Cellular Traffic Offloading onto Network-Assisted Device-to-Device Connections

Sergey Andreev[†], Alexander Pyattaev, Kerstin Johnsson, Olga Galinina, and Yevgeni Koucheryavy

February 12, 2014

Abstract

While operators have finally started to deploy fourth generation broadband technology, many believe it will *still* be insufficient to meet the anticipated demand in mobile traffic over the coming years. Generally, the natural way to cope with traffic acceleration is to reduce cell size; and this can be done in many ways. The most obvious method is via pico-cells, but this requires additional capital (CAPEX) and operational (OPEX) investment to install and manage these new base stations. Another approach, which avoids this additional CAPEX/OPEX, involves offloading cellular traffic onto direct device-to-device (D2D) connections whenever the users involved are in proximity. Given that most client devices are capable of establishing concurrent cellular and WiFi connections today, we expect the majority of immediate gains from this approach to come from the use of the unlicensed bands. However, despite its huge commercial success, WiFi-based direct connectivity may suffer from stringent session continuity limitations, excessive user contention, and cumbersome manual setup/security procedures.

In this article, we detail our vision of integrating managed D2D communications into current cellular technology to overcome the limitations of WiFi. We also quantify the estimated network performance gains from offloading cellular traffic onto D2D connections. Our analysis is based on an advanced system-level simulation toolkit which captures the relevant details of the network environment and on a detailed characterization of dynamic D2D communications based on stochastic geometry. We conclude that D2D communications provide a significant boost to network capacity as well as user energy efficiency and quality of service perception.

Introduction

Industry has recently completed the fourth generation (4G) of mobile broadband standards offering decisive improvements in all aspects of wireless system design. However, with the predicted explosion in both types and numbers of wireless devices [1] it is commonly believed that despite novel 4G technologies, mobile broadband networks will still face a capacity crunch in the near future.

The most expedient way of boosting network capacity is by increasing cell densities across the network (i.e. shrink cell sizes and increase their numbers in congested areas). This improves network capacity by increasing the frequency reuse per unit area and the average data rate per transmission [2] (smaller cells yield shorter radio links and thus higher data rates). However, greater cell densities imply increased interference management complexities and CAPEX/OPEX for the mobile broadband operator. Hence, industry will not be able to leverage the full potential of the "small cell revolution" until changes are made to the way we approach wireless content delivery.

Currently, the lion's share of expected mobile traffic growth comes from peer-to-peer (P2P) services that commonly involve clients in close proximity. This presents an excellent opportunity for clients to offload their traffic onto direct device-to-device (D2D) radio links (which are generally shorter and lower-to-the-ground than standard "small cell" connections). If the mobile broadband operators were to encourage this form of offloading by providing assistance with device discovery, D2D connection establishment, and service continuity, it could reduce the network load *without* the cost of additional infrastructure while creating the potential for new service revenue.

From the client's perspective the benefits are clear; D2D communication promises higher data rates, lower transfer delays, and better power efficiency [3]. These potential benefits along with the growing number of services and applications that could leverage user proximity have led academia and industry to aggressively pursue research and standardization of D2D communications over the past couple of years. The potential applications of D2D in cellular networks are numerous [4] and include local voice service (offloading calls between proximate users), multimedia content sharing, gaming, group multicast, context-aware applications, and public safety.

However, depending on client mobility patterns, some services are better suited for proximity-based communication than others [5]. For example, if D2D peers are non-stationary, the quality of the link may change dramatically over short periods of time, thus making it difficult to guarantee service. In these cases, the best candidates for network offloading are delay-tolerant services, i.e. those whose traffic can be queued until either the D2D link recovers or a path switch to the infrastructure network completes (e.g. video-on-demand or file transfers). However, if both clients are (semi-)stationary, many other P2P services, such as cooperative streaming and social gaming, can be offloaded onto D2D links with good results.

We recently completed an advanced, in-depth characterization of D2D communication with the goal of fully understanding the performance gains and resolving any impediments to them. In this article, we reveal our most important findings starting with a technical discussion of several options for network-assisted D2D and their potential implementation within 3GPP's Long Term Evolution (LTE) technology. We continue by quantifying the predicted gains of these solutions via both system-level simulations and mathematical analysis based on stochastic geometry.

Technology Alternatives for D2D

Driven by a wealth of potential use cases and suitable applications, the concept of licensed band D2D communication as an *underlay* to a cellular network has been developed and described at length in [6] and numerous subsequent works. The D2D underlay operates on the same bands as the cellular network, thus D2D users must adjust their transmit powers to avoid interfering with cellular users. While licensed band D2D communication constitutes a well explored area, the corresponding standardization efforts are still in the early stages, and the complexity and variety of potential solutions suggest a long standardization process. Thus, "time to market" is expected to be several years.

By contrast, unlicensed band D2D protocols are already standardized and available on client devices today. Thus, it makes sense to leverage their market availability. Unfortunately, for most of these protocols device discovery consumes too much energy, connection establishment is cumbersome, there is no service continuity, and radio resource management is inefficient [7]. Since all of these shortcomings can be improved if not eliminated with the use of network assistance, we propose that devices with mobile broadband connectivity receive help from their operator networks to manage their unlicensed band D2D connections.

D2D as a Cellular Network Underlay

D2D connectivity represents a cost-effective means of achieving enhanced radio resource utilization in the cellular network. However, the distributed nature of D2D communications creates inefficiency at every stage of the protocol [8]. Consequently, there is a significant amount of literature describing various levels of network management for D2D communications ranging from minimal involvement, such as in Aura-net, to fully controlled solutions where the network controls/schedules each D2D link. Clearly, the latter is more challenging, since the network must manage radio resources (e.g. channel, power, and rate selection) across all links, both infrastructure and D2D.

When clients are in proximity, there are several radio resource alternatives for their P2P data: (i) the cellular infrastructure, (ii) a direct link reusing resources reserved for cellular use, and (iii) a direct link reusing free resources not allocated for cellular use. The choice between these alternatives is known as transmission mode selection, and research on this topic encompasses a range of optimization targets from signal to interference plus noise ratio (SINR) and throughput to energy efficiency, data delay, fairness, and outage probability. Beyond the choice of optimization target(s), the literature primarily differs in the numbers and types of communicating entities (base stations, cellular and D2D users), the emphasis on uplink (UL) vs. downlink (DL) communication and the resulting interference, orthogonal vs. non-orthogonal resource sharing, the degree of available network assistance, and network/D2D duplexing mode.

In summary, many aspects of licensed band D2D have been thoroughly evaluated, including the design of D2D-aware multiple-input and multiple-output (MIMO) schemes, application of network coding, successive interference cancellation, and even wireless video distribution over D2D [9]. However, given the significant changes required to implement licensed band D2D (particularly on the physical layer) the 3GPP standardization process is slow going. Thus, the immediate attention of many industrial players has shifted towards first implementing D2D over the unlicensed bands.

Leveraging Unlicensed Spectrum for D2D

In the unlicensed bands, no entity has exclusive use of the spectrum (i.e. there is no single entity managing radio resources). As a result, radio access technologies designed for use on the unlicensed bands must be robust to random interference (i.e. unscheduled channel access), and therefore are typically revolving around the notion of "random access". Based on IEEE 802.11 standards, WiFi is currently the predominant technology employed in wireless local area networks (WLANs), whether used solely between devices or in conjunction with infrastructure access points. Since WiFi operates over shorter links and higher frequencies, it achieves greater spatial reuse, higher data rates, and better energy efficiency than cellular technologies (e.g. LTE).

Given that WiFi technologies reside on the unlicensed bands, they cause little (or no) interference to licensed band LTE networks. But while this makes WiFi a great choice for D2D underlay in cellular networks from the operator's perspective, from the client's perspective, this may not always be the case. For example, WiFi technology lacks a fast and resource efficient method of device discovery [10] (i.e. of notifying clients when/if they are in D2D range). Thus, if a client searches for a specific peer who happens to be out of range for a long period of time, it will suffer significant battery drain. There are other issues as well, such as the cumbersome D2D connection establishment procedure, lack of service continuity, and inefficient radio resource management. However, all of these issues can be addressed with proper management from the network.

If clients are connected to the LTE network, it knows their most recent cell associations (and tracking areas if they are in idle mode), and if Location Services are enabled, it knows their geographic locations to within a few meters. This information enables the network to quickly and without significant overhead determine if and when clients are in proximity (i.e. potentially within D2D range) and inform them accordingly. Once proximity is detected and clients decide to connect, the network can assist with D2D connection establishment, speeding up the procedure and reducing the amount of required signaling. Then, when clients are engaged in D2D communication, the network can provide robust session continuity in case the D2D link fails. In fact, the network can assist with many aspects of D2D communication including mode selection (i.e. when to offload onto D2D vs. fall back to LTE), power control, and transmission format (modulation and coding rates, MIMO transmission mode, etc.).

D2D Management in 3GPP LTE

Understanding the significant market potential of D2D, 3GPP SA1 (the "Services" group) began a study item on Proximity Services (ProSe) in 2010 with the goal of defining relevant usage models and deriving technical requirements for D2D within 3GPP LTE networks. This study item finished in early 2013 and included network assistance for D2D discovery and communication (for both licensed and unlicensed band D2D).

In late 2012, the ProSe standardization work moved to SA2 (the "Architecture" group), and thus far they have agreed to numerous "solutions for further study" on a range of ProSe topics including Discovery (both direct discovery and EPC detection/reporting of device proximities), Direct Communication (one-to-one and one-to-many), Relays (UE-to-UE and UE-to-network relays), Identifiers, and EPC Support for WLAN Direct Communication. Since SA2 is still in the study phase, none of these solutions are part of the specification as of yet. They have simply been accepted for consideration into the

specification. Given the scale of ProSe topics under consideration, SA2 recently decided that some level of prioritization of features is necessary in order to meet the deadline for Release 12. The remaining ProSe features will be considered for Release 13.

Given the inherent complexity of implementing D2D on the licensed bands, the 3GPP RAN (Radio Access Network) groups began a feasibility study on "LTE-Direct" two years ago. This Stage 1 work was recently completed, and Stage 2 (logical analysis) work has begun. However, given the slow pace of this standardization effort, LTE Direct is not expected on the market for several years. Hence, in our work we focus on a network-assisted D2D solution based on the existing WLAN D2D protocol, WiFi Direct, which implements the IEEE 802.11-2012 standard.

Neighbor Discovery and D2D Connection Establishment

To enable D2D communication, two primary steps are required: device discovery and D2D connection establishment. Both can be accomplished in a distributed manner, but there are benefits from network assistance for both.

Since the network is capable of tracking client locations, it can significantly reduce the amount of time a client spends in discovery by informing them when they are in proximity. Essentially, this allows clients to keep their D2D radios in idle until the devices are close enough for D2D communication, resulting in significant radio resource and battery savings. Moreover, since the network has secure access to all of its clients (as well as other networks' clients through interworking with other network operators), it can provide secure discovery of "stranger" devices, which opens the door to content/service discovery. In other words, instead of being limited to traditional discovery, where a client must either search for a *known* device that has its desired content/services or connect to a series of *stranger* devices in search of the desired content/services, with network assistance a client can simply enter a search for the desired content/services, and the network will inform it if/when there is an *authenticated* device in proximity offering the desired content/services.

Another benefit of network assistance is that it enables client anonymity during discovery and D2D communication by masking their permanent device IDs (i.e. clients identify themselves via their application layer IDs, but use temporary link layer IDs on the D2D channel and remain anonymous to everyone except the clients they are currently communicating with).

Network-assisted device discovery (whether for a specific device or content/services) can be implemented in a number of ways, but the most logical and efficient solution leaves P2P content/services management on 3rd party P2P servers and D2D discovery/communication management on the 3GPP operator network, while enabling interworking between them.

Let us observe the participants of a typical P2P session in Figure 1a. Most P2P applications have some sort of content tracker (i.e. the 3rd party application server), which is a trusted entity in the Internet to which all registered users have access. The content tracker logs all available (i.e. offered) user content, authenticates users, and authorizes content access. In conventional cloud-based services, this content tracker functionality is coupled with a content delivery network that acts as a relay between users for content exchange. Nearly all social networking applications work this way, examples include Facebook (uses Akamai for content delivery), YouTube (which uses separate sets of servers for the UI and data storage), and many more. In essence, they provide their users with P2P connectivity via a huge persistent cache, yet the idea matches the more conventional BitTorrent networks quite closely.

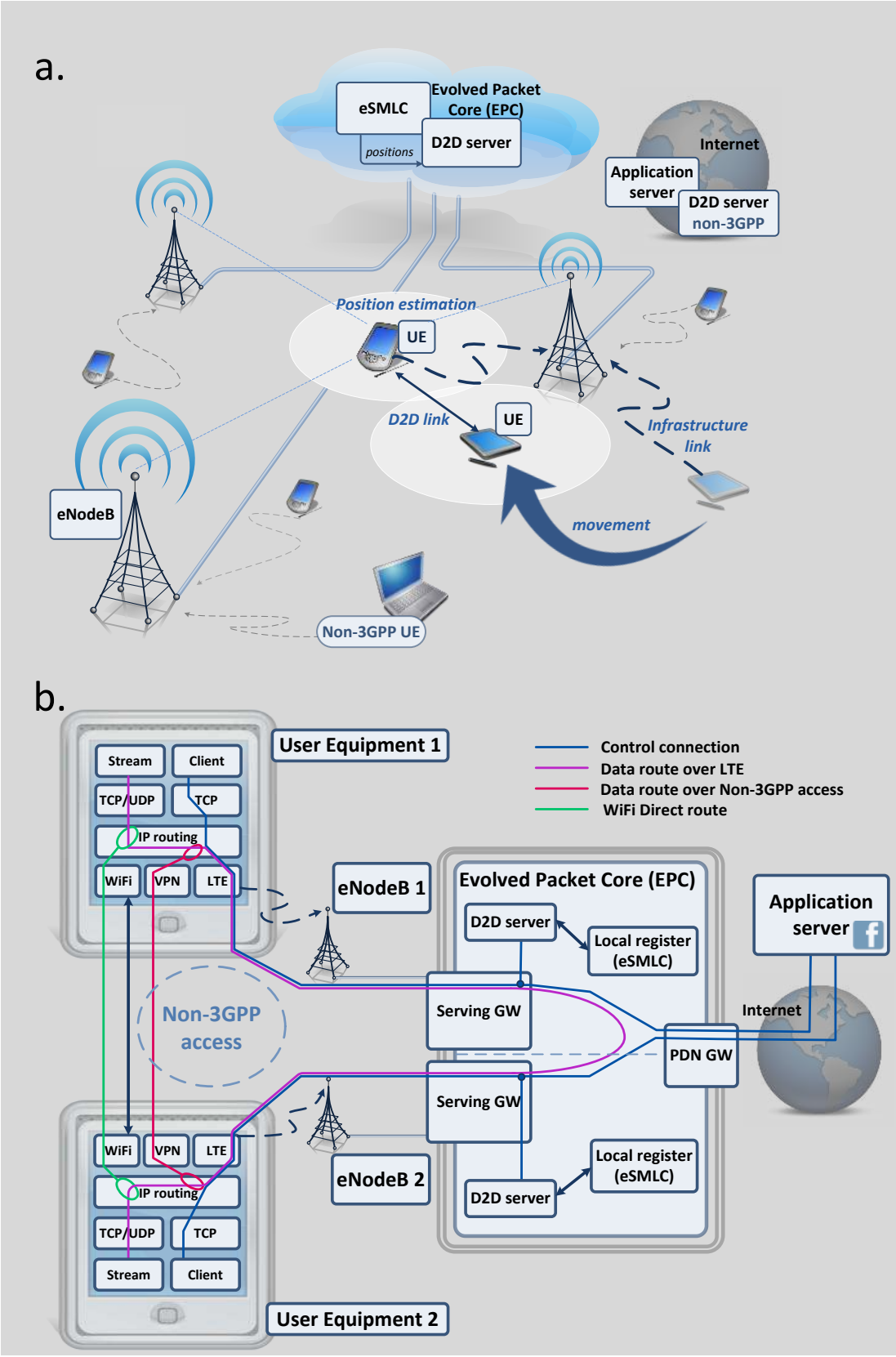


Figure 1: Architecture of the envisioned D2D offload system

To facilitate P2P sharing over D2D links, content trackers must store the locations of all offered P2P content/services from their registered users. For example, YouTube's content tracker would store the addresses of all Internet servers offering a specific video as well as the identifiers of clients offering to share the video via infrastructure and/or D2D. The content tracker would then provide the requesting user with alternative download sources by encoding the video's locations as URIs. In our proposed concept of network-assisted D2D communications, content trackers provide the following:

- Unique application-specific user IDs in the form of username@domain (we refer to this as the appID).
- Means to authorize a 3rd party to perform actions on behalf of the user (e.g. with oAuth).
- Tracking of P2P content and permissions to access it.

Note: all of the above are already provided by social networking applications.

Unfortunately, a D2D connection that is not yet established cannot be represented or managed in any conventional way. Our proposed solution uses a new network entity called a D2D server to circumvent this problem. The D2D server is a globally visible entity with a static domain name, normally located in the network operator's EPC for 3GPP devices and in the ISP's network for non-3GPP devices. The D2D server acts as a trusted connection manager for devices engaged in D2D discovery and/or communication. The D2D server performs the following functions:

- Maps client device identifiers to their users' appIDs;
- Tracks client device positions based on available positioning services:
  - Manual configuration;
  - GPS or Assisted-GPS (if available);
  - WiFi AP information and/or fingerprinting (for devices with WLAN interface);
  - Cell-ID reported by pico-eNB base stations (it would be too inaccurate for macro deployments);
  - Network positioning using OTDoA mechanisms through PRS or SRS in E-UTRAN (for 3GPP devices only, this requires cooperation with eSMLC);
- Provides clients with temporary link layer IDs to enable anonymous discovery;
- Automates D2D connection establishment (including security key exchange);
- Manages active D2D connections (e.g. initiating fall-back to infrastructure to guarantee service continuity, providing guidance for improved radio resource management, etc.).

In the case of 3GPP client devices, placing the D2D server in the core network enables additional benefits. First, should a client behave in a way considered detrimental to D2D performance, it can be banned from network-assisted D2D based on its hardware address (IMSI-code). Second, network assistance can be sent via the 3GPP control channels guaranteeing short round-trip times between the D2D server and client devices even under heavy traffic conditions. Finally, in the EPC a D2D server can leverage the network's Location Services to track client locations. Next, we outline the operation of the proposed solution during D2D connection setup, highlighting the interactions between entities.

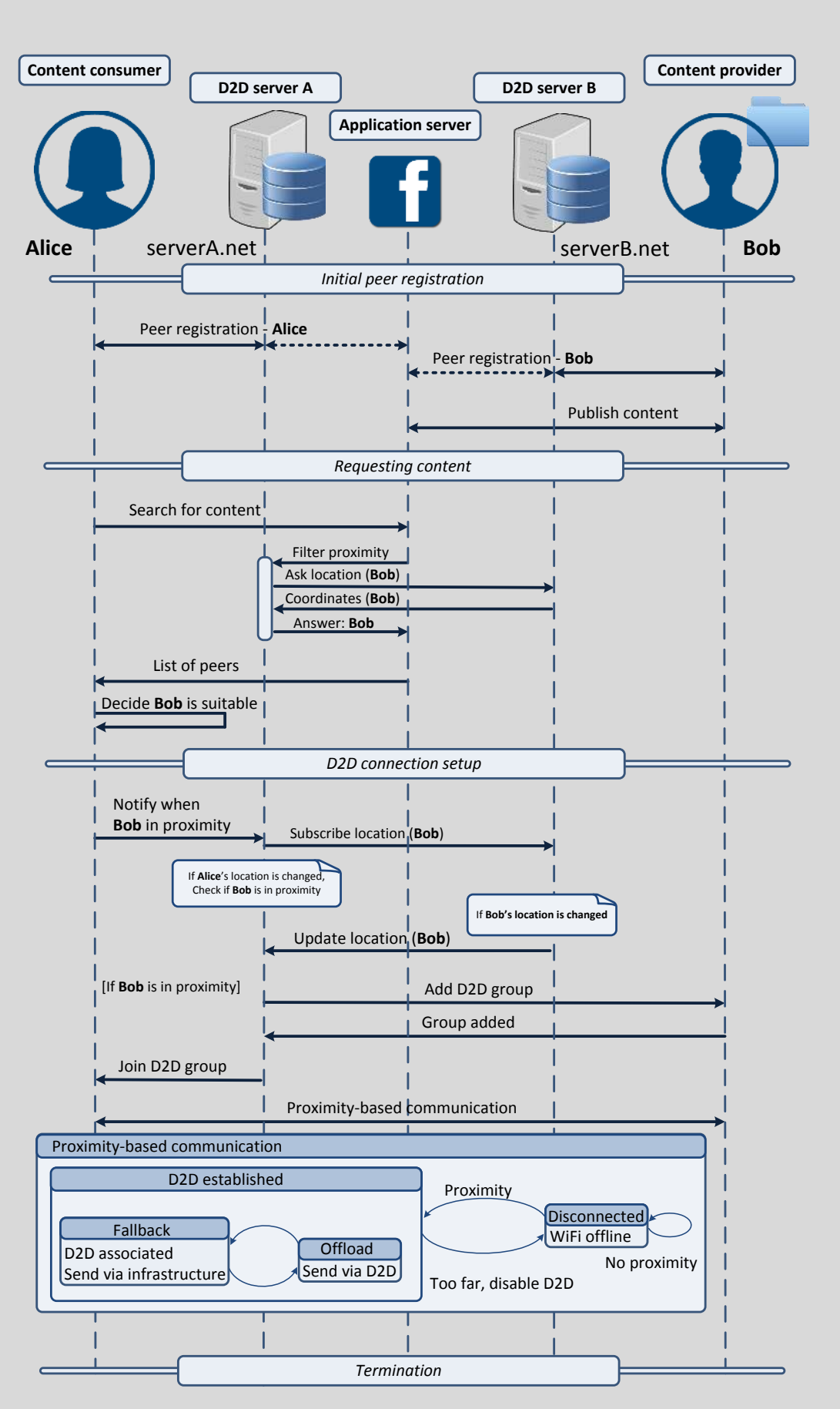


Figure 2: Proposed signaling for ProSe procedure

Communication Between Entities

Network-Assisted D2D Enablement during Network Registration During network registration (see Figure 2), the user authorizes (via the oAuth protocol) his UEs D2D server to enter the "server domain name" into his profile on the application servers where he is registered. This list of registered application servers may be stored in the UEs system settings. Later, if/when the UE is assigned to a new D2D server, these updates are performed again.

Publishing and Searching for P2P Content/Services When a user wants to access P2P content/services, he logs into the relevant 3rd party application server (i.e. content tracker) and searches (becoming a consumer) for the locations of the desired P2P content/service. Similarly, if a user wants to make specific P2P content/services available to its peers (becoming a provider), he will log into the relevant 3rd party application server and publish that information; optionally, the user can also upload the content to the application server's cloud storage (not shown on the diagram).

When a user logs into the application server and searches for specific P2P content/services, the content tracker may filter the list of content/service locations based on conditions given by the user. For example, if the user indicates that he wants to access the P2P content via unlicensed band D2D, the tracker may filter the list of content locations to only those belonging to peers with the appropriate D2D capability and within proximity of the requesting user. To perform this filtering, the application server must first request information from the requesting user's D2D server regarding the D2D capabilities and proximities of the requesting user and potential content providers. If a potential content provider (i.e. peer) is not managed by the same D2D server, the requesting user's D2D server will contact the peer's D2D server to determine its position and D2D capability. In essence, contact with D2D servers enables application servers to determine if two users (identified via their appIDs) are in proximity without having to know their exact geographic locations. This is important since many users prefer to keep their location information private.

Establishing/Terminating D2D Connections When a user clicks on the content link provided by the application server, and that content link references a peer with D2D capability, the user's device contacts its D2D server for assistance with D2D connection establishment. Then, depending on the information provided by the network's location services, the D2D server decides which path will be used for the P2P session (i.e. infrastructure or D2D).

While the P2P session is active on the D2D link, the D2D server monitors the users' locations (and potentially their D2D link) to determine when/if the session should be moved back to the infrastructure. A simple state machine in the D2D server allows for some hysteresis in the decision to offload onto D2D or fallback to the infrastructure, preventing the client devices from powering interfaces on and off due to small channel or position changes. The offloading procedure can be implemented through route injection (see Figure 1b) or more conventional mechanisms such as mobile IP (MIP).

During the termination stage the client devices and their respective D2D servers are all informed of the D2D connection termination. The particular details may be implementation specific.

Prototyping a D2D Offloading System

Based on the protocol described above, we developed a network-assisted D2D offloading prototype. The prototype includes a content tracker which allows clients to publish their offered content in the form of content links: `d2d://user@domain@prose\_server/http:8080/path`, which can be resolved by the D2D server into client identifiers and URIs (uniform resource identifiers, e.g. a shared directory, webcam, etc.). The content tracker is a conventional web server running a PHP-based application to construct the content links and an SQL database to store the information.

In addition, a D2D server is deployed in the Internet (instead of the 3GPP network's EPC, as this would require permission from a mobile network operator) to assist with D2D connection establishment and traffic offloading. In the prototype, the D2D server is implemented as a web-based application in Python allowing HTTP-based communications with both clients and content trackers. The clients are D2D-capable Android-4 devices equipped with a D2D network assistance service that enables them to communicate with the D2D server and change the kernel's routing table for data sessions with peers for whom the user has requested D2D (so that when the D2D path becomes available those data sessions are routed via D2D).

D2D offloading is transparent to the user, i.e. all operations are performed in the background. When a user logs into the content tracker, he searches for the desired content. The content tracker then returns all available content locations (specifying the content provider's connectivity capabilities, i.e. infrastructure only or infrastructure + D2D), and the user chooses a content location. When the user chooses a content location (i.e. provider), its UE contacts the D2D server to request the current IP address of the content provider in order to establish a transport connection. The D2D server responds with the IP address and then informs the application on the content provider to begin serving incoming connections. If the user has requested D2D, and the content provider is D2D capable and in proximity, the D2D server will also assist the clients in establishing a D2D connection.

During evaluation of the prototype, the quality of video streaming was significantly increased by using D2D offloading even at large inter-device distances (up to 20 meters indoors, 50 meters outdoors). While infrastructure path delays were commonly around 100 ms, which is unacceptable for many applications such as gaming where tolerable round-trip times are generally below 50 ms, D2D path delays in office environments were mostly below 5 ms.

System-Level View of Direct Communications

With network-assisted D2D, users can efficiently determine when they come into D2D range and offload their P2P sessions from infrastructure to D2D links. This represents the potential for significant gains in network capacity and client device throughput and energy efficiency. In this section, we demonstrate these gains using 3GPP LTE traffic offloading onto WFD (WiFi Direct) as our baseline.

For this purpose, we developed an advanced system-level simulator (SLS) based on up-to-date LTE evaluation methodology and current IEEE 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/WFD infrastructure and client deployment choices (hexagonal vs. square cells, environment with or without wrap-around, uniform vs. clustered client distribution, etc.). With its help, we demonstrate the potential performance gains (from both network

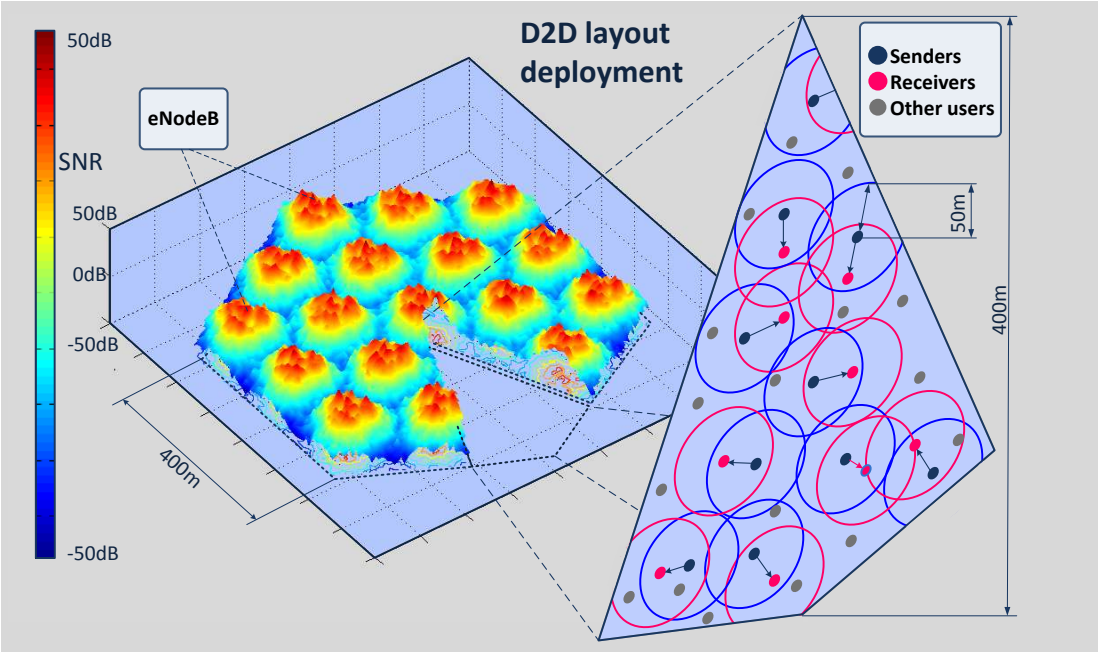


Figure 3: Integrated cellular and D2D layout

and client perspectives) from network-assisted D2D communications.

Representative D2D Scenario

In order to estimate the benefits of LTE-assisted WFD, we construct a sample scenario based on modern urban conditions. In particular, we implement the urban microcell environment defined by ITU/3GPP and combine that with a relatively high user density. We choose this type of dense deployment in order to recreate conditions where D2D would be most needed, i.e. where the cellular network would have difficulty supporting the offered traffic load. In this environment, the client density is such that each UE has a high probability of being within D2D range of at least one other UE (which is not necessarily the peer it wants to connect to).

Instead of modeling P2P content "supply and demand" from clients explicitly, we assume that a certain percent, x, of clients requesting P2P content are within D2D range of the clients providing the specific content. This approach allows us to explore the system without narrowing down to a particular model for content distribution and discovery, and thus compare the achievable performance against the standard LTE (without D2D benefits). In other words, we assume that all clients are engaged in P2P communication with a peer in the network, but that only a percentage of P2P pairs are within D2D range.

The LTE infrastructure network is comprised of 19 hexagonal cells supporting 3GPP LTE Release 10 technology (see Figure 3), and the distance between neighboring eNBs (inter-site distance) is 200 meters resulting in a cell radius of approximately 110 meters. A wrap-around technique is used to improve precision of the simulation at the edges of the deployment area. The system works over two 10 MHz bands for FDD operation (for UL and DL), shared by all cells with 3 sectors in each, resulting in a 1x3x1 reuse pattern.

Clients have both LTE and WiFi interfaces and are capable of engaging in LTE and WFD communications concurrently. They are uniformly distributed across the cellular environment and associate with eNBs based on the best DL SINR (resulting in 20 client devices per sector). Clients are stationary throughout the simulation run. Channels are modeled to incorporate all relevant source, destination, and environment characteristics.

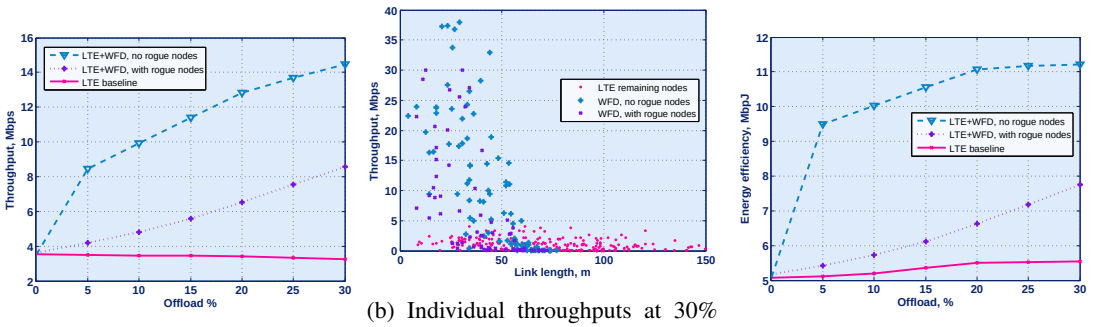
Each eNB is connected to the core network, providing cellular connectivity to all clients associated with it. Every client has its own traffic generator, enabling a variety of traffic patterns across the cellular deployment. For simplicity, in the examples below all client traffic is modeled as full buffer with packets of 1500 bytes each.

For more details on the configuration of the reference LTE network, the interested reader is directed to Table 1 and relevant standardization documents (e.g. 3GPP TR 36.814-900 and ITU-R M.2135-1). For performance verification purposes, we also implemented a calibration scenario from 3GPP TR 36.814-900, Table A-2.1, and ran the corresponding tests. Our simulation results fall well within the required limits for both cell-center and cell-edge spectral efficiency targets.

Table 1: Evaluation scenario parameters

Parameter	Value/Source
Core parameters	
UE Tx power limit	23 dBm IRP per interface
Observation period	10 seconds
LTE	
Propagation model	ITU-R M.2135-1, Tbl. A.2.2-1, A1-3
Shadowing model	ITU-R M.2135-1, 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, Tbl. 8-4
Antenna configuration	1x2 (diversity reception at eNB)
WiFi	
Propagation model	Empirical
Shadowing model	Correlation only
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 figure 7 dB, noise floor -95 dBm
Antenna configuration	1x1 (single antenna)

When clients are engaged in WFD communications, they not only have to contend with interference from other WFD links but also from devices engaged in regular WLAN communications with their WLAN APs. We assume these WLAN clients are not associated with the cellular network, thus their activity on the unlicensed WiFi bands cannot be monitored or managed by the LTE network. Hence, we refer to them as "rogue" clients. Rogue clients also have full buffers with packets of 1500 bytes, but their traffic always travels to their associated APs (i.e. they never engage in WFD). To simplify the evaluation methodology, we do not model WiFi AP DL traffic. Instead, we adjust the number of rogue clients to obtain the desired level of competition on the WiFi bands.



(a) Cell throughput

(b) Individual throughputs at 30% offload

(c) Cell energy efficiency

Figure 4: Performance results for saturated cellular + D2D network

Without loss of generality, our study assumes that all WiFi connections (conventional 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 (i.e. following the 802.11 protocol). We also assume that all APs and their respective clients (i.e. rogues) run the same version of the technology as WFD clients, namely IEEE 802.11-2012. To mimic realistic deployments, rogue devices are clustered around their associated APs. APs may be located anywhere inside the deployment area, recreating hot-spots similar to those in cafes, transportation hubs, etc. For more details on the configuration of WLAN deployments the reader is referred to Table 1. For calibration purposes, we employ reliable results from publications on ad-hoc WLAN deployments.

Understanding Performance Results

The results for total cell throughput are presented in Figure 4a. In these curves, the throughputs from LTE and WFD data sessions are totaled per cell, based on the requesting client's cell association. One can easily see that offloading LTE traffic onto WFD links results in a significant boost in cell throughput, actually increasing the throughput by the *factor of four* at the 30% offload level (and the more we offload, the higher is the gain).

If interfering rogue clients are present (with on average 5 devices per AP), throughput gains are more modest, but they are still around the factor of 2.5 at the 30% offload level. As our results show, D2D links perform best when the offloading percentage is low, and their performance degrades as the number of offloaded traffic sessions grows. This happens primarily due to increased contention between D2D links, but also due to rising overall noise levels.

Figure 4b presents more detailed throughput comparisons for WFD vs. LTE users in the same deployment. As the figure shows, WFD link performance varies significantly with length, and is naturally affected by the level of contention from the rogue nodes. Nevertheless, over standard WFD ranges (i.e. below 50 m), WFD generally achieves several times better throughput than LTE. Given this clear performance advantage, it is obvious that clients should use WFD whenever they are in range unless the link is significantly degraded due to fading or interference.

Since energy efficiency is measured in bits per Joule, it is agnostic to the particular technology involved. We compute energy efficiency based on the 100 mW circuit power, 200 mW RX, and 100 mW + transmit power for TX. Figure 4c shows that WiFi communication is significantly more energy

efficient than LTE. This is essentially due to WiFi's larger channel bandwidth. However, in addition to this, LTE clients are allocated smaller frequency chunks across multiple time slots, thus their transceiver circuitry has to stay active for extended periods of time while their amount of channel access is actually relatively low.

The WiFi MAC, on the other hand, activates the transceiver only when it is actually accessing the channel. When WiFi users are forced to defer their channel access due to RTS or CTS messages, they can sleep during those periods of time. When they 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.

As this study shows, there is significant potential for both network and client performance improvement from cellular network offloading onto WFD in urban environments. Since much of the predicted growth in social media traffic will be generated between clients in close proximity, ignoring this network offloading mechanism represents a significant loss in network capacity and user satisfaction. Whereas in this section we concentrated on the static offloading scenario and resulting performance limitations, in the following section we continue by assessing the performance of network-assisted D2D in a dynamic environment.

Applying Stochastic Geometry to D2D

The load on a given cellular network varies significantly both in time and location, thus it is important to capture network dynamics when evaluating system performance [11]. Unfortunately, dynamic systems are complex to model and time-consuming to simulate, thus in the next section we assess *flow-level* network performance *analytically* taking into account both user and traffic dynamics.

In our methodology, traffic sessions are initiated at random and leave the system after being served either by the cellular network (in the licensed bands) or by D2D links (in the unlicensed bands). As an example, we consider real-time sessions with a particular target bitrate and identical, independent holding times, which are characteristic of multimedia traffic.

Capturing Spatial Randomness

As demonstrated by our system-level simulations, the locations of network clients relative to each other highly impact system performance. Given that users are not spaced a regular distance apart, there may be a high degree of spatial randomness among them which needs to be accounted for. We thus adopt a random spatial model where user locations are drawn from a particular realization of a random process. When this topological randomness is coupled with the system dynamics, standard analytical methods of characterizing user signal power and interference no longer apply. Fortunately, the field of *stochastic geometry* provides us with a rich set of powerful results and analytical tools that can capture the network-wide performance of a random user deployment.

The use of stochastic geometry (i.e. statistical modeling of spatial relationships) has become increasingly popular over the last decades to analyze network performance averaged over multiple spatial realizations. It has also been useful in characterizing many important aspects of current cellular technology, from conventional macro-cell deployments to hyper-dense heterogeneous and small cell networks [12]. The application of stochastic geometry typically features a spatial *point process* to statistically model e.g. user locations yielding insights on the impacts of user density, transmit power, and path loss.

In the absence of information about user locations, the simplest statistical model is a uniform distribution, which in the two-dimensional plane corresponds to a homogeneous (stationary) Poisson Point Process (PPP). The PPP assumes that the points are independently distributed with some density in a unit area and that their positions are uncorrelated [13]. Other more realistic, but also significantly more complex point processes are binomial process spawning a fixed number of users in a given area and Poisson cluster process allowing users to cluster in certain locations. Finally, there is also hard core point process which is a thinning of the PPP such that the users have a guaranteed minimum separation. For more theory behind point processes, the interested reader is referred to a comprehensive tutorial in [14].

While the independence assumption may appear somewhat unrealistic, the Poisson model is surprisingly tractable and provides a reasonable first-order understanding of random deployments [15]. Assuming that transmitters and receivers are randomly scattered on a plane, the SINR due to varying path losses and transmit powers can be well modeled with a spatial distribution. The PPP models have thus been extensively used in the past, however, our approach in this article is different in that it targets joint characterization of spatial randomness together with dynamic user population and traffic load.

In particular, we model user locations as a PPP in R^3 treating time as another component of the vector space. In doing so, we arrive at space-time formulation in which space- and time-related variables are easier to decouple. However, extended formulations are also possible where user location models go beyond the stationary Poisson distribution.

Dynamic Model for Traffic Offloading

In what follows, we use stochastic geometry to characterize cellular traffic offloading onto network-assisted D2D. We look at a cellular network residing on the licensed bands co-located with a D2D "network" (i.e. collection of D2D links) employing the unlicensed bands. Both cellular and D2D links serve real-time UL user traffic. Due to their non-overlapping frequency bands, transmissions on the two networks do not interfere with each other, and every transmitter may send its data to a dedicated receiver via either the cellular network (infrastructure path) or D2D (direct path). In addition, we assume proper network planning reduces cellular link interference to "noise", while the D2D network is inherently interference-limited.

When a new user session arrives (per PPP) to the system, the cellular network checks if there is a valid D2D link available for the session (i.e. this link must meet the predefined offloading policy), and if so offloads the session onto the D2D link. The session is then served by the D2D link without interruption until it successfully completes and leaves the system. If the offloading policy is not met, the cellular network checks if the session meets the admission criteria (e.g. minimum signal quality). If the admission criteria are met, the cellular network serves the session until it successfully completes and leaves the system. If neither the offloading policy nor the cellular admission criteria are met, the session is considered blocked and leaves the system. At this stage, we are primarily interested in evaluating session blocking probability, when a new user session is admitted to neither the D2D nor the cellular network.

The session blocking probability P_{block} may be established as follows:

$$P_{block} = 1 - \left[P_a + (1 - P_a)\,(1 - P_b)\right],$$

where P_a is the D2D "network" acceptance probability and P_b is the cellular network blocking probability. In other words, a newly arriving user data session is said to be in "outage" with P_{block}.

To satisfy the offloading policy, a direct link must achieve a specified target bitrate (e.g. running at full power), while not significantly deteriorating the performance of already existing D2D connections. In particular, the interference from either client on the direct link to any receiver in the network cannot exceed a predefined threshold (may be technology-specific).

Cellular admission control of a new user session also depends on the specified target bitrate as well as the estimated impact of its admission into the network on UL transmission rates of existing users. When a new user session is admitted, the cellular network assigns the user a transmit power level (not to exceed the maximum power) and a dedicated fraction of the cellular network's time-frequency resources. The cellular network assigns new scheduling and power control levels for all active users every time a user data session enters or exits the system.

Analyzing Acceptance/Outage Probabilities

In order to verify our stochastic geometry analysis with extensive SLS evaluations, we implement a dynamic traffic offloading scenario from 3GPP LTE onto WFD. Our scenario concentrates on a so-called *area of interest* where co-located cellular and D2D networks cover a limited region with many users requiring service (i.e. shopping mall, business center, etc.). In this area, the users need to exchange small multimedia fragments with a given target bitrate. However, a particular transmitting user may either be successfully accepted by the D2D network, or rejected and need to demand LTE service instead. If cellular resource is insufficient to admit this user, it is blocked permanently.

In Figure 5, we compare the probabilities of session to be successfully accepted by 3GPP LTE vs. corresponding probabilities for 3GPP LTE *with* WFD offloading. The figure shows how the session acceptance probability of 3GPP LTE *with* WFD offloading, $1 - P_{block}$, the D2D acceptance probability, P_a, and the session acceptance probability of 3GPP LTE, $1 - P_b$, evolve with increasing session load.

These results indicate that when the cellular network is low loaded, it accepts all new user data sessions as long as they meet the minimum bitrate requirement. However, as the load increases, the network only admits user data sessions with high link quality that can be accommodated without significantly degrading existing sessions. As a result, we see a larger percentage of high bitrate connections at heavy traffic loads. Similarly, at low loads the D2D network only blocks new user data sessions due to insufficient bitrate (resulting from long links). However, unlike the cellular network, as the load grows the primary reason for session blocking on the D2D network is interference.

In summary, our methodology accurately models dynamic interworking between 3GPP LTE and WFD technologies. However, the main derivations are more general and can be extended to accommodate, for example, D2D operation in the licensed bands. Moreover, the proposed approach makes it possible to characterize other important system performance metrics such as area spectral efficiency and energy consumption of a typical data session. It can also be extended to a wider variety of practical offloading scenarios, network selection algorithms, quality of service measures, and advanced wireless technologies.

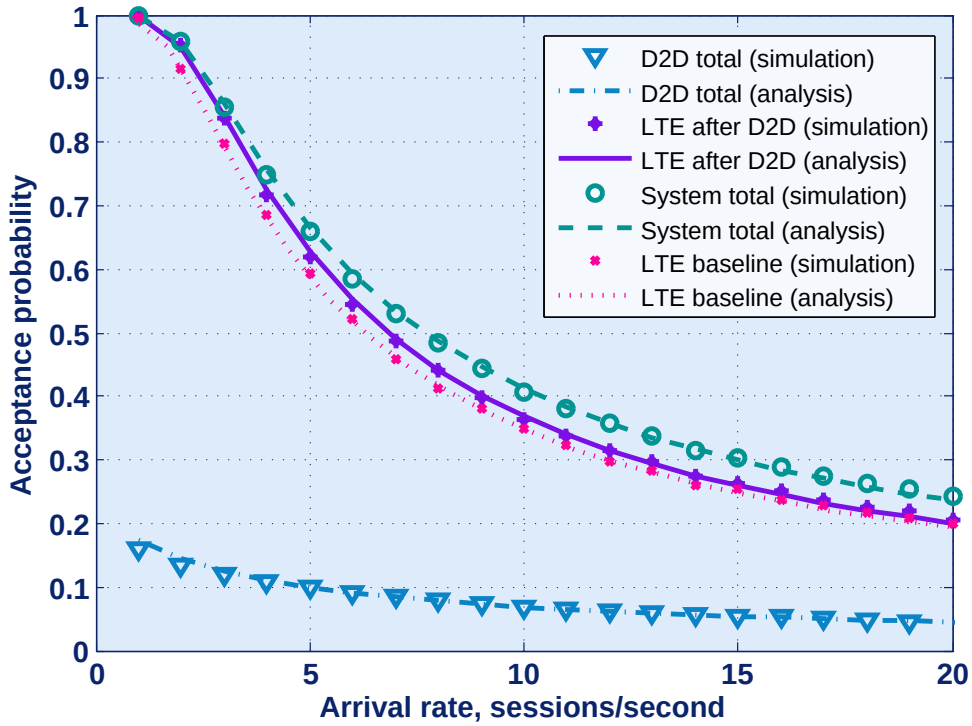


Figure 5: Acceptance probabilities for LTE and WFD networks

Conclusions

As this study reveals, network-assisted D2D has the potential to significantly improve both network and client performance. However, these gains depend on (i) establishing the proper offloading criteria (e.g. the D2D link must achieve the session's QoS requirements and inflict minimal harm to existing D2D links) and (ii) having enough source/destination pairs that meet said criteria. For example, in the case of 30% offloading, cell throughput increases over four times, while mean client energy efficiency improves by the factor of two. Without the proper offloading criteria, D2D links may fail to meet their session QoS requirements and/or cause excessive contention to existing D2D links resulting in lengthy packet delays and client battery drain.

This study also demonstrates the enormous range in D2D link qualities (even for links of the same length) compared to those of 3GPP LTE. Unlike 3GPP LTE where the link length does not significantly affect the bitrate (assuming fair scheduling), the D2D network's best links have to be carefully selected in order to meet the session's QoS requirements. However, with network-assisted D2D scheduling, the cellular operator can manage the resources of these D2D links to improve fairness, increase overall data rates by moving slow or highly-interfering D2D links back to 3GPP LTE, etc. Resource scheduling for D2D links can be done with varying granularity, the results of which we leave for future publications.

All in all, network assistance provides benefits to the cellular network and its clients on many levels. Because of its authorization/authentication capabilities, network assistance can provide secure D2D connectivity between P2P users that are currently outside each other's social spheres. It can also enable fast and energy efficient discovery of such peers. Finally, as was demonstrated in the last sections, it can

provide significant capacity and session acceptance probability improvements to the cellular network, as well as offer better throughputs and energy efficiencies to clients if/when they are in a position to meet the predefined offloading criteria.

Acknowledgment

This work is supported by Intel Corporation, GETA, TISE, and the IoT SRA program of Digile, funded by Tekes.

Bibliography

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017*, February 2013.

[2] S.-P. Yeh et al., "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Comm.*, vol. 18, pp. 32–38, 2011.

[3] G. Fodor et al., "Design aspects of network assisted device-to-device communications," *IEEE Comm. Mag.*, vol. 50, pp. 170–177, 2012.

[4] L. Lei et al., "Operator controlled device-to-device communications in LTE-Advanced networks," *IEEE Wireless Comm.*, vol. 19, pp. 96–104, 2012.

[5] C. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Comm. Mag.*, vol. 50, pp. 46–53, 2012.

[6] K. Doppler et al., "Device-to-device communication as an underlay to LTE-Advanced networks," *IEEE Comm. Mag.*, vol. 47, pp. 42–49, 2009.

[7] L. Al-Kanj et al., "Energy-aware cooperative content distribution over wireless networks: Design alternatives and implementation aspects," *IEEE Comm. Surveys & Tutorials*, vol. 15, pp. 1736–1760, 2013.

[8] B. Kaufman et al., "Spectrum sharing scheme between cellular users and ad-hoc device-to-device users," *IEEE Trans. on Wireless Comm.*, vol. 12, pp. 1038–1049, 2013.

[9] N. Golrezaei et al., "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Comm. Mag.*, vol. 51, pp. 142–149, 2013.

[10] A. Vigato et al., "Joint discovery in synchronous wireless networks," *IEEE Trans. on Comm.*, vol. 59, pp. 2296–2305, 2011.

[11] H. Kim and G. de Veciana, "Leveraging dynamic spare capacity in wireless systems to conserve mobile terminals' energy," *IEEE/ACM Trans. on Networking*, vol. 18, pp. 802–815, 2010.

[12] H. Dhillon et al., "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. on Sel. Areas in Comm.*, vol. 30, pp. 550–560, 2012.

[13] J. Andrews et al., "A primer on spatial modeling and analysis in wireless networks," *IEEE Comm. Mag.*, vol. 48, pp. 156–163, 2010.

[14] H. ElSawy et al., "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Comm. Surveys & Tutorials*, vol. 5, pp. 996–1019, 2013.

[15] J. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Comm. Mag.*, vol. 51, pp. 136–144, 2013.

Biographies

Sergey Andreev (sergey.andreev@tut.fi) is a Senior Research Scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the Specialist degree (2006) and the Cand.Sc. degree (2009) both from St. Petersburg State University of Aerospace Instrumentation, St. Petersburg, Russia, as well as the Ph.D. degree (2012) from Tampere University of Technology. Sergey (co-)authored more than 80 published research works on wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Alexander Pyattaev (alexander.pyattaev@tut.fi) is a Ph.D. Candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received his B.Sc. degree from St. Petersburg State University of Telecommunications, Russia, and his M.Sc. degree from Tampere University of Technology. Alexander has publications on a variety of networking-related topics in internationally recognized venues, as well as several technology patents. His primary research interest lies in the area of future wireless networks: shared spectrum access, smart RAT selection and flexible, adaptive topologies.

Kerstin Johnsson (kerstin.johnsson@intel.com) is a Senior Research Scientist in the Wireless Communications Laboratory at Intel, where she conducts research on MAC, network, and application layer optimizations that improve the mobile client experience while reducing wireless operator costs. She graduated from Stanford with a Ph.D. in Electrical Engineering and has more than 10 years' experience in the wireless industry. She is the author of numerous publications and patents in the field of wireless communications.

Olga Galinina (olga.galinina@tut.fi) is a Ph.D. Candidate in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. She received her B.Sc. and M.Sc. degrees in Applied Mathematics from Department of Applied Mathematics, Faculty of Mechanics and Physics, St. Petersburg State Polytechnical University, Russia. Her research interests include applied mathematics and statistics, queueing theory and its applications; wireless networking and energy efficient systems, machine-to-machine and device-to-device communication.

Yevgeni Koucheryavy (yk@cs.tut.fi) is a Full Professor in the Department of Electronics and Communications Engineering at Tampere University of Technology, Finland. He received the Ph.D. degree (2004) from Tampere University of Technology (TUT). Prior to joining TUT, Yevgeni spent five years in industry with R&D LONIIS in St. Petersburg, Russia, where he held various technical and managerial positions. Yevgeni actively participates in national and international research and development projects. He has authored or co-authored over 100 papers in the field of advanced wired and wireless networking and communications.