Network-Assisted D2D Over WiFi Direct

Alexander Pyattaev, Olga Galinina, Kerstin Johnsson, Adam Surak, Roman Florea, Sergey Andreev and Yevgeni Koucheryavy

1 Introduction

D2D communications have been an active topic for the last years, and appeared many times under different names: Ad-hoc and mesh networking, Cooperative communications (client) relay networks, even to some extent under the notion of cognitive radio. Fundamentally, the proximity of user devices promises higher data rates, lower transfer delays, and better power efficiency [1]. More broadly, employing client devices within the integral network infrastructure is envisioned as the logical next step to improve spatial reuse toward the vision of 1000x capacity [2] by the year 2020 in 5G systems. Consequently, over the past few years, D2D communications have received significant attention, both in industry and academia, due to the growing number of services and applications that could leverage proximity. The prospective applications of D2D connectivity in cellular networks are numerous and include, to name a few, local voice service (offloading calls between proximate users), multimedia content sharing, gaming, group multicast, context-aware applications, and public safety.

However, the glaring absence of the practical D2D solutions on the market is alarming. Next to none of the intricate concepts created over the years were cheap and usable enough for actual deployment. Some required new types of radio that have not yet been invented, others needed unfeasible scheduling mechanisms that would resolve contention and interference. On the other hand there are real D2D solutions on the market (e.g., Apple's Airdrop, WiFi Direct, Bluetooth). Made by engineers, not scientists, they rely on well-known, existing technology, as a consequence, most of them operate in unlicensed bands and are driven by enterprise cloud services. In this chapter we present an overview of the D2D-friendly

A. Pyattaev (✉) · O. Galinina · A. Surak · R. Florea · S. Andreev · Y. Koucheryavy
Tampere University of Technology, Tampere, Finland
e-mail: alexander.pyattaev@tut.fi

K. Johnsson
Intel Labs, Pittsburgh, USA

S. Mumtaz and J. Rodriguez (eds.), *Smart Device to Smart Device Communication*, DOI: 10.1007/978-3-319-04963-2_7,

technologies that exist on the market today, as well as a new look on their potential capabilities when combined with proper management. In particular, we will discuss how existing cellular network infrastructure can be leveraged to improve the performance of existing short-range radio technologies, as well as approaches to modeling and analysis of such networks.

1.1 Motivation

Overall, there is a distinct niche for a D2D solution that is based on existing technology and acts as a transition agent between current state of the art where D2D is next to impossible, and potential future where D2D communications are natural. The primary role of current D2D solutions thus is to enable applications that now rely on clouds for data transfer to transition toward proximity-based communication. However, although existing networks are advanced and diverse, they do not easily link into a solid system. For instance, if two mobiles are meters away from each other, it may be easier to transfer files between them using an SD card or a QR code, rather than using WiFi, as the latter might require e.g., manual security pairing between the devices. Therefore, before we can deploy novel proximity-based applications, we need to develop an infrastructure to support them, mostly in the directions of discovery and authentication (for more detail on this particular priorities reader is referred to publication [3]). However, the demand for supporting infrastructure can not be satisfied within the proximity protocols themselves due to their ad-hoc nature: most of them revolve around idea of external security (like WPA passphrases or Bluetooth PIN numbers). In what follows, we discuss how current cellular networking technology can be augmented to accommodate this demand, and present a potential design paradigm that could enable large-scale deployment of proximity-aware services based on existing solutions, in particular WiFi Direct (WFD) and Long-Term Evolution (LTE).

In this chapter, we focus on a subset of the possible solutions, aiming to deliver the following:

- to the application developers—a way to implement an open and secure system to provide device discovery and security contexts for D2D links (irrespective of the actual D2D technology);
- to the end-users and—a way to interact with the application services (such as social networks) that can make use of D2D, as well as represent D2D connections;
- to the operators—a scalable solution that allows an operator to assume a degree of control over what is happening in the unlicensed bands.

1.2 Background and Previous Works

The potential applications for D2D connectivity in cellular networks are many [4]. They range from local voice services (offloading calls between proximate users) to proximity-based data services, such as content sharing, gaming, local multi-casting, context-aware applications, and public safety [5]. More broadly, the term "device" refers to more than just user equipment (UE); it also applies to "machines" (i.e., Machine-Type Communications or MTC). Thus, D2D enables a plethora of emerging MTC-related applications and services as well [6].

1.2.1 D2D in Licensed Bands

Licensed spectrum continues to be scarce and expensive, and while there are efforts to make additional bands available for mobile communications, they are not enough to meet the expected capacity demand. Instead, mobile broadband providers need to find new ways to boost capacity on their existing cellular bands [7]. One promising method is network-assisted D2D, as evidenced by the rich amount of literature on this topic, covering a *range* of network assistance levels. At one end, network assistance is as simple as providing synchronization for communicating devices (e.g., Aura-net [8]). At the other, the network manages each D2D connection, enabling them to act as an underlay tier in the cellular network [9].

Interference management, including proper admission and power control [10], is required to support multiple D2D connections in the same coverage area. Recent publications (see e.g., [11, 12] or [13]) propose interference mitigation techniques that employ inputs such as channel state information (CSI), exact user location information, etc. The interference mitigation scheme instructs D2D connections to either (i) share licensed band resources with standard cellular transmissions (those between users and base stations), (ii) use dedicated resources, or (iii) remain on the cellular infrastructure network. This decision-making process, also known as transmission *mode selection*, has attracted a great deal of research focusing on various optimization targets from signal to interference plus noise ratio (SINR) [14] and throughput [12] to energy efficiency [15], data delay [16], fairness, and outage probability [17, 18].

In general, published D2D studies differ in terms of the number of communicating nodes (base stations, cellular users, and D2D users), the emphasis on uplink (UL) versus downlink (DL) cellular transmissions, orthogonal versus nonorthogonal resource sharing, the amount of available network assistance, and the network/D2D duplexing mode. Most of them attempt to integrate D2D into LTE technology by Third Generation Partnership Project (3GPP) ([19, 20]). However, some papers address legacy cellular systems as well [21]. More recently, the FlashLinQ technology was proposed in [22] and analyzed in [23], offering a distributed D2D communications technology in the licensed bands that uses the cellular network for synchronization purposes only.

Given the current focus on LTE networks (and their impending capacity crunch), many performance improvement techniques have already been evaluated for licensed band D2D, such as the design of D2D-aware multiple-input/multiple-output (MIMO) schemes [24, 25], network coding [26], successive interference cancellation [27], and wireless video distribution over D2D [28]. With the recent introduction of comprehensive D2D frameworks in [12] and [13], this research direction is essentially concluded.

In response to the excitement around D2D, 3GPP began a feasibility study on LTE Direct [29]—a synchronous system operating in licensed spectrum under the control of the operator—approximately 2 years ago. This work was recently completed and Stage 2 work has begun. However, given the many technical challenges and disjoint opinions of 3GPP member companies, "product" is not expected for several years, thus the immediate attention of industrial players is on D2D in the unlicensed bands.

1.2.2 D2D in Unlicensed Bands

An operator may not claim exclusive use of any unlicensed spectrum, such as that associated with the industrial, scientific, and medical (ISM) bands. As a result, these bands can experience significant uncontrolled interference, which requires a robust wireless technology which can cope with random interference. The Bluetooth and WiFi technologies are designed with this in mind and have therefore become increasingly popular in wireless personal and local area networks (WPANs/WLANs). Based on the IEEE 802.11 standards, WiFi is currently the predominant solution (both with and without support from the infrastructure access points) for user device connectivity [30].

Unfortunately, in conventional WLANs, access points have no means of managing resources used by *ad hoc* user connections, which contend for unlicensed band channels in a distributed fashion. Thus, WiFi is often criticized for delivering an unsatisfactory QoS experience [31]. However, WiFi generally promises users higher data rates and energy efficiencies than competing wireless technologies [30], and ad hoc connections can, in principle, be made to deliver stable performance results without assistance of the access point [32]. In the recently introduced WFD technology [33], user devices connect and communicate without help of the infrastructure by assigning one device as "Group Owner" and the others as "Clients". Each WFD Group Owner provides synchronization to all WFD Clients connected to it, allowing them to efficiently discover and page one another.

Since many user devices already support WFD (and this is only expected to increase), and WLAN access points continue to proliferate, interference on the unlicensed bands is expected to grow quickly. Thus, future WLAN users could benefit from some form of radio resource management and support from a central entity such as the cellular infrastructure [1]. The cellular network can provide node synchronization, resource management, and assisted device/service discovery.

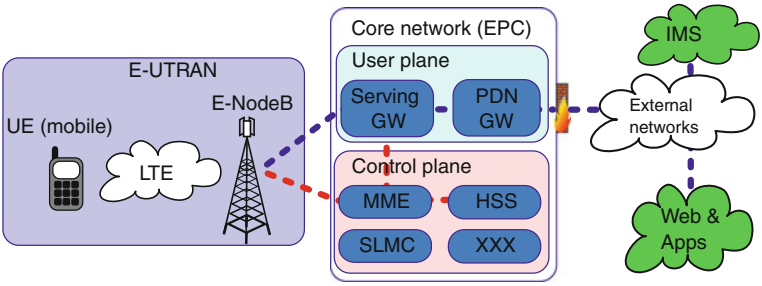

Fig. 1 3GPP core network architecture

If user devices are continually associated with the cellular network, it can also help with radio selection (LTE/WFD), power control, medium access control, and transmission format (modulation and coding rates, MIMO transmission mode, etc.). In addition, with support of the cellular network, device authentication and D2D link security can be automated. In other words, we recommend some degree of cellular network assistance for D2D connections in the unlicensed bands, which is similar to the loosely controlled D2D mode proposed in [4] (as opposed to fully controlled D2D in the licensed bands).

2 Current Technology

2.1 Cellular Networks

In this section, we consider the existing wireless technologies that are deployed today, and focus on the features that are influencing the deployment of D2D solutions and, more specifically, network-assisted D2D. We will later consider a novel architecture, in which all of the mentioned components work together to compose a fully functional system, satisfying our motivation demands.

2.1.1 3GPP Cellular Architecture

3GPP has been defining the architecture of the cellular networks and operator's core networks for the past decade. And even though there have been attempts to consider alternatives (such as WiMAX), 3GPP solutions are still predominant in current cellular networks. Therefore, in order to deploy a scalable network-assisted D2D solution, one needs to cooperate and integrate with the existing operator infrastructure.

The 3GPP infrastructure is a highly scalable system aimed at providing transparent connectivity for a large number of roaming users. Its schematic diagram as of LTE Release 8 is shown in Fig. 1. Its main components are the radio access

network (RAN), which can in fact be just as well an old 3G network, evolved packet core (EPC) and the integration with external IP multimedia services, which can also be under operator's control. Let us briefly detail the functionality of this architecture to identify which parts of it would be involved in coordinating network-assisted D2D activities of the UE's.

[e]-UTRAN UTRAN (short for "universal terrestrial radio access nework") was introduced as a concept together with 3G networks, and is now deployed in practically every city on the planet. Its main goals are handling radio resource distribution, synchronization of the base stations, and other similar low-level tasks. The services it provides are not directly visible to the mobile, except for the capability to communicate with the EPC. In addition, vast majority of the UTRAN components are located inside existing base stations, which makes it extremely difficult to modify them. However, there are quite a few pieces of important information that UTRAN collects and that could be useful for D2D.

First and foremost, UTRAN tracks the positions of the mobiles. This is required to provide roaming service, and therefore such information is collected at all times. Position information for the UE's may be published through special subsystem in the EPC, which would make it available for mobiles themselves as well as external services. The accuracy of this positioning data is far from GPS, but it has one important advantage—the device does not need to have GPS to be tracked. Moreover, most of the inaccuracy is caused by the multipath, and therefore devices that are close to each other would appear to be in similar locations irrespective of the absolute value of positioning error. As far as estimation of these values goes, interested reader is directed to e.g., works by Signal processing for positioning lab in Universitat Autonoma de Barcelona (http://spcomnav.uab.es/).

In addition, UTRAN provides UE with an encrypted data channel, that has almost ubiquitous availability. This important feature enables us to design services that employ 3GPP access network as secure communication medium, without having to go through security context establishment first.

EPC stands for Evolved Packet Core (as an evolution of the core network in 3G), and provides the devices with the capability to send traffic outside the network using IP. In particular, Packet Data Network Gateway (PDNgw) entity acts as a NAT device of sorts, that translates connections from user devices to the outgoing sessions toward the Internet.

In addition to that, EPC provides UE's with IP addresses and also hosts a variety of registers that hold information about subscribers. Most of those registers take their roots in the GSM networks, such as Home Location Register, Visitor Location Register, and so forth. What is important here, is that Core Network (or EPC in context of LTE), acts not just as a gateway for data and, naturally, voice communications, but it is also a huge database, that may know alot more about mobiles than mobiles know about themselves. Unfortunately, most of this information is well hidden behind the firewalls and is never made available to the third-party application developers.

2.1.2 LTE Concepts and Their Effect on Services

The introduction of LTE has been a decisive step for 3GPP and mobile world as a whole. LTE primarily extends an existing 3G infrastructure in the UTRAN part, but it also brings in completely new philosophy: everything is packet data in LTE. As a result, LTE core network, the EPC, deals only with IP streams, and not with voice streams as such. However, LTE also reuses a significant portion of current infrastructure. For instance, all the registers, most control plane mechanisms and the management functions of 3G core networks are still present in the EPC. LTE MAC layer enables one to establish dedicated signaling channels and send IP packets to the individual UE's without activating the "data connection", thus enabling VoIP telephony to look like the conventional one, and haul control messages over IP. The all-IP concept of LTE, therefore, makes it significantly easier to introduce new entities in the EPC, as well as new signaling to support communication with those entities.

In our work, we primarily utilize the IP connectivity provided by LTE, but its physical layer brings some interesting specifics compared to 3G. First of all, it is significantly better at handling high-speed connections. Unlike 3G, LTE is suitable for practically any kind of streaming service imaginable, including Full-HD video streaming (which "only" requires 10 Mbps connection speed), even in uplink. While building an architecture for D2D communications, we employ this capability to potentially address one of the key issues with D2D—service continuity.

2.2 Unlicensed Band Radios

Unlicensed (also known as ISM) radio bands, and especially those around 2.4 and 5 GHz, have been extensively utilized for short-range communications ever since short-range communications became necessary. Indeed, it is somewhat natural to use those bands for short range, low-power radio because of its nature—short range means that many people can reuse the same bands without much difficulty. Same approach would not work for longer waves at higher powers, as those would propagate for kilometers, jamming everyone in their wake. However, the proliferation of short-range radio technologies also created a lot of problems, such as various interference that has to be dealt with somehow. As a result, the technologies that are in use today have a variety of mechanisms that make them good at surviving harsh interference conditions, and may in many cases compromise energy efficiency for more interference tolerance.

2.2.1 WiFi

IEEE 802.11 MAC [34, 35], and the WiFi protocol based on it, are one of the major occupants in the 2,4 GHz band, and most of the mobile devices made today

support them. WiFi is, in fact, so popular, that it is hard to imagine a mobile data device without it. Some attempts have been made to make cellular-only devices, but the simple math is that in the short range under 20 m nothing quite beats WiFi in throughput or energy efficiency. However, WiFi has a huge management overhead, and setting up security associations can be quite tricky. In addition, ad-hoc mode in WiFi is nowhere near suitable for any scalable deployment it lacks security, it has extremely poor energy efficiency and so forth. Finally, one of the key limitations of WiFi is the fact that a given device can have only one role in WiFi. In particular, it can be:

- an access point, thus providing others with capability to connect,
- or an end-user device, thus associating with exactly one access point.

Therefore, in a complex D2D topology, where a single device may be a consumer and a server at the same time, the only way WiFi can work is in the ad-hoc mode. Problem is, in ad-hoc mode all the devices share the same network ID, and thus same security context. While this may be acceptable in some cases, it is not a generally applicable solution for D2D.

In the end, WiFi remains a technology for Internet access, rather than for D2D, primarily since its software part is unable to provide the necessary flexibility in topology. On top of that there are, of course, issues of spectrum efficiency, but those are not key constrains as unlicensed spectrum is free, and thus its inefficient usage is not a major issue for operators.

2.2.2 WiFi Direct

As we have discussed, WiFi is not a suitable solution for D2D straight away. However, recently some advances have been made to change that. In particular, the new WFD [33] protocol delivers new features that would enable WiFi devices to perform discovery and association much faster and in a more efficient manner. Moreover, it also allows a device to host multiple access points for others while also being connected somewhere, which in turn means more flexibility for devices to set up and drop D2D connections. In general, a WFD device can establish or accept as many independent connections as necessary (assuming link capacity and firmware allow it), and thus is very suitable as a MAC layer for D2D.

However, even though WFD provides the signaling packets for device discovery, it remains vastly inefficient at it. More specifically, the device publishing a service has to keep broadcasting information about the service, even if there are no clients nearby, while device searching for a service has to keep listening, even though there may be nobody around providing the necessary service. In the worst-case scenario, the battery drain in either case would never get rewarded, thus making the technology quite repulsive to the end-users.

2.2.3 Bluetooth and Bluetooth-LE

Bluetooth, and especially Bluetooth-Low Energy take a different approach to the energy problems of scanning outlined before. Originally designed as ultra-low power technology, Bluetooth radio in a smartphone can stay operational the entire day without changing its power consumption profile significantly, while in laptops and tablets it is barely noticeable compared to the screen, CPU, and GPU. So, Bluetooth devices can afford to stay awake and broadcast some information around them for discovery purposes. However, Bluetooth itself can not be utilized for the D2D file transfers or HD video streaming it just does not have the bitrate for it. In addition, Bluetooth lacks the necessary flexibility in terms of accepting connections while establishing new ones, thus suffering from the same problem as WiFi it is too reliant on simple access point paradigm to implement topologies necessary for D2D. Indeed, scatternets (networks where a Bluetooth master device is also a slave in a different network) are standardized, it is still difficult for a single device to perform both scanning and establishing of the connection at once.

Combined with WFD, Bluetooth provides an interesting solution for D2D communications, as it can greatly reduce the cost of scanning for services, while letting WiFi's high bitrate sine when the proximity with service is confirmed. However, such hybrid system requires two short-range radios to stay online, while also interfering between each other. Therefore, when WiFi is used for data, Bluetooth link can not be active, and if a single connection is started proximity detection, as well as interaction with conventional Bluetooth peripherals, become difficult.

2.2.4 ZigBee and Proprietary Technologies

Low-power WPAN based on IEEE 802.15.4 protocol is one of the most common starting points for sensor networks today. Zigbee is just one of the names, and one of the most comprehensive of the entire family. Originally designed for sensor networks, ZigBee is remarkably efficient and simple, but it has one downside as D2D technology it is not designed to work in fast-changing topologies. ZigBee normally optimizes itself to transfer occasional packets in a multi-hop network, not for transfer of huge amounts of multimedia between devices. Therefore, like Bluetooth, it is limited to discovery support only.

On top of the above, Zigbee is not available in most mobiles, and therefore utilizing it today is hardly possible. Same holds true for a variety of proprietary D2D technologies like Qualcomm's FlashLinQ [22, 29]. Despite all their potential benefits, their practical deployment is complicated by legal and licensing issues, as well as closed nature of the protocol specifications. Same argument applies even to more practical solutions like AirDrop which rely on existing WiFi chips: as the protocol is closed and proprietary, only Apple devices can use it, which limits the user base to the market share of a single vendor.

Table 1 Comparison of short-range radios for D2D

Technology	Bitrate	Efficient discovery	Range (m)	D2D topology support
WiFi	~30 Mbps	Not available	30–50	In ad-hoc mode with no security
WiFi direct	~30 Mbps	Not available	30–50	Supported
Bluetooth	~2 Mbps	Built-in	20–40	Supported
Bluetooth-LE	~250 Kbps	Designed for it	10–20	Not supported
ZigBee	~250 Kbps	Built-in	20–50	Supported, but insufficient bitrate
FlashLinQ	~50 Mbps	Designed for it	50–500	Supported, but no chips available

2.2.5 Summary

As a summary, let us compare the performance of different short-range radio technologies that we have discussed in the section.

One can easily see from Table 1 that the choice of technology heavily favors either Bluetooth or WFD, with similar ranges and also similar services available. However, WiFi with its significantly higher data rate remains the most attractive option that is currently on the market, if only the discovery procedures could be improved. In the following section, we will discuss how exactly WFD can be augmented to deliver the necessary discovery performance without compromising its data rate advantages.

3 Assisted D2D Architecture

3.1 Generic D2D Service Architecture

As we have discussed earlier, it is highly desirable for the operators to enable cellular traffic offloading onto D2D connections in unlicensed bands, as current multi-radio user devices are already capable of establishing concurrent LTE and WiFi links. However, WiFi lacks fast and efficient way of device/service discovery, and has no way to implement long-term D2D communications due to its short-range nature. To overcome these limitations, this section details our proposal for implementing a network-assisted D2D architecture.

Cellular traffic offloading in current 4G networks presents an interesting challenge in protocol design and we outline a complete standards-compliant solution attempted to enable seamless D2D connectivity experience to the end user. The solution was proven by assembling a demo that runs on the proposed architecture. In what follows we will rely on WFD as link-layer technology for proximal D2D connections for the reasons discussed in Sect. 1.2, yet challenges faced during the design phase are universal to all link-layer protocols, and the proposed solutions are easily extensible toward prospective LTE Direct and other potential technologies.

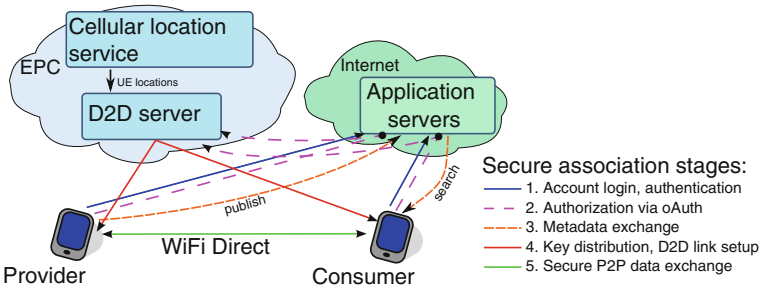

Fig. 2 Assisted D2D link establishment via dedicated D2D server

Naturally, it is very difficult to guarantee particular (good) conditions on D2D links, and the quality of such links may vary significantly over time and with movements of peers. Therefore, delay-tolerant services such as distributed caching and cooperative download (multicast) have often been considered as the prime candidates for offloading. However, if the peer devices are reasonably close to each other and the link can be predicted to remain stable, many demanding P2P services may become feasible, for instance, video streaming (remote sight), social multiplayer games, and many more. However, to enable these promising advantages, the critical design requirement is to give the clients some way to know when to set up the D2D links and how to do it exactly.

The prime issue is that a connection that is not yet established cannot be represented or managed in any conventional way. After careful consideration for network assistance possibilities, we have reached the conclusion that there cannot be a single entity that would handle the tracking of content and security as well as the link management: the content tracking needs scalability and rich functionality (which are available at the service infrastructure level), while link management requires real-time decisions based on position and radio resource availability (such information is only collected by operators for the the access network management). Therefore, a proper separation has to be made, and the most natural point, it appears, is between the features specific to the link management (managed by what we call a D2D server) and the features specific to content tracking (managed by what we call an application server). Our proposed solution is illustrated in Fig. 2 and works as follows:

(1) Each UE uses the application-layer credentials to authenticate itself with the application server (e.g., Facebook). This allows it to perform operations with content as well as authorize third-party access.
(2) The UEs authorize the D2D server of their operator to represent them on the application server when D2D connections are concerned. The D2D server never gets access to content, just to the user's profile, but this is enough to verify that the device indeed belongs to the owner of a particular application ID. The D2D server thus allows resolution between application-layer names and actual physical devices, including cases when there are multiple devices.

(3) The UE may publish or search for content links on the application server, and those links will refer to a specific content from specific user (but not a device).
(4) At the next stage, the consumer UE asks the D2D server to facilitate in establishment of D2D connection, thus resolving the application-layer link provided by the application server into an actual link-layer connection and IP address to which sockets can be bound.
(5) Finally, the P2P data exchange may commence. Note that the application server is not involved at this point, and does not track the P2P exchanges directly, only making sure that the links it gives follow the security model, but not taking part in micromanagement. The D2D server, however, may monitor and adjust the properties of the D2D link as necessary, as it only serves a small set of users.

Clearly, the proposed scheme for network assistance is not straightforward and must be justified appropriately. To show that it fits the requirements, let us go through the most important features:

- The scheme allows to maintain the current security and permission model already employed by the application services: no changes have to be made there; and if the content is only supposed to be visible to a certain group of users, only those users would get the appropriate D2D links to access it. This means that malicious users that wish to retrieve restricted content would need to crack the application server rather than having direct access to the UE hosting the content via a D2D connection, adding an extra layer of security.
- Neither of the UEs have to broadcast discovery information, or listen to discovery requests. In fact, they may keep their D2D radio interfaces off until appropriate activation command is received. This is extremely important as active D2D radios tend to consume significant amounts of energy even if no data is being sent or received.
- Anonymous D2D sharing is possible. Since the actual device ID is not broadcast at any point, one may create them on per-session basis at D2D server, thus making sure that the content provider remains anonymous on the link layer (e.g., its real device identification is never sent, even though it is actually sharing content). Such operation is not possible with distributed discovery schemes, as a permanent link-level ID that has to be broadcast there in order to identify devices.
- The operator has the capability to monitor what is going on in its "D2D network". Although the direct connections may be running in the unlicensed ISM band, it may still be extremely useful to know how much content is being shared exactly, and where. This information could be useful for future network planning, as well as for coordination with existing infrastructure WiFi networks to mitigate interference.
- The users do not have to compromise their privacy, as the system does not allow any of the parties in the network to see the complete picture: the application server does not know if a certain D2D link is ever actually used, and where it is happening, while the D2D server does not have any information about the

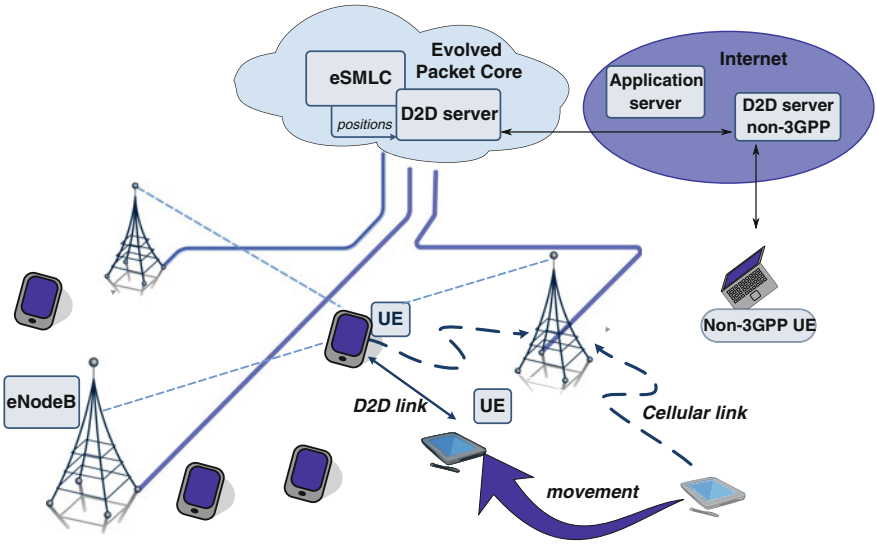

Fig. 3 Proposed D2D services layout

content that is being shared, only which devices are involved and where they are. It is worth noting that both entities (application server and D2D server) would have the exact same knowledge even without our D2D system: as network operator anyway knows where the users are, and application service provider knows which content they have, sometimes even if they are not sharing it.

Additional benefits of the proposed network-assisted D2D scheme may be identified compared to the conventional distributed solutions. One of the most notable advantages compared to the centralized solutions, however, is the ease of integration into the existing 3GPP LTE architecture. Figure 3 summarizes how such integration may be performed naturally. Indeed, the layout of our proposed scheme can be mapped to the 3GPP entities almost exactly, with the only extra entity, the D2D server, residing somewhere in the EPC of the network. This position allows the D2D server to communicate with the location center (SMLC) to learn the UE positions, while also allowing it to interact with the outside world application servers effectively.

3.2 Technology Mapping

In this section, we provide several key design details that make the implementation of our network-assisted D2D offloading system a reality within current Web and Internet. As of today, neither of those are a part of any specification or standard, yet their simplicity serves as the proof of concept for the architecture presented above.

Android and IP networks Android as any Linux-based system[1] already allows to have simultaneous connections with more than one radio, yet even if both 3GPP LTE and WiFi interfaces are active the UE has only one default gateway for sending its traffic outside of the directly connected networks. At this state, it is possible to reach the other peer on a D2D link only when the destination address of an IP packet is the WFD address of the peer. Due to lack of spare IP addresses, however, WFD link has to use private address range, which means that if the D2D link is ever disconnected the peer becomes unreachable, even if there is an alternative path present (because private range packets are not forwarded). For this reason, it is desirable to be able to reach peer's public IP address of 3GPP LTE interface through the WFD link.

One of our goals is thus to create a solution that would be transparent for already existing applications and this way ease the adoption of the proposed technology. For this reason, changes on the physical layer do not bring the desired results as it is heavily vendor specific. The similar situation is with the link layer: putting rerouting logic into the existing hardware would be next to impossible, and creating virtual interfaces causes overheads. Since applications heavily rely on existing transport layer protocols, any modifications there are not possible either. On network layer, IP addresses are in a way bound to the physical interfaces, but the forwarding decisions are made independently of the interfaces. This allows us create an interface-independent solution without the need for modifications at the upper layers involving mobile IP/virtualization.

The default configuration of an Android system allows having multiple gateway routes, but gateways are inserted into routing table with different costs. This way no load-balancing is performed and only one gateway route is used at a time. In the case of LTE (or any cellular) interface and WiFi, the LTE gateway route is preferred when the Internet connectivity is expected, whereas the WiFi link and especially WFD do not guarantee Internet connectivity at all. Hence, changing the cost of gateway route would cause unreachability of the Internet for all the applications in the mobile device.

We, therefore, propose here the *route injection* solution, which is based on allowing the mobile device to route IP traffic as usual and then inject the routing table with custom route for a particular peer. Owing to the Linux kernel in Android system, it is possible to enable routing by modifying the system value *net.ipv4.conf.all.forwarding* from 0 to 1. After this change, the Android mobile device has routing capabilities of a conventional router, and thus can forward packets from one interface to another. It allows sending IP packets with the source address equal to LTE interface public IP address and the destination address equal to peer's WFD private IP address, and the IP layer of Android system will send them through a WiFi link.

Since we only want to offload communication with a single peer, we can insert a route into the routing table saying that the peer's LTE public IP address is

[1] http://www.openhandsetalliance.com/android_overview.html

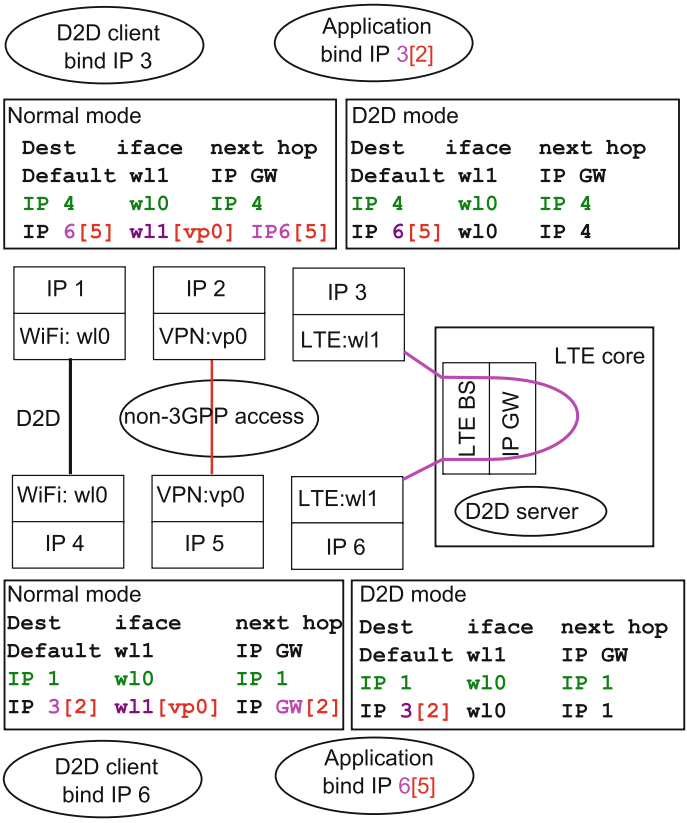

Fig. 4 Route injection example

reachable through his WFD private IP address. The insertion is performed by the command.

ip route add `PEER_LTE_IP/32` via `PEER_WD_IP`.

Once done, all the traffic with destination IP address `PEER_LTE_IP` will be forwarded to peer's WFD interface, with the shortest path going through the WiFi link, as intended. When not needed anymore, the route can be removed by running.

ip route del `PEER_LTE_IP/32` via `PEER_WD_IP`.

The insertion and/or removal is performed by the client-side application that is running as system service. Both operations are performed as commanded by the D2D server, but the removal can also be based on link conditions (for instance, poor RSSI).

When reaching a particular signal-strength threshold, the route can be inserted or removed as needed. An example of the routing table during injection in case of WFD is presented in Fig. 4. One may see that the traffic may be steered not only through WFD, but also through other forms of non-3GPP access, e.g., a campus networks.

The described routing table injection does not have to be performed on both devices participating in the D2D offloading, in fact, the return traffic can continue traveling through the cellular interface, which may be useful in some scenarios. Also the injection is not limited to a single peer: as many peers as necessary may have their own custom routes provided that they are all allocated different private IP addresses (the D2D server can ensure this).

Infrastructure and platform support Our technology prototype for network-assisted cellular traffic offloading relies on the three components:

1. Client-side service, running in background on the UEs;
2. Content database, which holds content links and handles the access control lists for content;
3. The D2D server, which we also refer to as Proximity Services (ProSe) server, which handles discovery and coordinates the connection setup.

The content database is a web platform providing data sharing services to its users, good examples are social networks like Facebook, Google+, and YouTube. The ProSe server is designed to be run by the network operator or Internet service provider (ISP). It is worth noting that alternatively to the suggested architecture layout, similar service can be delivered if both entities are run by an operator, except that getting users to use it could be more challenging.

As proof of concept, all server side functions were deployed within our cloud infrastructure. Two virtual machines were setup to act as the content register and the D2D server respectively. Content register was implemented as PHP script served by Apache web server with MySQL as database backend. In our implementation, application is a regular website that gives a registered user the possibility to post its intent to share some content, or to search for shared data records (acting like a torrent tracker). The user posts only the information required to access the data rather than its location, i.e., the sharing protocol and port number, while the IP connection setup is assisted by the D2D (ProSe) server. We have introduced a new addressing scheme required by D2D connections, as well as new protocol identifier in the URI: *"d2d://"*. End-user devices can be easily configured to interpret this protocol type as a request to start client service, that in turn is capable of communicating with the ProSe server to resolve the username of the serving peer into an actual connection to one of his devices.

ProSe server was implemented in Python as a standalone application using HTTPS as transport for the control messages. The system assumes the UE's mobile data link to be up during the service usage. Our solution shows that content register and ProSe server(s) are easily integrated in a seamless fashion into the existing web serving infrastructure, and do not require any obscure design tricks that would not fit into the well-defined web paradigm.

The current implementation uses Sony Xperia ZL phones provided for the project by Sony Mobile. Android, as open source platform provides the needed flexibility in configuration and available tools to fulfill the requirements demanded from the user devices. One of the main features required by the solution is that the

UEs should have both cellular and WiFi connections up simultaneously. Due to the energy consumption constraints, most of the systems avoid such operation, so it was necessary to bypass the native Android service controlling WiFi, and interact with the WiFi driver directly. Similarly to GNU Linux, Android provides the needed tools: *wpa_supplicant* interface controlled via *wpa_cli* utility. Unfortunately, no stock firmware allows access to those utilities, even for developers, and therefore an aftermarket firmware Cyanogenmod maintained by the FreeXperia group (http://www.cyanogenmod.org) was used, which can be deployed by anyone who owns an appropriate UE and unlocked device.

One of the key requirements for the end-user mobile platform is to be capable of receiving incoming P2P connections through the cellular data link. Considering the fact that most operators use the private IPv4 address pool to assign to the user devices, and provide Internet connectivity through cone NAT and firewalls, the access to the services running on the user's device from outside its local link is not trivial. One of the possible solutions to overcome this issue would be using IPv6, but the mobile operators do not rush toward IPv6 support, as replacing the existing infrastructure is extremely costly. And even then it is unlikely they would get rid of the firewalls. The simplest tested option for a technology demonstration is encapsulating the mobile data link of both communicating devices inside a VPN tunnel to a common VPN server, thus moving both devices into the same IP subnet. However, due to excessive complication of the solution and large tunnel overhead, this approach is not scalable. Discussing this issue with local operators, we were able to negotiate with TeliaSonera Finland Oyj for an APN that provides the user devices with a publicly routable IPv4 address. Later on similar agreement has been reached with AT&T in the US.

In an actual deployment, however, it will be critical to come up with a way to bypass operator's firewall, as there are not enough IPv4 addresses to allocate to every mobile device. The most reasonable solution, it appears, is to allow the D2D server to negotiate firewall policies just before the actual connection is set up to provide NAT traversal functions. Such solution would allow the devices to set up direct connection without having to resort to VPN, with no overhead during data transmission, as well as very small security impact.

Results of QoE evaluation The evaluation of the D2D technology prototype was performed in two directions. First, system-level simulations have been used to make sure that the solution will be scalable in practical network deployments. More details on this work will be given in the following section. The second evaluation direction has been to assess the actual end-user experience while using our D2D prototype under different link conditions. This was done with the Sony Xperia devices and various multimedia applications as benchmarks. The performance of the video streaming as well as that of the connection management procedures have been thus evaluated.

We have conducted several test with video-on-demand based on the idea that a user is likely to be sharing a popular short clip (e.g., from YouTube), which could be needed in his proximity. The video clips for the testing have been selected with

various bitrates from 300 kbit/s (very poor quality) and up to 5 Mbit/s (HD quality). The duration of the clips was chosen to be 5 min. The networks used for testing were TeliaSonera Finland (LTE, DC-HSPA), AT&T US (HSPA+, HSPA), and T-Mobile US (LTE, DC-HSPA, HSPA). D2D links have been tested in office environments with campus networks (on university and large company campuses), in open-air with close to no interference, and in urban environment of a medium-sized city.

The testing was performed under different conditions of cellular and WiFi networks, with the following results:

Poor cellular conditions (HSPA, HSPA+)

- Video over cellular is not possible;
- Signaling messages delayed significantly (order of seconds);
- Attempts to use cellular for data make system unusable;
- WiFi streaming works fine, but connection establishment is noticeably slowed down by cellular access times.

Good conditions 3G (DC-HSPA)

- Speed over cellular is sufficient for low resolution videos, but random stops are probable, caching is necessary;
- Signaling messages are delayed, but not significantly enough to make any noticeable difference;
- HD video is sufficiently overloading the system making it unusable just like in poor conditions case.

Excellent 4G (LTE)

- HD video streaming is possible and does not require buffering;
- Significant transfer delays are noticeable even on low-bandwidth transmissions, those have nothing to do with capacity and are caused by the nature of LTE access:
  - Measurements indicate delays of approximately 50 ms between two peers in the same cell, and up to 80 ms for peers in the same area but with different operators;
  - For comparison, WFD delays seldom exceed 5 ms.
- Signaling messages are handled in a timely matter, no matter the load.

Further, when it comes to WFD, one would expect interference to play a major role resulting in poor indoor performance. However, our measurements from the user's point of view indicate, that even in a highly populated area (e.g., busy office) the link length has a much stronger effect on the transmission quality, making it next to useless at distances of approximately 80 m. We have not observed strong enough WiFi congestion such that HD video streaming would not be possible at all, even in university campus environment with massive amounts of interfering access points on all frequencies.

Our conclusion is that depending on the quality of the cellular link the usefulness of the D2D connections may vary. However, even with the state-of-the-art LTE technology, the MAC transfer delays are up to 10 times higher than those with WiFi, and with 3G cellular technologies HD video streaming is not even possible due to capacity limits of real deployments. This means that D2D over WFD is extremely competitive.

3.3 Implementation Prospects

For mobile network operators, D2D connectivity is becoming vital to enable traffic offloading from the core network and to realize efficient support of social networking through localization. Along these lines, our network-assisted D2D technology prototype has been implemented to identify the major challenges and potential gains of enabling direct connectivity between proximal mobile devices. Below we summarize our most important findings and lessons learned.

1. The successful integration of the D2D connectivity with the existing 3GPP LTE architecture shows that there are no technical issues that would prevent the application service providers from enabling D2D communications for their clients. Moreover, some of them could benefit significantly by using this new infrastructure to design new services that were not possible before without continuous GPS tracking.
2. Successful deployment of the network-assisted D2D service on the Android platform indicates that the OEMs will be easily able to implement the necessary control protocols. Certain platforms, that so not implement standards-compliant networking stack, may face some difficulties with the route injection procedure required to steer traffic.
3. The quality of D2D experience by far exceeds the best cellular connections within reasonable ranges between peers (under 50 m). WFD enables HD video streaming as well as real-time applications easily and with decent energy efficiency.
4. The lack of external connectivity capabilities and firewall policies deployed by mobile operators significantly limit the availability of D2D connections for as long as IPv4 remains the predominant addressing solution, as there are just more mobile devices out there than are there IPv4 addresses.
5. The success of the entire proximal D2D concept relies heavily on the operator's support for cell tower-based positioning, as well as some willingness for cooperation between the operators. This should become reality once the appropriate standards are completed by 3GPP.

Overall, we are confident that the challenges identified during the implementation of our D2D traffic offloading prototype will be resolved within extremely short time, with first services supporting the new capabilities shortly after. As the

pressure from both the services and the capacity points of view is rising, it is just a matter of time before market solutions are deployed, and the architecture described here will likely be the foundation for them.

4 Performance Evaluation

4.1 Simulation Study of Assisted D2D

Depending on client mobility patterns, some services are better suited for proximity-based network offloading than others. For example, if D2D peers are non-stationary, the quality of the link may change dramatically over short periods of time [36], thus making it difficult to guarantee service. In these cases, the best candidates for network offloading are delay-tolerant services, i.e., those that can be queued until the D2D link recovers or the data session can be moved back to the infrastructure network (e.g., video-on-demand or file transfers). However, if both clients are stationary, many other P2P services, such as cooperative streaming and social gaming, can be offloaded onto D2D links with good results. In all cases, in order to justify offloading from the client's perspective, the D2D link must provide improved throughput, delay, and/or power performance compared to the infrastructure path. In this section, our goal is to understand how network-assisted WFD performs relative to LTE (i.e., the direct vs. infrastructure path). From the network perspective, we are interested in system capacity; from the user's perspective, we care about throughput, medium access time, and power efficiency. Since these questions are difficult to address analytically, we first perform extensive system-level simulations to mimic the behavior of D2D and infrastructure communications between client source/destination pairs and compare their performance.

4.1.1 Evaluation Methodology

This subsection introduces the network entities and respective mechanisms required to enable network-assisted WFD. In particular, it describes our evaluation methodology, which is able to accommodate a wide variety of prospective D2D technologies and P2P usage models.

Network entities In our study, we consider a heterogeneous wireless network composed of multiple communicating entities with diverse capabilities comprising a variety of radio technologies (see Fig. 5 for entity diagram). First, there is an underlying 3GPP LTE network represented by E-UTRAN Node B (eNB) base stations. Each eNB is connected to the core network, providing cellular connectivity to all wireless clients associated with it.

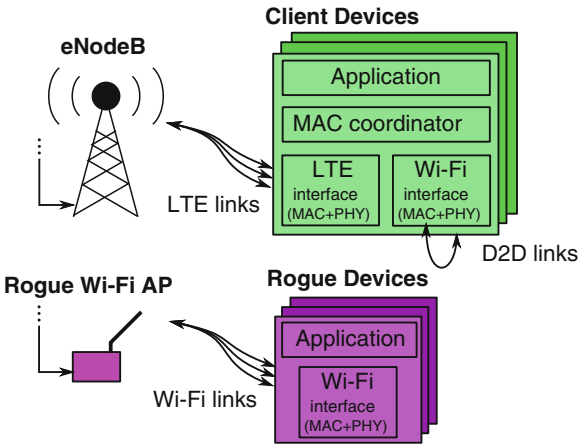

Fig. 5 Network entity diagram

Each eNB is accessed by a number of multi-radio client devices capable of communicating over LTE and/or WiFi. Each client runs applications that use the device's MAC coordination function to determine which wireless technology to use. The MAC coordinator can be regarded as a layer 2.5 entity implemented in hardware or middleware, but it can also be implemented in software as a virtual network interface. Depending on the recommendation of the MAC coordinator, a client may direct data flows onto the LTE or WiFi interface.

We also account for interference on the unlicensed bands from devices engaged in regular WLAN communications with neighboring WiFi Access Points (APs). These devices compete for channel resources with multi-radio clients. Since we assume they are not associated with the cellular network, their activity on the unlicensed bands cannot be monitored or managed by the LTE network, hence we refer to them as "rogue" devices.

Traffic flows and network loading In our methodology, according to the recommendations in [37], we assume that random number N of LTE clients placed uniformly across the deployment area. All clients have an LTE and a WiFi interface, and they are capable of engaging in LTE and WFD communications concurrently. The client density is high enough that each client is within D2D range of at least one other client. However, only 50 % of clients are data "sources", i.e., have data to send. Their traffic loads are modeled as full buffers with packets of 1,500 bytes each.

Instead of modeling content distribution and demand among clients explicitly, we assume that a certain percent, x, of source clients are within D2D range of their P2P "destination" clients. For simplicity, we assume that P2P communication is uni-directional, i.e., there is only one source and one destination client in any given P2P session. However, since destination clients are chosen randomly from within D2D range of source clients, two source clients in close proximity could be randomly given each other as destinations, effectively creating bi-directional P2P

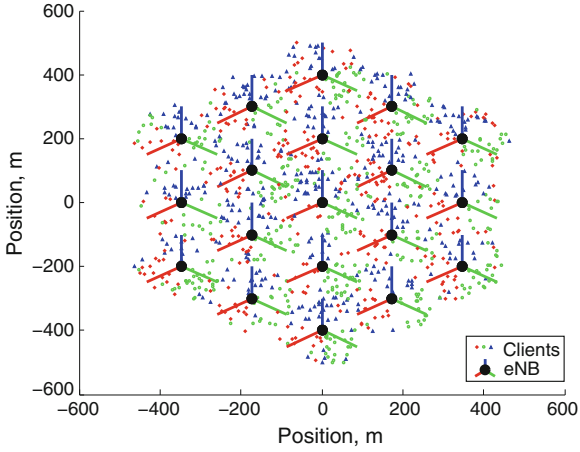

Fig. 6 Network deployment

communication. Nevertheless, in the analysis, this would still be two separate P2P sessions.

Rogue devices also have full buffers with packets of 1,500 bytes, but their traffic always travels to the APs they are associated with. To simplify the evaluation methodology, we do not model WiFi AP downlink traffic. Instead, we adjust the number of rogue clients to mimic the desired level of competition on the unlicensed bands.

4.1.2 Example Scenario

Cellular deployment In order to estimate the benefits of network-assisted WFD, we construct an example scenario based on modern urban conditions. The LTE infrastructure network comprises 19 hexagonal cells of 3 sectors each (see Fig. 6). Each eNB supports LTE Release 10 technology, and the distance between neighboring eNBs (inter-site distance) is 200 m, resulting in a cell radius of approximately 110 m. A wraparound technique is used to improve precision of the simulation at the edges of the deployment area [38].

All cells share the same 60 MHz bandwidth, which is split into three pairs of 10 MHz bands for FDD operation. Every cell is divided into three sectors, and each sector is allocated a pair of 10 MHz bands, resulting in a $1 \times 3 \times 3$ frequency reuse pattern. 3GPP LTE clients associate with eNBs based on the best downlink SINR, with a handover threshold of 1 dB. For more details on the configuration of the reference LTE network, the interested reader is directed to Table 2 and relevant standardization documents (e.g., 3GPP TR 36.814-900 and ITU-R M.2135-1). For performance verification purposes, we implement a calibration scenario from 3GPP TR 36.814-900, Table A-2.1, and run the corresponding tests. Our simulation results fall well within the required limits for both cell-center and cell-edge spectral efficiency targets (see Fig. 7).

Table 2 Baseline simulation parameters

Parameter	Value/source
Core parameters	
Client Tx power limit	23 dBm IRP per interface
Mobility model	Random direction, 3 km/h speed
Observation period	10 s
LTE	
Propagation model	ITU-R M.2135-1 [40], Tables A.2.2-1, A1-3
Shadowing model	ITU-R M.2135-1 [40], Sect. 1.3.1.1
Medium access	Round-robin scheduling
Power and rate control	Closed-loop SINR target at 15 dB
Frequency resources	10 + 10 MHz FDD in each sector, short CP
Signaling mode	2 out of 20 special subframes, 10 ms frame
RF equipment	ITU-R M.2135-1 [40], Table 8-4
Antenna configuration	1 × 2 (diversity reception at eNB)
WiFi	
Propagation model	Empirical, based on [41]
Shadowing model	Correlation only, based on [42]
Medium access	CSMA/CA, -76 dBm yielding threshold
Power and rate control	Open-loop SINR target at 25 dB
Frequency resources	20 MHz TDMA
Signaling mode	Green-field, control rate 18 Mbps, RTS/CTS
RF equipment	Noise Fig. 7 dB, noise floor-95 dBm
Antenna configuration	1 × 1 (single antenna)

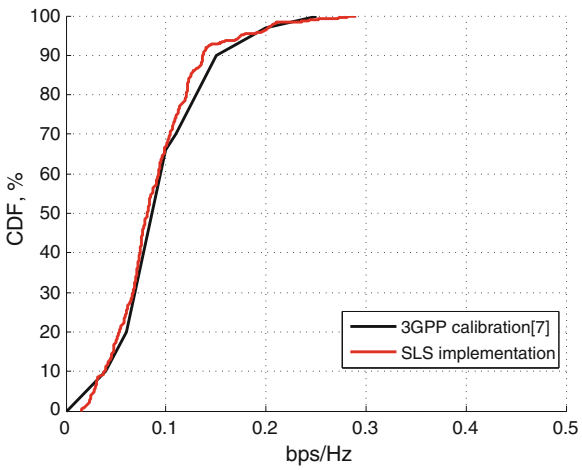

Fig. 7 LTE calibration, spectral efficiency

Conventional WLAN deployment We assume that all APs and their respective clients (i.e., rogue devices) run the same version of the technology, namely IEEE 802.11-2012 [35]. To mimic realistic deployments, rogue devices are positioned

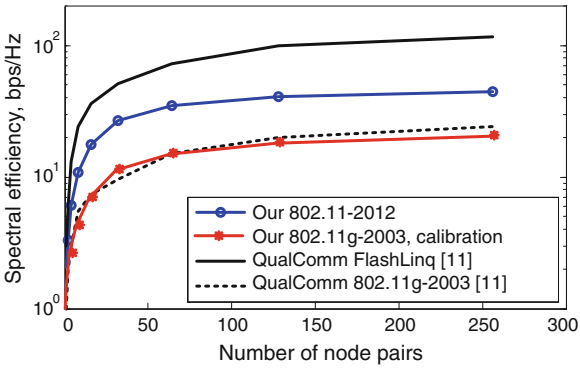

Fig. 8 WiFi calibration, spectral efficiency in 1×1 km area

around their respective APs. APs may be located anywhere inside the deployment area, recreating hot-spots similar to those in cafes, transportation hubs, etc. A rogue's distance to its AP is constrained by the maximum tolerable path loss. APs and rogues do not move during the simulation, thus handover is not considered.

Our study assumes that all WiFi connections (AP and D2D) use the same frequency bands and have to yield to any active transmission for which the received power exceeds the designated threshold. For more details on the configuration of WiFi networks the reader is referred to Table 2 and Atheros driver documentation available online [39]. For calibration purposes, we employ reliable results from publications on ad-hoc WLAN deployments. Calibrating against WiFi performance results in [22], we achieve near perfect alignment (see Fig. 8), and reasonable coherence with FlashLinQ technology.

Additional D2D functionality On top of the above technologies, we deploy our new WFD devices, that are in most ways similar to WiFi AP's except for their traffic destination. While conducting this simulation study, we have developed an advanced system-level simulator (SLS) based on the LTE evaluation methodology described in TR 36.814-900 and current 802.11 specifications. This simulator is a flexible tool designed to support diverse deployment strategies, traffic models, channel characteristics, and wireless protocols. It models all of the conventional LTE infrastructure and client deployment choices (hexagonal vs. square cells, environment with or without wraparound, uniform versus clustered client distribution, etc.).

Every client has its own dedicated traffic generator, enabling a variety of data patterns across the deployment. Channels are modeled to incorporate all relevant source, destination, and environment characteristics, and each client is capable of supporting multiple radio interfaces, which actively interact up and down the stack. This simulator was not made to be task-specific, but rather an extensible "sandbox" suitable for supporting different D2D scenarios and infrastructure deployments. It should be noted that it is difficult to find an off-the-shelf solution to simulate assisted D2D, which motivated us to build our own.

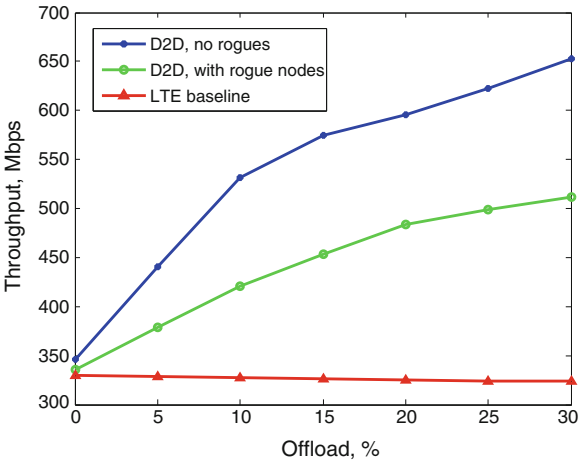

Fig. 9 Total cell throughput (sum rate WiFi + LTE)

Table 3 Normalized energy expenditure

Operation mode	LTE	WiFi
Offline	0	0
Idle/circuit power	0.1	0.1
Energy sensing	N/A	0.25
Data reception	0.5	0.5
Data transmission	1.0	1.0

4.1.3 Simulation Experiments and Results

For a complete picture of the benefits to network and client from offloading onto WFD, we analyze the performance of network-assisted WFD under a variety of interference conditions (i.e., with and without WiFi APs and associated rogue devices). We do not model any particular type of client traffic, but instead consider different client densities in order to observe how network offloading onto WFD performs under different load conditions.

We also vary the percentage of approved WFD connections (i.e., those that outperform their alternative infrastructure path) from 0 to 30 %. Based on current P2P traffic statistics and client behaviors, we consider it unlikely that more than 30 % of clients will be within D2D range of their peers, but this could change in the future.

The results for total cell throughput are presented in Fig. 9. In these curves, the throughputs from LTE and WFD data sessions are totaled per cell, based on the source client's cell association. One can easily see that offloading LTE traffic onto WFD links results in a significant boost in cell throughput, actually doubling it at the 30 % offload level. However, if interfering rogue devices are present, throughput gains are more modest, but they are still nearly 50 % at the 30 % offload level.

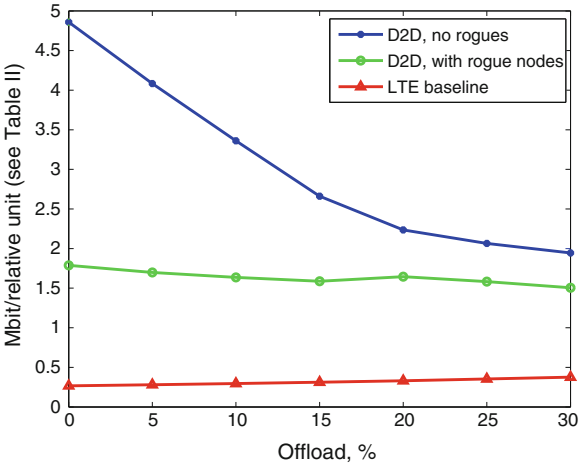

Fig. 10 Energy efficiency

Energy efficiency is typically measured in bits per Joule and is therefore agnostic to the particular technology involved. Since device energy consumption figures are generally vendor specific, we use the power coefficients from Table 3, which are not based on any particular implementation.

The energy efficiency curves in Fig. 10 clearly indicate that communication over WiFi is significantly more power efficient than over LTE. This is in large part due to WiFi's higher data rates. In addition, LTE clients are allocated small frequency chunks across multiple time slots, thus their transceiver circuitry has to stay active for extended periods of time, while the actual data rate is relatively low. By comparison, the WiFi MAC activates the transceiver only when it is actually accessing the channel.

Even when WiFi users are forced to delay their channel access due to RTS or CTS messages, they can sleep during those periods of time. Then, when they finally do get access to the channel, they utilize the **entire** bandwidth. As a result, only a handful of WiFi interfaces across the deployment are powered on at any given time, and those are all either transmitting or receiving data.

One of the known issues with the IEEE 802.11 MAC is its excessive medium access time in the presence of heavy traffic. However, this understanding is based on legacy IEEE 802.11g-2003 [34] behavior. Our study models the latest version of the standard, IEEE 802.11-2012. With this latest version, the MAC transfer times (i.e., the time a packet spends in the MAC layer and below) of WFD clients in the absence of rogue devices are sometimes shorter than those of LTE (see Fig. 11). This is primarily because in LTE data rates are significantly lower. When rogue clients are present, the situation benefits LTE more, yet no considerable degradation can be observed.

As this simulations show, there is significant potential for both network and client performance improvement from network offloading onto WFD in urban environments. For example, in case of 30 % offloading, cell throughput can be

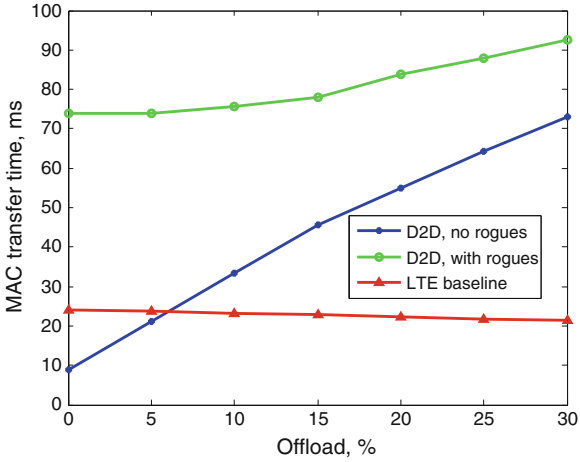

Fig. 11 Average MAC transfer time

nearly doubled, while energy efficiency can be improved by as much as six times. Similar results can be easily obtained for other deployments, channel models, versions of WiFi and so forth.

4.2 Analysis of Assisted D2D

The simulation approach presented in Sect. 4.1 has apparent limitations. Although it is flexible enough to accommodate next to every possible scenario, the simulation models may be unsuitable, when optimization tasks have to be performed. Either to approach optimization problems with D2D networks or to estimate fine-grained statistics (such as e.g., system blocking probability) by analysis, it is important to formulate an analytical model that couples a cellular network in licensed bands and a D2D network in unlicensed bands. In what follows, we give an example where a joint D2D/cellular system serves real-time flows of data from one user to another (termed *sessions*), which adhere to certain time-spatial process.

More specifically, we propose a *general methodology* for modeling assisted offloading of cellular licensed bands user sessions onto D2D connections in the unlicensed spectrum. The proposed methodology is flexible enough to accommodate various offloading scenarios, radio selection algorithms, user performance characteristics, and advanced wireless technologies (e.g., WFD and LTE). We are primarily interested in evaluating session *blocking* and *reject* probabilities, which are when a user session is not admitted by the D2D network, cellular network, or both. However, given the increasing importance of energy efficiency for mobile battery-driven user devices [43], we are also interested in characterizing the *energy expenditure* of a typical data session based on the power model from [44].

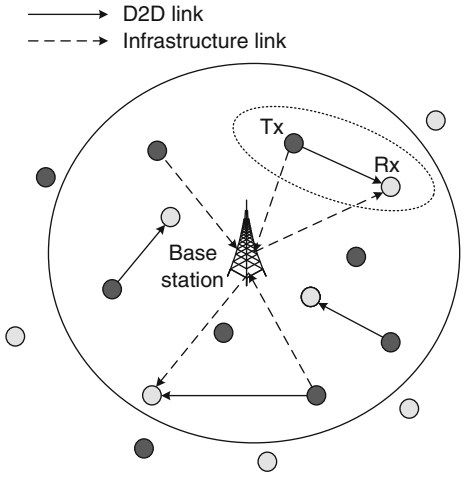

Fig. 12 Assisted offloading of cellular traffic

In our work, sessions are initiated according to a *Poisson point process*—one of the fundamental ingredients of stochastic geometry. Such processes have been used extensively to characterize the coexistence of cellular and mobile ad-hoc networks [45], study device discovery aspects of FlashLinQ [46], assess the performance of multi-tier heterogeneous cellular systems [47], and capture the distributions of transmit power and SINR in D2D networks [48]. The application of stochastic geometry makes it much easier to model spacial randomness of user sessions with respect to different session characteristics such as SINR and rate. However, the existing literature fails to provide a unified framework for modeling the intricate interactions between a cellular network in the licensed bands and a D2D network in the unlicensed bands under dynamic load.

4.2.1 General Analytical Model

We concentrate on a cellular network in the licensed bands coupled with a D2D network in the unlicensed bands both serving uplink data. In particular, we focus on traffic within a single cell of cellular network, where R is radius of the cell. The considered traffic corresponds to real-time sessions with the certain target bitrate r. For every session i, we differentiate between the data originator T_i termed *transmitting user* and the respective destination termed R_i *receiving user*. Transmissions on the two networks do not interfere with each other due to non-overlapping frequency bands. Further, we assume that each and every T_i may send its data to R_i via either the cellular network (*infrastructure path*) or the D2D network (*direct path*) as shown in Fig. 12. For the sake of simplicity, we disregard any communication that is not directed at a particular D2D partner.

To explicitly model system randomness, we employ the following stochastic processes that facilitate such analysis. To this end, we make two principal assumptions.

Assumption 1 *The transmitting users are spatially distributed as a Poisson point process (PPP) into the three-dimensional space, which includes time component and two-dimensional location component. We assume that time and location are independent, so that density may be split into stationary component* λ *and* $f(x), x \in R^2$. *For the sake of simplicity we assume that* $f(x)$ *is homogeneous within the cell of radius R and* $f(x) = 0$, *otherwise.*

The first assumption implies that the locations of transmitting users are distributed uniformly within the same circle R [49]. Moreover it maybe easily proven that arrivals onto the time axis form Poisson process of rate λ.

Assumption 2 *For a transmitting user* T_i, *the corresponding receiving user* R_i *arrives simultaneously with* T_i, *such that the location of* R_i *is distributed uniformly within a circle of a particular radius R.*

We further assume that the duration of a real-time session by each T_i is exponentially distributed with mean $\frac{1}{\mu}$.

4.2.2 Analytical Model for Cellular Network

We consider an isolated cell of a centralized network, which is exempt from inter-cell interference. This formulation implies interference-free communication, as user transmissions are orthogonal by network design. Here, we only address the uplink component of the infrastructure path, that is, from a user to the cellular base station (BS). This refers to an assumption that downlink channel is typically more reliable and has more resources than the uplink. Hereinafter, we exclusively focus on the transmitting users.

Assumption 3 *We model the channel propagation according to the standardization documents* [40] *and assume for tractability that for the session i the signal to noise ratio (SNR) per power unit* γ_i *is expressed as:*

$$\gamma_i = \frac{G^k}{d_i^k}, \tag{1}$$

Where d_i the distance between the BS and the transmitting user T_i, k is the propagation exponent, and G is the propagation constant.

Without loss of generality, we further assume that the data rate is continuous, and the power/rate mapping is given by well-known Shannon's formula.

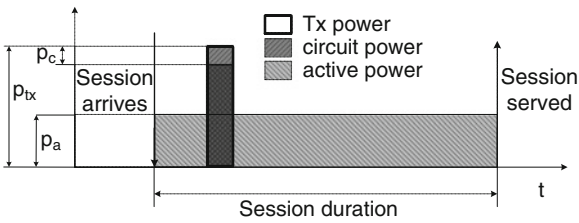

Fig. 13 Power model of a transmitting user

Assumption 4 *The transmit power p_i of a user i and its transmit rate r_i are coupled by Shannon's capacity theorem for interference-free environment:*

$$r_i = w \log(1 + \gamma_i p_i), \tag{2}$$

where p_i is the output power of the radio frequency (RF) power amplifier, γ_i is the SNR per power unit in (1), and w is the spectral bandwidth.

We employ a realistic but also analytically tractable power model (see Fig. 13) with different *power levels* for every transmitting user and take into account antenna efficiency η, which is set to one without loss of generality. This model is similar to that in [44] and includes

- dynamic transmit power $p_{tx} = \frac{1}{\gamma_i \eta}\left(e^{r_i/w} - 1\right) + p_c$, which is incurred whenever the user is transmitting ($r_i > 0$), with p_c being some constant circuit power;
- active power p_a, which is consumed whenever the user does not transmit but waits for a transmission opportunity.

The BS governs the network by applying *transmission policies*. A particular policy generally decides on user admission, scheduling, and transmit power. Whenever admitted, a transmitting user occupies a fraction of the time frame resource and sets its power as commanded by the BS to achieve the data rate given by (2). The BS makes a new decision on scheduling allocations and transmission power for all active users at every new arrival or when an existing session is served and leaves the system.

For the *Maximum Rate* (MR) policy, we assume that a user sends its data at the maximum allowed transmit power level. In the absence of interference, this ensures that the data rate of each user is thus maximized. Given the relationship in (2), the instantaneous data rate for the session i is determined by the maximum transmit power $p_{\max}$ as:

$$r_i^{\max} = w \log(1 + \gamma_i p_{\max}). \tag{3}$$

Consequently, the system admits a newly arrived session if it still has sufficient resources to serve it. In other words, each ongoing session i has to occupy exactly $r/r_i^{\max}$-fraction of time frame duration, while for all sessions it holds the following:

$$\sum_{all\ sessions} \left(\frac{r}{r_i^{\max}}\right) \leq 1. \tag{4}$$

With the MR policy, even for the increasing arrival rate, the system is under-utilized in the sense that there is always a (vanishingly small) portion of time frame resource that is unused by the active sessions.

As an alternative, the *Full Utilization* (FU) policy ensures that the system time is always used completely. More specifically, each admitted session is allocated an equal portion of the frame duration, i.e., $\frac{r}{r_i} = \frac{1}{n}$, and users adjust their transmit power to match the required target bitrate. Clearly, in case of n active sessions, it holds the following:

$$\frac{r}{r_i} = \frac{1}{n}, r_i = rn, \forall i = \overline{1,n}. \tag{5}$$

Therefore, in order to admit a new session, the BS has to increase the power of already running transmissions, such that they would fit into the smaller allocations. If it is not possible for at least one of n active sessions (or the new session), that is, $r_i^{\max} = w\log(1+\gamma_i p_{\max}) < (n+1)r$, a newly arrived session cannot be admitted by the system. Otherwise, the system time is re-allocated for n+1 sessions and users employ other (higher) transmit power levels:

$$p_i = \frac{1}{\gamma_i}\left(e^{(n+1)r/w} - 1\right). \tag{6}$$

As a summary, the MR and FU policies offer a flexible choice between more system capacity (resulting also in higher power consumption) and better network resource utilization (enabling some transmit power savings). By considering both policies, we ensure that the system may support good balance between network and user side performance.

4.2.3 Analytical Model for D2D Network

As D2D network resides in the unlicensed bands, several transmission sessions can be activated simultaneously. Therefore, the D2D system is inherently interference-limited and this interference has to be accounted for explicitly by the analysis. As previously, the channel gain $\gamma_{i,j}$ between the transmitter T_i and the receiver R_j depends on the distance $d_{i,j}$ between them similar to (1).

By contrast to our cellular network model, we make the following assumption on the power/rate mapping.

Assumption 5 *The transmit power p_i of a user and its transmit rate r_i are coupled by Shannon's capacity theorem for interference-limited environment:*

$$r_i = w\log(1+SINR_i) = w\log\left(1+\frac{p_i\gamma_{i,i}}{N_0+I}\right), \tag{7}$$

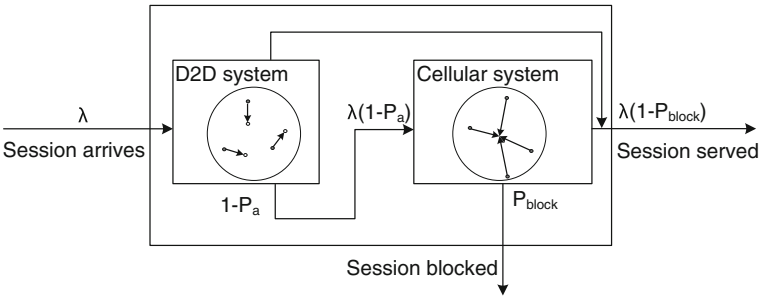

Fig. 14 Considered system operation

where p_i is the output power of the RF power amplifier, $\gamma_{i,i}$ is the channel gain between the transmitter and the receiver belonging to session i, w is the spectral bandwidth, N_0 is the fixed noise power level, and I is the level of interference.

As previously, we consider two different power levels for every transmitter in the system: (i) transmit power consumption (including the circuit power p_c) and (ii) active power consumption. The transmit power is assumed fixed at its maximum $p_{\max}$.

We impose that the noise plus interference power does not exceed some network-wide threshold $N_0 + I \leq KN_0$ (see related discussion in Sect. 4.2.5). Further, it is assumed that the D2D network of n-1 active users admits a new session n if for the set $\{T_j\}_{j=1}^{n}$ of transmitters the following conditions hold at each receiver R_i, $i = \overline{1,n}$:

$$\frac{p_{\max}\gamma_{i,i}}{KN_0} \geq e^{\frac{r}{w}-1} \text{ and } p_{\max}\gamma_{j,i} \leq N_0, , i \neq j, \tag{8}$$

where the value of K is fixed throughout the D2D network. These conditions imply that the required bitrate r can be achieved on each link i and the interference on R_i produced by T_j does not exceed the given threshold N_0.

4.2.4 System Operation and Metrics

When a new data session arrives into the system, we assume the following consecutive service. First, cellular network attempts to offload the newly arrived session onto the D2D network by performing an admission control procedure (8). In case the session is accepted, it is served by the D2D network without interruption until when it successfully leaves the system. Otherwise, the cellular network attempts to serve this session given MR or FU admission criteria. Finally, if the session cannot be admitted by the cellular network as well, it is considered blocked and permanently leaves the system. General system operation is illustrated in Fig. 14.

We remind that the arrival rate on the D2D network is λ (see discussion after Assumption 1). Due to the Poisson property of thinned flow, the arrivals on the cellular network (those rejected by the D2D network) also follow a Poisson process of density $\lambda(1 - P_a)$, where P_a is the D2D network *accept* probability. Abstracting away the point locations for analytical tractability, we assume that the arrivals on the cellular network are also uniformly distributed within the circle of radius R. Formally, it is not true due to selective thinning of the arrival flow by the D2D network, extensive simulation confirms that such approximation is very precise.

Consequently, the system *blocking* probability P_{block} may be established as follows:

$$P_{\text{block}} = 1 - [P_a + (1 - P_b)(1 - P_a)], \tag{9}$$

where P_a is the D2D network accept probability and P_b is the cellular network blocking probability.

Another important metric considered by this work is the energy consumption ϵ of a typical session given that it satisfies the bitrate requirement. This follows from the Little's law and the definition of the average energy consumption as:

$$E[\epsilon] = \frac{E[P]}{\lambda P_a}. \tag{10}$$

Here, P_a for the D2D network may be replaced by $1 - P_b$ for the cellular network.

4.2.5 Applicability of Analytical Model

Below we discuss how our methodology corresponds to the practical wireless technologies. As per Assumption 1, the proposed model can actually mimic the dynamic interworking between 3GPP LTE and WFD. However, the main derivations are more general and may very well be extended to e.g., accommodate D2D operation in licensed bands. More specifically, our assumption about the exponential holding times for new data sessions is only made for the sake of clarity. All our derivations can be generalized for an arbitrary session length distribution. To explicitly model interactions between LTE and WFD, we need to assume that the system users are multi-radio terminals and have capability of using both wireless technologies. We further require that a user constantly maintains a (signaling) connection with the BS, which controls the offloading procedure.

According to Assumptions 2 and 3, Shannon's capacity theorem is used as the power/rate mapping. We have recently shown that it alone may serve a reasonable approximation of current wireless networks [50]. However, to make our model even more realistic, we apply several additional restrictions imposed by the modulation and coding schemes of LTE and WFD. In particular, transmitting users are not allowed to exceed some maximum feasible data rate $r_{\max}$ (of around

60 Mbps for LTE and 56 Mbps for WFD) by limiting the maximum usable SINR value on the receiving end:

$$r_{\max} = w \log(1 + \mathrm{SINR}_{\max}). \tag{11}$$

Hence, γ_i and $\gamma_{i,j}$ cannot grow infinitely as d_i or $d_{i,j} \to 0$, and after some $\gamma_{\max} = \frac{1}{p_{\max}} \left(e^{\frac{r_{\max}}{w}} - 1\right)$ the data rate would not increase any further.

For the D2D network, the admission control procedure in (8) determines if a particular session may be accepted. With perfect D2D planning, the power level of each transmission would be selected individually, as to maximize e.g., total throughput of the network. However, actually performing such planning for a practical network is infeasible due to prohibitive overheads. Therefore, we employ a simplification following the ideas used by the IEEE 802.11 protocols. We assume that (i) the transmit power is fixed and (ii) the background noise never exceeds some fixed threshold N_0. Each time this condition does not hold on the receiving end (or would not hold on one of the other receivers), the link backs off from transmission and leaves the D2D system.

This procedure essentially matches the carrier sensing mechanism of WFD and also guarantees that the interference caused by a particular transmission on any given receiver will never exceed N_0. However, what it does not guarantee is that the sum of many interferences from all running transmissions combined does not exceed N_0. To account for cumulative interference, we also introduce a specific link budget reserve factor K in (8). The practical value of K can be estimated as the maximum number of potentially interfering links in the vicinity of the receiver. Our study shows that $K = 6$ provides sufficient protection against aggregate background interference.

4.2.6 Analysis of D2D Network

Below we provide a summary of our analytical findings to evaluate the primary D2D-related performance metrics. The most important results are formulated as theorems, while auxiliary derivations are presented as propositions.

Stochastic model We begin introducing our generic analytical approach by example of D2D network and discuss its applicability. This approach is employed for the cellular system in what follows.

Accordingly, the D2D network is observed at the particular moments t of session (user) arrivals and departures. System behavior is represented by a stochastic Markov process $S(t)$, where the future process evolution is determined solely by the set of ongoing sessions that are served by the network. Provided that the state of the process depends on the set of current sessions, it is represented by a vector with the variable number of elements (see process diagram in Fig. 15), which makes the number of states uncountable. For convenience, we let $(\omega_1, \ldots, \omega_n)$ denote the vector of abstract transmission characteristics for the set of pairs Rx-Tx of size n (e.g., ω_i is locations of Tx T_i and Rx Ri).

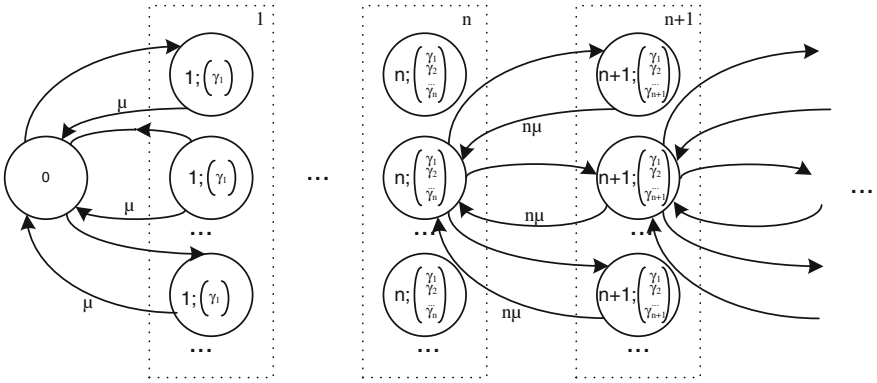

Fig. 15 General state diagram

Formally, our model is an extension of a simpler model in [44], where only the constant parameter γ has been considered. However, the set of possible states of the Markov process in our generalization is uncountable. Due to this fact, the solution constitutes a far more challenging task.

Let the number of already running sessions be equal n. We denote the probability of an event when the newly arrived session is rejected by the system as Q_{n+1}. Then, transitions from the state $(n; \omega_1, \ldots, \omega_n)$ to the state $(n+1; \omega_1, \ldots, \omega_n, \omega_{n+1})$ and backwards have the following rates:

$$\lambda(1 - Q_{n+1}) \text{ and } (n+1)\mu. \tag{12}$$

Steady-state distribution Due to uncountable number of states in the considered system, it is complicated to derive the steady-state distribution straightforwardly (however, not impossible). We note, that the corresponding Markov process *S(t)* may be simplified by employing the *state aggregation* technique. Hence, we aggregate states $\{(n; \omega_1, \ldots, \omega_n)\}_{\omega \in \Omega}$ by n (where Ω is the space of all possible vectors $(\omega_1, \ldots, \omega_n)$, $n \in N$). Here, we replace the original system by a system, where at every state, locations' coordinates are random and do not depend on locations at the previous state. Therefore, we obtain a new continuous Markov chain, where the current state is represented by the number of ongoing sessions and does not depend on the history of the process. Basing on that assumption, we may treat the considered process as a Birth-Death Process (BDP) and then formulate the following proposition.

Proposition 1 *The steady-state distribution* $\{\pi_i\}_{i=0}^{\infty}$ *for the considered process* $S(t)$ *with the transitions in* (12) *can be closely approximated by:*

$$\pi_n = \pi_0 \frac{\lambda^n}{\mu^n} \frac{\prod_{i=1}^{n} (1 - Q_n)}{n!}, \tag{13}$$

where

$$\pi_0 = \left(\sum_{i=0}^{\infty} \frac{\lambda^n}{\mu^n} \frac{\prod_{i=1}^{n} (1 - Q_n)}{n!} \right)^{-1},$$

and Q_{n+1} *is the reject probability on the transition from the state n to the state* $n+1$.

The average number of sessions in service may be calculated as:

$$E[N_{\text{sessions}}] = \sum_{n=0}^{\infty} n\pi_n, \tag{14}$$

where $\pi_n, n \geq 0$ *are the steady-state probabilities.*

Proof The above expressions follow from steady-state distribution of BDP, which can be easily found in any corresponding literature on elementary queuing theory. □

Here, we emphasize that the key assumption is that we disregard the history of the process from the perspective of the ongoing sessions, i.e., at each point we examine the arbitrary set of respective random variables.

We can easily obtain the sought steady-state distribution by using (13), if the reject probabilities Q_{n+1} are known. Therefore, further we concentrate on calculating the value of Q_{n+1}. Our result is summarized by Theorem 1, which exploits the distributions of random variables $\gamma_{i,j}$ and $d_{i,j}$. The latter can be derived after massive but straightforward transformations which are omitted here due to space limitations.

First, we consider D2D admission control as it has been described in Sect. 4.2.3. If n sessions already exist in the network, then, for all $i = \overline{1,n}$ we require the following target data rate condition to hold:

$$r \leq w \log\left(1 + \frac{p_{\max}\gamma_{i,i}}{KN_0}\right) \Leftrightarrow p_{\max}\gamma_{i,i} \geq KN_0\left(e^{\frac{r}{w}} - 1\right). \tag{15}$$

Then, the following theorem can be formulated.

Theorem 1 *If admission control in D2D network is performed according to (8) and, in particular, accounting for (15), then the reject probabilities* Q_{n+1} can be closely approximated by:

$$\begin{aligned} Q_{n+1} &= 1 - \Pr\{accepted | arrived\} = \\ &= 1 - \left[F_\gamma\left(\frac{N_0}{p_{\max}}\right)\right]^{2n}\left[1 - F_\gamma\left(\frac{\theta_0}{p_{\max}}\right)\right], \end{aligned} \tag{16}$$

where $\theta_0 = KN_0\left(e^{\frac{r}{w}} - 1\right)$ *and the cumulative distribution function (CDF) for SNR per power unit* γ *is given as:*

$$F_\gamma(\gamma) = 1 + \frac{G^4\gamma^{-\frac{4}{k}}}{8R^4} - \frac{G^2\gamma^{-\frac{2}{k}}}{R^2}\ln 2, \text{if } \frac{G^k}{(2R^2)^{\frac{k}{2}}} \le \gamma \le \gamma_{\max},$$

$$F_\gamma(\gamma) = 1 - \frac{1}{R^2}\left(\frac{G^4\gamma^{-\frac{4}{k}}}{8R^2} + G^2\gamma^{-\frac{2}{k}}\ln\frac{4R^2\gamma^{\frac{2}{k}}}{G^2}\right),$$

$$\text{if } \frac{G^k}{(2R)^k} \le \gamma \le \frac{G^k}{(2R^2)^{\frac{k}{2}}}; \gamma_{\max} = \frac{KN_0}{p_{\max}}\left(e^{\frac{r_{\max}}{w}} - 1\right).$$

Proof The proof is based on sequential calculation of distributions of random variables d (distance between Rx R_i and Tx T_j) and its function γ. The distribution of distance to the center of the cell may be easily obtained, since the locations follow uniform distribution within a circle. Then, we write down quite precisely the approximation for the random variable $z = d^2 = d_i^2 + d_j^2 - 2\cos(\alpha_i - \alpha_j)d_i d_j$, where $\alpha_{i/j}$ and $d_{i/j}$ are spherical coordinates of Rx/Tx. Using the estimate for the distribution of z and the transform (1), we may estimate the distribution of SNR per unit of power.

Further, knowing the necessary distributions, we take into account conditions (8) and find the acceptance probability at the state$n$:

$$\begin{aligned}\Pr\{\text{accepted} \mid \text{arrived}\} &= \Pr\{p_{\max}\gamma_{j,i} \le N_0, \forall i, j = \overline{1, n+1}, i \ / \\ &= j | p_{\max}\gamma_{j,i} \le N_0, \forall i, j = \overline{1, n}, i \ne j\}.\end{aligned}$$

$$\begin{aligned}\Pr\{p_{\max}\gamma_{i,i} \ge \theta_0, &= \overline{1, n+1} | p_{\max}\gamma_{i,i} \ge \theta_0, \forall i = \overline{1, n}\} \\ &= \frac{\left[\Pr\{\gamma_{j,i} \le \frac{N_0}{p_{\max}}\}\right]^{(n+1)n}}{\left[\Pr\{\gamma_{j,i} \le \frac{N_0}{p_{\max}}\}\right]^{n(n-1)}} \cdot \frac{\left[\Pr\{\gamma_{i,i} \ge \frac{\theta_0}{p_{\max}}\}\right]^{n+1}}{\left[\Pr\{\gamma_{i,i} \ge \frac{\theta_0}{p_{\max}}\}\right]^{n}}\end{aligned}$$

$$= \left[\Pr\{\gamma_{j,i} \le \frac{N_0}{p_{\max}}\}\right]^{2n} \cdot \left[\Pr\{\gamma_{i,i} \ge \frac{\theta_0}{p_{\max}}\}\right] = \left[F_\gamma\left(\frac{N_0}{p_{\max}}\right)\right]^{2n}\left[1 - F_\gamma\left(\frac{\theta_0}{p_{\max}}\right)\right],$$

which leads to the sought expression.

Power and energy consumption Using the results of Theorem 1, we can obtain the average power and energy consumption for a typical data session. The expected value of the user power consumption can be calculated as:

$$E[p_{total}] = \sum_{n=1}^{\infty} p^{(n)}\pi_n, \tag{17}$$

where $p^{(n)}$ is the average power consumption in the state n and π_n are the probabilities given by the steady-state distribution as obtained above. We note that the power consumption in the state $n = 0$ is 0 since we only focus on the ongoing sessions.

The power consumption of the system in the state n may be easily estimated as:

$$p^{(n)} = \left(\frac{p_{\max}}{\eta} + p_c\right) \sum_{i=1}^{n} \frac{r}{r_i^{\max}} + \sum_{i=1}^{n} \left(1 - \frac{r}{r_i^{\max}}\right) p_a. \tag{18}$$

Then, the average power consumption in the state n is given by:

$$E\left[p^{(n)}\right] = \left(\frac{p_{\max}}{\eta} + p_c - p_a\right) \sum_{i=1}^{n} E\left[\frac{r}{r_i^{\max}} | n \text{ sessions}\right] + np_a, \tag{19}$$

where $E\left[\frac{r}{r_i^{\max}} | n \text{ sessions}\right]$ is the expected value of a random variable $\frac{r}{r_i^{\max}}$ conditioning on the fact that n sessions are already accepted.

Further, we can establish the total energy consumption of a typical session in the D2D network by using (10):

$$E[\epsilon] = \frac{\sum_{n=1}^{\infty} p^{(n)} \pi_n}{\lambda P_a}, \tag{20}$$

where the D2D network accept probability P_a is determined by the law of total probability:

$$P_a = 1 - \sum_{n=0}^{\infty} \Pr\{\text{rejected}|\text{arrived}\} \pi_n = 1 - \sum_{n=0}^{\infty} Q_{n+1} \pi_n. \tag{21}$$

The probability of the D2D network network rejection is $\sum_{n=0}^{\infty} Q_{n+1} \pi_n = 1 - P_a$.

4.2.7 Analysis of Cellular Network

Below we concentrate on the steady-state distribution and related performance metrics of the cellular network. Generally, this analysis follows similar methodology as the respective D2D network analysis and we only highlight important differences below.

Stochastic model Recall that the flow of points on the cellular network is assumed to constitute a P_a-thinned Poisson process (see Fig. 14) and has the rate of $\lambda(1 - P_a)$. Similarly to the above, the system behavior can be described by a Markov process $S(t)$ at the moments t of session arrivals and departures. State transition rates are defined as they have been given by (12), but with the corresponding cellular network parameters. We note that for the cellular system, the characteristics of a data session are fully defined by the transmitting user location.

Steady-state distribution Aggregating the states of the corresponding Markov chain similarly to Sect. 4.2.6 and substituting the cellular arrival rate $\lambda(1-P_a)$, we can obtain the steady-state distribution $\{\pi_i\}_{i=0}^{\infty}$ by using (13).

Further, we base on the distributions of random variables γ_i and d_i, which may be derived by taking into account the uniform distribution of locations similarly to our calculations for the D2D network. In order to establish the steady-state distribution, we find the reject probabilities Q_{n+1} for a particular admission control discipline.

For that reason, we formulate Theorem 2 for the MR policy and Theorem 3 for the FU policy allowing us to establish exact solution for all reject probabilities Q_{n+1} in the former case and approximation for the latter. Due to the space limitations, we omit the full proofs of the theorems and only point out the main reasoning.

Theorem 2 *For the MR policy, the reject probabilities Q_1, Q_2,..., Q_{1+1} can be obtained as follows:*

a. *For n = 0, we directly use the distribution function of the random variable $q = \frac{r}{w\log(1+p_{\max}\gamma)}$ and establish:*

$$Q_1 = 1 - \frac{\left(Gp_{\max}^{\frac{1}{k}}\right)^2}{R^2}\left(e^{\frac{r}{wq_0}} - 1\right)^{-\frac{2}{k}}, \tag{22}$$

where $q_0 = \min(q_{\max}, 1)$.

b. *For the reject probability when a session arrives in the state 1, we have:*

$$Q_2 = 1 - \frac{\left(Gp_{\max}^{\frac{1}{k}}\right)^2}{R^2}\int_0^{q_0} \frac{e^{\frac{r}{wq_1}}\left(e^{\frac{r}{w\min(q_{\max},1-q_1)}} - 1\right)^{\frac{2}{k}}}{q_1^2\left(e^{\frac{r}{wq_1}} - 1\right)^{\frac{k+2}{k}}}dq_1. \tag{23}$$

c. *The probabilities $Q_{n+1}, n > 2$ are closely approximated by:*

$$Q_{n+1} = 1 - \frac{\left(Gp_{\max}^{\frac{1}{k}}\right)^2}{\phi_0}\int_0^{q_0'}\left(\int_0^{b'} \frac{e^{-\frac{(z_{n-1}-m_{n-1})^2}{2\sigma_{n-1}}}e^{\frac{r}{wq_n}}}{{}_n^2\left(e^{\frac{r}{wq_n}} - 1\right)^{\frac{k+2}{k}}}dq_n\right)dz_{n-1}, \tag{24}$$

where $\phi_0 = R^2\sqrt{2\pi(n-1)}\sigma_{n-1}\left[\Phi\left(\frac{(n-1)q_{\max}-m_{n-1}}{\sigma_{n-1}}\right) - \Phi\left(\frac{-m_{n-1}}{\sigma_{n-1}}\right)\right]$, $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{x^2}{2}}dx$, *and other parameters (m_{n-1}, σ_{n-1} $q_{\max}$ q_0' b') are given below.*

Proof We may find the reject probabilities as $\Pr\left\{\sum_{i=1}^{n+1}\frac{r}{c_i^{\max}} > 1 \middle| \sum_{i=1}^{n}\frac{r}{c_i^{\max}} \le 1\right\}$ by applying sequential transforms. For $n = 0$ we use directly the distribution function for the random variable $q = \frac{r}{c_i^{\max}}$ and

$$\Pr\left\{\frac{r}{c_i^{\max}} > 1\right\} = 1 - F_q(1), \tag{25}$$

where $F(q)$ may be easily found as a distribution of function $q(d)$ and d is the distance to the BS, $f_d(d) = \frac{2d}{R^2}$. The same has been done in the case of $n = 1$ when the sum $z_2 = \sum_{i=1}^{2} \frac{r}{c_i^{\max}}$ is considered. For $n > 1$ we propose using the normal distribution fitting.

In particular, for the state n we need to consider the distribution of a random variable $z_n = \sum_{i=1}^{n} \frac{r}{r_i^{\max}}$. Due to the complexity of straightforward derivation, we do not calculate the convolutions of random variables $q_i = \frac{r}{r_i^{\max}}$ and approximate $z_n, n > 2$ as follows. We split the sum z_n into two components $z_n = z_{n-1} + q_n$, where q_n is the random variable corresponding to a new session. Then, we approximate z_{n-1} by the random variable distributed according to the truncated normal distribution over $[0, \min(q_{\max}, 1)]$ with mean m_{n-1} and variance σ_{n-1}. Here, the variable $q_{\max} = \frac{r}{w \log(1 + p_{\max}(G/R)^k)}$ is the maximum value of the random variable q_i on the edge of the cell.

We continue with the FU policy and recall that the maximum data rate $r_i^{\max}, \forall i = \overline{1, n}$ is defined as:

$$r_i^{\max} = \min[w \log(1 + \gamma_i p_{\max}), r_{\max}] \geq r_i,$$

where $r_{\max}$ is the maximum feasible data rate restricted by a particular wireless technology in (11).

Theorem 3 *For the FU policy, the reject probabilities Q_{n+1} can be calculated from the distribution of the random variable $r_i^{\max} = w \log(1 + \gamma_i p_{\max})$ as follows:*

$$Q_{n+1} = 1 - \frac{G^2 p_{\max}^{\frac{2}{k}}}{R^2} \cdot \left[\frac{e^{\frac{m}{w}} - 1}{\left(e^{\frac{rn+n}{w}} - 1\right)^2}\right]^{\frac{2}{k}}, \tag{26}$$

where the distribution of the random variable $r_i^{\max}$ follows from the distribution of γ_i.

Proof We calculate the transitions in a similar way as before with only difference that we do not use any approximations here. After simple transforms we may obtain the following:

$$\begin{aligned}(1 - Q_{n+1}) &= \left(\Pr\left\{\frac{r}{r_i^{\max}} \leq \frac{1}{n+1}, = 1, (n+1) \middle| \frac{r}{r_i^{\max}} \leq \frac{1}{n}, = 1, n\right\}\right) \\ &= \Pr\{r_{n+1}^{\max} \geq r(n+1)\} \cdot \left(1 - \frac{\Pr\{rn \leq r_i^{\max} < r(n+1)\}}{\Pr\{r_i^{\max} \geq rn\}}\right)^n.\end{aligned}$$

Thus, using the transformation for $c^{\max}$ and the approach exploited above we can obtain the PDF for the maximum instantaneous rate. Here, for the simplification we denote $r_i^{\max}$ as c and further study the random variable c. Therefore, we may obtain:

$$Q_{n+1} = 1 - (1 - F_c(r(n+1)) \cdot \left(\frac{1 - F_c(r(n+1))}{1 - F_c(rn)}\right)^n.$$

Here, we take into account the condition $r_i \leq r_i^{\max}$ by considering the limitation on the number of ongoing sessions, i.e., $n \leq \frac{1}{q_{\max}}$, where $q_{\max}$ has been given above. We also highlight that for the FU policy we do not offer an approximation, as the obtained solution is exact for the considered model.

Power and energy consumption As in Sect. 4.2.6, the expected value of the user power consumption is given by (17). For a particular set of n ongoing sessions, the power consumption can be calculated as:

$$p^{(n)} = \sum_{i=1}^{n} \frac{r}{r_i}\left(\frac{p_i}{\eta} + p_c\right) + \sum_{i=1}^{n}\left(1 - \frac{r}{r_i}\right)p_a,$$

where p_i is the transmit power of the user i.

The average power consumption $p^{(n)}$ in the state n for the MR policy is given as:

$$E[p^{(n)}] = \left(\frac{p_{\max}}{\eta} + p_c - p_a\right)E\left[\sum_{i=1}^{n}\frac{r}{r_i^{\max}} | \sum_{i=1}^{n}\frac{r}{r_i^{\max}} \leq 1\right] + np_a, \tag{27}$$

where the component $E\left[\sum_{i=1}^{n}\frac{r}{r_i^{\max}} | \sum_{i=1}^{n}\frac{r}{r_i^{\max}} \leq 1\right]$ is given in the Sect. 4.2.8.

The average consumed power in the state n for the FU policy is the following:

$$\begin{aligned} E[p^{(n)}] &= \frac{1}{\eta}E\left[p_i | \frac{r}{r_i} \leq \frac{1}{n}\right] + p_c + (n-1)p_a \\ &= \frac{1}{\eta}\left(e^{\frac{m}{w}} - 1\right) \cdot E\left[\frac{1}{\gamma_i} | \frac{1}{\gamma_i} \leq \max\left(\frac{R^k}{G^k}, \frac{p^{\max}}{e^{\frac{m}{w}} - 1}\right)\right] + p_c + (n-1)p_a, \end{aligned} \tag{28}$$

where the expression for $E[.]$ is given below.

The total energy consumption of a typical session in the cellular network is given by (20) using the corresponding value of session blocking probability as:

$$P_b = \sum_{n=0}^{\infty} Q_{n+1}\pi_n. \tag{29}$$

4.2.8 Auxiliary Calculations for all Systems

In what follows, we provide necessary explanations and details on auxiliary variables introduced in Theorems 1–3 for the steady-state distribution as well as expressions for the energy/power consumption calculation. We note that for all three different systems (cellular network under MR, FU policies and D2D network) we use the following approach to estimate the reject probabilities, as well as power consumption in certain states.

Knowing the distribution of user locations within a cell, we can obtain the distribution of distances between the transmitting and the receiving user (or the BS). Then, following the different in all cases limitations of admission control, we find the conditional expectations of corresponding random variables if there are currently n sessions in service. For the sake of brevity, we only summarize the final expressions for the required variables below.

D2D network: For the further calculations, we obtain the CDF of $z = d_{i,j}^2$ as follows (the corresponding probability density function can be trivially found via differentiation):

$$F_z(z) = -\frac{z^2}{8R^4} + \frac{z}{R^2}\ln 2, \left(\frac{p_{\max}G^k}{K\theta\left(e^{\frac{r_{\max}}{w}}-1\right)}\right)^{2/k} \quad 0 \le z \le 2R^2, F_z(z)$$
$$= \frac{1}{R^2}\left(\frac{z^2}{8R^2} + z\ln\frac{4R^2}{z}\right), 2R^2 \le z \le 4R^2.$$

We also introduce the following additional notation:

$$y_{\min} = \frac{r}{w\log(1+\mathrm{SINR}_{\max})}, z_1 = \frac{N_0}{p_{\max}} \text{ and } z_0 = \frac{\theta_0}{p_{\max}}.$$

Then, using the distribution of random variable d_i^2 we calculate the average power consumption for D2D network as:

$$E\left[\frac{r}{r_i}|n \text{ sessions}\right] = [F_z(z_1)]^{1-n}\left[1 - F_z(z_0)\int\limits_{y_{\min}}^{1} y\psi(y)f_z(z(y))dy\right]^{-1}.$$

Here, the expression for function $z(y)$ is given as:

$$z = \left[\frac{p_{\max}G^k}{K\theta\left(e^{\frac{r}{wy}}-1\right)}\right]^{2/k} = \left[\frac{p_{\max}G^k}{K\theta}\right]^{2/k}\left(e^{\frac{r}{wy}}-1\right)^{-2/k},$$

and introduced for simplicity auxiliary function $\psi(y)$ is:

$$\psi(y) = \left[\frac{p_{\max}G^k}{K\theta}\right]^{2/k}\frac{2}{k}\frac{r}{w}\left(e^{\frac{r}{wy}}-1\right)^{-\frac{2}{k}-1}e^{\frac{r}{wy}}\frac{1}{y^2}.$$

Cellular network: MR policy Here, we provide description of all parameters and auxiliary expressions that we need for calculations in Theorem 3 and user power consumption. For simplicity, we denote the upper bound for possible values of q_i defined by the distance R as $q_{\max}$ and lower bound defined by the maximum level of SNR as $q_{\min}$, letting $q_0 = \min(q_{\max}, 1)$:

$$q_{\max} = \frac{r}{w \log\left(1 + \frac{G^k}{R^k} p_{\max}\right)}, q_{\min} = \frac{r}{w \log(1 + \mathrm{SNR}_{\max})}.$$

The CDF for the random variable q and the conditional probability density function are defined as:

$$F_q(q) = \frac{\left(Gp^{\frac{1}{k}}\right)^2}{R^2\left(e^{\frac{r}{wq}} - 1\right)^{\frac{2}{k}}}, \; q \le q_{\max}, F_q(q|q \le 1) = \frac{\left(Gp^{\frac{1}{k}}\right)^2}{R^2\left(e^{\frac{r}{wq}} - 1\right)^{\frac{2}{k}}} \cdot \frac{1}{F_q(1)}$$
$$= \frac{\left(e^{\frac{r}{w}} - 1\right)^{\frac{2}{k}}}{\left(e^{\frac{r}{wq}} - 1\right)^{\frac{2}{k}}}, \; q \le 1,$$

where the probability density function $f_q(q|q \le 1) = dF_q(q|q \le 1)/dq$.

For derivation of all transition probabilities as well as energy consumption in a certain state, we obtain the conditional expectation $E[z_n|z_n \le 1]$, $n \ge 1$ calculated in what follows. For $n = 1, 2$ we calculate straightforwardly:

$$E[q|q \le 1] = \frac{1}{1 - Q_1}\left[C_0 q_{\min} + q_0 \frac{G^k}{R^k} p_{\max}\left(e^{\frac{r}{wq_0}} - 1\right)^{-\frac{2}{k}}\right]$$
$$- \frac{G^k p_{\max}}{R^k(1 - Q_1)}\left[q_{\min}\left(e^{\frac{r}{wq_{\min}}} - 1\right)^{-\frac{2}{k}} - \int_{q_{\min}}^{q_0}\left(e^{\frac{r}{wq}} - 1\right)^{-\frac{2}{k}} dq\right],$$

where $C_0 = \Pr\{q \ge q_{\min}\} = 1 - \frac{G^2 p^{\frac{2}{k}}}{R^2}\left(e^{\frac{r}{wq_{\min}}} - 1\right)^{-\frac{2}{k}}$.

Knowing the distribution for the random variable q, we find the first moment:

$$E[q|q \le 1] = q_{\min} F_{q|q \le 1}(q_{\min}) + \int_{q_{\min}}^{1} q f_{q|q \le 1}(q) dq$$
$$= q_{\min} \frac{\left(e^{\frac{r}{w}} - 1\right)^{\frac{2}{k}}}{\left(e^{\frac{r}{wq}} - 1\right)^{\frac{2}{k}}} + \frac{2r\left(e^{\frac{r}{w}} - 1\right)^{\frac{2}{k}}}{kw} \int_{q_{\min}}^{1} \frac{e^{\frac{r}{wq}}}{q\left(e^{\frac{r}{wq}} - 1\right)^{\frac{k+2}{k}}} dq.$$

The second moment of q may be obtained as:

$$E[q^2|q \le 1] = q_{\min}^2 F_{q|q\le 1}(q_{\min}) + \int_{q_{\min}}^{1} q^2 f_{q|q\le 1}(q)dq$$
$$= q_{\min}^2 \frac{\left(e^{\frac{r}{w}}-1\right)^{\frac{2}{k}}}{\left(e^{\frac{r}{wq}}-1\right)^{\frac{2}{k}}} + \frac{2r\left(e^{\frac{r}{w}}-1\right)^{\frac{2}{k}}}{kw} \int_{q_{\min}}^{1} \frac{e^{\frac{r}{wq}}}{\left(e^{\frac{r}{wq}}-1\right)^{\frac{k+2}{k}}} dq.$$

$$E[q^2] = \frac{1}{C_0}\left[q_{\max}^2 - q_{\min}^2 \frac{G^k}{R^k} p_{\max}\left(e^{\frac{r}{wq_{\min}}}-1\right)^{-\frac{2}{k}}\right]$$
$$-\frac{1}{C_0}\left[2\frac{G^k}{R^k} p_{\max} \int_{q_{\min}}^{q_{\max}} q\left(e^{\frac{r}{wq}}-1\right)^{-\frac{2}{k}} dq\right],$$

where $q_{\min} = r/w/\log(1+\mathrm{SNR}_{\min})$ and

The corresponding variance for the random variable q can be obtained as:

$$\sigma_q^2 = E[q^2|q \le 1] - (E[q|q \le 1])^2.$$

The expressions for the conditional expected value $E[z_2|q \le 1]$ and conditional blocking probability $Q_2 = Pr\{z_2 > 1|z_1 = q \le 1\}$ can be summarized as follows:

$$E[q_1+q_2|q_1+q_2 \le 1] = \frac{\phi(q_{\min})}{a} \int_0^{q_{\min}} \frac{e^{\frac{r}{wq_1}}}{q_1^2\left(e^{\frac{r}{wq_1}}-1\right)^{\frac{k+2}{k}}} dq_1$$
$$+\frac{1}{a}\int_{q_{\min}}^{q_0} \phi(q_1) \frac{e^{\frac{r}{wq_1}}}{q_1^2\left(e^{\frac{r}{wq_1}}-1\right)^{\frac{k+2}{k}}} dq_1,$$

where $q_0 = \min(q_{\max}, 1)$ and $a = \Pr\{q_1+q_2 \le 1\}$, and:

$$\phi(q) = (q+q_{\min})(1-C_0) + \int_{q_{\min}}^{b} (q+q_2) \frac{e^{\frac{r}{wq_2}}}{q_2^2\left(e^{\frac{r}{wq_2}}-1\right)^{\frac{k+2}{k}}} dq_2, a$$
$$= \int_0^{q_0} \left(\int_0^{b} \frac{e^{\frac{r}{wq_2}}}{q_2^2\left(e^{\frac{r}{wq_2}}-1\right)^{\frac{k+2}{k}}} dq_2 \right) \frac{e^{\frac{r}{wq_1}}}{q_1^2\left(e^{\frac{r}{wq_1}}-1\right)^{\frac{k+2}{k}}} dq_1,$$

and $b = min(q_{\max}, 1-q_1, 1-q_{\min})$.

For the proposed distribution approximation in case $n > 2$:

$$E[z_{n-1}+q_n|z_{n-1}+q_n\le 1]=\frac{\phi_{n-1}(q_{\min})}{a'}\int_0^{(n-1)q_{\min}} e^{-\frac{(z_{n-1}-m_{n-1})^2}{2\sigma_{n-1}}}dz_{n-1}$$
$$+\frac{1}{a'}\int_{(n-1)q_{\min}}^{q_0'}\phi_{n-1}(z_{n-1})e^{-\frac{(z_{n-1}-m_{n-1})^2}{2\sigma_{n-1}}}dz_{n-1},$$

where, $a^{'}=\Pr\{z_{n-1}+q_n\le 1\}$, $q_0^{'}=\min((n-1)q_{\max},1)$, and:

$$\phi_{n-1}(q)=(q+q_{\min})(1-C_0)+\int_{q_{\min}}^{b}(q+q_n)\frac{e^{\frac{r}{wq_n}}}{q_n^2\left(e^{\frac{r}{wq_n}}-1\right)^{\frac{k+2}{k}}}dq_n,a$$
$$=\int_0^{q_0'}\left(\int_0^{b'}\frac{e^{\frac{r}{wq_n}}}{q_n^2\left(e^{\frac{r}{wq_n}}-1\right)^{\frac{k+2}{k}}}dq_n\right)e^{-\frac{(z_{n-1}-m_{n-1})^2}{2\sigma_{n-1}}}dz_{n-1},$$

and $b'=min(q_{\max},1-z_{n-1},1-(n-1)q_{\min})$. Using the same logic, we obtain the second moments $E[z_2^2|z_2\le 1]$, $E[z_n^2|z_n\le 1]$ and variances σ_2^2, σ_n^2. Therefore, parameters of distribution m_n and σ_n^2 can be found from the integral expressions using the calculations from one step before.

Cellular network: FU policy For the FU policy we need to calculate the distribution of the random value $c=c_i^{\max}$:

$$f_c(c)=\frac{2\left(Gp_{\max}^{\frac{1}{k}}\right)^2e^{\frac{c}{w}}}{kwR^2(e^{\frac{c}{w}}-1)^{\frac{k+2}{k}}},c\ge c_{\min},F_c(c)=1-\frac{G^2p_{\max}^{\frac{2}{k}}}{R^2(e^{\frac{c}{w}}-1)^{\frac{2}{k}}},c\ge c_{\min},$$

where $c_{\min}$ is the lower border for possible values of c defined by the distance R:

$$c_{\min}=w\log\left(1+\frac{G^k}{R^k}p_{\max}\right).$$

For the calculation of average power consumption, we derive the following expression:

$$E\left[\frac{1}{\gamma}|\frac{1}{\gamma}\le\frac{p_{\max}}{(e^{\frac{nr}{w}}-1)}\right]=\frac{k}{k+2}\frac{p_{\max}}{e^{\frac{nr}{w}}-1}\left(\frac{e^{\frac{nr}{w}}-1}{e^{r_{\max}/w}-1}\right)^{\frac{2}{k}+1}+\frac{2}{k+2}\frac{p_{\max}}{e^{\frac{nr}{w}}-1}.$$

Where $r_{\min}$ is the lower border for possible values of r_i defined by the distance R:

$$r_{\min}=w\log\left(1+\frac{G^k}{R^k}p_{\max}\right).$$

We note that due to the space limitations we omit all proofs and distribution derivations. Beyond that, this paragraph contains all auxiliary variables required for system performance metrics estimation.

4.2.9 Simulation Backup for the Analysis

We remind that in this study we have developed an advanced SLS based on the 3GPP LTE evaluation methodology and current IEEE 802.11 specifications. This simulator is a flexible tool designed to support dynamic deployment strategies, user radio interface models, channel characteristics, and wireless protocols [51, 52]. To further optimize its performance, here we make several simplifications of realistic wireless systems, yet we attempt to mimic the most important mechanisms and dependencies explicitly.

As suggested by our evaluation methodology, we use 3GPP LTE and IEEE 802.11 for infrastructure and D2D transmissions, respectively. For the LTE system, the simulation captures the following practical features (as opposed to the above analytical model): data frame structure, bandwidth requests, and scheduling by the BS. For the D2D system, the simulation is largely based on IEEE 802.11 medium access control procedure with carrier sensing. However, to match the capabilities of the analytical model (see Sect. 4.2.3), the following modifications were applied to the real system.

1. The medium access procedure assumes that the channel quality between all users is known in advance. This assumption is feasible given the network-assisted operation where the BS can act as database for such information.
2. For simplicity, we assume that a user transmission reserves the channel for its entire duration, unlike in the real protocol where a reservation is made only when there is data to be sent. Such reservation protocol may serve as a pessimistic performance estimate, but it also guarantees that whenever a D2D connection is established it can reliably serve the target bitrate within its capacity limit.
3. When the connection is established, the entire transmission is fragmented into fixed-size packets (we use 1,000 bytes) and those fragments are sent with regular intervals, which are adjusted to match the required bitrate for as long as the session is active.

4.2.10 Evaluation Scenario for Analytical Study

Here we summarize our test scenario that mimics LTE-assisted offloading of user sessions from cellular onto WFD. This scenario concentrates on an *area of interest* [53], in which co-located cellular and D2D networks cover a limited region with many users requiring service (e.g., shopping mall, business center, etc.). In particular, we consider an isolated circle cell of radius $R = 100$ m and disregard

Table 4 Simulation parameters

Notation	Cellular network	D2D network
R, m	100	100
r, Mbps	4.8	4.8
μ^{-1}, s	3	3
k	5	6.5
G	197.43	2.4
w, MHz	10	20
η	0.5	0.5
N_0, dB	−60	−70
$r_{\max}$, Mbps	60	56
$p_{\max}$, mW	0.20	0.20
$p_{c,}$ mW	1.53	0.15
p_{a}, mW	0.05	0.005

interference coming from the neighboring cells as assumed in Sect. 4.2.2. In this area, the users need to exchange small multimedia fragments with the required bitrate of $r = 4.8$ Mbps. As session duration is distributed exponentially with mean of $s^{-1} = 3$ s, an average transmission carries about 2 mb of information.

As assumed in Sect. 4.2.1, the session inter-arrival times are exponential with λ new sessions arriving every second and requesting service. All sessions have specific destinations within the considered area of interest. However, a particular transmitting user may either be successfully accepted by the D2D network, or rejected and need to attempt the LTE BS instead. If cellular resource is insufficient to admit this user, it is blocked permanently. For clarity, below we only consider the MR transmission policy on LTE when all users transmit at their maximum power levels. The other power-related parameters are specified in [54] for WiFi and in [55] for LTE, whereas the rest of the system settings are summarized in Table 4.

Results and discussion Following the description of the system operation in Sect. 4.2.4, we model the integrated system as shown in Fig. 14 and compare its performance against the cellular baseline without any D2D support.

One of the primary metrics of interest in our system is its capacity, as in how many sessions can be served at the same time (14). Figure 16 contrasts the LTE baseline against the D2D-enhanced network to confirm the considerable benefits (about 20 % improvement) provided by D2D connections. Hereinafter, continuous lines indicate simulation data (S), whereas symbols correspond to analytical values (A). Clearly, the overall trend is the increase in the expected number of running links, up to the saturation point which depends on the deployment, scheduling, and multiplexing methods used.

In close connection with the capacity goes the blocking probability (see Fig. 17), or the proportion of service requests that cannot be served by the network. We demonstrate how system blocking probability P_{block} (9), D2D reject probability (21), and blocking probability by the LTE baseline (29) evolve with increasing load on the network.

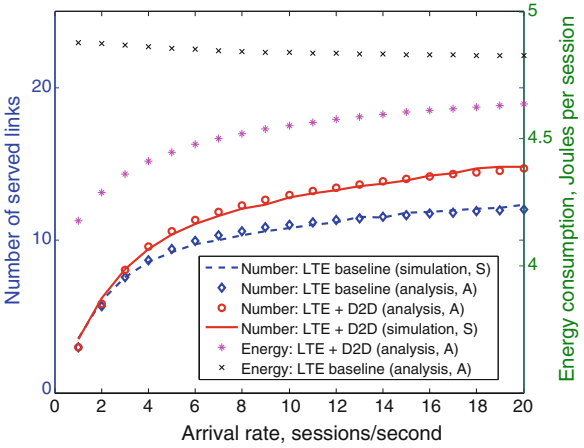

Fig. 16 System capacity (*in number of sessions*) and user energy consumption

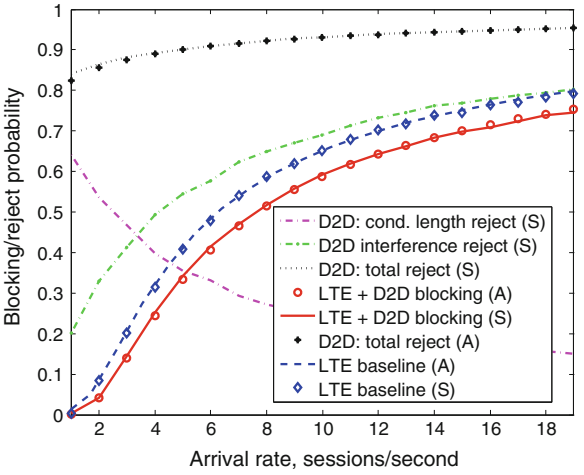

Fig. 17 Session blocking/reject probabilities

We also remind that a cellular session is blocked if it cannot fit into the schedule at the time of arrival, whereas for the D2D network we differentiate between session rejections due to (i) prohibitive interference from the existing transmissions and (ii) excessive link length to support the required bitrate (given that the interference constraint has been satisfied). It is important to analyze the structure of the blocking processes for both systems. For the D2D system, at low loads the blocking is primarily caused by the link length, whereas as the load increases the probability of a blocking due to interference becomes dominant.

Contrary to the intuition, in the LTE system the blocking is not a hard-threshold like one would expect of a scheduled system. In fact, the cellular system never reaches the 100 % blocking in the given scenario. This is explained by the fact that instead of discarding all the links it simply accepts those still fitting into the

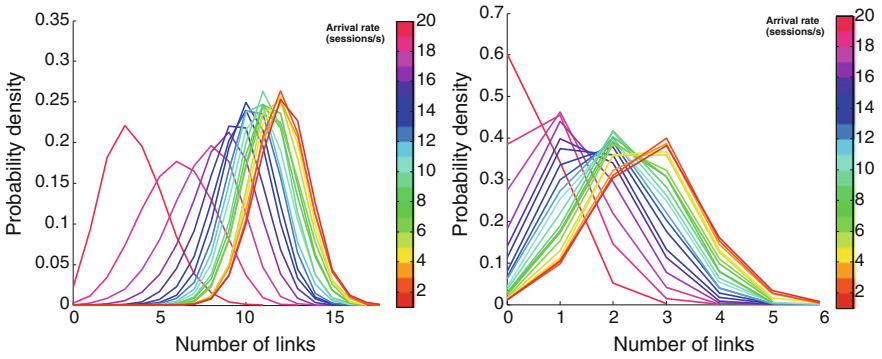

Fig. 18 Distribution of the number of links, LTE (*left*) and D2D (*right*)

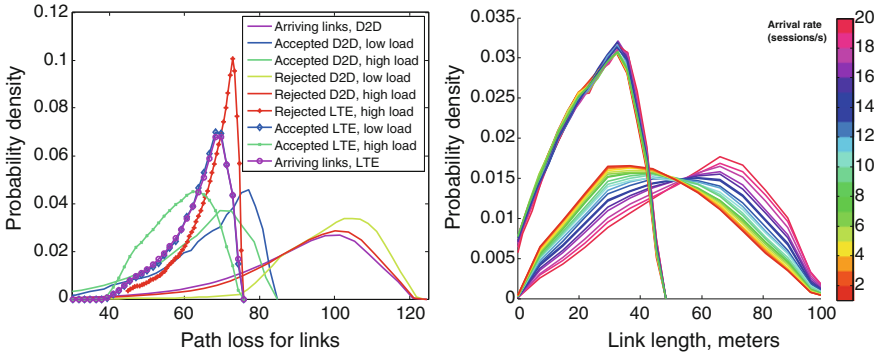

Fig. 19 Link quality (*left*) and length distribution (*right*)

schedule, thus giving priority to higher-rate links under high loads (see Fig. 18 for details).

In order to detail the effects we have just noted, let us take a look at the quality of the links in our system. When the cellular system is empty, it can afford accepting all links, no matter the quality. Under such conditions, the link quality for arrivals and accepted links is similar, and there are almost no discards (see Fig. 19, left). When the cellular system gets loaded, however, we see that it takes only shorter links in—as those have significantly better chances to fit into the schedule (refer to Fig. 19, right).

An empty D2D system cannot afford such luxury—the links are overall much worse, and it has to be very selective to ensure connectivity. One can see that irrespective of the arrival rate, the D2D system consistently remains highly selective to the links based on their length, with almost identical distributions for both empty and overloaded conditions. The reason for this is that the survival of a D2D link is primarily determined by its interference at higher loads. Indeed,

shorter links have somewhat better chances of not getting blocked, but combined with other effects it does not reflect in the final statistics.

Finally, one can observe user energy consumption (20) in the LTE and the integrated LTE-D2D system also in Fig. 16. It can be clearly seen that at low arrival rates the D2D connections have very high impact on the energy efficiency of the system, improving it by up to 14 %. However, as the system gets loaded, the D2D can no longer take over any significant portion of the links, and the energy savings become less significant.

Of course, the energy consumption reduction effects are largely dependent on the specific parameters of the transmitter. Therefore, our analytical approach may be extremely useful when it comes to evaluation of the energy consumption, as the solutions for arbitrary power models can be obtained quickly and over a large range of arrival rates. We generally conclude that network-assisted offloading of LTE data onto WFD D2D connections may significantly improve session blocking probabilities, as well as boost energy efficiency of wireless transmitters.

4.3 Evaluation Summary

Based on the presented evaluation, one can see that the assisted D2D is far from impossible to analyze, and can be studied through simulation, given the appropriate tools. On the other hand, the analysis presented here has severe limitations on the cellular network side, being limited to a single cell. It also makes strong assumptions on the interference in the network and abstracts away significant portion of the D2D protocol. Simulation data allows us to complement it by deepening our understanding of the performance of network-assisted D2D.

5 Conclusions

Probably, the most important take-away for this chapter is the fact that it is indeed possible to construct a working D2D system utilizing unlicensed bands radio. Moreover, it is not just possible, it is in fact rather easy to do, even though the existing platforms are almost deliberately designed against that.

It is also worth noting, that the proposed architecture is almost unavoidable in a proximity-based service, even if it does not explicitly claim network assistance as one of its features. Indeed, providing integration with the social networks and web ecosystem is a key requirement for D2D applications, and it is difficult to arrange without persistent network access.

5.1 Future of D2D in Unlicensed Bands

We believe that in the future D2D communications will become part of our daily life just like QR codes are now becoming part of the posters and advertisements. It is convenient and natural for people to rely on proximity when communicating, and therefore D2D communications will eventually become popular for general public, and not just among IT geeks. Unlicensed bands will likely play a significant role in this process, allowing free medium to be utilized when the devices are close, and thus adding monetary value to the proximity-based services for the users and operators alike.

It is hard to predict if later D2D will migrate to licensed bands. From the technical point of view, it is quite difficult to manage the D2D spectrum efficiently, as transmitter's locations are not known exactly, and network load keeps changing. Therefore, it is questionable if licensed band would bring any significant boost to the capacity, but it is certain that it will increase the costs.

5.2 Interesting Research Directions

As far as research is concerned, D2D is still a largely understudied area. For example, should the network be given the ability to control which D2D links are established, it could avoid offloading onto D2D links that degrade network and/or user performance. Similarly, if the network can control when certain D2D links transmit, it could potentially establish scheduling zones when groups of non-competing D2D links are allowed to communicate, thus potentially significantly reducing contention and improving throughput and energy efficiency of D2D links (here the reader is referred to works [10, 56, 57]). Of course, advanced power control options also become available when network assists D2D communications [58].

User interaction models are a completely different side of the future D2D research. As technologies evolve, new opportunities appear for the people to integrate them into their daily life, and they affect each other heavily. For the successful integration of D2D into the existing social models, a lot of work would have to be done to make the solutions reliable, user-friendly, and safe to use.

In this work, we have only offered a summary of the first steps into the attractive space of unlicensed bands, network-assisted D2D. One could easily proceed with looking at operation of D2D communications when network assistance becomes unavailable or harmful for some reason, or how to represent the connections to the applications in a better way. Opportunities here are endless, but it is critical to move quickly, as within 1 or 2 years we may see a significant shift in scale and overall approach to D2D in general, with operators looking at it as a "must-have" technology rather than a quirky prototype.

References

1. G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, Z. Turanyi, Design aspects of network assisted device-to-device communications. IEEE Commun. Mag. **50**(3), 170–177 (2012)
2. J. Andrews, *Can cellular Networks Handle 1000x the Data?* (Technical Report, UT Austin, 2011)
3. S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, Y. Koucheryavy, Cellular Traffic Offloading onto network-assisted device-to-device connections. IEEE Commun. Mag. (2014)
4. L. Lei, Z. Zhong, C. Lin, X. Shen, Operator controlled device-to-device communications in LTE-Advanced networks. IEEE Wirel. Commun. **19**(3), 96–104 (2012)
5. T. Doumi, M. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, D. Flore, LTE for Public Safety Networks. IEEE Commun. Mag. **51**(2), 106–112 (2013)
6. G. Wu, S. Talwar, K. Johnsson, N. Himayat, K. Johnson, M2M: from mobile to embedded internet. IEEE Commun. Mag. **49**(4), 36–43 (2011)
7. S.-P. Yeh, S. Talwar, G. Wu, N. Himayat, K. Johnsson, Capacity and coverage enhancement in heterogeneous networks. IEEE Wirel. Commun. **18**(3), 32–38 (2011)
8. M. Corson, R. Laroia, J. Li, V. Park, T. Richardson, G. Tsirtsis, Toward proximity-aware internetworking. IEEE Wirel. Commun. **17**(6), 26–33 (2010)
9. K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, K, Hugl. Device-to-device communication as an underlay to LTE-advanced networks. IEEE Commun. Mag. **47**(12), 42–49 (2009)
10. M. Belleschi, G. Fodor, A. Abrardo, M. Belleschi, G. Fodor, A. Abrardo, Performance analysis of a distributed resource allocation scheme for D2D communications. In *Proceedings of GLOBECOM Workshops* (2011), 358–362
11. P. Janis, C.-H. Yu, K. Doppler, C. Ribeiro, C. Wijting, K. Hugl, O. Tirkkonen, V. Koivunen, Device-to-device communication underlaying cellular communications systems. Int. J. Commun. Netw. Syst. Sci. **2**(3), 169–247 (2009)
12. C.-H. Yu, K. Doppler, C. Ribeiro, O. Tirkkonen, Resource sharing optimization for device-to-device communication underlaying cellular networks. IEEE Trans. Wirel. Commun. **10**(8), 2752–2763 (2011)
13. B. Kaufman, J. Lilleberg, B. Aazhang, Spectrum sharing scheme between cellular users and ad-hoc device-to-device users. IEEE Trans. Wireless Commun. **12**(3), 1038–1049 (2013)
14. C.-H. Yu, O. Tirkkonen, K. Doppler, C. Ribeiro, Power optimization of device-to-device communication underlaying cellular communication. IEEE Int. Conf. Commun. **4**, 1–5 (2009)
15. K. Doppler, C. Ribeiro, J. Kneckt, Advances in {D2D} communications: energy efficient service and device discovery radio. In *Proceedings of the International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE)* (2011)
16. J. Seppala, T. Koskela, T. Chen, S. Hakola, Network controlled device-to-device ({D2D}) and cluster multicast concept for {LTE} and {LTE-A} networks. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)* (2011), 986–991
17. H. Min, W. Seo, J. Lee, S. Park, D. Hong, Reliability improvement using receive mode selection in the device-to-device uplink period underlaying cellular networks. IEEE Trans. Wirel. Commun. **10**(2), 413–418 (2011)
18. H. Wang, X. Chu, Distance-constrained resource-sharing criteria for device-to-device communications underlaying cellular networks. Electron. Lett. **48**(9), 528–530 (2012)
19. T. Chen, G. Charbit, S. Hakola, Time hopping for device-to-device communication in {LTE} cellular system. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)* 2010
20. Y. Xu, R. Yin, T. Han, G. Yu, Dynamic resource allocation for device-to-device communication underlaying cellular networks. Int. J. Commun. Syst. **38**, (2012)

21. M. Jung, K. Hwang, S. Choi, Joint mode selection and power allocation scheme for power-efficient device-to-device (D2D) communication. In *Proceedings of the IEEE Vehicular Technology Conference (VTC-Spring)* (2012)
22. X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, A. Jovicic, FlashLinQ: a synchronous distributed scheduler for peer-to-peer ad Hoc networks. In *Forty-Eighth Annual Allerton Conference* (2010)
23. A. Vigato, L. Vangelista, C. Measson, X. Wu, Joint discovery in synchronous wireless networks. IEEE Trans. Commun. **59**(8), 2296–2305 (2011)
24. P. Janis, V. Koivunen, C. Ribeiro, K. Doppler, K. Hugl, *Interference-Avoiding MIMO Schemes for Device-to-Device Radio Underlaying Cellular Networks* (In Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2009), pp. 2385–2389
25. A. Mukherjee, A. Hottinen, Energy-efficient device-to-device {MIMO} underlay network with interference constraints. In *Proceedings of the International ITG Workshop on Smart Antennas (WSA)* (2012), 105–109
26. A. Osseiran, K. Doppler, C. Ribeiro, M. Xiao, M. Skoglund, J. Manssour, Advances in device-to-device communications and network coding for IMT-advanced. In *ICT-Mobile Summit 2009 Conference Proceedings* (2009)
27. C.-H. Yu, O. Tirkkonen, Device-to-device underlay cellular network based on rate splitting. In *Proceedings of the IEEE Wireless Communications and Networking Conference: PHY and Fundamentals* (2012), 262–266
28. N. Golrezaei, A. Molisch, A. Dimakis, G. Caire, Femtocaching and device-to-device collaboration: a new architecture for wireless video distribution. IEEE Commun. Mag. **51**(4), 142–149 (2013)
29. LTE Direct: The Case for Device-to-Device Proximate Discovery, *Technical report, Qualcomm Research*. (2013)
30. L. Al-Kanj, Z. Dawy, E. Yaacoub, Energy-aware cooperative content distribution over wireless networks: design alternatives and implementation Aspects. *IEEE Communications Surveys & Tutorials*, TBD:TBD (2013)
31. C.B. Sankaran, Data offloading techniques in 3GPP Rel-10 networks: a tutorial. IEEE Commun. Mag. **50**, 46–53 (2012)
32. Bo Xing, Karim Seada, Nalini Venkatasubramanian, An experimental study on Wi-Fi ad-Hoc mode for mobile device-to-device video delivery. IEEE INFOCOM Workshops **2009**, 1–6 (2009)
33. WiFi Alliance, Wi-Fi Peer-to-Peer (P2P) Specifications, v1.2
34. IEEE. Std 802.11 g-2003*, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications* (2003)
35. IEEE. Std 802.11-2012*, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*. (2012)
36. A. Berl, H.D. Meer, Integration of mobile devices into popular peer-to-peer networks.In *Proceedings of the 5th Euro-NGI conference on Next Generation Internet Networks* (2009)
37. Evolved Universal Terrestrial Radio Access (E-UTRA), *3GPP Technical Report (TR) 36.814-900*. (2010)
38. J. Zhuang, L. Jalloul, R. Novak, J. Park, IEEE 802.16 m evaluation methodology document (EMD) (2009), http://ieee802.org/16/tgm/docs/80216m-08_004r2.pdf
39. Atheros 9 k drivers
40. ITU-R M.2135, *Guidelines for evaluation of radio interface technologies for IMT-Advanced. Technical report*. (2009)
41. K. Konstantinou, S. Kang, C. Tzaras, A measurement-based model for mobile-to-mobile UMTS links. In *Vehicular Technology Conference* (2007)
42. M. Gudmundson, Correlation model for shadow fading in mobile radio systems. Electron. Lett. **27**, 2145–2146 (1991)

43. G. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, S. Xu, Energy-efficient wireless communications: tutorial, survey, and open issues. IEEE Wirel. Commun. **18**(6), 28–35 (2011)
44. H. Kim, G. de Veciana, Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy. IEEE/ACM Trans. Netw. **18**(3), 802–815 (2010)
45. K. Huang, V. Lau, Y. Chen, Spectrum sharing between cellular and mobile ad Hoc networks: transmission-capacity trade-off. IEEE J. Sel. Areas Commun. **27**(7), 1256–1267 (2009)
46. F. Baccelli, N. Khude, R. Laroia, J. Li, T. Richardson, S. Shakkottai, S. Tavildar, X. Wu, On the design of device-to-device autonomous discovery. In *Proceedings of the International Conference on Communication Systems and Networks (COMSNETS)* (2012)
47. H. Dhillon, R. Ganti, F. Baccelli, J. Andrews, Modeling and analysis of {K}-tier downlink heterogeneous cellular networks. IEEE J. Sel. Areas Commun. **30**(3), 550–560 (2012)
48. M. Erturk, S. Mukherjee, H. Ishii, H. Arslan, Distributions of transmit power and SINR in device-to-device networks. IEEE Commun. Lett. **17**(2), 273–276 (2013)
49. F. Baccelli, B. Blaszczyszyn, Stochastic Geometry and Wireless Networks: Volume I Theory. Now Publishers Inc. *Foundations and Trends in Networking***3**(3–4), 249–449 (2009)
50. O. Galinina, A. Trushanin, V. Shumilov, R. Maslennikov, Z. Saffer, S. Andreev, Y. Koucheryavy, Energy-efficient operation of a mobile user in a multi-tier cellular network. In *Proceedings of the International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA)* (2013), 198–213
51. A. Pyattaev, K. Johnsson, S. Andreev, Y. Koucheryavy, Proximity-based data offloading via network assisted device-to-device communications. In *Proceedings of the IEEE Vehicular Technology Conference (VTC-Spring)* (2013)
52. A. Pyattaev, K. Johnsson, S. Andreev, Y. Koucheryavy, 3GPP LTE traffic offloading onto WiFi direct. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC)* (2013)
53. J. Gebert, R. Fuchs, Probabilities for opportunistic networking in different scenarios. In *Proceedings of the Future Network & Mobile Summit (FutureNetw)* (2012)
54. Qualcomm Atheros, *AR4100 System in Package 802.11n—General Availability*. (2012)
55. A. Jensen, M. Lauridsen, P. Mogensen, T. Sorensen, and P. Jensen, LTE UE power consumption model: for system level energy and performance optimization. In *Proceedings of the IEEE Vehicular Technology Conference (VTC-Fall)* (2012)
56. H. Wang, X. Chu, Distance-constrained resource-sharing criteria for device-to-device communications underlaying cellular networks. Electron. Lett. **48**(9), 528 (2012)
57. T. Chen, G. Charbit, S. Hakola, Time hopping for device-To-device communication in LTE cellular system. In *Wireless Communications and Networking Conference (IEEE WCNC)* (2010)
58. J. Gu, S. Jae Bae, B.-G. Choi, M.Y. Chung, Dynamic power control mechanism for interference coordination of device-to-device communication in cellular networks. In *Proceedings of Third International Conference on Ubiquitous and Future Networks ({ICUFN})* (2011)