Wireless Device-to-Device Caching Networks: Basic Principles and System Performance

Mingyue Ji, *Student Member, IEEE*, Giuseppe Caire, *Fellow, IEEE*, and Andreas F. Molisch, *Fellow, IEEE*

***Abstract*—As wireless video is the fastest growing form of data traffic, methods for spectrally efficient on-demand wireless video streaming are essential to both service providers and users. A key property of video on-demand is the *asynchronous content reuse*, such that a few popular files account for a large part of the traffic but are viewed by users at different times. Caching of content on wireless devices in conjunction with device-to-device (D2D) communications allows to exploit this property, and provide a network throughput that is significantly in excess of both the conventional approach of unicasting from cellular base stations and the traditional D2D networks for "regular" data traffic. This paper presents in a *tutorial and concise form* some recent results on the throughput scaling laws of wireless networks with caching and asynchronous content reuse, contrasting the D2D approach with other alternative approaches such as conventional unicasting, *harmonic broadcasting*, and a novel *coded multicasting* approach based on caching in the user devices and network-coded transmission from the cellular base station only. Somehow surprisingly, the D2D scheme with spatial reuse and simple decentralized random caching achieves the same near-optimal throughput scaling law as coded multicasting. Both schemes achieve an unbounded throughput gain (in terms of scaling law) with respect to conventional unicasting and harmonic broadcasting, in the relevant regime where the number of video files in the library is smaller than the total size of the distributed cache capacity in the network. To better understand the relative merits of these competing approaches, we consider a holistic D2D system design incorporating traditional microwave (2 GHz) and millimeter-wave (mm-wave) D2D links; the direct connections to the base station can be used to provide those rare video requests that cannot be found in local caches. We provide extensive simulation results under a variety of system settings and compare our scheme with the systems that exploit transmission from the base station only. We show that, also in realistic conditions and nonasymptotic regimes, the proposed D2D approach offers very significant throughput gains.**

***Index Terms*—Device-to-device communication, millimeter-wave communication, wireless caching networks, throughput-outage tradeoff, system design.**

I. Introduction

WIRELESS data traffic has dramatically increased over the past few years. Mainly driven by on-demand video streaming, it is expected to further grow from today's level by almost two orders of magnitude in the next five years [1]. Traditional approaches for coping with this growth are increasing spectral resources (bandwidth), spectral efficiency (modulation, coding, MIMO), or spatial reuse (density of base stations). However, these methods either provide only limited throughput gains in practical conditions [2], [3] or are expensive to implement. In particular, while heterogeneous networks with a large number of small cells can provide high area spectral efficiency [4], the necessity of high-speed backhaul makes this option prohibitively expensive.

It is noteworthy that current methods for on-demand video streaming treat video like individual data sources with (possibly) adaptive rate. Namely, each video streaming session is handled as a unicast transmission, where users successively download video "chunks"[1] as if they were web-pages, using HTTP, with possible adaptation the video quality according to the conditions of the underlying TCP/IP connection (e.g., Microsoft Smooth Streaming and Apple HTTP Live Streaming [5]–[7]) This approach does not exploit one of the most important properties of video, namely, a constrained request pattern. In other words, the same video is requested by different users, though the requests usually occur at different times. For example, video services such as Amazon or Netflix provide a finite (albeit large) library of video files to the users and, in some cases, may shape the request pattern by making some videos available free of charge. It should also be noted that *naive multicasting* by overhearing, i.e., by exploiting the broadcast nature of the wireless medium, is basically useless for wireless video on-demand. In fact, while the users' requests exhibit a very significant *content reuse* (i.e., the same popular files are requested over and over), the asynchronism between such requests is so large that the probability that two users are streaming the same file at the "same time" (i.e., within a relative delay of a few seconds) is basically zero. We refer to this very peculiar feature of video on-demand as the *asynchronous content reuse*.

Over the years, a number of other suggestions have been made to make better use of constrained request patterns: [8]–[11] considers the case that users want the same video at the same time (e.g., in a live streaming service) but with a different channel quality or requested video quality. In this case, scalable video coding can be coupled with some form of broadcast channel coding [12]. Specifically, scalable rateless codes [13]–[15] to support heterogeneous users in a broadcast

Manuscript received May 20, 2013; revised November 18, 2013; accepted May 17, 2015. Date of publication July 6, 2015; date of current version December 15, 2015. This paper was supported in part by a grant from Intel/Cisco/Verizon through their Video-Aware Wireless Network (VAWN) Program and in part by the U.S. National Science Foundation under Grants CCF-1423140 and CNS-1457340.

The authors are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: mingyuej@usc.edu; caire@usc.edu; molisch@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2015.2452672

[1]Typically a video chunk corresponds to 0.5 s to 1 s of encoded video, i.e., to a group of pictures (GOP) between 15 and 30 frames for a typical video playback rate of 30 frames per second.

channel scenario are considered in [8], [10], [11]. Another set of recent works considers the case where neighboring wireless users want the *same* video at the same time, and collaborate in order to improve their aggregate downlink throughput. In particular, [16] suggests that different users download simultaneously different parts of the same video file from the serving base station and then share them by using device-to-device (D2D) communications.

The above approaches are suited for synchronous streaming of live events (e.g., live sport events) but yield no gain in the presence of asynchronous content reuse, characteristic of on-demand video streaming. On the other hand, treating each user request as independent data yields a fundamental bottleneck: in conventional unicasting from a single serving base station, the per-user throughput decreases linearly with the number of users in the system. In [17]–[20], a coding scheme referred to as *harmonic broadcasting* is introduced. This scheme can handle asynchronous users requesting the *same* video at different times, such that each user can start playback within a small delay from its request time. With harmonic broadcasting, a video encoded at rate R requires a total downlink throughput of $R\log(L/\tau)$, where L is the total length of the video file and τ is the maximum playback delay. For $\tau \ll L$ (as it is required in on-demand video streaming), the bandwidth expansion incurred by harmonic broadcasting can be very significant.

Recent work by the authors, as well as by other research groups, has shown that one of the most promising approaches relies on *caching*, i.e., storing the video files in the users' local caches and/or in dedicated helper nodes distributed in the network coverage area. From the results of [21]–[24], we observe that caching can give significant (order) gains in terms of throughput. Intuitively, caching provides a way to exploit the inherent content reuse of on-demand video streaming while coping with asynchronism of the requests. Also, caching is appealing since it leverages the wireless devices storage capacity, which is probably the cheapest and most rapidly growing network resource that, so far, has been left almost untapped.

One possible approach consists of "Femtocaching," i.e., of deploying a large number of dedicated "helper nodes," which cache popular video files and serve the users' requests through local short-range links. Essentially, such helper nodes are small base stations that use caching in order to replace the backhaul, and thus obviate the need for the most expensive part of a small cell infrastructure [21]. Another recently suggested method combines caching of files on the user devices with a common multicast transmission of network-coded data [25]. We refer to this approach as *coded multicasting*. The third approach, which is at the center of this paper, combines caching of files on the user devices with short-range device-to-device (D2D) communications [22]. In this way, the caches of multiple devices form a *common virtual cache* that can store a large number of video files, even if the cache on each separate device is not necessarily very large. Both coded multicasting and D2D caching have a common interesting feature: the common virtual cache capacity grows linearly with the number of users in the system. This means that, as the number of users in the network grows, also their aggregate cache capacity grows accordingly. We shall see that, qualitatively, this is the key property that allows for significant gains with respect to the other methods reviewed here, where the content is only stored in the network infrastructure.

The purpose of this paper is two-fold. On one hand, we provide a tutorial overview of the schemes and recent results on wireless on-demand video streaming summarized above, in terms of their throughput vs. outage probability tradeoff, in the regime where both the number of users in the system and the size of the library of video files grow large. While the results presented in Section II are not new, they have been established mostly in individual papers with different assumptions and notations; the tutorial summary presented in Section II is intended to allow a fast and fair comparison under idealized settings. On the other hand, looking at throughput-outage tradeoff scaling laws for idealized network models does not tell the whole story about the relative ranking of the various schemes. Hence, in this work we present a detailed and realistic model of a single cell system with n users, each of which has a cache memory of M files, and place independent streaming requests to a library of m files. Requests can be served by the cellular base station, and/or by D2D links. We make realistic assumptions on the channel models for the cellular links and the D2D links, assuming that the former uses a 4th generation cellular standard [26] and the latter use either microwave or mm-wave communications depending on availability [27], [28]. By means of extensive simulations, this paper relaxes some restrictive assumptions of the theoretical scaling laws analysis based on the "protocol model" of [29], and provides more in-depth *practical* results with the goal of assessing the true potential of the various methods in a realistic propagation environment, where the actual transmission rate of each link depends on physical quantities such as pathloss, shadowing, transmit power and interference. Furthermore, we study how the use of short-range mm-wave links can influence the overall capacity. Such links can provide very high rates but suffer from high outage probability in some environments such as office environment (see Section IV). We investigate a composite scheme that combines robust microwave D2D links with high-capacity mm-wave links in order to achieve, opportunistically, excellent system performance. We also show that the type of environment in which we operate, while irrelevant for the asymptotic scaling laws analysis, plays a major role for the actual system throughput and outage probability. Eventually, we shall show that, in such realistic conditions, the D2D caching scheme largely outperforms all other competing schemes both in terms of per-user throughput and in terms of outage probability.

The paper is organized as follows. Section II presents a literature review of the recent results on wireless caching networks, where the system model and the main theoretical results are summarized. Then the system design approach is presented in Section III and the simulation results are given in Section IV. Conclusion are pointed out in Section V.

II. Literature Review

In this section we review the most important recent results on the throughput of wireless caching networks. The emphasis

lies on results that use caching in combination with D2D communications, though we also review results for caching combined with BS-only transmission, as well as pure D2D communication (without caching).

A. Conventional Scaling Laws Results of Ad Hoc Networks and D2D Communications With Caching

The capacity of conventional ad hoc networks, where source-destination pairs are drawn at random with uniform probability over the network nodes, has been studied extensively. Under the protocol model (see Section II-B) and a decode-and-forward (i.e., packet forwarding) relaying strategy, the throughput per user of such networks scales as $\Theta(\frac{1}{\sqrt{n}})$, where n denotes the number of nodes (users) in the network. While the conclusions for realistic physical models including propagation pathloss and interference are more variegate [29]–[34], we can conclude that practical relaying schemes are limited by the same per-user throughput scaling bottleneck of $\Theta(\frac{1}{\sqrt{n}})$ which holds for the protocol model. Notice that this result assumes that the traffic generated by the network is $\Theta(n)$, i.e., constant requested throughput per user. This does not take into account the intrinsic content reuse of video on-demand streaming. In other words, when treating each session as independent data, the per-user throughput vanishes as the total demanded throughput increases.

Fortunately, the video-aware networks, i.e., networks designed to support video on-demand, can behave in a much better way. For this purpose, it is useful to consider another measure of network performance called transport capacity [30], which is the sum over each link of the product of the throughput per link times the distance between source and destination. It is known that the transport capacity of ad-hoc dense networks (i.e., networks of fixed area $O(1)$ with node density that scales as $\Theta(n)$), under the protocol model, or under a physical model with decode and forward relaying, scales as $\Theta(\sqrt{n})$. For random source-destination pairs, at distance $O(1)$, the throughput per link scales again as $\Theta(\frac{1}{\sqrt{n}})$ as mentioned before. On the other hand, if we can reduce the distance between the source (requested file) to the destination (requesting user) to the minimum distance between nodes ($\Theta(\frac{1}{\sqrt{n}})$), which corresponds to one hop, then a constant throughput per user can be achieved. The reason is that many short distance links can co-exist by sharing the same spectrum, which can be used more and more densely as the density of the network grows. In another word, by caching the files into the network such that request can be satisfied by short-range links, the spectrum spatial reuse of the network increases linearly with the number of users. Based on this observation, it is meaningful to consider a system design based on one-hop D2D transmission and caching of the video files into the user devices.

B. Network Model and Problem Definitions

In this section, we introduce the formal network model and the detailed problem definition for the uncoded D2D caching networks. We consider a network formed by user nodes $\mathcal{U} = \{1, \ldots, n\}$ placed on a regular grid on the unit square, with minimum distance $1/\sqrt{n}$ (see Fig. 1(a)).[2] Each user $u \in \mathcal{U}$ makes a request for a file $f \in \mathcal{F} = \{1, \ldots, m\}$ in an i.i.d. manner, according to a given request probability mass function $P_r(f)$. Communication between user nodes obeys the protocol model [29][3]: namely, communication between nodes u and v is possible if their distance $d(u, v)$ is not larger than some fixed range r, and if there is no other active transmitter within distance $(1 + \Delta)r$ from destination v, where $\Delta > 0$ is an interference control parameter. Successful transmissions take place at rate C_r bit/s/Hz, which is a non-increasing function of the transmission range r [21]. In this model we do not consider power control (which would allow different transmit powers, and thus transmission ranges). Rather, we treat r as a design parameter that can be set as a function of m and n.[4] All communications are single-hop. We assume that the request probability mass function $P_r(f)$ is the same for all users and follows a Zipf distribution with parameter $0 < \gamma_r < 1$ [35], i.e., $P_r(f) = \frac{f^{-\gamma_r}}{\sum_{i=1}^{m} \frac{1}{i^{\gamma_r}}}$. These model assumptions allow for a sharp analytical characterization of the throughput scaling law.

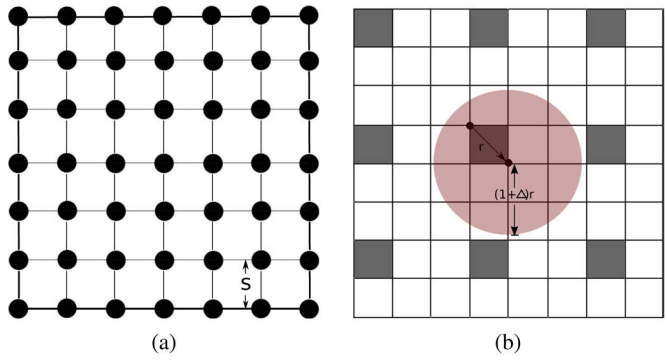


Fig. 1. a) Grid network with $n = 49$ nodes (black circles) with minimum separation $s = \frac{1}{\sqrt{n}}$. b) An example of single-cell layout and the interference avoidance TDMA scheme. In this figure, each square represents a cluster. The grey squares represent the concurrent transmitting clusters. The red area is the disk where the protocol model allows no other concurrent transmission. r is the worst case transmission range and Δ is the interference parameter. We assume a common r for all the transmitter-receiver pairs. In this particular example, the TDMA parameter is $K = 9$, which means that each cluster can be activated every 9 transmission scheduling slot durations.

We consider a simple "decentralized" random caching strategy, where each user caches M files selected independently from the library $\mathcal{F}$ with probability $P_c(f)$. On the practical side, video streaming is obtained by sequentially sending "chunks" of video, each of which corresponds to a fixed duration. The transmission scheduling slot duration, i.e., the duration of the physical layer slots, is generally two to three orders of magnitude shorter than the chunk playback duration (e.g., 2 ms versus 0.5 s [36]). Invoking a time-scale decomposition, and

[2]For some of the later simulations, we will also consider the case that nodes are uniformly and randomly distributed in a square region.

[3]In the simulations of Section IV, we relax the protocol model constraint and take interference into consideration by treating it like noise.

[4]Since the number of possibly requested files m typically varies with the number of users in the system n, and r can vary with n, r can also be a function of m.

provided that enough buffering is used at the receiving end, we can always match the average throughput per user (expressed in information bit/s) with the average source coding rate at which the video file can be streamed to a given user. Hence, while the chunk delivery time is fixed, the "quality" at which the video is streamed and reproduced at the user end depends on the user average throughput. Therefore, in this scenario, we are concerned with the ergodic (i.e., *long-term average*) throughput per user.

Referring to Fig. 1(b), the network is divided into clusters of equal size, denoted by $g_c(m)$ (number of nodes in each cluster) and independent of the users' requests and cache placement realization. A user can look for the requested file only inside its own cluster. If a user can find the requested file inside the cluster, we say there is one *potential link* in this cluster. We use an *interference avoidance* scheme for which at most one transmission is allowed in each cluster on any time-frequency slot (transmission resource). A system admission control scheme decides whether to serve potential links or ignore them. The served potential links in the same cluster are scheduled with equal probability (or, equivalently, in round robin), such that all admitted user requests have the same average throughput $\mathbb{E}[T_u] = \overline{T}_{\min}$, for all users u, where expectation is with respect to the random user requests, random caching, and the link scheduling policy (which may be randomized or deterministic, as a special case). To avoid interference between clusters, we use a time-frequency reuse scheme [37, Ch. 17] with parameter K as shown in Fig. 1(b). In particular, we can pick $K = (\lceil\sqrt{2}(1+\Delta)\rceil + 1)^2$, where Δ is the interference parameter in the protocol model.

Qualitatively (for formal definition see [22]), we say that a user is in outage if the user cannot be served by the D2D network. This can be caused by: (i) the file requested by the user is not in the user's own cluster, (ii) that the system admission control decides to ignore the request. We define the outage probability p_o as the average fraction of users in outage. At this point, we can define the throughput-outage tradeoff as follows:

Definition 1 (Throughput-Outage Tradeoff): For a given network and request probability mass function $\{P_r(f) : f \in \mathcal{F}\}$, an outage-throughput pair (p, t) is *achievable* if there exists a cache placement scheme and an admission control and transmission scheduling policy with outage probability $p_o \le p$ and minimum per-user average throughput $\overline{T}_{\min} \ge t$. The outage-throughput achievable region $\mathcal{T}(P_r, n, m)$ is the closure of all achievable outage-throughput pairs (p, t). In particular, we let $T^*(p) = \sup\{t : (p, t) \in \mathcal{T}(P_r, n, m)\}$. ◊

Notice that $T^*(p)$ is the result of the optimization problem:

$$\begin{aligned} &\text{maximize} && \overline{T}_{\min} \\ &\text{subject to} && p_o \le p, \end{aligned} \tag{1}$$

where the maximization is with respect to the cache placement and transmission policies. Hence, it is immediate to see that $T^*(p)$ is non-decreasing in p, since for given outage probability constraint p_1, p_2, the policies satisfying $p_2 > p_1$ are a superset of the policies satisfying p_1. The range of feasible outage probability, in general, is an interval $[p_{o,\min}, 1]$ for some $p_{o,\min} \ge 0$. Whether $p_{o,\min} = 0$ or strictly positive depends on the model assumptions.

C. Key Results for D2D Networks With Caching

The following results are proved in [22] and yield scaling laws of the optimal throughput-outage tradeoff under the clustering transmission scheme defined above. First, we characterize the optimal random caching distribution P_c:

Theorem 1: Under the model assumptions and the clustering scheme, the probability that any user $u \in \mathcal{U}$ finds its requested file inside its own cluster is maximized by the caching distribution

$$P_c^*(f) = \left[1 - \frac{\nu}{z_f}\right]^+, \quad f = 1, \ldots, m, \tag{2}$$

where $\nu = \frac{m^*-1}{\sum_{f=1}^{m^*} \frac{1}{z_f}}$, $z_f = P_r(f)^{\frac{1}{M(g_c(m))-1}}$, $m^* = \Theta\Big(\min\{\frac{M}{\gamma_r} g_c(m), m\}\Big)$ and $[\Lambda]^+ = \max[\Lambda, 0]$. □

From (2), we observe a behavior similar to the water-filling algorithm for the power allocation in point-to-point communication [37]: if $z_f > \nu$, file f is cached with positive probability $(1 - \frac{\nu}{z_f})$. Otherwise, file f is not cached.

Although the results of [22] are more general, here we focus on the most relevant regime of the scaling of the file library size with the number of users, referred to as "small library size" in [22]. Namely, we assume that $\lim_{n\to\infty} \frac{m^\alpha}{n} = 0$, where $\alpha = \frac{1-\gamma_r}{2-\gamma_r}$. Since $\gamma_r \in (0, 1)$, we have $\alpha < 1/2$. This means that the library size m can grow even faster than quadratically with the number of users n. In practice, however, the most interesting case is where m is sublinear with respect to n. An example of such sublinear scaling is provided by the following simple model: suppose that user 1 has a set m_0 of highly demanded files, user 2 highly demanded files overlap over $m_0/2$ files with the set of user 1, and consists of $m_0/2$ new files, user 3 requests overlap for $2m_0/3$ over the union of user 1 and user 2, and contributes with $m_0/3$ new files and so on, such that the union of all highly demanded files of the users is $m = m_0 \sum_{i=1}^{n} 1/i \approx m_0 \log n$. Remarkably, any scaling of m versus n slower than $n^{1/\alpha}$ is captured by the following result:

Theorem 2: In the small library regime, the outage-throughput tradeoff achievable by one-hop D2D networks with random caching and clustering transmission scheme behaves as:

$$T^*(p) \ge \begin{cases} \frac{C_r}{K}\frac{M}{\rho_1 m} + \delta_1(m), & p = (1-\gamma_r)e^{\gamma_r - \rho_1} \\ \frac{C_r A}{K}\frac{M}{m(1-p)^{\frac{1}{1-\gamma_r}}} + \delta_2(m), & p = 1 - {\gamma_r}^{\gamma_r}\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma_r}, \\ \frac{C_r B}{K} m^{-\alpha} + \delta_3(m), & 1 - {\gamma_r}^{\gamma_r} M^{1-\gamma_r} \rho_2^{1-\gamma_r} m^{-\alpha} \\ & \le p \le 1 - a(\gamma_r) m^{-\alpha} \\ \frac{C_r D}{K} m^{-\alpha} + \delta_4(m), & p \ge 1 - a(\gamma_r) m^{-\alpha} \end{cases} \tag{3}$$

where $a(\gamma_r), A, B, D$ are some constant depending on γ_r and M, which can be found in [22], and where ρ_1 and ρ_2 are positive

parameters satisfying $\rho_1 \geq \gamma_r$ and $\rho_2 \geq \left(\frac{1-\gamma_r}{\gamma_r^{\gamma_r} M^{1-\gamma_r}}\right)^{\frac{1}{2-\gamma_r}}$. The cluster size $g_c(m)$ is any function of m satisfying $g_c(m) = \omega(m^\alpha)$ and $g_c(m) \leq \gamma_r m/M$. The functions $\delta_i(m)$, $i = 1, 2, 3, 4$ are vanishing for $m \to \infty$ with the following orders $\delta_1(m) = o(M/m)$, $\delta_2(m) = o\left(\frac{M}{m(1-p)^{\frac{1}{1-\gamma_r}}}\right)$, $\delta_3(m)$, $\delta_4(m) = o(m^{-\alpha})$. □

The dominant term in (3) can accurately capture the system performance even in the finite-dimensional case, as shown through simulations in [22]. Notice that the first two regimes of (3) are the most relevant ones in practice, providing the throughput for small outage probability. The reason for the different behaviors in these two regimes is that the first regime is achieved by a large cluster size $g_c(m)$, yielding $m^* = m$ in the optimal caching distribution. In this case, all files are stored in the common virtual cache with positive probability. In the second regime, $m^* < m$ if $g_c(m) < \gamma_r m/M$. The third and fourth regimes in (3) correspond to the large outage probability regimes, where the outage probability asymptotically goes to 1 as $m \to \infty$. These regimes are not interesting in practice, and are included here for completeness.

In [22], we show that the throughput-outage scaling laws of Theorem 2 are indeed tight, in the sense that an upper bound on the throughput-outage tradeoff that holds for any one-hop scheme under the protocol model yields the in the same order of the dominant terms with (slightly) different constants.

D. Coded Multicasting From the Base Station

In this section, we review the recent work on coded multicast by the base station proposed in [25]. This scheme is based on a deterministic cache placement with sub-packetization, where each user cache contains a fraction M/m of packets from each of the files in the library. The scheme is designed to handle arbitrary requests. Therefore, its outage probability (under the ideal protocol model where only the base station transmits and all nodes can receive the same rate with zero packet error rate) is zero. Here we start with some simple example.

We consider the case of $n = 2$ users requesting files from a library of $m = 3$ files denoted by A, B and C. Suppose that the cache size of each user is $M = \frac{3}{2}$ file. Each file is divided into three packets, A_1, A_2, B_1, B_2 and C_1, C_2, each of size $\frac{1}{2}$ of a file. Each user u caches the packets with index containing u. For example, user 1 caches A_1, B_1, C_1. Suppose that user 1 requests A, user 2 requests B. Then, the base station will send the packets $\{A_2 \oplus B_1\}$, where "$\oplus$" denotes a modulo 2 sum over the binary field), of size $\frac{1}{2}$ files, such that all requests are satisfied. Clearly, the scheme can support (with the same downlink rate) any arbitrary request. For example, suppose that user 1 wants B and user 2 wants C, then the base station will send $\{B_2 \oplus C_1\}$, which again results in transmitting $\frac{1}{2}$ files.

The scheme is referred to as "coded multicasting" since the base station multicasts a common message to all the users, formed by linear combinations of the packets of the requested files. The term "coded" refer to the fact that sending linear combinations of the messages is a instance of linear network coding [38], [39].

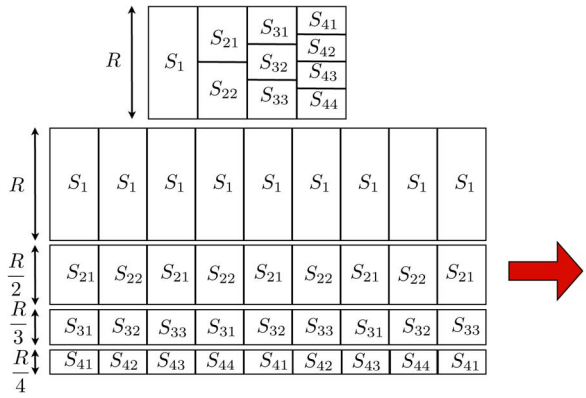


Fig. 2. A video file encoded at rate R is split into blocks $S_{ij} : j = 1, \ldots, i$, for $i = 1, \ldots, 4$, such that the size of S_{ij} is τ/i chunks. Each i-th set of blocks is periodically transmitted in a downlink parallel channel of rate R/i. Any user tuning into the multicast transmission can start its playback after at most τ chunks.

By extending this idea to general n, m and M, and letting $N_{TX}(n, m, M)$ denote the number of equivalent file transmissions from the base station, [25] proves the following result:

Theorem 3: For any m, n, M and arbitrary requests, for $\frac{Mn}{m} \in \mathbb{Z}^+$ and $M < m$,

$$N_{\mathrm{TX}}(n, m, M) = n\left(1 - \frac{M}{m}\right)\frac{1}{1 + \frac{Mn}{m}}, \tag{4}$$

is achievable. For $\frac{Mn}{m} \notin \mathbb{Z}^+$, the convex lower envelope of the points with coordinates $(n, m, M, N_{\mathrm{TX}}(n, m, M))$ for integer $\frac{Mn}{m}$ is achievable. □

Through a compound channel (over the requests) and cut-set argument, [24] proves that the best possible caching and delivery scheme transmitting form the base station requires a number of transmissions not smaller than 1/12 of (4). This means that, within a bounded multiplicative gap not larger than 12, the coded multicasting scheme of [24] is information theoretically optimal, under the arbitrary request and zero outage probability constraint.

Letting C_{r_0} denote the rate at which the base station (BS) can reach any point of the unit-square cell, the corresponding order-optimal per-user throughput achieved by the coded multicast scheme is:

$$T^*_{\mathrm{BS,coded}} = \frac{C_{r_0}}{N_{\mathrm{TX}}(n, m, M)}. \tag{5}$$

Obviously, since the scheme is designed to handle *any* user request, the outage probability of this scheme is $p_o = 0$.

E. Harmonic and Conventional Broadcasting

In brief, harmonic broadcasting works as follows: fix the maximum waiting delay of τ "chunks" (from the time a streaming session is initiated to the time playback is started), and let L denote the total length of the video file, expressed in chunks. In harmonic broadcasting, the video file is split into successive blocks such that for $i = 1, \ldots, \lceil L/\tau \rceil$, there are i blocks of length τ/i (see Fig. 2). Then, each i-th set of blocks of

length τ/i is repeated periodically on a (logical) subchannel of throughput R/i, where R is the transmission rate (in bit/s) of the video playback (see again Fig. 2). Users receive these channels in parallel.

In this way, each file requires a downlink rate of $R\log(L/\tau)$. Hence, the total number of files that can be sent in the common downlink stream is $m' = \min\left\{\frac{C_{r_0}}{R\log(L/\tau)}, m\right\}$, yielding an average throughput per user of $R(1-p_o)$ with outage probability $p_o = \sum_{f=m'+1}^{m} P_r(f)$, since all requests to files not included in the common downlink stream are necessarily in outage.

Finally, the conventional approach of today's technology in cellular and WiFi networks consists of handling on-demand video streaming requests exclusively at the application layer. Then, the underlying wireless network treats these requests as independent individual data. In this case, the average throughput per user is $\Theta\left(\frac{C_{r_0}}{n(1-p_o)}(1-p_o)\right) = \Theta\left(\frac{C_{r_0}}{n}\right)$ for a system whose admission control serves a fraction $1-p_o$ of the users, and denies service to a fraction p_o of users (outage users).

F. Summary: Comparison Between Different Schemes

In this section, we compare the schemes reviewed before in terms of theoretical scaling laws. We focus on case where $Mn \gg m$, M is a constant and $m, n, L \to \infty$. As indicated in Section I, we consider a single cell of fixed area containing one base station and n user nodes (dense network), and take into account adjacent cell interference into the noise floor level.

For the conventional unicasting where no D2D communication is possible and the users don't cache files, the system serves users' requests as if they were independent messages from the BS. Hence, we are in the presence of an information-theoretic broadcast channel with independent messages, whose per-user throughput is known to scale as $\Theta(1/n)$, i.e., even significantly worse than the ad-hoc networks scaling law. As an intermediate system, we may consider the case of conventional caching (e.g., using prefix caching as advocated in [40]), where users can cache M files, but the system does not handle D2D communication. In the prefix caching, users requesting files with index larger than the cutoff index $\hat{m}$ are not served and are in outage. The users that are served, need to download a fraction $(1-\lambda_f)$ for each file f, with index $f < \hat{m}$, where $\lambda_f \leq 1,\ \forall f$ and $\sum_{f=1}^{m} \lambda_f = M$. Thus, the fundamental scaling behavior of this case is again $\Theta(1/n)$ in small outage regime.

In the case of harmonic broadcasting, as mentioned in Section II-E, if we constrain the maximum waiting time to be τ chunks, then the throughput per user of harmonic broadcasting scales as $\Theta\left(\frac{1}{m'\log\frac{L}{\tau}}\left(1-\sum_{f=m'+1}^{m} P_r(f)\right)\right)$, where $m' \leq m$.

Next, we examine the scaling laws of the throughput for the uncoded D2D scheme for arbitrary small outage probability. By using the first line of (3), the average per-user throughput scales as $\Theta\left(\frac{M}{m}\right)$, which is very attractive, since the throughput increase linearly with the size of the user cache.

Finally, from (5) we observe that the throughput of coded multicasting scales also $\Theta\left(\frac{M}{m}\right)$. This indicates that by one-hop communication (either D2D or multicasting from the base station), the fundamental limit of the throughput in the regime of small outage probability is $\Theta\left(\frac{M}{m}\right)$.[5]

As a conclusion of this section we observe that, in the regime of $Mn \gg m$, where the total network storage is larger than the library size, both the uncoded D2D caching scheme and the coded multicasting scheme have an unbounded gain with respect to conventional unicasting as $m, n \to \infty$. Harmonic broadcasting yields also a constant throughput with respect to the number of users n. The gain of the caching schemes over harmonic broadcasting depends critically on the system parameters, L, τ and m' for harmonic broadcasting, and M, m, for the caching schemes. According to the above model, uncoded D2D caching and coded multicasting are equivalent in terms of throughput scaling laws.[6] However, several other factors play a significant role in determining the system throughput and outage in realistic conditions. For example, the availability of D2D links may depend on the specific models for propagation at short range and may significantly differ depending on the frequency band such links operate in. Also, coded multicasting requires to send a common coded message to all the users in the cell. Multicasting at a common rate incurs the worst-case user bottleneck, since in practice users have different path losses and shadowing conditions with respect to the base station. Hence, in order to appreciate the performance of the various schemes reviewed in this paper in realistic system conditions, beyond the scaling laws of the protocol model, in the next sections we resort to a holistic system optimization and simulation.

III. System Design

We assume that devices can operate in multiple frequency bands. For transmission from the BS to mobile stations (MS), we assume operation at 2.1 GHz carrier frequency, corresponding to one of the standard *long-term-evolution* (LTE) bands.[7] We furthermore assume that D2D communication can occur at 2.45 GHz carrier frequency (specifically in the Industrial, Scientific and Medical (ISM) bands), as well as in the unlicensed mm-wave band at 38 GHz. Note that the 2.45 and 38 GHz bands are not suitable for BS-to-MS communications due to propagation conditions as well as transmit power restrictions imposed by frequency regulators. The 38 GHz band provides the possibility for very high data rates at very short range, due to the large available bandwidth at that frequency and the large pathloss.

A. Holistic Multi-Frequency D2D System Design

In this case, we try to use all the resources in the network. As discussed above, file delivery is most efficiently achieved

[5]Notice that, in practice, also coded multicasting is subject to outages, due to the shadowing of the channel between the base station and the users. Since a common transmission rate has to be guaranteed for all the users, then some users will be in outage if the channel capacity between these users and base station is less than the common transmission rate.

[6]Interestingly, in the recent work [41], the authors showed that the gain of spatial reuse from uncoded D2D caching scheme and the gain of the coded multicasting do not accumulate in the order sense.

[7]Due to the non-universal availability of sub-1 GHz bands for LTE, we do not consider it further in this paper.

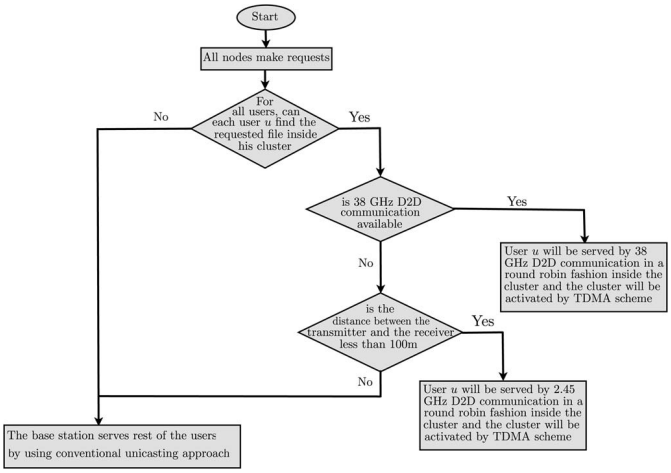


Fig. 3. The flow chart of the delivery algorithm for the combination of D2D communications and multicast by the base station.

if it involves a short range communication link, for which the mm-wave frequency band is ideally suited. However, such connections are not robust since the mm-wave can be easily blocked by walls or even human body. Hence, if the mm-wave link is not available, the next best option is then D2D communication in the 2.45 GHz band. Finally, if even this band is not available or if the requested file is not present within the range of the D2D connections, the file may be served from the BS using the cellular downlink, depending on the admission control decisions. For the D2D links, we use clusterization, i.e., D2D communication is possible within a cluster, but not between clusters. For the cache placement, an independent and randomized cache placement of complete files is used as described in Section II-C. The clustering and caching placement algorithm is summarized in *Algorithm* 1,[8] in which we focus on the regime of small outage, where all potential links (requests found in the cluster) are served.

The flow chart of the delivery algorithm is shown in Fig. 3.

Algorithm 1 Clustering and uncoded caching placement

1: According to the given outage probability of D2D communications (the probability that any user is not served by the D2D networks), decide the cluster size $g_c(m)$ by using (3).
2: **for all** $u \in \mathcal{U}$ **do**
3: Node u randomly caches M files independently according the probability distribution given in (2).
4: **end for**

B. Conventional Unicasting, Coded Multicasting and Harmonic Broadcasting Approaches

- Conventional Unicasting Approach: For the cache placement, we restrict to the case where we do prefix caching and each node caches a fraction of M/m of each file such that the local caching gain $(1 - \frac{M}{m})$ can be obtained. As our baseline approach, we consider a fairness constraint subject to all the users having the same outage probability. Consider a link between BS and user u, with log-normal shadowing χ_σ and deterministic distance-dependent pathloss $PL(d_u)$ not including the shadowing,[9] where d_u denotes the distance between the BS and user u, let C_u denote the required individual downlink rate for user u such that for a fixed outage probability p_o, $\forall\, u \in \{1, \cdots, n\}$, we have

$$\mathbb{P}\left(\log\left(1 + \frac{\mathsf{SNR}}{\chi_\sigma PL(d_u)}\right) \leq C_u\right) = p_o, \tag{6}$$

where SNR (Signal-to-Noise Ratio) denotes the ratio between the transmit power and the noise power spectral density at the receiver. From (6), the user u downlink rate C_u can be determined as a function of p_o and d_u, given the log-normal distribution of χ_σ.

Now, given C_u, $\forall\, u$, let ρ_u denote the fraction of downlink transmission resource dedicated to serve user u. In order to maximize the minimum user rate (our reference performance metric, see (1)), the downlink transmission resource allocation is the solution of:

$$\begin{aligned} \text{maximize} \quad & \min_{u\in\mathcal{A}} \rho_u C_u \\ \text{subject to} \quad & \sum_{u\in\mathcal{A}} \rho_u \leq 1, \end{aligned} \tag{7}$$

where $\mathcal{A}$ is the set of users that are not in outage. It is immediate to obtain the solution of (7) as $\rho_u = \frac{\frac{1}{C_u}}{\sum_{u'\in\mathcal{A}} \frac{1}{C_{u'}}}$. Thus, let 1_u be the indicator that user u is not in outage, we obtain

$$\begin{aligned} \overline{T}_{\min} &= \mathbb{E}[\rho_u C_u 1_u] = \mathbb{E}\left[\frac{\frac{1}{C_u}}{\sum_{u'\in\mathcal{A}} \frac{1}{C_{u'}}} C_u 1_u\right] \\ &= \mathbb{E}\left[\frac{1_u}{\sum_{u'\in\mathcal{A}} \frac{1}{C_{u'}}}\right] = \mathbb{E}\left[\frac{1_u}{\sum_{u'=1}^{n} \frac{1}{C_u'} 1_{u'}}\right]. \end{aligned} \tag{8}$$

Therefore, for any given p_o, we obtain the set of points $(\overline{T}_{\min}, p_o)$ that yields the throughput-outage tradeoff achievable by conventional unicasting. It is immediate to see that $\overline{T}_{\min} = \Theta(1/n)$ for any target p_o in (0, 1).

- Coded Multicasting Approach: Since common multicast messages have to be decoded by all the served users, the downlink rate R is possible if

$$R = \frac{C_{r_0}}{N_{\mathrm{TX}}(n, m, M)} < \frac{\log\left(1 + \frac{\mathsf{SNR}}{\chi_\sigma PL(d_u)}\right)}{N_{\mathrm{TX}}(n, m, M)}, \tag{9}$$

where $N_{\mathrm{TX}}(n, m, M)$ is given by (4). Since for all channel models used in our results, the receiver SNR of users at

[8] In reality, there is a chance that the same file is selected to be cached multiple times in the same cache. Although irrelevant for the sake of the throughput scaling laws, this case should be avoided by practical caching algorithms. We do not further consider this aspect since its impact on the overall system performance is negligible in our simulation setting.

[9] In Section IV-B, *PL* is defined to include the log-normal shadowing χ_σ for the ease of presentation.

larger distance from the BS is stochastically dominated by the receiver SNR of users at smaller distance from the BS, the most stringent outage condition is imposed on the worst case users at largest distance r_0 from the BS, namely,

$$\mathbb{P}\left(\frac{\mathsf{SNR}}{\chi_\sigma PL(r_0)} > 2^{C_{r_0}} - 1\right) < \mathbb{P}\left(\frac{\mathsf{SNR}}{\chi_\sigma PL(d_u)} > 2^{C_{r_0}} - 1\right) \quad (10)$$

for $d_u < r_0$. Hence, we have

$$\overline{T}_{\min} = R\left(1 - \mathbb{P}\left(\chi_\sigma PL(r_0) \geq \frac{\mathsf{SNR}}{2^{C_{r_0}} - 1}\right)\right) \quad (11)$$

The outage probability is given by

$$p_o = \frac{1}{n}\sum_u \mathbb{P}\left(\chi_\sigma PL(d_u) \geq \frac{\mathsf{SNR}}{2^{C_{r_0}} - 1}\right). \quad (12)$$

Note that C_{r_0} is the only control parameter, for this scheme. Hence, using (11) and (12), the throughput-outage tradeoff of coded multicasting can be obtained by varying the common downlink rate C_{r_0}.

- Harmonic broadcasting approach: In this case, the common multicasting messages also have to be decoded by all the served users. Let $m' \leq m$, the downlink rate reliable R is possible if

$$R = \frac{C_{r_0}}{m' \log \frac{L}{\tau}} < \frac{\log\left(1 + \frac{\mathsf{SNR}}{\chi_\sigma PL(d_u)}\right)}{m' \log \frac{L}{\tau}}. \quad (13)$$

Two events can cause outage for harmonic broadcasting: i) the physical channel is not good enough to support the video encoding rate; ii) the requested file is not included in the set of files broadcasted by the BS. Since the two outage events are independent, we have

$$\overline{T}_{\min} = R\left(1 - \sum_{f=m'+1}^{m} P_r(f)\right) \cdot \left(1 - \mathbb{P}\left(\chi_\sigma PL(r_0) \geq \frac{\mathsf{SNR}}{2^{C_{r_0}} - 1}\right)\right). \quad (14)$$

The outage probability is given by

$$p_o = \frac{1}{n}\sum_{u=1}^{n}\left(1 - \left(1 - \sum_{f=m'+1}^{m} P_r(f)\right) \times \left(1 - \mathbb{P}\left(\chi_\sigma PL(d_u) \geq \frac{\mathsf{SNR}}{2^{C_{r_0}} - 1}\right)\right)\right). \quad (15)$$

Also in this case, by varying the common downlink rate C_{r_0}, by (14) and (15) we can obtain the throughput-outage tradeoff of harmonic broadcasting.

IV. Simulations and Discussions

In this section, we provide the simulation results and some discussions. First, we describe the environments and discuss the channel models for the three types of links mentioned in Section III. We then present our simulation results and discuss implications for deployment.

A. Deployment Environments

We perform simulations in two types of environments: (i) office environments and (ii) indoor hotspots. More specifically we assume a cell of dimensions 0.36 km^2 (600 m $\times$ 600 m) that contains buildings as well as streets/outdoor environments. We assume $n = 10000$, i.e., on average, there are $2 \sim 3$ nodes, every square 10×10 meters for the grid network model. We will also investigate the effect of user density later.

For the office environment, we assume that the cell contains a Manhattan grid of square buildings with side length of 50 m, separated by streets of width 10 m. Each building is made up of offices; of size 6.2 m $\times$ 6.2 m.[10] Corridors are not considered, as they would lead to a further complication of the channel model.

For the "indoor hotspot" model, which describes big factory buildings or airport halls, we also assume that the cell is filled with multiple buildings. The size of these buildings are squares with side length of 100 m and distributed on a grid with street width of 20 m. There are no internal partitions (walls) within the building.

Within the cell, users (devices) are distributed at random according to a uniform distribution. Due to our geometrical setup, each node is assigned to be outdoors or indoors, and (in the case of the office scenario) placed in a particular office.[11] This information is exploited to determine which channel model (e.g., indoor-to-indoor or outdoor-to-indoor) is applicable for a particular link. The use of such a virtual geometry is similar in spirit to, e.g., the Virtual Deployment Cell Areas (VCDA) of the COST 259 microcellular model [42].

B. Channel Models

Corresponding to the three types of transmissions (cellular, microwave D2D, millimeter-wave D2D), we use three types of channel models. We only consider pathloss and shadowing, since the effect of small scale fading can be eliminated by frequency/time diversity over the bandwidth and timescales of interest.

The channel models are mostly obtained from the Winner II channel models [43]. We note that although these channels are not explicitly defined for device-to-device, the range of parameter validity includes device height of about 1.5 m, which is typical for a user-held device.

[10]The motivation for this size stems from the line-of-sight (LOS) model (see below); we choose the office size such that it results in a LOS probability of 0.5 if two devices are half the office dimensions apart from each other.

[11]Note that the division of buildings into offices is only used to determine the "wall penetration loss," while the basic pathloss and LOS probability are determined by the purely distance-dependent model (see below for details).

TABLE I
THE LOS PROBABILITY MODELS

indoor office	$P^W_{\rm LOS} = \begin{cases} 1, & d \le 2.5 \\ 1 - 0.9(1 - (1.24 - 0.61\log_{10}(d))^3)^{\frac{1}{3}}, & d > 2.5 \end{cases}$
indoor hotspots	$P^W_{\rm LOS} = \begin{cases} 1, & d \le 10 \\ \exp\left(-\frac{d-10}{45}\right), & d > 10 \end{cases}$
outdoor-to-outdoor	$P^W_{\rm LOS} = \min(18/d, 1)(1 - \exp(-d/36)) + \exp(-d/36)$
outdoor-to-indoor	$P^W_{\rm LOS} = 0$

1) LOS Probability: One of the key parameters for any propagation channel is the existence of a line-of-sight (LOS) condition: all channel characteristics, including path loss, delay spread, and angular spread, depend on this issue. It is obvious that the existence of a LOS is independent of the carrier frequency; it thus seems straightforward to simply apply the LOS model of Winner at all frequencies. However, as we will see in the following, there are subtleties that depend on the carrier frequency and greatly impact the overall performance. By denoting the distance between each transmitter-receiver pair as d, the LOS probability ($P^W_{\rm LOS}$) models by Winner [43] are summarized in Table I.

The LOS probability given in the literature (including the Winner model) usually refers to a LOS connection between *users*, not necessarily between the *antennas* on the devices held by the users. In other words, there are situations where a transmit and receive antenna nominally have LOS (according to the model definition), because there are no *environmental obstacles* between them; however, the bodies of the users and/or the device casings might prevent an actual LOS. We will henceforth refer to this situation as "body-obstructed LOS" (BLOS). It is especially critical at mm-wave frequencies, which to which the human body is essentially impervious. For microwaves, the human body can be taken into account by introducing an additional shadowing term.

Let us first consider the case of mm-wave propagation. Considering the way smartphones are usually held in front of the body, approximately, we assume that each user has a degree of $\frac{360}{\sqrt{2}} = 250$ "free" sector, then half the cases of "nominal" LOS are actually BLOS, while the rest is "true" LOS:

$$P_{\rm BLOS} = P_{\rm LOS} = \frac{1}{2} \cdot P^W_{\rm LOS}, \tag{16}$$

For the case of BLOS, alternative propagation paths such as reflections by walls, can sustain links, but the resulting path loss and related parameters are different from the "true" LOS; thus separate parameterization has to be used. The case of non-line-of-sight (NLOS),[12] clearly occurs with probability $1 - P^W_{\rm LOS}$. In the case of mm-wave communications, walls constitute an insurmountable obstacle, i.e., penetration of radiation into neighboring rooms, and between inside/outside the building, is negligible.

For microwave propagation, the effect of body shadowing is better explained by an additional lognormal fading. In contrast to the "standard" shadowing that describes shadowing by environmental obstacles and that changes as users move laterally, body shadowing variations are created by rotation of the users—resulting in the highest attenuation when they are standing back-to-back. In [44], it is shown that the body shadowing attenuation $\chi_{\sigma_{Lb}}$ follows log-normal distribution with mean 0 and standard deviation σ_{Lb}. For D2D communication, we use the hand-to-hand model (HH2HH) as shown in [44].

2) Device-to-Device Channels at 38 GHz: In this case, the pathloss is given by

$$PL(d) = 20\log_{10}\left(\frac{4\pi d_0}{\lambda}\right) + 10\alpha\log_{10}\left(\frac{d}{d_0}\right) + \chi_\sigma, \tag{17}$$

where $d_0 = 5$ m is the free-space reference distance, λ is the wavelength, α is the average pathloss, χ_σ is the shadowing parameter with mean 0 and standard deviation σ. We assume that no 38 Hz communication is possible when $d > 80$ m. From [45], [46], the system parameters are given by: $\alpha_{LOS} = 2.21$, $\alpha_{NLOS} = 3.18$, $\sigma_{LOS} = 9.4$ and $\sigma_{NLOS} = 11$.

3) Device-to-Device Channels at 2.4 GHz: For this case we can directly use the Winner II channel model, although we assume that no communication is possible for a distance larger than 100 m.[13] Since 2.4 GHz communication can penetrate walls, we have to account for different scenarios, which are indoor communication (Winner model A1), outdoor-to-indoor communication (B4), indoor-to-outdoor communication (A2), and outdoor communication (B1).

We illustrate the case of the indoor (A1) communication, where the path loss model for both LOS and NLOS is given by [43],

$$PL(d) = A_1\log_{10}(d) + A_2 + A_3\log_{10}(f_c[\text{GHz}]/5) + X + \chi_\sigma, \tag{18}$$

where f_c is the carrier frequency. A_1 includes the path loss exponent. A_2 is the intercept and A_3 describes the path loss frequency dependence. $X = 5n_w$ is the (light) wall attenuation parameter, where n_w is the number of walls between transmitter and receiver. χ_σ is the shadowing parameter assumed to be a log-normal distribution with mean 0 and standard deviation σ, where $\sigma_{\rm LOS} = 3$ and $\sigma_{\rm NLOS} = 6$. Note that according to our discussion above, we add the body shadowing loss to Eq. (18), where for LOS, $\sigma_{L_b} = 4.2$ and for NLOS, $\sigma_{L_b} = 3.6$. All the other parameters for the indoor pathloss channel model in 2.4 GHz are summarized in Table II.

For the other three cases, namely outdoor (B1), indoor-to-outdoor (A2) and outdoor-to-indoor (B4), we similarly directly use the respective Winner II channel models with antenna heights of 1.5 m, probabilistic LOS, and with the consideration of body shadowing.

4) Channel Between the Base Station and Devices: In this case the Winner II channel model can also be used directly. In particular we use the urban macro-cell (C2) model for outdoor to outdoor communications and the urban macro outdoor to indoor (C4) model for outdoor to indoor communication; the only modification is the addition of the rotational body shadowing $\chi_{\sigma_{L_b}}$. As model for the rotational body shadowing,

[12] One example of NLOS transmission is that the transmitter and the receiver are in different rooms, with walls between them.

[13] This is a conservative assumption motivated by the fact that at low SNR it is difficult for a D2D link to acquire beacon signals and discover other D2D devices.

TABLE II
THE CHANNEL PARAMETERS FOR 2.4 GHZ D2D COMMUNICATIONS

	$A_1(LOS)$	$A_2(LOS)$	$A_3(LOS)$	$A_1(NLOS)$	$A_2(NLOS)$	$A_3(NLOS)$
indoor office	18.7	46.8	20	36.8	43.8	20
indoor hotspot	13.9	64.4	20	37.8	36.5	23

TABLE III
THE PARAMETERS FOR THE THREE TYPES OF TRANSMISSIONS

	B	f_c	P_{TX}	G_t	G_r	K
mm-wave D2D transmissions	800 MHz	38 GHz	20 dBm	9 dB	9 dB	4
microwave D2D transmissions	20 MHz	2.45 GHz or 2.1 GHz	20 dBm	12 dB	0	4
cellular transmissions	20 MHz	2.1 GHz	43 dBm	12 dB	0	3

we use the access point to handheld device model (AP2HH [44]: for the case of LOS, $\sigma_{L_b} = 2.3$ dB, while for NLOS, it is $\sigma_{L_b} = 2.2$ dB.)

For example, for the urban macro-cell (C2) channel model, the pathloss for LOS of the Winner model (i.e., without body shadowing) is given by

$$PL_{\text{LOS}}(d) = \begin{cases} A_1 \log_{10}(d) + A_2 + A_3 & \\ \cdot \log_{10}\left(f_c[\text{GHz}]/5\right) + \chi_{\sigma_1}, & 10\,\text{m} < d < d'_{\text{BP}} \\ 40\log_{10}(d) + 13.37 - 14\log_{10}\left(h'_{\text{BS}}\right) & \\ -14\log_{10}\left(h'_{\text{MS}}\right) + 6 & \\ \cdot \log_{10}\left(f_c[\text{GHz}]/5\right) + \chi_{\sigma_2}, & d'_{\text{BP}} < d < 5000\,\text{m}, \end{cases} \tag{19}$$

where $d'_{\text{BP}} = 4h'_{\text{BS}}h'_{\text{MS}}f_c/c$ and $h'_{\text{BS}} = h_{\text{BS}} - 1$ and $h'_{\text{MS}} = h_{\text{MS}} - 1$. We pick $h_{\text{BS}} = 25$ m and $h'_{\text{MS}} = 1.5$ m. χ_{σ_1} and χ_{σ_2} are shadowing attenuations, which are lognormally distributed with mean 0 and standard deviation $\sigma_1 = 4$ and $\sigma_2 = 6$. For NLOS, we have

$$PL_{\text{NLOS}}(d) = \left(44.9 - 6.55\log_{10}(h_{\text{BS}})\right)\log_{10}(d) + 34.46 + 5.83\log_{10}(h_{\text{BS}}) + 23\log_{10}\left(f_c[\text{GHz}]/5\right) + \chi_{\sigma}, \tag{20}$$

where 50 m $< d <$ 5000 m. The shadowing χ_{σ} is zero-mean and has standard deviation $\sigma = 8$. Similarly, the urban outdoor to indoor (C4) channel model can be found in [43].

Moreover, to simulate the realistic scenario, we also assume a frequency reuse factor K in this case to avoid the interference between cells [37].

5) Link Capacity Computation: Given all the system parameters, the link capacity for a transmitter-receiver pair is given by

$$C = B \cdot \log_2(1 + \text{SINR}) \tag{21}$$

where $\text{SINR} = P_{\text{signal}}/(P_{\text{noise}} + P_{\text{interference}})$ (Signal to Interference plus Noise Ratio), and B denotes the signal channel bandwidth. Specifically, on a dB scale, P_{signal} is given by

$$P_{\text{signal,dB}} = P_{TX} + G_t + G_r - PL(d) \tag{22}$$

where the P_{TX} is the transmit power, G_t and G_r are the transmit and receive antenna gains. $P_{\text{interference}}$ is the sum of the all the interference to a receiver.[14]

On a dB scale, the noise power is given by

$$P_{\text{noise,dB}} = 10\log_{10}(k_B T_e) + 10\log_{10} B + F_N, \tag{23}$$

where $k_B T_e = -174$ dBm/Hz is the noise power spectral density and $F_N = 6$ dB is a typical noise figure of the receiver. We assume this model to hold at all frequencies.[15] The parameters of the three types of transmissions are summarized in Table III.

C. Results and Discussions

In this section, we will present the simulation results. If not stated otherwise, we will use the following settings: the number of users is $n = 10000$; the users are uniformly and indepdently distributed in the cell (It can be shown that a negligible difference between regular grid and random distribution by simulation (not shown here); we thus henceforth show only results for the random node distribution). The number of files in the library is $m = 300$, which is representative of the library size of a video on-demand service.[16] The user cache size is $M = 20$ files unless specifically mentioned, which even with high definition (HD) quality requires less than the (nowadays) ubiquitous 64 GByte of storage space. We let each user independently make a request by sampling from a Zipf distribution with $\gamma_r = 0.4$; this value is at the lower edge of the range of values that have been measured in practice [35]; note that the advantages of caching would be *more* pronounced for larger γ_r. The interference between concurrent D2D links sharing the same frequency band is treated as noise. For the harmonic broadcasting, we chose a video file size of $L = 5400$ chunks and $\tau = 10$ chunks, then the number of blocks is $\lceil \frac{L}{\tau} \rceil = 540$ [6].

1) Throughput-Outage Tradeoff: In Fig. 4, we plot the performance of all the discussed schemes separately, where a 2.45 GHz D2D only scheme is implemented. From Fig. 4, we can see that the throughput of the D2D scheme is markedly

[14]The model for mm-wave communication is considered to be interference free ($P_{\text{interference}} = 0$) since the angle of arrival (AOA) is very narrow (less than 10 degree).

[15]While for the same cost, receivers at 2 GHz might provide a better noise figure due to better-established fabrication processes, the impact of this effect on the system performance is low, and will be neglected henceforth.

[16]In practice, the library of titles in such a service would be refreshed every few days.

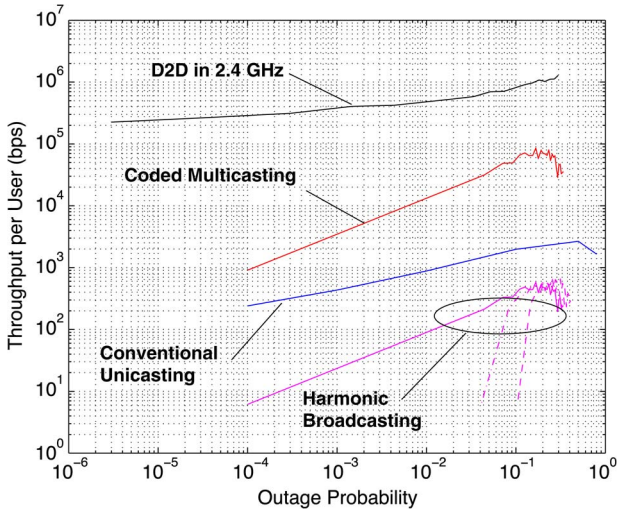


Fig. 4. Simulation results for the throughput-outage tradeoff for conventional unicasting, coded multicasting, harmonic broadcasting and the 2.45 GHz D2D communication scheme under indoor office channel models. For harmonic broadcasting with only the m' most popular files, solid line: $m' = 300$; dash-dot line: $m' = 280$; dash line: $m' = 250$. We have $n = 10000$, $m = 300$, $M = 20$ and $\gamma_r = 0.4$.

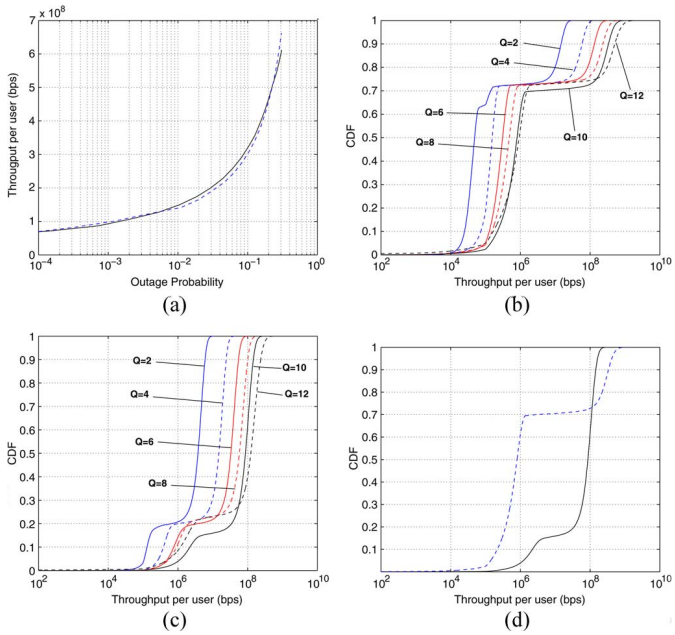


Fig. 5. (a). Simulation results for the throughput-outage tradeoff by holistic system design. Black Solid lines: indoor office; blue dashed lines: indoor hotspot. (b). The CDF of the throughput for different outage probabilities (cluster size of $600^2/Q^2$) under indoor office model. (c). The CDF of the throughput for different outage probabilities (cluster size of $600^2/Q^2$) under indoor hotspot model. (d). The CDF of the throughput for the cluster size of 100 m $\times$ 100 m under indoor office and indoor hotspot channel model. Solid lines: indoor office; dashed lines: indoor hotspot.

(orders of magnitude at low outages) higher than the conventional unicasting, harmonic broadcasting and even coded multicast scheme. This shows that in practical situations, the "scaling law" is not the only aspect of importance. Rather, the higher capacity of the short-distance links plays a significant role, and a good throughput-outage tradeoff can be achieved even without the use of a BS connection as "backstop". The main reason lies in the fact that for the coded multicasting or harmonic broadcasting scheme, outage is determined by bad channel conditions, and no diversity is built into the system. For D2D, even though the outage in our scheme is caused by both physical channel and the lack of the requested files in the corresponding cluster, the channel diversity plays a more importance role. Moreover, although not shown in Fig. 4 for the ease of presentation, the behaviors of all schemes hold for both the indoor office and the indoor hotspot environment.[17]

In Fig. 6(a), for the hotspot, we furthermore obtain the interesting result that the throughput-outage tradeoff is non-monotonous if we use the (theoretically derived) cluster size. This behavior is caused by a higher LOS probability when the cluster size becomes small: there is an appreciable probability that the useful signal is NLOS but there exist some LOS interferers. From Fig. 6(a), (b) and (c), a similar phenomenon can also be observed for the case of the indoor office model but for different parameter settings. Of course this does not mean that the optimum throughput-outage tradeoff in practice is non-monotonous; rather it is a consequence of using a cluster size derived under one specific model in the deployment by using a different model.

Besides the performance advantage of the D2D approach (compared to coded multicast), it also has the advantage of a simpler caching placement and delivery. The coded multicasting approach in [25] constructs the cache contents and the coded delivery scheme in a combinatorial manner which does not scale well with n. For example, in our network configuration, it requires a code length larger than $\binom{10000}{600}$, which is larger than 10^{15}.

2) Holistic Multi-Frequency D2D System Performance: In this section, we investigate the performance of the proposed D2D system given by Fig. 3 in Section III-A.

Fig. 5(a) shows that the average throughput per user increases significantly due to the help of the 38 GHz D2D communications. Consider the CDF (Cumulative distribution function) of the throughput for different outage probabilities shown in Fig. 5(b), in this way, we can see on average, how many users will be served with a throughput that is less than the minimum required rate for video streaming, for example, 100 Kbps.

Fig. 5(b) and (c) show the throughput as a function of the cluster size. Intuitively, a large cluster size corresponds to a small outage probability, as the probability is high that the desired file is found within a cluster. This is reflected by the throughput CDF: a small cluster size results in a small minimum throughput, but a large maximum throughput (compare, e.g., the red-dotted line in Fig. 5(c); this is similar to the effect we have observed in the previous subsection. For example, if we pick a cluster size of 100 m $\times$ 100 m, then the number of users whose rate is less than 100 Kbps only around 250. Moreover,

[17]In fact, our D2D scheme performs better in the hotspot scenario than in the indoor office case. This is mainly due to the low probability of LOS from interferers and the high probability of LOS for useful signal (note that the LOS probability in the hotspot is unity up to distances of 10 m and decreases exponentially for larger distances). However, for the coded multicast transmission, the performance in indoor hotspot is actually worse than the indoor office model; this is due to the larger size of the buildings so that the pathloss caused by d_{in} in the urban macro outdoor-to-indoor model (C4) [43] is very significant, where d_{in} is the distance from the wall to the indoor terminal.

about 30% of users are served with a data rate larger than 2 Mbps) due to help by 38 GHz D2D communications.

From Fig. 5(a), (b) and (c), we observe similar behavior under the indoor hotspot model, where interestingly, the performance of our holistic multi-frequency design in terms of average throughput is not very different from that of the indoor office model. The reason is that the variance of the throughput for the indoor model is much larger than that for the indoor hotspot model, which is due to the fact that fewer users can be served by the 38 GHz D2D communications.[18] Both CDFs for the case of 100 m $\times$ 100 m cluster size are shown in Fig. 5(d), In this case, almost no users have a service rate less than 100 Kbps and about 90% of users can get HD quality services. Moreover, we notice that the role of base station in this scenario is to reduce the outage probability. For example, when cluster size is 100 m $\times$ 100 m, the base station can serve $400 \sim 500$ users in the indoor office model.

3) Effects of the Density of Nodes: From Theorem 2, we expect that the throughput-outage tradeoff does not depend on the number of users or user density as long as n and m are large and $Mn \gg m$. However, the throughput-outage scaling behavior was obtained under the simplified protocol model, where the relation between the link rate and the link range (source-destination distance) is not specified. In practice, if we want to obtain a high communication rate, the D2D communication distance cannot be very large due to the large pathloss, which reduces the per-link capacity. This is especially true for 38 GHz communications under the indoor office environment. Therefore, the user density is also an important parameter for the system performance. In this section, we investigate the system behavior for different user densities by focusing on the case of 2.45 GHz D2D communications only.

In Fig. 6(a), we observe that there exists a tradeoff between the user density and the throughput, which is because that the impact of the user density on the link rate is twofold: on one hand, a higher user density allows a smaller cluster size, in turn resulting in shorter links and higher SINR. On the other hand, a small cluster size increases the probability for having LOS interference, which can degrade the performance of the system significantly.

4) Effects of the Storage Capacity and the Library Size: As already observed in Section II-C, in the regime $nM \gg m$ the D2D system yields a linear dependence of the throughput on the user storage capacity M. This means that such a system can directly trade cache memory for throughput. Since storage memory is a cheap and rapidly growing network resource, while bandwidth is scarce and very expensive, the attractiveness of the D2D approach is self-evident. The result also holds true in practice, as demonstrated by the simulation results in Fig. 6(b). We observe that, when the outage is small, the average throughput per user increases even faster than linearly with M. This is because in practice the link rate C_r is a decreasing function of the link range. Therefore, when M becomes large, we can decrease the D2D cluster size (and therefore the average link range) while maintaining a constant outage probability.

[18] We serve the users in a round robin fashion in one cluster even for 38 GHz communications to avoid interference.

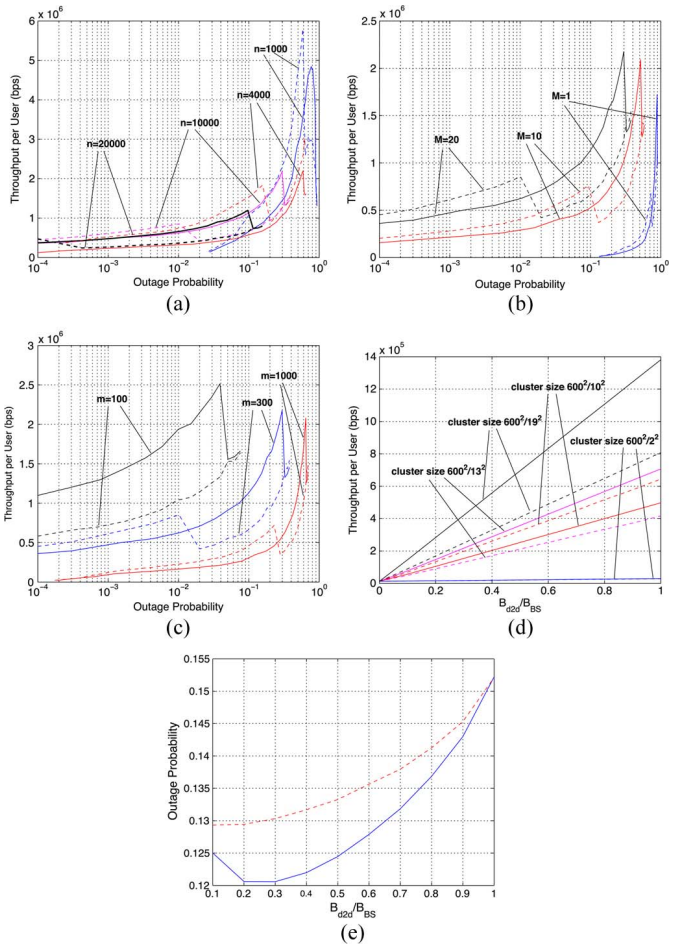


Fig. 6. Solid lines: indoor office; dashed lines: indoor hotspot. (a) The throughput-outage tradeoff for different user densities. (b) The throughput-outage tradeoff for different user storage capacity. (c) The throughput-outage tradeoff for different library size of files. (d) The throughput v.s. bandwidth division between 2.1 GHz communication and the base station under different cluster size, where B_{d2d} is the bandwidth by 2.1 GHz communications and B_{BS} is the bandwidth by the cellular base station. $B_{\mathrm{d2d}} + B_{\mathrm{BS}} = B = 20$ MHz. (e) The outage v.s. bandwidth division between 2.1 GHz communication and the base station for the cluster with size $600^2/19^2$.

Fig. 6(c) shows the throughput-outage tradeoff for different library size. As expected from Theorem 2, in this case we notice that the throughput decreases roughly inversely proportional to the library size m, for fixed cache capacity M.

5) 2.1 GHz in Band Communications: Sometimes, the D2D communications and the cellular communications by the BS have to share the same spectrum, which raises the question of how to divide the bandwidth for each type of communications. Obviously, this depends on the channel realizations for each type of communications. From our simulations, we obtain that under our channel model (either indoor office or indoor hotspot), the base station cannot support more than about 1000 users in the best case if the minimum video coding rate is 100 Kbps, while, for 2.1 GHz D2D communication, it is very easy to support a much larger number of users at a certain playback rate requirement. Therefore, it is intuitive that there is no tradeoff between the bandwidth division and the *average* throughput by fixing the cluster size (outage probability in Theorem 2). The simulation results are shown in Fig. 6(d), which confirm our intuition.

On the other hand, if we care more about the outage probability, then there is a clear tradeoff between the outage probability and the division of the bandwidth, especially for the small cluster size. This occurs because the BS is capable of satisfying "costly" links that normally would either increase outage probability or would enforce an increase in cluster size.

In Fig. 6(e), under the office channel model, when the area of the cluster is $600^2/19^2$, the best bandwidth division is when $B_{\mathrm{d2d}}/B_{\mathrm{BS}} = 0.2$, which means that we need to only allocate 20% of the bandwidth to the D2D communication to obtain the minimum outage probability. Similar behavior can also be observed for the hotspot model with the difference that now $B_{\mathrm{d2d}}/B_{\mathrm{BS}} = 0.1$ is the best bandwidth division, which is because that the link rate under the hotspot channel model is better than that for the indoor office model.

V. Conclusion

In this paper we have reviewed in a tutorial fashion some recent results on base station assisted D2D wireless networks with caching for video delivery, recently proposed by the authors [22], as well as some competing conventional schemes and a recently developed scheme based on caching at the user devices but involving only coded multicasting transmission from the base station. We reviewed the throughput-outage scaling laws of such schemes on the basis of a simple protocol model which captures the fundamental aspects of interference and spatial spectrum reuse through geometric link conflict constraints. This model allows a sharp characterization of the throughout-outage tradeoff in the asymptotic regime of dense networks. This tradeoff shows the superiority of the D2D caching network approach and of the coded multicasting approach over the conventional schemes, which can be regarded as today current technology.

In order to gain a better understanding of the actual performance in realistic environments, we have developed an accurate simulation model and some guidelines for the system design. In particular, we have considered a holistic system design including D2D links at 38 GHz and 2.45 (or 2.1) GHz, and the cellular downlink at 2.1 GHz, representative of an LTE network.

We compared the schemes treated in the tutorial part of the paper on the basis of their throughput-outage tradeoff performance, and we have put in evidence several interesting aspects. In particular, we have shown the superiority of the D2D caching network even in realistic propagation conditions, including all the aspects that typically are expected to limit D2D communications, such as NLOS propagation, limited link range, environment shadowing and human body shadowing. The D2D caching network shows very competitive performance with respect to the other schemes. In particular, the proposed system is able to efficiently trade the cache memory in the user devices for the system throughput. Since the former is a rapidly growing, cheap and yet untapped network resource, while the latter is known to be scarce and very expensive, the interest in developing and deploying such D2D caching networks becomes evident. This fact is even more remarkable if we consider the fact that the D2D network requires simple decentralized caching and does not require any sophisticated network coding technique to share the files over the D2D links.

References

[1] [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white/paper/c11-520862.html

[2] S. Annapureddy, A. Barbieri, S. Geirhofer, S. Mallik, and A. Gorokhov, "Coordinated joint transmission in WWAN," in *IEEE Commun. Theory Workshop*, 2010.

[3] R. Irmer *et al.*, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[4] J. G. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[5] Y. Sanchez *et al.*, "Improved caching for HTTP-based video on demand using scalable video coding," in *Proc. IEEE CCNC*, 2011, pp. 595–599.

[6] Y. Sánchez de la Fuente *et al.*, "iDASH: Improved dynamic adaptive streaming over http using scalable video coding," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst.*, 2011, pp. 257–264.

[7] A. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 1: Streaming protocols," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 54–63, Mar./Apr. 2011.

[8] Y. Li, E. Soljanin, and P. Spasojević, "Three schemes for wireless coded broadcast to heterogeneous users," *Phys. Commun.*, vol. 6, pp. 114–123, Mar. 2013.

[9] S. Jakubczak and D. Katabi, "Softcast: One-size-fits-all wireless video," in *Proc. ACM SIGCOMM Comput. Commun. Rev.*, 2010, vol. 40, pp. 449–450.

[10] S. Aditya and S. Katti, "Flexcast: Graceful wireless video streaming," in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw.*, 2011, pp. 277–288.

[11] O. Y. Bursalioglu, M. Fresia, G. Caire, and H. V. Poor, "Lossy multicasting over binary symmetric broadcast channels," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3915–3929, Aug. 2011.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley-interscience, 2006.

[13] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.

[14] M. Luby, M. Watson, T. Gasiba, T. Stockhammer, and W. Xu, "Raptor codes for reliable download delivery in wireless broadcast systems," in *Proc. 3rd IEEE CCNC*, 2006, vol. 1, pp. 192–197.

[15] O. Oyman, J. Foerster, Y. Tcha, and S. Lee, "Toward enhanced mobile video services over wimax and lte [wimax/lte update]," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 68–76, Aug. 2010.

[16] L. Keller *et al.*, "Microcast: Cooperative video streaming on smartphones," in *Proc. ACM MobiSys*, 2012, pp. 57–70.

[17] L.-S. Juhn and L.-M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Trans. Broadcast.*, vol. 43, no. 3, pp. 268–271, Sep. 1997.

[18] J.-F. Pâris, S. W. Carter, and D. E. Long, "Efficient broadcasting protocols for video on demand," in *Proc. IEEE MASCOTS*, 1998, pp. 127–132.

[19] L. Engebretsen and M. Sudan, "Harmonic broadcasting is bandwidth-optimal assuming constant bit rate," *Networks*, vol. 47, no. 3, pp. 172–177, May 2006.

[20] J.-F. Pâris, S. W. Carter, and D. E. Long, "A low bandwidth broadcasting protocol for video on demand," in *Proc. IEEE 7th Int. Conf. Comput. Commun. Netw*, 1998, pp. 690–697.

[21] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[22] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," to be published, arXiv preprint:1302.2168.

[23] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[24] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.

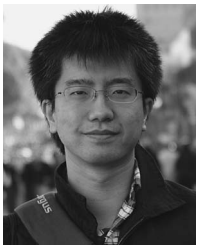
[25] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[26] S. Sesia, I. Toufik, and M. Baker, *LTE: The Long Term Evolution-From Theory to Practice*. New York, NY, USA: Wiley, 2009.

[27] Y. Azar *et al.*, "28 ghz propagation measurements for outdoor cellular communications using steerable beam antennas in new york city," in *Proc. IEEE ICC*, 2013, pp. 1–5.

[28] R. C. Daniels, J. N. Murdock, T. S. Rappaport, and R. W. Heath, "60 GHz wireless: Up close and personal," *IEEE Microw. Mag.*, vol. 11, no. 7, pp. 44–50, Dec. 2010.

[29] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.
[30] F. Xue and P. R. Kumar, *Scaling Laws for Ad Hoc Wireless Networks: An Information Theoretic Approach*. Delft, The Netherlands: Now Publisher, 2006.
[31] S. R. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1041–1049, Jun. 2004.
[32] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1009–1018, Mar. 2007.
[33] A. Ozgur, O. Lévêque, and D. N. C. Tse, "Hierarchical cooperation achieves optimal capacity scaling in Ad Hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3549–3572, Oct. 2007.
[34] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: Information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3413–3424, Aug. 2009.
[35] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, 1999, vol. 1, pp. 126–134.
[36] X. Wu *et al.*, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," in *Proc. IEEE Allerton Conf.*, 2010, pp. 514–521.
[37] A. F. Molisch, *Wireless Communications*, 2nd ed. New York, NY, USA: IEEE Press, 2011.
[38] S.-Y. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 371–381, Feb. 2003.
[39] T. Ho *et al.*, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2006.
[40] S. Sen, J. Rexford, and D. Towsley, "Proxy prefix caching for multimedia streams," in *Proc. 88th Annu. Joint Conf. IEEE Comput. Commun. Soc. INFOCOM*, 1999, vol. 3, pp. 1310–1319.
[41] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," to be published, arXiv preprint:1405.5336.
[42] L. M. Correia, *Wireless Flexible Personalized Communications*. London, U.K.: Wiley, 2001.
[43] "Winner 2 channel models," Eur. Union, Brussels, Belgium, Winner 2, Deliverable d1, 2007.
[44] J. Karedal, A. J. Johansson, F. Tufvesson, and A. F. Molisch, "A measurement-based fading model for wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4575–4585, Nov. 2008.
[45] T. S. Rappaport *et al.*, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.
[46] J. Kim, Y. Tian, S. Mangold, and A. F. Molisch, "Quality-aware coding and relaying for 60 ghz real-time wireless video broadcasting," in *Proc. IEEE ICC*, 2013, pp. 5148–5152.



Mingyue Ji (S'09) received the Bachelor's degree in communication engineering from Beijing University of Posts and Telecommunications, Beijing, China, and the M.S. degree in electrical engineering from the Royal Institute of Technology, Stockholm, Sweden. He is currently working toward the Ph.D. degree with Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA, USA. His adviser is Prof. Giuseppe Caire, and he also frequently collaborates with Prof. Andreas F. Molisch during his Ph.D. study. His research interests include information theory, coding theory, concentration of measure and statistics with the applications of caching networks, wireless communications, distributed storage, and (statistical) signal processing. Prior to USC, he worked as a Research Engineer and finished his Master thesis at the Access Technologies and Signal Processing Group, Ericsson, Stockholm, Sweden. He received the Best Student Paper Award in the 2010 IEEE European Wireless Conference and the Best Paper Award at the 2015 IEEE International Conference on Communications.

Giuseppe Caire (S'92–M'94–SM'03–F'05) was born in Turin, Italy, in 1965. He received the B.Sc. degree in electrical engineering from Politecnico di Torino, Turin, in 1990; the M.Sc. degree in electrical engineering from Princeton University, Princeton, NJ, USA, in 1992; and the Ph.D. degree from Politecnico di Torino in 1994. From 1995 to 1995, he was a Postdoctoral Research Fellow with the European Space Agency (ESTEC), Noordwijk, The Netherlands; an Assistant Professor of telecommunications with Politecnico di Torino; an Associate Professor with the University of Parma, Parma, Italy; and a Professor with the Department of Mobile Communications, Eurecom Institute, Sophia-Antipolis, France. He is currently a professor of electrical engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA, and an Alexander von Humboldt Professor with the Department of Electrical Engineering and Computer, Technical University of Berlin, Berlin, Germany. His main research interests include communications theory, information theory, and channel and source coding with particular focus on wireless communications. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as an Officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. He served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS in 1998–2001 and as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY in 2001–2003. He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society and Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, and the Vodafone Innovation Prize in 2015.

Andreas F. Molisch (S'89–M'95–SM'00–F'05) received the Dipl. Ing., Ph.D., and Habilitation degrees from the Technical University of Vienna, Vienna, Austria, in 1990, 1994, and 1999, respectively. He subsequently was with AT&T (Bell) Laboratories Research, USA, Lund University, Lund, Sweden, and Mitsubishi Electric Research Labs, USA. He is currently a Professor of electrical engineering and the Director of the Communication Sciences Institute, University of Southern California, Los Angeles, CA, USA. He is the author or coauthor or the editor of four books, among them the textbook *Wireless Communications*, (Wiley–IEEE Press); 16 book chapters; some 170 journal papers; 250 conference papers; and more than 80 patents and 70 standards contributions. His current research interests include measurement and modeling of mobile radio channels, ultrawideband communications and localization, cooperative communications, multiple-input–multiple-output systems, wireless systems for healthcare, and novel cellular architectures. He has served as a General Chair, Technical Program Committee Chair, or Symposium Chair of multiple international conferences, and Chair of various international standardization groups. He has received numerous awards, among them the Donald Fink Prize of the IEEE, and the Eric Sumner Award of the IEEE. He is a Fellow of the American Association for the Advancement of Science and the Institution of Engineering and Technology. He is an IEEE Distinguished Lecturer and a member of the Austrian Academy of Sciences.