case. Consequently, the probability of finding requested content unit(s) in a neighbour device is higher according to (14) and hence D2D service rate is improved. Our algorithm EPDC achieves slightly better rate and correspondingly, slightly larger energy consumption compared to alternative algorithms as shown in Fig. 9. As already explained in the previous subsection, the high power level of the BS is the key factor in determining the overall network energy consumption and EPDC algorithmchookes the BS selection rather than increasing the D2D mode selection. Hence, no significant D2D usage change is observed between different policies.

Fig. 10: The total energy consumed for retrievals and local hits (B: of base layer, E: of enhancement layer, S: successful, F: fail).
We also inspect the total service rate and energy expenditure of the network with respect to varying D2D transmission radius $R_{D2D}$. For all policies, the total network service rate rises with increasing $R_{D2D}$ from 80 m to 120 m. With further increase, it becomes saturated in all policies. For all caching algorithms, we observe an increase in the total network service rate for the greatest D2D service improvement rate ($R_{D2D}$ from 80 m to 120 m) that is not canceled out by the BS service rate drop. This is because the D2D service is more crucial with its larger rate compared to the other service factors in determining total network service rate. With further increasing $R_{D2D}$ values, we again observe recession in BS service rates and this time they cancel the increase in D2D service rates. This is due to the slower D2D service improvement. As already mentioned, in all caching techniques the BS energy expenditure decreases, the D2D energy consumption rises whereas the local expenditure is stagnant with increasing $R_{D2D}$’s. The BS services take longer service durations with higher power levels in contrast to D2D mechanism. Thus, the BS manifests itself with its great energy consumption impact. Overall, for all caching strategies in Fig. 10, the total energy expense of the network decreases with increasing $R_{D2D}$’s.

For fixed $R_{D2D}$ values, the total network service rate does not change evidently among different caching techniques. However, as shown in Fig. 10 for all $R_{D2D}$’s, the overall energy consumption of the network by EPDC is less than other algorithms. This energy expenditure improvement of EPDC rises with increasing $R_{D2D}$. This is basically due to the difference in BS usage and corresponding energy consumption shown in Fig. 8. The greatest decrease in the total network energy consumption of EPDC compared to LRU (PDC / SXO) is for $R_{D2D} = 240$ m with 9.5% from 5.38 to 4.87 kJ (12.4% from 5.56 to 4.87 kJ / 10.6% 5.45 to 4.87 kJ). When we compare our proposal EPDC with the optimal counterpart, it slightly performs worse with at most 2.4% performance degradation in the total network energy consumption at $R_{D2D} = 240$ m. In spite of this fall back, we observe improvement of EPDC over the compared algorithms for $R_{D2D} \geq 120$ m. As already mentioned, no evident change in total network service rate is observed between different caching techniques for fixed $R_{D2D}$ values. Hence, especially for large $R_{D2D}$’s, EPDC algorithm becomes more energy efficient than the alternative ones.

**VII. CONCLUSIONS**

D2D networking is an attractive technique for improving system energy efficiency. Similarly, caching is a beneficial scheme used for boosting system capacity and reducing energy consumption in wireless networks. To this end, we study the cache replacement management in cellular D2D networks in this work. We formulate and solve the optimal caching problem minimizing the energy consumption in cellular D2D network for the video content consumption scenario. Besides, we propose an energy prioritized D2D caching algorithm considering the prospective energy consumption with the aim of reducing the overall system energy burden. For rigorous system analysis, we inspect the service rate and energy consumption of different service modes with respect to varying device cache capacity and D2D transmission range. Based on our evaluations, our proposed algorithm EPDC provides network-wide energy efficiency gains compared to alternative algorithms, especially for larger D2D transmission ranges.

**ACKNOWLEDGMENTS**

This work was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under grant number 116E245.

**REFERENCES**

[1] Cisco. “Cisco annual internet report (2018–2023) white paper,” 2020.
[2] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, “In-network caching effect on optimal energy consumption in content-centric networking,” 2012 IEEE International Conference on Communications (ICC), June 2012, pp. 2889–2894.
[3] Y. An and X. Luo, “An in-network caching scheme based on energy efficiency for content-centric networks,” IEEE Access, vol. 6, pp. 20184–20194, 2018.
[4] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, “An energy-efficient distributed in-network caching scheme for green content-centric networks,” Computer Networks, vol. 78, pp. 119 – 129, 2015.
[5] D. Prerna, R. Tekchandani, and N. Kumar, “Device-to-device content caching techniques in 5G: A taxonomy, solutions, and challenges,” Computer Communications, vol. 153, pp. 48 – 84, 2020.
[6] Y. Pan, M. Chen, N. Huang, and Y. Wang, “Power-efficient d2d relaying cellular network access,” Transactions on Emerging Telecommunications Technologies, vol. 30, no. 3, p. e3576, 2019, e3576 et.3576. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3576.
[7] Z. Zhang, L. Wang, D. Liu, and Y. Zhang, “Peer discovery for D2D communications based on social attribute and service attribute,” Journal of Network and Computer Applications, vol. 86, pp. 82 – 91, 2017.
[8] G. Gür and S. Kafioglu, “Layered content delivery over satellite integrated cognitive radio networks,” IEEE Wireless Communications Letters, vol. 6, no. 3, pp. 390–393, June 2017.
[9] S. S. Kafioglu, G. Gür, and F. Alagöz, “Multidimensional content modeling and caching in D2D edge networks,” 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Sep. 2019, pp. 1–6.
[10] Z. Chen, N. Pappas, and M. Kontoulis, “Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal,” IEEE Communications Letters, vol. 21, no. 3, pp. 584–587, March 2017.
[11] C. Jarray and A. Giovanidis, “Successful file transmission in mobile D2D networks with caches,” Computer Networks, vol. 147, pp. 162 – 179, 2018.
[12] A. S. Ali, K. R. Mahmoud, and K. M. Naguib, “Optimal caching policy for wireless content delivery in D2D networks,” Journal of Network and Computer Applications, vol. 150, p. 102467, 2020.
[13] N. Zhao, X. Liu, Y. Chen, S. Zhang, Z. Li, B. Chen, and M. Alouini, “Caching D2D connections in small-cell networks,” IEEE Transactions on Vehicular Technology, vol. 67, no. 12, pp. 12326–12338, Dec 2018.
[14] G. Gür, “Energy-aware cache management at the wireless network edge for information-centric operation,” Journal of Network and Computer Applications, vol. 57, pp. 33 – 42, 2015.
[15] M. Lee and A. F. Molisch, “Caching policy and cooperation distance design for base station-assisted wireless D2D caching networks: Throughput and energy efficiency optimization and tradeoff,” IEEE Transactions on Wireless Communications, vol. 17, no. 11, pp. 7500–7514, Nov 2018.
[16] M.-C. Lee, M. Ji, A. Molisch, and N. Sastry, “Throughput-outage analysis and evaluation of cache-aided D2D networks with measured popularity distributions,” IEEE Transactions on Wireless Communications, vol. PP, pp. 1–1, 2019.
[17] B. Chen and C. Yang, “Caching policy for cache-enabled D2D communications by learning user preference,” IEEE Transactions on Communications, vol. 66, no. 12, pp. 6586–6601, Dec 2018.
[18] Z. Li, J. Chen, and Z. Zhang, “Socially aware caching in D2D enabled fog radio access networks,” IEEE Access, vol. 7, pp. 84293–84303, Dec 2019.
[19] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, “An overview of sustainable green 5G networks,” IEEE Wireless Communications, vol. 24, no. 4, pp. 72–80, Aug 2017.
For the calculation of the prospective overall expected energy consumption $E(u)$ of any content unit $u$ that is mapped to some content unit tuple $\{i,j,k\}$, we utilize content unit availability probabilities. The availability probability in the local cache of a requester of some arbitrary content unit $u$ (with the one-to-one and onto correspondence content unit tuple $\{i,j,k\}$) is defined in (10) as $p_{loc}^{(u)}$. The multiplication of the $s^\text{th}$ content request probability $p_s$, the $j^\text{th}$ chunk request probability $p_j$ and the $k^\text{th}$ layer request probability $p_k$ gives request probability of content unit tuple $\{i,j,k\}$ (content unit $u$). Aside from the request characteristic, the requester device storage capability is eminently prominent for the calculation of the content unit availability probability at a requester device. For the integration of this aspect, the device capacity $C_{Dev}$ is normalized over the total size of all content space ($\sum_{all} s(u)$) by $F_{fit}$ function in (11).

\begin{equation}
p_{loc}^{(u)} := p_s \cdot p_j \cdot p_k \cdot F_{fit}^{loc}
\end{equation}

\begin{equation}
F_{fit}^{loc} := \frac{C_{Dev}}{\sum_{all} s(u)}
\end{equation}

The availability probability of some arbitrary content unit $u$ in the BS cache $p_{BS}^{(u)}$ in (12) is calculated similarly. The corresponding normalization function $F_{BS}$ that outputs the BS storage capability is given in (13).

\begin{equation}
p_{BS}^{(u)} := p_s \cdot p_j \cdot p_k \cdot F_{BS}^{fit}
\end{equation}

\begin{equation}
F_{BS}^{fit} := \frac{C_{BS}}{\sum_{all} s(u)}
\end{equation}

$N_{ngh}$ is defined as the total number of devices that are at most $R_{D2D}$ away from a requester. The availability probability of some content unit $u$ in at least one of the neighbour devices of a requester $p_{D2D}^{(u)}$ is defined in (14). $(1-p_{loc}^{(u)})^{N_{ngh}}$ gives the probability of a requested content unit $u$ not being available in any neighbouring device of a requester. By the complement, we obtain the availability probability of some content unit $u$ being in at least one of the neighbours of a requester device.

\begin{equation}
p_{D2D}^{(u)} := 1 - (1 - p_{loc}^{(u)})^{N_{ngh}}
\end{equation}