Optimal Resource Allocation for Content Delivery in D2D Communications

Can Güven*, Suzan Bayhan‡, Gürkan Gür†, and Salim Eryigit§
†TETAM, *Department of Computer Engineering, Bogazici University, Turkey
E-mail: {can.guven, gurgurka}@boun.edu.tr
‡Technische Universität Berlin, Germany, e-mail: bayhan@tkn.tu-berlin.de
§ AirTies Research Division, Istanbul, Turkey, e-mail: salim.eryigit@airties.com

***Abstract*—Future wireless networks face a great challenge in spectral resource management to meet an overwhelming demand for network capacity. In 5G systems, Device-to-Device (D2D) communications is a key technology to alleviate this "capacity crunch" while content consumption is the key usage mode. In this work, we devise a resource allocation problem for a cellular network that facilitates the delivery of requested contents to its users via either BS mode or D2D mode. By solving the formulated optimization problem, we investigate the interplay between D2D transmissions, cache characteristics, and mode selection preference. Our numerical results suggest that while enabling D2D operation improves delivery performance significantly, how much a network can deliver in D2D mode is determined by both the network density and cache capacity.**

I. Introduction

The disparity between the relatively-stable increase in network capacity and almost exponential surge in traffic demand has ignited several remedial approaches. One approach known as *information-centric networking* (ICN) is motivated by the fact that current use of the Internet is heavily content-oriented, e.g., video accounts for a big share of the current Internet traffic. To improve resource efficiency of content delivery, ICN proposes to drastically alter the underlying design principles of current networks; an ICN decouples the content from its location and each network node storing a copy of a content can act as a provider for this content. Such an approach is in stark contrast with current Internet where every communication is abstracted as an interaction between two end points.

Another approach, known as the fifth generation (5G) networks, is composed of a family of solutions to address the *1000x challenge*, which posits that current mobile networks must be improved 1000x to meet the demand of future networks [1]. This is elementally driven by the fact that wireless and mobile devices are expected to generate two-thirds of total IP traffic by 2020 [4]. In that regard, one key component of 5G is device-to-device communications (D2D). The main rationale behind D2D is that spectral reuse efficiency can be improved by decreasing the distance between the transmitter-receiver pairs. Thus, D2D proposes to use short-range radio with a restricted interference domain between peer devices rather than a long range link with a high-power transmitter.

Despite seeming drastically different, ICN and D2D share the following important merits: **1- Exploiting locality:** As user interests are highly correlated throughout time and space, known as *temporal locality* [9] and *spatial locality* [10] respectively, a user can retrieve the content it requests in D2D mode from a nearby node who might have already fetched it from the network. Similarly, in an ICN, a router in the core network or a node at the edge can search and fetch the requested content from its vicinity. **2- Decreasing network traffic:** As a result of increased local transmissions, the distance between the content consumer and its provider decreases, which in turn translates into lower network traffic and overhead. For an ICN, this decrease may be reflected as lower cost for a network provider. Similarly, for a D2D setting, the backhaul traffic between the Base Station (BS) and the core network of the provider is expected to be lower.

978-1-5386-3531-5/17/$31.00 © 2017 IEEE

While ICN and D2D share these traits in their kernels, there is a disconnect between ICN and D2D-based solutions. In this work, we aim to alleviate this divergence by modelling the content delivery in a cellular network using an ICN approach. Particularly, our key contributions are as follows:

- We model a resource allocation problem at a BS that facilitates the delivery of requested contents to its users via either BS mode or D2D mode. Our problem is content-aware in the sense that users share their cache information with the BS, which then aims at utilizing the D2D links as much as possible.
- We investigate the interplay between D2D transmissions and cache characteristics. This insight is crucial to co-design efficient D2D and caching policies for future networks employing these technologies.
- We illustrate the operational regime of the system regarding mode preference (BS mode vs. D2D transmissions) in our resource allocation problem. This weight-based framework is flexible and facilitates policy design for "D2D intensity" in content-centric 5G networks.

II. System Model

We consider a cellular network consisting of a BS and N user devices (Fig. 1). We denote the users by u_i with their indices ranging from 1 to N. Each device is capable of both receiving the cellular signal from the BS and transmitting/receiving in D2D mode. We assume that the BS has F frequencies. User devices are equipped with local storage and each content takes up a single unit space in the local storage, i.e., caches, with capacity of W content items. We assume that each user device follows some caching policy (LRU) to decide on which

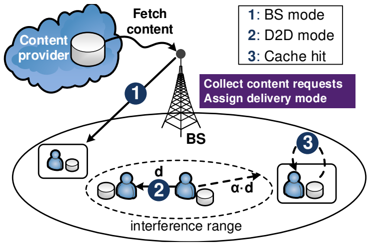


Fig. 1. System model.

item to cache or evict in the event of competition for the cache space.

Content Provider (CP) has a set of K diverse content objects. Each content object c_k is defined by its *content popularity* p_k, which represents the probability of a request to be for this particular content. We assume that this is time-invariant for the system time-scale during the algorithm's operation. A user can be a CP for an object available in its local cache or a content requester if it consumes that content. Additionally, BS is the content provider for all content objects, i.e. acts as universal source by retrieving the requested content from the CP. For notational brevity, we denote content requester by u_i and content provider by u_j, and u_0 represents the BS.

We denote the content requests of users by *demand matrix* $\mathbf{D} = [D_{ik}]$ where $D_{ik} = 1$ indicates u_i requests c_k. Similarly, *content provider matrix* $\mathbf{C} = [C_{jk}]$ stores whether u_j stores c_k locally or not. This information of $\mathbf{C}$ is generated/updated at the BS after being notified of the changes in users' cache. Each user's D2D range is d_{tx} and thereby it can only retrieve content from the users which are in that range. We assume that each time slot is sufficiently long to retrieve the intended content. Moreover, BS must assign resources such that users' communications do not interfere with other users in its transmission range. Let $\mathbf{T} = [T_{i,j}]$ be the transmission range indicator matrix where $T_{i,j} = 1$ means u_i is within transmission range of u_j. For $j = 0$ (BS), all other users are in transmission range, i.e., $T_{i,0} = 1$, $\forall i$. Let $\mathcal{R}_{ij}$ be the set of users that are within the interference range of u_j when it transmits to u_i. We assume that both the nodes and the BS have power control capability. The interference range is a multiple α of the transmission distance between the transmitter and the receiver $d_{i,j}$. More formally, we define $\mathcal{R}_{ij}$ as $\mathcal{R}_{ij} \coloneqq \{\, u_m \mid d_{j,m} \leq \alpha d_{i,j} \,\}$.

There are four possibilities regarding the user's content request. First, the user has already cached the content in its cache and retrieves it from its local storage. We call this event as *cache hit* (i). Otherwise, the user sends its content request to the BS at the beginning of next time slot. At the start of each time slot, BS collects content requests from each user, and determines the appropriate content delivery mode for each user based on the available spectrum resources, content availability and node locations, considering interference limitations. The mode assignment is broadcast to users in the coverage area. BS may directly deliver the content to the device (ii), or it may assign a peer device from which this user can retrieve the content (iii), or it may skip content delivery (iv) for this time slot for the requesting user in case of insufficient spectrum resources. In *BS mode* (conventional cellular transmission), users retrieve the content directly from the BS using the assigned frequency. However, it is also possible for the users to directly receive the content from their peers using short-range radio interface through the assigned frequency (referred to as *D2D mode*). Short-range transmission is possible only if the other peer (u_j) has the content that is requested by u_i.

III. Weight-Based Optimal Frequency Assignment

In each time slot, the BS allocates its resources ensuring (i) users do not interfere with each other, and (ii) high user satisfaction in terms of content delivery success. The resource assignment has two components: (i) BS mode assignment and (ii) D2D mode assignment. To have a flexible resource assignment scheme, we define ω_{BS} and ω_{D2D} as the weights assigned to each delivery mode. Based on the ratio ω_{D2D}/ω_{BS}, our scheme can favor one mode and can use that transmission mode, e.g., D2D mode, more than the other. Please note that $\omega_{D2D} = 0$ means that resource allocation is done as in conventional cellular networks. In this general setting, we seek for the optimal resource allocation (F frequencies) to N users for the delivery of requested content among K content objects.

Let X_{ijkf} a binary decision variable denoting whether u_i retrieves content k from u_j at frequency F. We formulate resource allocation problem, so called *Weight-Based Optimal Frequency Assignment WOFAS*, as an integer linear programming (ILP) problem as follows:

$$\max \sum_{i=1}^{N} \Big(\omega_{BS} \sum_{k=1}^{K} \sum_{f=1}^{F} X_{i0kf} + \omega_{D2D} \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{f=1}^{F} X_{ijkf}\Big) \tag{1}$$

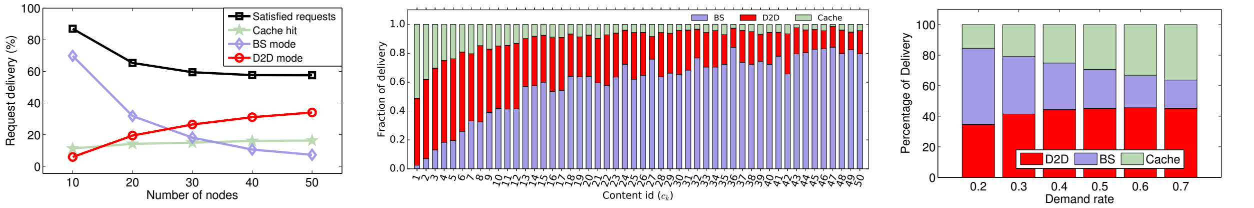
$$s.t. \quad \sum_{j=0}^{N} \sum_{k=1}^{K} \sum_{f=1}^{F} X_{ijkf} + \sum_{j=1}^{N} \sum_{k=1}^{K} \sum_{f=1}^{F} X_{jikf} \leq 1, \; \forall i \tag{2}$$

$$\sum_{f=1}^{F} X_{ijkf} \leq D_{ik} C_{jk} T_{ij}, \qquad \forall i, j, k \tag{3}$$

$$\sum_{i=1}^{N} \sum_{m \in \mathcal{R}_{ij}} \sum_{k=1}^{K} X_{imkf} \leq (1 - \sum_{k=1}^{K} X_{ijkf}), \forall i, j, f \tag{4}$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} X_{i0kf} \leq 1 \qquad \forall f {=} \{1, \cdots, F\} \tag{5}$$

Objective in (1) shows the weighted total resource assignment to users. Const. (2) is due to half-duplex operation of devices, i.e., a device can either receive or transmit at a time. Const. (3) states the conditions of short-range content delivery: (i) assigned content provider must have the content, (ii) the user requests the content and (iii) they are within transmission range of each other. Interference-free transmission is guaranteed by (4) which states that if u_j transmits any content to u_i at frequency F, then no node within the interference range of u_j can transmit on the same channel. Finally,



(a) Percentage of requests delivered with increasing number of nodes.

(b) Content popularity vs. fraction of successful delivery.

(c) Demand rate vs. percentage of successful delivery.

Fig. 2. Numerical results based on various parameters and metrics.

TABLE I
SYSTEM PARAMETERS.

Parameter	Value	Explanation
r	200 m	Cell radius
tr	100 m	Transmission range of each node
α	2	Interference range multiplier
N	30	Numer of Nodes
W	5	Cache size
F	3	# of frequencies for downlink
β	0.7	Zipf parameter
K	50	# of unique content objects
λ	0.4 reqs per sec	Request generation (demand) rate
TTL	5 timeslots	Time-to-live for a request
C	3 contents	Cache size of a node

Const. (5) ensures that BS can only serve one user on a given channel.

IV. Numerical Results and Discussion

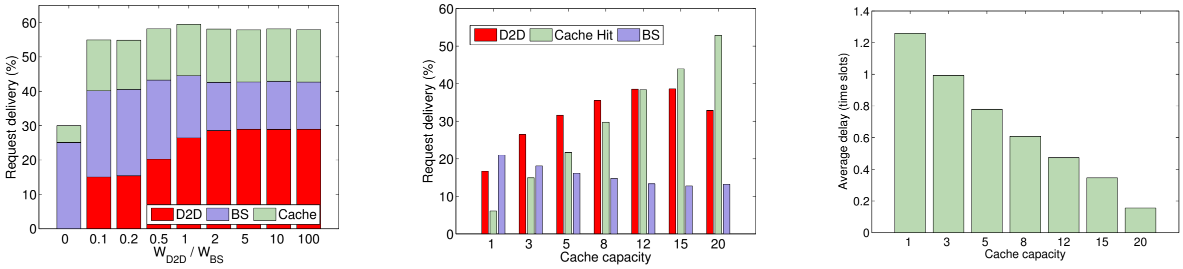
As finding the optimal solution for the considered content delivery problem is both time consuming and resource hungry, we consider a relatively small-scale network in our experiments. Our system consists of randomly deployed nodes in a cell. The parameters are listed in Table I. We increase the network size, i.e., number of nodes, gradually to see the impact of node density and number of neighbors on content delivery.

First, we evaluate how the system performance changes with increasing node density and which transmission modes are used for content delivery. We set $\omega_{D2D} = \omega_{BS} = 1$ in this setting. As Fig.2(a) shows, the system can successfully deliver between 59% to 89% of the requests generated by the nodes in general. Since $F = 3$ is fixed while the node density increases, the system can deliver only some of the requests generated by the nodes. However, the density increase vs. satisfied request decrease is not linear due to the increasing D2D capacity of the network. Around 11% to 17% of these requests are delivered from the devices' own caches, which can store 6% of the contents. The minor relative increase in cache hits is due to the decrease in total delivery success. The fraction of cache hit depends on the Zipf parameter β [2]; a larger Zipf parameter refers to a scenario where most of the requests are for only a few content objects which can mostly be found in a node's cache. Hence, we expect higher cache hits for larger β.

As the number of nodes in the system increases, so does the node density in the area and the average number of neighbors. More neighbors facilitate increased opportunity for finding a content within D2D transmission range, and the results reflect this insight. The fraction of successful delivery is dominated by BS mode for 10 nodes, where each node has 0.4 neighbors on average. At this point, D2D mode performs slightly worse than a device's own cache. As the number of nodes increases, D2D mode starts to take over content delivery, reaching 35% of all transmissions when the number of nodes are 50 and each node has approximately 11 neighbors on average. Since the BS can deliver up to three contents in a time slot, the fraction of delivery by BS mode decreases as network density increases. In a system of limited resources, when there are devices within transmission range with sufficient content coverage and the demand follows a Zipf distribution, the results demonstrate the contribution potential of D2D mode to content delivery. Fig.2(a) depicts the fraction of successful delivery in a network under heavy load where each node has a demand rate of $\lambda = 0.4$. For $\lambda = 0.1$ (not plotted), where the request delivery percentage is above 99% in the network, the general principles and trends are similar. As the number of nodes accessible per node increases, request delivery via D2D mode increases and BS mode delivery decreases while cache hits for the same Zipf parameter stay almost the same.

To understand how each content is delivered, i.e., BS mode, D2D mode, or directly from cache, Fig.2(b) shows the fraction of delivery mode per content. We have a moderately dense system of 30 nodes for this scenario. In the x-axis, as the content id increases, content popularity decreases. From this figure, we can gain the following insights: (i) BS mode is the most dominant mode in content delivery for the majority of contents and the fraction of successful delivery via this mode increases as the content popularity decreases. (ii) For delivering moderately high to most popular content, D2D mode is ideal. (iii) Using an LRU caching model, devices can retrieve the most popular content using their caches. As the popularity decreases, there is a sharp decline in the fraction of successful content retrieval via a device's own cache.

Fig.2(c) shows the changes in how the content is delivered with increasing demand rate for the same setting. Apparently, as the demand increases, content is delivered more via cache hits and D2D mode, thus less with BS mode. When the demand rate of each node in the system increases from 0.2 to 0.7, content delivery via D2D mode rises from 34.6% to 45.2% and cache hits from 15.5% to 36.2% whereas BS mode delivery drops from 49.9% to 18.6%. In line with the behavior in Fig.2(b), this means that moderate to highly popular content constitute a larger portion of deliveries under increased load in the network. We should also note that with increasing demand, the percentage of total satisfied requests drops from 81.6% to 49.2%, largely due to the resource limitation in BS mode transmissions. Meanwhile, D2D



(a) Impact of weights on delivery modes.

(b) Impact of cache size on delivery modes.

(c) Impact of cache size on delivery delay.

Fig. 3. Impacts of system settings on D2D mode and content delivery.

transmissions drop only from 28.4% to 22.3%, with the amount of D2D mode deliveries actually increasing by almost 2.7 times. This behavior shows that D2D mode is robust for increasing demand and can scale its contribution to content delivery under heavy load as long as devices have accessible neighbors in their vicinity.

A. *Impact of D2D and BS mode weights*

In Fig.3(a), we can see the effect of various weights for D2D mode in the objective function. In case of $\omega_{D2D} = \omega_{BS}$, our solution attains the global maximum in terms of total network throughput. When $\omega_{D2D} > \omega_{BS}$, however, our system is operating at a sub-optimal point in terms of total network throughput, e.g., a local maximum, for preferring the amount of content delivery with D2D transmissions. For $\omega_{D2D} < \omega_{BS}$, the sub-optimal operating point conversely favors BS mode transmissions. We can see in the figure that, tipping the balance towards D2D, starting with $\omega_{D2D} = 2$, raises the fraction of deliveries made via D2D mode from 26% to over 28.6%. However, further increasing the weight does not have a noticeable impact on D2D mode delivery, since favoring D2D at the expense of new content transmissions from BS into the network affects the system in future time slots. BS delivering new content and thus increasing diversity in the device caches is as important as utilizing D2D transmission opportunities. There is also the fact that the maximum D2D capacity of the system in given simulation conditions is reached. On the other hand, when $\omega_{D2D} < 1$, favoring BS mode, there is a decline in the fraction of successful delivery via D2D mode. However, introducing D2D to the system, even with a small weight, means that an extra 15% of content delivery can be done via D2D mode, given that each node has sufficient neighbors and content variety to utilize D2D capacity.

When the system does not have D2D mode (i.e., $\omega_{D2D} = 0$), another result that can be observed is that cache hits are 4.92% of satisfied content requests. With D2D mode enabled, cache hits can deliver 14.8% of requests. This result directly follows from the fact that the D2D-enabled system can deliver 25 to 30% more content and distribute popular content among neighbors, thus improving cache hits. The figure also illustrates the fact that the scenario where D2D mode and BS mode having equal weights achieves the highest throughput possible for the system. Looking at the fraction of BS mode transmissions in content delivery, it can be observed that increasing the weight of D2D in the objective reduces the ratio of BS mode deliveries. Therefore, favoring D2D mode can have a significant impact on reducing the amount of backhaul traffic driven by the BS. To summarize, the weighted objective renders the upper and lower bounds of D2D mode content delivery for the adopted setting with 30 nodes and 5.3 neighbors per node on average, each having LRU caches. The other observation is on the effects of having an emphasis on either BS or D2D mode transmissions.

B. *Impact of cache size*

For evaluating the impact of cache size, we set cache capacity for each node to $\{2, 6, 10, 16, 24, 30, 40\}$ percents of the content catalogue. Fig.3(b) plots the effect of cache capacity on the three transmission modes—BS, D2D mode delivery, and cache hits. The first observation is that until a certain cache size (e.g., $W = 8$), the ratio of delivery via D2D mode increases with increasing cache capacity. Similarly, content delivery via cache hits increases with larger cache capacity. The linearity of ascending trend is in accordance with the content popularity distribution. However, the fraction of successful transmission by D2D mode benefits most from increases in cache size when it is under $W = 8$, only to begin saturating above the point where a node's neighborhood contains over 50% of all available content. This happens due to the D2D potential of the system being limited by the availability of frequencies to deliver content. The channels are used both to transmit new content from the BS and from the neighbor nodes, subject to interference range constraints. Furthermore, for cache sizes over $W = 15$, the system starts to transmit less via D2D since each node's own cache starts to host more than 30% of content at this point, and cache hits start to overtake. It should be noted that the content coverage of a node's all accessible neighbors on average for the respective cache capacities are 8.4%, 21.4%, 32.8%, 47.8%, 64.7%, 74.4% and 84.7%. Finally, we can remark that with a linear increase in cache size, we observe a linear decrease in the BS mode deliveries and hence backhaul traffic.

Fig.3(c) plots average delay of successfully delivered content with increasing cache size. As cache size increases, the delay on content delivery decreases. The decline in delay is inversely related to the increase in D2D mode deliveries and cache hits observed in Fig. 3(b). The results show that the node caches and their counterparts in neighboring nodes can be effectively utilized to reduce delay along with reducing the backhaul traffic and spectrum congestion.

V. Related Work

The potential of caching for resource efficiency in mobile networks has been widely explored, e.g. [3], [11]. While caching closer to the users, e.g., at the eNBs, is more desirable for quicker content access, its potential for cache hits is lower since eNBs are limited in their storage and serve a limited number of users. Putting the content in the upper layers in the mobile network hierarchy, e.g., network core, improves the resource efficiency at the expense of longer latency. Motivated by these pros and cons of each tier, [11] overviews caching at different tiers, e.g., at the access points, the mobile core, in a mobile network, and proposes to optimally assign which objects to cache at each tier to minimize the data access delay under cache capacity constraints. Our work differs from [11] in that we consider D2D caching along with the resource allocation problem in a BS cell coverage.

There is also some research, e.g., [3], [6], [8] on pushing the caching towards the edge of the network. The most relevant research to ours is [3] which proposes to employ some mobile nodes as data carriers (named as *helpers*) and use these nodes as mobile caches which are strictly controlled by the BS. The BS manages content resolution after receiving requests from users in its coverage and actively pushes popular content to the helpers. Moreover, the BS selects some nodes as helpers among the candidate nodes based on their mobility profiles, i.e., nodes with higher probability to spend their time in point-of-interests longer are more likely to be selected as helpers determined by their mobility patterns.

In a similar spirit, [6] explores the dynamics of D2D content delivery with a focus on user mobility as data transmission may abruptly terminate due to the mobility of either peer. Authors provide a theoretical model for successful D2D transmission considering the duration of contact period of a transmitter-receiver pair, which is a function of node mobility and file size. Our work distinguishes itself from [6] in several ways. First, we account for the multi-user interference while calculating the D2D transmission capacity which is neglected in [6]. Second, we formulate an optimal resource allocation scheme, i.e., mode selection, with a goal of utilizing D2D links as much as possible. This objective is in line with 5G and ICN, both aiming at decreasing redundant traffic at the core and also at the edge of the network. Correlatively, [5] considers mode selection and content popularity/caching with similar aims, however is scope-limited to satellite integrated cognitive radio networks and explore mainly a state-based system model. Similar to our proposal, [7] aims at identifying the optimal mode for content delivery among the three delivery options, directly from the BS, D2D direct, and D2D opportunistic multi-hop delivery. Assuming that BS knows the trajectory of every mobile node, it can create a time-expanded graph representing all connections between nodes and next solves a max-flow problem on this graph to optimally select one of the delivery modes. While the presented analysis driven from the proposed approach provides insight on D2D potential of a cellular network, it does not elaborate on the effect of content dynamics. Our work differs from [7] mainly in this aspect.

VI. Conclusion

We have proposed a resource allocation scheme for content delivery in D2D-supported networks using a linear integer problem model. Our model supports D2D mode delivery, weighted if desired, under an omnidirectional interference constraint. We have showed that the inclusion of a D2D mode can have significant impact on the amount of content delivered in a network, even with interference-limited channel resources. As the weight given on D2D increases, the system caps out at a certain point where performing more D2D is impossible due to the competition for the spectral resources with new content delivery from BS and the limited-capacity device caches. Another result deduced from analyzing various scenarios was the usage of delivery mode based on content popularity, where highly popular content are delivered mostly from cache or via D2D mode and the less popular content are delivered by the BS (creating backhaul traffic). This phenomenon lights the way for designing clever caching schemes or D2D mode selection for systems that are also time-bound for calculating which content to transmit and what mode to use.

Acknowledgment

This work was supported in part by the Scientific and Technical Research Council of Turkey (TUBITAK) under Grant 116E245, and in part by the Turkish State Planning Organization (DPT) through the TAM Project under Grant 2007K120610.

References

[1] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran. 5G roadmap: 10 key enabling technologies. *Computer Networks*, 106:17–48, 2016.

[2] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: evidence and implications. In *IEEE INFOCOM '99*, volume 1, pages 126–134 vol.1, Mar 1999.

[3] G. Chandrasekaran, N. Wang, and R. Tafazolli. Caching on the move: towards D2D-based information centric networking for mobile content distribution. In *IEEE Conference on Local Computer Networks (LCN)*, pages 312–320, 2015.

[4] Cisco. Cisco visual networking index: Forecast and methodology; 2015–2020, 2016.

[5] G. Gür and S. Kafiloğlu. Layered content delivery over satellite integrated cognitive radio networks. *IEEE Wireless Communications Letters*, 6(3):390–393, June 2017.

[6] C. Jarray and A. Giovanidis. The effects of mobility on the hit performance of cached D2D networks. *arXiv:1603.02927*, 2016.

[7] Y. Li, Z. Wang, D. Jin, and S. Chen. Optimal mobile content downloading in D2D communication underlaying cellular networks. *IEEE Transactions on Wireless Communications*, 13(7):3596–3608, 2014.

[8] D. Liu, B. Chen, C. Yang, and A. F. Molisch. Caching at the wireless edge: design aspects, challenges, and future directions. *IEEE Communications Magazine*, 54(September):22–28, 2016.

[9] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini. Temporal locality in today's content caching: why it matters and how to model it. *ACM SIGCOMM CCR*, 43(5):5–12, 2013.

[10] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini. Unravelling the Impact of Temporal and Geographical Locality in Content Caching Systems. *ArXiv e-prints*, Jan. 2015.

[11] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. Leung. Cache in the air: exploiting content caching and delivery techniques for 5G systems. *IEEE Communications Magazine*, 52(2):131–139, 2014.