Network-Assisted Outband D2D-Clustering in 5G Cellular Networks: Theory and Practice

Article *in* IEEE Transactions on Mobile Computing · December 2016

DOI: 10.1109/TMC.2016.2621041

Arash Asadi
Technische Universität Darmstadt
43 PUBLICATIONS 2,635 CITATIONS
SEE PROFILE

Vincenzo Mancuso
Madrid Institute for Advanced Studies
142 PUBLICATIONS 3,971 CITATIONS
SEE PROFILE

Network-assisted Outband D2D-clustering in 5G Cellular Networks: Theory and Practice

Arash Asadi, *Member, IEEE,* and Vincenzo Mancuso, *Member, IEEE*

Abstract—We introduce a channel-opportunistic architecture that enhances the user experience in terms of throughput, fairness, and energy efficiency. Our proposed architecture leverages D2D communication and it is built on top of the forthcoming D2D features of 5G networks. In particular, we focus on outband D2D where cellular users are allowed to exploit both cellular (i.e., LTE-A) and WLAN (i.e., WiFi Direct) technologies to establish a D2D connection. In this architecture, cellular users form clusters, in which only the user with the best channel condition communicates with the base station on behalf of the entire cluster. Within the cluster, the unlicensed spectrum is utilized to relay traffic. In this article, we provide analytical models for the proposed system and study the impact of several payoff distribution methods commonly adopted in the literature on coalitional game theory. We then introduce an operator-controlled relay protocol based on the D2D features of LTE-A and WiFi Direct, and demonstrate the feasibility and the advantages of D2D-assisted cellular communication with our SDR prototype.

Index Terms—5G, WiFi Direct, Device-to-Device (D2D), Clustering, Protocol, Software Defined Radio (SDR).

1 Introduction

The emergence of Device-to-Device (D2D) communications has set off numerous proposals in industry and academia to improve the performance of future generation of cellular networks. As of today there are not only several proposals for cellular relaying, cellular offloading, and content distribution leveraging D2D [1], but also entire system architectures based on D2D to *complement* cellular-based services in a scalable way, with new types of applications to be supported in future networks, e.g., in 5G cellular networks.

In 3GPP's definition [2], D2D is a flexible paradigm for direct communication which is open to use cellular platforms (i.e., inband D2D) as well as IEEE 802.11-based WLAN platforms (i.e., outband D2D) [1]. The latter is gaining momentum due to its intrinsic value for optimizing heterogeneous networks, as well as due to regulatory issues arising for inband D2D [3] and thanks to the recent sprouting of fast D2D technologies in the IEEE 802.11 family over 2.4 GHz and 5 GHz, and up to the 60 GHz band [4]. D2D is a key 5G feature to support a variety of use-cases such as network offloading, public safety, Internet of Things (IoT), and Vehicle-to-Everything (V2X) communication. In particular, these use-cases highly benefit from D2D for grouping/clustering techniques. Analytical and simulation results indicate that clustering cellular users to relay traffic to each other leads to lower signaling overhead, higher spectral efficiency, and better energy efficiency than in legacy cellular systems [5]–[7]. Thus, clustering appears to suit the forthcoming 5G technologies such as IoT and V2X where the network density is high and energy/spectrum efficiency is of utmost importance. Moreover, operators can incentivize regular users to help each other to improve the overall system performance in return for a reward proportional to their contributions.

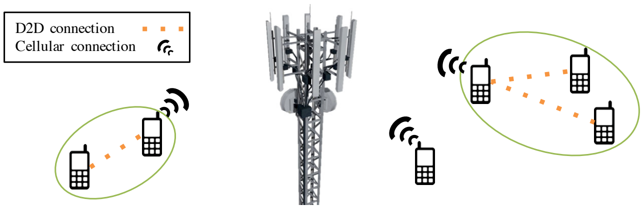


Fig. 1: Example scenario of D2D-clustering.

In this manuscript, we analytically and experimentally investigate an outband D2D clustering scheme that *opportunistically* leverages the flexibility of D2D communication in cellular networks for improving network performance. As described in Fig. 1, in our vision, cellular devices can form clusters using WiFi Direct [8] (and in the future 5G also with mm-Wave IEEE 802.11ad [9], [10]) and the cluster member with the highest channel quality will act as relay for other cluster members. The *opportunism* in our proposal is twofold: first, the relay changes over time based on the signal quality of the users within the same cluster; second, the throughput gain obtained due to clustering is shared within the cluster according to the contribution of each member. In the proposal detailed in this article, we aim to maximize the efficiency of cellular resource utilization, although we also account for the impact of per-user performance in general and for cluster formation policies in particular. We go beyond *theory* (used here to model throughput, energy consumption, and sharing of opportunistic gain) and

- *A. Asadi is at the Technische Universität Darmstadt with department of computer science. Email: arash.asadi@seemoo.tu-darmstadt.de*
- *V. Mancuso is with IMDEA Networks Institute, Madrid, Spain. Email: vincenzo.mancuso@imdea.org*

protocol design (presented here in the effort to cast our proposal in a 3GPP-compliant D2D framework), as we also *prototype* the first SDR-based experimental platform for a realistic analysis of D2D proposals. Our evaluation results indicate that joining a cluster is beneficial for all members, not just for average system performance. Indeed, our experiments reveals that our proposal enhances system capacity by up to 76% with clusters comprising as few as five users.

The rest of the article is organized as follows. Section 2 discusses the state of the art. Section 3 presents the system model. Section 4 models throughput and energy consumption of opportunistic D2D schemes. Section 5 provides the means for computing cluster revenue distribution by leveraging coalitional game theory. In Section 6, we complement the analysis by designing a protocol to integrate our proposal in LTE-A and WiFi Direct. In Section 7, we evaluate the performance of our proposal via extensive numerical and packet simulations, and by means of an SDR-based platform. Section 8 concludes the article.

2 Related Work

A comprehensive survey on D2D communication can be found in [1]. Here we focus on proposals that use clustering and D2D communication to improve the network performance.

In [11], the authors have shown that cluster formation in inband D2D increases system capacity specially for multicast. The numerical analysis and simulations show that their proposal achieves up to 66% throughput gain in comparison to legacy cellular system when only 20% of users have D2D opportunities.

Zhou *et al.* [5] propose an optimal resource utilization for multicast relaying with D2D clusters. They provide a closed-form expression for the probability distribution function (pdf) of the optimal number of relays in a cluster, and an intra-cluster retransmission scheme. They also show via numerical simulations that their proposed scheme achieves up to 40% gain in terms of resource utilization efficiency.

Andreev *et al.* study different aspect of outband D2D communication in [3], [12], [13]. In [12] and [3], the authors develop an analytical model for D2D offloading scenarios using stochastic geometry. Next, they evaluate the potential gain of outband D2D systems by using both system-level and mathematical analysis. They show that with as low as 30% cellular offloading, the aggregate cell throughput and the energy efficiency of UEs increases by a factor of four and two, respectively. Finally the authors study the implementation challenges of a network-assisted D2D setup with special focus on social networking in [13]. Moreover, they leverage an existing LTE experimental testbed [14] to implement their proposed D2D system and demonstrate its practicality in terms of delay constraints and user satisfaction.

This article offers an extensive research summary behind cellular-assisted opportunistic D2D communication, detailing the enabling technology and its implementation, security challenges, and user experience observations from large-scale deployments. Differently from our work, prior studies on clustered-D2D communication do not take advantage of channel-opportunistic gain in the clusters (i.e., choosing the cluster heads opportunistically). Moreover, the existing literature either does not specify the platform(s) used for relaying or it does not explain how relaying is adapted into a protocol structure of the intended platform. Finally, we are the first to verify the advantages of D2D clustering using an SDR experimental setup.

3 System model

In our proposal, users can form clusters and receive downlink traffic through the *cluster head*, i.e., the user enabled to exchange data with the LTE base station (eNB), see Fig. 1. Each user is a potential candidate to act as a cluster head. A cluster consists of several users that form a group and all intra-cluster communications occur over a WLAN in unlicensed spectrum (e.g., WiFi Direct or mm-Wave IEEE 802.11ad). Since the eNB controls clustering decisions, whenever a packet is destined to a cluster member, the eNB simply sends it to the cluster head. This maximizes the throughput at that scheduling epoch. Thereby, the eNB schedules clusters as if they were regular users. From a modeling perspective, a cluster can be considered a user whose Signal to Noise Ratio (SNR) is the highest of the SNR values of cluster members. As for intra-cluster resource sharing, unless otherwise specified, we assume that the extra throughput gained from clustering is equally distributed among users.

We model transmissions in a single OFDMA cell operating in FDD mode (like in LTE/LTE-A). There are N mobile users in the cell. Since we are interested in network capacity and fairness under heavy load conditions, we study the case of fully backlogged downlink flows in the analysis. The uplink case is similar and can be derived from our downlink analysis. The total per-frame capacity is denoted by S_{tot}, and we assume that the D2D link uses an IEEE 802.11 technology, and does not become a bottleneck in the data flow. In fact, considering the short-range nature of D2D communication, the available IEEE 802.11 capacity exceeds per-cluster achievable throughput over the cellular network's capacity. It is also assumed that all mobile users belong to the same operator. The channel of mobile user i is characterized by stationary Rayleigh fading. Therefore, the SNR can be described as a r.v. C_i with average SNR γ_i, so that the Cumulative Distribution Function (CDF) of the SNR is given by $F_i(z) = 1 - e^{-\frac{z}{\gamma_i}}, z \geq 0, \forall i \in \{1 \dots N\}$.

We assume that user channels are independently distributed but not identically, and the Channel State

Information (CSI) is available at the eNB. Transmissions occur at different rates according to M available Modulation and Coding Schemes (MCSs). We assume that the MCS for user i is a function of the instantaneous SNR, i.e.:

$$\begin{gathered} MCS_i = k \iff C_i \in [th_k; th_{k+1})\,,\ k = 1 \ldots M; \qquad (1) \\ th_1 = 0; th_p < th_q \iff p < q; th_{M+1} = \infty. \end{gathered}$$

Thus, the probability that scheduled transmissions to user i are encoded with the kth MCS is:

$$\pi_k^{(i)} = \int_{th_k}^{th_{k+1}} dF_i(z) = e^{-\frac{th_k}{\gamma_i}} - e^{-\frac{th_{k+1}}{\gamma_i}}. \qquad (2)$$

The number of data bits transferred in one OFDM symbol with the kth MCS is denoted by b_k.

4 Throughput and Energy Analysis

In this section, we derive the analytical expressions for calculating throughput and energy consumption of legacy cellular users and D2D clusters.

4.1 Throughput model

Opportunistic schemes commonly result in unfairness which is often resolved at the cost of increased complexity (i.e., requires solving hard or NP-hard problems). However, the practicality of such schemes is often doubted due to high computation overhead imposed to the eNB. We intentionally opt for opportunistic schemes with low complexity to pave the way towards a practical proposal. Here, we resolve the unfairness issue by leveraging cooperative nature of D2D clusters instead of increasing the scheduling complexity. We consider the case in which the eNB schedules N_c clusters instead of N normal users. This means that the eNB decides which cluster has to be served, and then transmissions are managed by the current cluster head. Defining X_n as the SNR of cluster n (CL_n), we have: X_n=$\max\{C_j, j : u_j \in CL_n\},\ n \in \{1 \ldots N_c\}$,where u_j is user $j \in \{1 \ldots N\}$. Considering that the random variables C_j are all independent, CDF of X_n is computed as:

$$F_{X_n}(z) = \prod_{j \in CL_n} F_j(z) = \prod_{j \in CL_n} \Big(1 - e^{-\frac{z}{\gamma_j}}\Big)\,, z \geq 0. \quad (3)$$

The adopted MCS, for each transmission, only depends on the instantaneous SNR of the best channel in the scheduled cluster, i.e., it only depends on X_n at the scheduling epoch:

$$\pi_k^{(CL_n)} = \int_{th_k}^{th_{k+1}} f_{X_n}(z) dz. \qquad (4)$$

Round Robin (RR). This is a simple scheduler that equally distributes the available resources. Under RR, a user's throughput depends on the number of users in the system, the probability to transmit with a given MCS, and the total resources S_{tot}:

$$E[T_i] = \frac{1}{N} S_{tot} \sum_{k=1}^{M} \pi_k^{(i)} b_k, \forall i \in \{1 \ldots N\}. \qquad (5)$$

Proportional Fair (PF). This scheme is a priority-based opportunistic scheduler with fairness constraints. Under PF, scheduling priorities are determined by the ratio of feasible data rate to average throughput at each time instant t (i.e., $R_i(t)/\mu_i(t),\ \forall i \in \{1 \ldots N\}$). Since closed form expressions for PF scheduling are available only for homogeneous scenarios (see, e.g., [15], [16]), we use simulations to evaluate the performance of PF in the heterogeneous scenarios assessed in our work.

Cluster Weighted Round Robin (CL(WRR)). This scheme chooses the cluster member with the best channel quality as the cluster head and it schedules the cluster heads in a Weighted Round Robin (WRR) fashion. Hence, each cluster n receives a portion of airtime which corresponds to its weight w_n, $n$$\in$$\{1 \ldots N_c\}$. In this paper, the weight of CL_n is calculated using w_n=N_n/N, where N_n denotes the number of cluster members of CL_n. In other words, each cluster receives an amount of airtime which is proportional to its size.

In such a system, the per-cluster scheduling probability is exactly w_n, while the average symbol rate only depends on the selected MCS. Since S_{tot} is devided in a WRR manner, the average cluster and the per-user throughput are given by the Propositions 1 and 2, whose proofs are immediate. From a theoretical perspective, throughput always improves as the cluster size increases unless all cluster members have always identical channel conditions. In particular, if a new user joins a cluster, it will either benefit from the presence of current members with better channel qualities or the current members benefit from the new user because it has a better channel quality. Thus, clustering reduces inefficient transmissions and enhances the aggregate system throughput. This can be seen in Eqs. (3) and (4) where the CDF of SNR and pdf of MCS for a cluster is obtained from the $\max$ of a series of random variables. We call this effect the *clustering gain*. Note that although large clusters are desirable theoretically, in practice the clustering gain is limited by intra-clustering signaling. Thus, the effective cluster size is bounded by the signaling overhead inside the cluster. Nevertheless, this overhead is tolerable for small and medium cluster sizes (5 to 10 members) as we show later in Fig. 10(c).

Proposition 1. *Under CL(WRR), the average throughput received by cluster CL_n is*

$$E[T_{CL_n}] = w_n S_{tot} \sum_{k=1}^{M} \pi_k^{(CL_n)} b_k,\ \ n \in \{1 \ldots N_c\}. \qquad (6)$$

Proposition 2. *Under CL(WRR), the average throughput of user $i \in CL_n$ can be expressed as*

$$E[T_i] = \frac{S_{tot}}{N} \sum_{k=1}^{M} \pi_k^{(CL_n)} b_k, \; i \in CL_n, \; n \in \{1 \dots N_c\}. \quad (7)$$

The following proposition gives the probability that a user i is scheduled.

Proposition 3. *Under CL(WRR), a user $i \in CL_n$ is scheduled with probability*

$$P_h^{(i)} = w_n \sum_{k=1}^{M} \pi_k^{(CL_n)} \int_0^\infty [1 - F_i(z|MCS_i = k)] \, dF_{Y_i}(z), \quad (8)$$

where $Y_i = \max_{j \in CL_n \setminus \{i\}} \{C_j\}, \; i \in CL_n$.

The proof of Proposition 3 is reported in [6]. Note that, under Rayleigh fading, $F_i(z|MCS_i = k)$ is simply given by the following formula:

$$F_i(z|MCS_i = k) = \frac{F_i(\min(z, th_{k+1})) - F_i(th_k)}{\pi_k^{(i)}}, z \geq th_k. \quad (9)$$

MaxRate Between Clusters (CL(MR)). Here, the cluster heads are scheduled in a pure MaxRate (MR) fashion [17]. In this scheme, the frame resources S_{tot} are allotted to the cluster whose cluster head is experiencing the best SNR in the system. Propositions 4 and 5 express the cluster throughput and average per-user throughput achieved using CL(MR).

Proposition 4. *Under CL(MR), the average throughput received by cluster CL_n is*

$$E[T_{CL_n}] = S_{tot} \sum_{k=1}^{M} \Big[\pi_k^{(CL_n)} b_k \times \int_0^\infty [1 - F_{X_n}(z|MCS_{CL_n} = k)] \, dF_{Y_n}(z) \Big], \quad (10)$$

where $n \in \{1 \dots N_c\}$, and $Y_n = \max_{j \notin CL_n} \{C_j\}$.

The proof of Proposition 4 is reported in [6].

Proposition 5. *Under CL(MR), the average throughput received by user $i \in CL_n$ is*

$$E[T_i] = \frac{S_{tot}}{N_n} \sum_{k=1}^{M} \Big[\pi_k^{(CL_n)} b_k \times \int_0^\infty [1 - F_{X_n}(z|MCS_{CL_n} = k)] \, dF_{Y_n}(z) \Big], \quad (11)$$

where $Y_n = \max_{j \notin CL_n} \{C_j\}$.

The proof of Proposition 5 is similar to Proposition 4. The following is the probability that a user i is scheduled as a cluster head, which is proven in [6].

Proposition 6. *Under CL(MR), a user i is scheduled with probability*

$$P_h^{(i)} = \sum_{k=1}^{M} \pi_k^{(i)} \int_0^\infty [1 - F_i(z|MCS_i = k)] \, dF_{Y_i}(z), \quad (12)$$

where $Y_i = \max_{j \neq i} \{C_j\}$ and $F_i(z|MCS_i = k)$ is given by Eq.(9).

4.2 Energy analysis

We derive the power requirements of mobile users from the empirical power models proposed for LTE and IEEE 802.11 in [18] and [19]. These studies quantify the baseline power required to keep a radio up and running, and the dependency of energy consumption on transmission rates. Here, unlike the existing models, we account for practical details such as energy consumption of mobilles in active and idle periods, and differentiate between transmission and reception power. Before elaborating on power models, we want to differentiate between the average throughput $E[T]$ and data rate R of a user. $E[T]$ is the user-application local data received by a user via the cellular link or the relay, and it is computed via (7) and (11). R is the amount of data received by a user and it includes non-local traffic to be relayed.

4.2.1 Power saving in cellular networks and WLANs

LTE-A leverages from Discontinuous Reception (DRX) and Discontinuous Transmission (DTX) which are mechanisms allowing cellular users to enter idle mode for energy efficiency purposes [20]. In IEEE 802.11-based WLANs, users can turn off the wireless interface during idle periods and only switch it on to receive beacons [21]. In both LTE and IEEE 802.11, interfaces in power saving mode periodically wake up to transmit/receive control information even if there is no data traffic to handle. However, it has been shown that the periodic wake-up of power saving mechanisms in LTE and WiFi impacts at most 5% of the idle time [18]. Therefore, for simplicity, we ignore the periodic wake-up operation. We assume that wireless interfaces instantaneously switch to power saving mode in absence of packets to be tranceived. With the arrival of a new packet in the transmission queue, the interfaces switch back to active mode instantly.

4.2.2 Cellular consumption

Based on [18], the downlink energy consumption of user i in the cellular network consists of the sum of a baseline power and a term which is proportional to the transmission rate of the device. As mentioned earlier, we extend the existing model to account for active/idle periods. The probability that the cellular interface is in active mode is equivalent to the probability $P_h^{(i)}$ of being the cluster head, see Eqs. (8) and (12). Therefore, the power used by the cellular interface of a device can be expressed as follows:

$$W_{\text{cell}}^{(i)} = P_h^{(i)} \, \beta_{\text{cell}} + \left(1 - P_h^{(i)}\right) \beta_{\text{cell}}^{\text{idle}} + \alpha_{\text{rx}} \, R_{\text{rx}}^{(i, \text{cell})}, \quad (13)$$

where, β_{cell} and $\beta_{\text{cell}}^{\text{idle}}$ are the baseline powers in active and idle mode, respectively; α_{rx} is the energy consumption per Mbps per second, and $R_{\text{rx}}^{(i, \text{cell})}$ is the average transmission rate of user i over the LTE interface. $R_{\text{rx}}^{(i,\,\text{cell})}$ is provided by Propositions 7 and 8.

Proposition 7. *Using CL(WRR), the cellular data rate of user $i \in CL_n$ is given by*

$$\begin{aligned} R_{rx}^{(i,\,cell)} =& w_n S_{tot} \sum_{k=1}^{M} \pi_k^{(CL_n)} b_k \\ & \times \int_0^{\infty} \left[1 - F_i(z|MCS_i = k)\right] dF_{Y_i}(z), \end{aligned} \tag{14}$$

where $Y_i = \max_{j \in CL_n \setminus \{i\}}\{C_j\}$, $i \in CL_n$.

The proof of Proposition 7 is omitted due to its similarity to the proof of Proposition 3.

Proposition 8. *Using CL(MR), the cellular data rate of user $i \in CL_n$ is given by*

$$R_{rx}^{(i,\,cell)} = S_{tot} \sum_{k=1}^{M} \pi_k^{(i)} b_k \int_0^{\infty} [1 - F_i(z|MCS_i = k)]\, dF_{Y_i}(z), \tag{15}$$

where $Y_i = \max_{j \neq i}\{C_j\}$, $i \in CL_n$.

The proof of Proposition 8 is omitted due to its similarity to the proof of Proposition 6.

4.2.3 D2D consumption

We use the model for IEEE 802.11 proposed in [19] that accounts for the power required for packet processing as well as for transmission. We extend the model to include the probability that user i is in active mode $P_a^{(i)}$. The power required by the D2D interface is therefore:

$$\begin{aligned} W_{\text{D2D}}^{(i)} =& P_a^{(i)}\, \beta_{\text{D2D}} + \left(1 - P_a^{(i)}\right) \beta_{\text{D2D}}^{\text{idle}} + \zeta_{\text{tx}}\tau_{\text{tx}} + \zeta_{\text{rx}}\tau_{\text{rx}} \\ & + \kappa_{\text{tx}}\lambda_{\text{tx}} + \kappa_{\text{rx}}\lambda_{\text{rx}}, \end{aligned} \tag{16}$$

where β_{D2D} and $\beta_{\text{D2D}}^{\text{idle}}$ are the WLAN baseline power levels in active and idle mode, respectively; ζ_{tx} and ζ_{rx} represent the power required in transmission and reception, respectively; τ_{tx} and τ_{rx} are the fractions of time spent in transmission and reception, respectively (i.e., $\tau_{\text{tx}}^{(i)} = R_{\text{tx}}^{(i,\text{D2D})}/R_{\text{D2D}}$ and $\tau_{\text{rx}}^{(i)} = R_{\text{rx}}^{(i,\text{D2D})}/R_{\text{D2D}}$); κ_{tx} and κ_{rx} are the power levels required for packet processing in transmission and reception, respectively; eventually, λ_{tx} and λ_{rx} are the packet rates, respectively in transmission and reception.

The WLAN/D2D power-related parameters introduced in Eq. (16) are computed as follows: $\lambda_{\text{tx}}^{(i,\text{D2D})}$ is computed as the ratio between the rate $R_{tx}^{(i,\text{D2D})}$ and the average packet size L_p; and similarly, user i transmits $\lambda_{\text{rx}}^{(i,\text{D2D})} = R_{\text{rx}}^{(i,\text{D2D})}/L_p$ packets per second. It is assumed that the achievable D2D rate is independent from the cellular network status and its average value R_{D2D} is the same for all clusters (i.e., this is an input parameter for our problem). If the achievable D2D rate is larger than the intra-cluster traffic (i.e., $R_{\text{D2D}} > \sum_{i \in CL_n} R_{\text{rx}}^{(i,\text{D2D})} = \sum_{i \in CL_n} R_{\text{tx}}^{(i,\text{D2D})}$), then to evaluate the D2D power consumption, we should compute the D2D data rates $R_{\text{rx}}^{(i,\text{D2D})}$ and $R_{\text{tx}}^{(i,\text{D2D})}$, and the probability $P_a^{(i)}$ that the D2D interface of user i be active. $R_{\text{rx}}^{(i,\text{D2D})}$ and $R_{\text{tx}}^{(i,\text{D2D})}$ is computed using Proposition 9, whose proof is reported in [6].

Proposition 9. *The D2D data rate of user $i \in CL_n$ is given by the following expressions, which hold for the received and transmitted traffic, respectively:*

$$R_{tx}^{(i,D2D)} = (1 - \delta_i) \cdot R_{rx}^{(i,\,cell)}, \tag{17}$$

$$R_{rx}^{(i,D2D)} = \delta_i \cdot \sum_{j \in CL_n \setminus \{i\}} R_{rx}^{(j,cell)}, \tag{18}$$

where $\delta_i = \frac{E[T_i]}{E[T_{CL_n}]}$.

Finally, the probability $P_a^{(i)}$ (i.e, the D2D interface of user i is in active mode) is given by Proposition 10.

Proposition 10. *The D2D interface of user i is active with probability $P_a^{(i)}$ that is computed as:*

$$P_a^{(i)} = \frac{E[T_i] + (1 - 2\delta_i) R_{rx}^{(i,\,cell)}}{R_{\text{D2D}}}. \tag{19}$$

Proof: $P_a^{(i)}$ is the sum of two terms: the probability that user i is the cluster head, i.e., sends data to other cluster members, and the probability that user i is not a cluster head, i.e., receives data from the cluster head. Since such probabilities can be interpreted as the average fraction of time spent in either in reception or transmission over the D2D interface, we have $P_a^{(i)} = (1 - \delta_i) \frac{R_{\text{rx}}^{(i,\,\text{cell})}}{R_{\text{D2D}}} + \delta_i \frac{E[T_{C_n}] - R_{\text{rx}}^{(i,\,\text{cell})}}{R_{\text{D2D}}}$, which leads to the result. □

Total power. Combining the results for cellular and D2D consumptions, the resulting total power required to operate a clustered user is as follows:

$$\begin{aligned} W_{\text{tot}}^{(i)} =& \beta_{\text{cell}}^{\text{idle}} + \beta_{\text{D2D}}^{\text{idle}} + \left(\beta_{\text{cell}} - \beta_{\text{cell}}^{\text{idle}}\right) P_h^{(i)} \\ & + \left(\beta_{\text{D2D}} - \beta_{\text{D2D}}^{\text{idle}}\right) \frac{E[T_i] + (1 - 2\delta_i) R_{\text{rx}}^{(i,\,\text{cell})}}{R_{\text{D2D}}} \\ & + \alpha_{\text{rx}}\, R_{\text{rx}}^{(i,\,\text{cell})} \\ & + \left(\zeta_{\text{tx}} + \frac{\kappa_{\text{tx}}}{L_p}\right) (1 - \delta_i) \frac{R_{\text{rx}}^{(i,\,\text{cell})}}{R_{\text{D2D}}} \\ & + \left(\zeta_{\text{rx}} + \frac{\kappa_{\text{rx}}}{L_p}\right) \frac{E[T_i] - \delta_i R_{\text{rx}}^{(i,\,\text{cell})}}{R_{\text{D2D}}}. \end{aligned} \tag{20}$$

The first term in Eq. (20) is the baseline power required by cellular and D2D interfaces in idle mode; the second and third terms are the baseline power of the interfaces in active mode; the fourth term accounts for cellular downlink transmissions, while the fifth term is the power for D2D transmissions when the user acts as a cluster head; finally, the last term represents the power needed to receive D2D traffic from the cluster head.

5 Cluster Formation and Payoff Allocation

This section provides a simple model for the cluster formation process, and sheds light on the impact of clustering when users experience non-stationary channel qualities. The cluster formation in our proposed architecture is modeled using *coalitional game theory* [22]. Here, we treat cluster formation as a game in which a users decide to join or to leave a cluster depending on the achievable reward. We analyze different alternatives to share the clustering gain, i.e., the *revenue*, among participating users. The revenue can be expressed in terms of throughput, power, energy efficiency, and so on. We choose energy efficiency so that we can maximize the system capacity with respect to power requirements, which is a key issue in today's cellular networks. Our analysis illustrated that clustering increases the probability of transmitting with higher MCSs. Since higher MCSs consume more power, our proposal increments power consumption during transmission and shortens transmission time. Indeed, the throughput gain is much higher than the increment in power consumption, and the energy cost to pay for a finite load data transfer decreases. To quantify this gain, we define energy efficiency as the amount of data (bits) transferred to the final user per energy unit, e.g., for user i, the energy efficiency is given by $\eta_i = E[T_i]/W_{\text{tot}}^{(i)}$.

5.1 Definition of the game

In the following, $U{=}\{u_1, \dots, u_N\}$ denotes the set of users in the network and $S{=}\{S_1, \dots, S_l\}$ is a partition of U, i.e., $\bigcup_{i=1}^{l} S_n{=}U$ and $S_n \cap S_j{=}\emptyset$ if $n \neq j$. The utility function $\nu(.)$ defines the value of cluster S_n as:

$$\nu(S_n)=\begin{cases}\sum\limits_{u_i\in S_n}\eta_{u_i} & \text{if } d_{S_n}\le d_{\mathrm{m}} \ \&\ \eta_{u_i}^{(S_n)}\ge\eta_{u_i}, \forall i\in S_n;\\ 0 & \text{otherwise;}\end{cases} \tag{21}$$

where d_{S_n} and d_{m} are the distances between the two farthest users in cluster S_n, and the maximum allowable distance among cluster members, respectively; $\eta_{u_i}^{(S_n)}$ and η_{u_i} are the energy efficiencies of user i when it joins cluster S_n and when it is not clustered, respectively. In particular, d_{m} accounts for the D2D transmission range, and can be set to guarantee that any user inside a cluster can directly reach the rest of the cluster members. The constraint on the energy efficiency guarantees that users form a cluster only if energy efficiency increases. For example, the users will not form a cluster if the the WiFi spectrum is too congested because their achievable data rate and consequently their energy efficiency reduces.

5.2 Cluster formation algorithm

The problem of finding optimal coalitions is NP-complete because it requires evaluating all possible partitions of the set of users U in the network. Obviously, the existing eNBs with limited computational resources are not able to handle an NP-complete problem involving a few tens of users. Hence, we adapt the *merge and split* algorithm to solve the coalition formation problem [22], [23]. It has been shown that merge and split can achieve near-optimal performance without imposing high computational overhead to the system [24], [25]. The merge and split rules are defined as follows: (i) merge any set $\{S_{a_1}, .., S_{a_k}\}$ into a unique coalition (i.e., cluster), if $\sum_{i=1}^{k} \nu(S_{a_i}) < \nu\left(\cup_{i=1}^{k} S_{a_i}\right)$; (ii) if the previous inequality does not hold for a coalition that can be described as $\cup_{i=1}^{k} S_{a_i}$, then split it into its components. Refer to [22] for the proof of convergence and D_{hp}-stablity of this approach. In a real implementation, merge and split algorithm runs at the eNB. The eNB notifies the users of the decided coalition formation. This notification triggers the cluster formation and association process as specified in Section 6.

5.3 Payoff allocation

The *payoff* of a cluster member is defined as the amount of throughput which it receives from the total cluster throughput. Formally, let $G \in S$ be a cluster of size $|G|$, and $\bar{x} = \{x_1, \dots, x_{|G|}\}$ the payoff vector of members of G. A payoff vector is called *cost efficient* if $\sum_{i\in G} x_i = \nu(G)$ [26]. Of course, we are only interested in cost efficient payoff vectors. Here, we chose to compare three mechanisms proposed in the literature, namely equal share, weighted share [26], and Shapley [22]. These mechanisms allow us to illustrate how payoff allocation can impact clustering decisions made by the users.

Equal share. Here, the clustering gain is equally divided among members. The cost efficient payoff distribution with this method is as follows:

$$x_i = \frac{\nu(G) - \sum_{j\in G}\nu(\{j\})}{|G|} + \nu(\{i\}), \quad i \in G. \tag{22}$$

Weighted share. Here, the payoff distribution is computed based on positive weights ω_i:

$$x_i{=}\frac{\omega_i}{\sum\limits_{j\in G}\omega_j}\cdot[\nu(G){-}\sum_{j\in G}\nu(\{j\})]{+}\nu(\{i\}),\ i\in G. \tag{23}$$

Shapley share. This is an alternative payoff distribution method that accounts for marginal contribution of each cluster member. Shapley is known to maintain good fairness while considering the contribution of the users in the cluster [22]. The Shapley value of user i in cluster G is computed as follows:

$$x_i{=}\sum_{S\subseteq G\setminus\{i\}}\frac{|S|!\,(|G|-|S|-1)!}{|G|!}\left[\nu\left(S\cup\{i\}\right)-\nu\left(S\right)\right]. \tag{24}$$

As shown in [6], the clustering gain is mainly due to the presence of *good* users, whereas the channel state probability distribution of a cluster does not

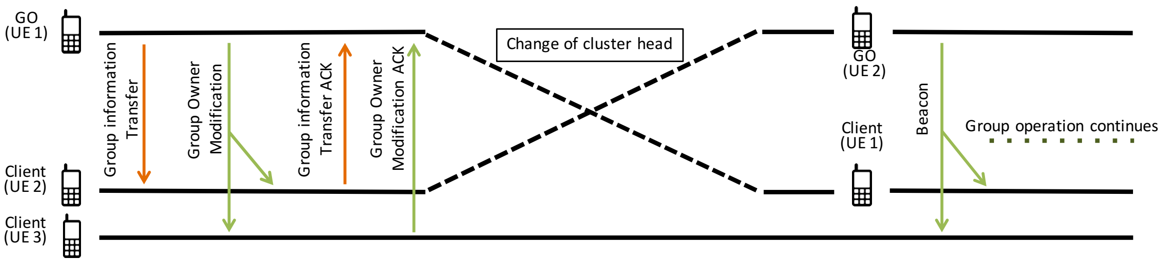

GO
(UE 1)
Client
(UE 2)
Client
(UE 3)
Group information Transfer
Group Owner Modification
Group information Transfer ACK
Group Owner Modification ACK
Change of cluster head
GO
(UE 2)
Client
(UE 1)
Beacon
Group operation continues

Fig. 2: Group ownership transfer in WiFi Direct between UE 1 and UE 2.

dramatically improve with the addition of a *poor* user (see Figure 2 in [6]). Hence, equal share may not strongly motivate *good* users to cluster with *poor* users. In contrast, by adjusting ω_i in Eq. (23), we can ensure that users with better channel quality are better incentivized. Specifically, in our numerical simulation, we use values of ω_i equal to the user's throughput achieved without clustering. Note that weighted share better motivates good users to join clusters but it may not achieve a fair payoff distribution (as Shapley share) and tuning ω_i for a complete fair payoff distribution can be challenging in real implementation. ω_i is not needed in Shapley because it is designed to distribute the payoffs based on the marginal contribution of each user. Moreover, Shapley ensures that all clustered users receive at least what they would have received without clustering. Therefore we do not need to add $\nu\{i\}$ in Eq. (24).

So far we evaluated our proposal analytically, however, the question remains: *Is it possible to implement a network-assisted opportunistic D2D system in real world with current cellular and WLAN technologies?* We answer this question in the next section.

6 Implementation of D2D clustering using WiFi Direct in LTE cells

In this section, we first propose a network-assisted protocol and position it with respect to the existing architecture of LTE-A and WiFi Direct. Afterwards, we use SDR to prototype our proposed D2D communication scheme. Specifically, this section shows how to adapt LTE and WiFi Direct to support our proposed D2D clustering scheme with minimal modification. We show how clusters form in WiFi Direct, register to LTE, obtain LTE connectivity and how the corresponding protocol stack for such a system looks like. In addition, other important procedures such as feedbacks, scheduling, security, and etc., are elaborated. Here, we refer to clusters as groups in the cluster formation procedure in order to have the coherent terminology with the WiFi Direct specification.

6.1 Cluster formation (WiFi Direct)

In our proposal, the first step is to form a cluster among users (LTE UEs) which are willing to use D2D communication. The cluster formation procedure is mostly coherent with that defined in the WiFi Direct specification [8]. The major changes to the existing specification are: (i) the users/cluster heads also announce their preferred payoff distribution method in the *Probe Requests*; (ii) the group ownership is transferable; and (iii) the cluster head receives the LTE ID from its client and shares this information in the form of a forwarding table that contains the LTE and WiFi Direct IDs of all members. We do not elaborate on *Search and discovery* and *Group ownership negotiation* as they are compliant with WiFi Direct specification (refer to [7] for more details). We briefly explain the remaining steps in the following.

LTE-WiFi mapping. Each group client sends an *LTE ID Notification* message to the Group Owner (GO) that contains its LTE identity. Then, the GO broadcasts the *WiFi-LTE ID Association Table* that includes LTE and WiFi Direct IDs of all cluster members.[1] This message can also include other group settings that are useful to quickly switch the GO when needed.

GO transfer. In WiFi Direct, the group ownership cannot be transferred. However, our proposal requires the GO to change dynamically. A GO transfer occurs when the eNB detects that another cluster member has a better cellular channel quality than the current GO (for details, refer to Section 6.4). We define two messages to enable GO transfer in WiFi Direct, as shown in Fig. 2. First, the GO sends the *Group Information Transfer* message to the provisioned GO. This message contains the updated list of members and their power saving parameters. Second, the GO sends the *GO Modification* broadcast message. Each group client should individually acknowledge this message before the GO transfer is completed.

6.2 Cluster registration in LTE

Once a cluster is formed via WiFi Direct, it registers at the LTE network. Fig. 3 shows the cluster registration procedure with the D2D-enabling modifications reported in red. This procedure consists of two phases: (i) cluster notification; and (ii) cluster verification.

Cluster notification. The cluster formation is reported to the eNB via *Cluster RRC Connection Management* message with *Request Cause* set to *connection*

[1] In our proposal, the cluster members should share their SAE-Temporary Mobile Subscriber Identity (S-TMSI) and Cell Radio Network Temporary Identifier (C-RNTI) with other cluster memebers.

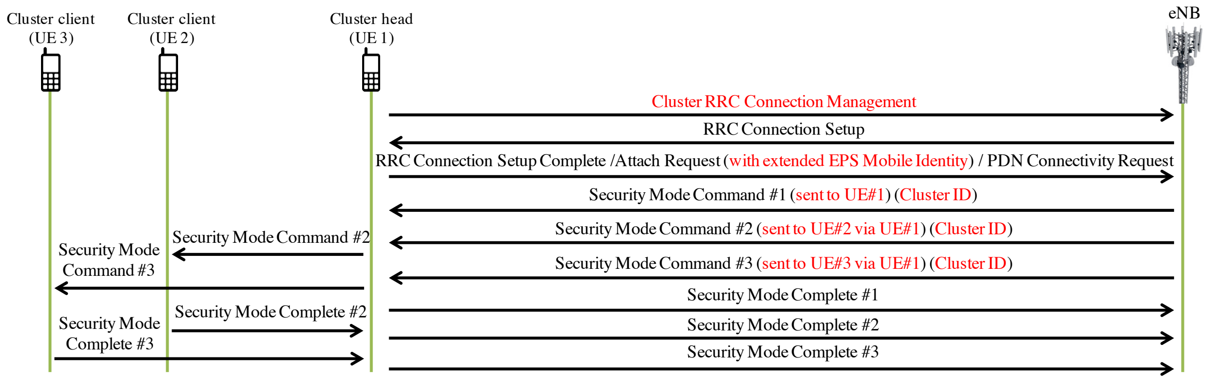

Cluster client
(UE 3)
Cluster client
(UE 2)
Cluster head
(UE 1)
eNB
Cluster RRC Connection Management
RRC Connection Setup
RRC Connection Setup Complete /Attach Request (with extended EPS Mobile Identity) / PDN Connectivity Request
Security Mode Command #1 (sent to UE#1) (Cluster ID)
Security Mode Command #2 (sent to UE#2 via UE#1) (Cluster ID)
Security Mode Command #3 (sent to UE#3 via UE#1) (Cluster ID)
Security Mode Complete #1
Security Mode Complete #2
Security Mode Complete #3
Security Mode Command #2
Security Mode Command #3
Security Mode Complete #2
Security Mode Complete #3

Fig. 3: Cluster registration procedure in LTE.

initiation. This message also contains information such as Identity of the members and their desired payoff allocation method (see Table 1). The eNB responds to the cluster notification with the *RRC Connection Setup* message. Next, the cluster head sends the *RRC Connection Setup Complete* to finish the RRC setup.

TABLE 1: Contents of Cluster RRC Connection Management

Information Elements	
Cluster Identity	Assigned by eNB
Cluster Head Identity	S-TMSI
Clients' Identities (All members are included for initiation. Otherwise, only departing/arriving member(s) are listed.)	S-TMSI of the Clients
Request Cause	CHOICE
	Connection Initiation
	Arrival
	Departure
Dedicated NAS Information (Attach Request)	

Cluster verification. Once the RRC connection is established, the eNB sends a *Security Mode Command* message to each cluster member via the cluster head. We propose to include the *Intent* value (i.e., average CQI) of each member in this message. Since the eNB knows the real CQIs, each member can verify the correctness of the values reported by others. If an anomaly is detected the member can send a negative response and leave the group. The clients send their response to the cluster head over WiFi and the cluster head forwards them to the eNB. By forcing the security verification to pass through the cluster head, the eNB ensures that all cluster clients are already members of the cluster over WiFi. This step is very important in terms of security because it ensures that any misreported value is detected.

In addition to the above, procedures such as bearer establishment and mobility should be considered in a real implementation. In our proposal, cluster heads use cluster specific bearers. The difference between cluster bearer and UE bearer is in resource provisioning. The allocated resources for a cluster bearer is equivalent to the total resources allocated to all cluster members. For more details refer to Section 5 in [7].

6.3 Data Plan Operation

Fig. 4 illustrates the adaptation of LTE and WiFi Direct data protocol stacks to our proposal. We choose to bridge, at the cluster head, the WiFi Direct MAC and LTE at Packet Data Convergence Protocol (PDCP) layer for three reasons: (i) LTE packets are ciphered and integrity-protected in the PDCP layer using keys which are only known to the client and the eNB. Therefore, other UEs cannot decipher the LTE packets traversing the WiFi network; (ii) the cluster head can further process PDCP Packet Data Unit (PDU)s in RLC layer for concatenation/segmentation according to its LTE physical link quality; and (iii) the WiFi Direct MAC provides a robust and secure transmission service, and natively allows to send frames to be relayed at MAC layer. Note that in our proposal each cluster member acknowledges its own packets so that the eNB is always aware of the transmission status. This approach also minimizes the impact of cluster head change in reliable packet delivery.

Uplink. As concerns uplink transmission requests, the clients send their Scheduling Request (SR) or Buffer Status Report (BSR) to the cluster head to be forwarded to the eNB. The eNB uses Downlink Control Information (DCI) to inform UEs regarding their downlink and uplink resource allocation. Since the cluster head is the only member which is listening to the LTE channel, it receives the DCI and updates the clients with the scheduling decision made by the eNB, using an 802.11 management frame with the same subtype value used by the UEs to encapsulate SR and BSR messages in the WiFi Direct frame. For data packets, the scheduled cluster clients encapsulate the LTE PDCP PDUs in WiFi frames and send them to the cluster head. The cluster head extracts the PDCP PDUs and forwards them to the eNB in the designated slot. The cluster head transmits the packets to the eNB with the client's C-RNTI address to simplify identifying the real source of the packets for the eNB.

Downlink. The eNB transmits the packets using the client's C-RNTI address but it selects the MCS according to the cluster head's channel quality. Since the cluster head is aware of scheduling plan for its

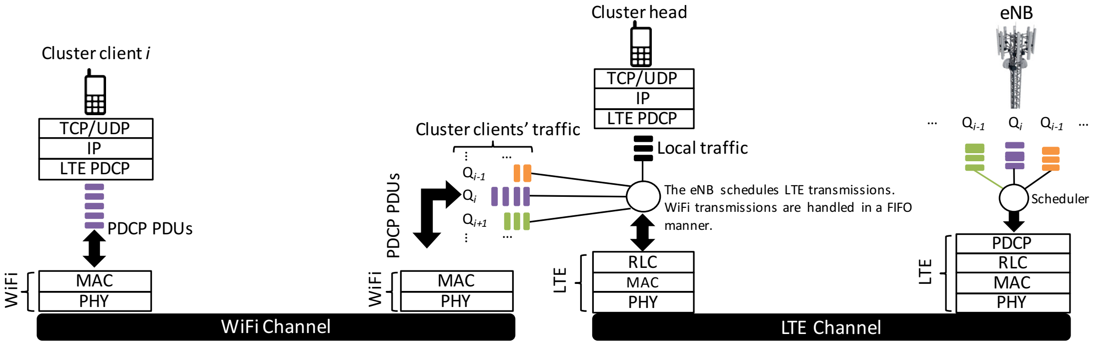


Fig. 4: Data flow between cluster client i and the eNB.

clients, it listens to the downlink channel to receive the packets belonging to all cluster members. Next, the cluster head encapsulates the PDCP PDUs in regular WiFi data frames that include the source and destination MAC addresses of cluster head and client, and the default MAC address of the eNB.

6.4 Adaptation of LTE Procedures

So far we introduced the required messaging to support our proposed architecture. Here, we elaborate on the adaptation of our proposal to other operations.

6.4.1 Channel State Information (CSI) reporting

In LTE, UEs send CSI reports to the eNB for scheduling purposes. In our proposal, the cluster head receives the CSI reports from all cluster members. This creates some flexibility to reduce the CSI-related signaling overhead, which does not exist in the standard LTE operations. We explain this in Section 6.4.6.

6.4.2 Cluster head selection

The eNB selects the cluster head among the cluster members based on the reported CSIs. We propose to add an extra field to the DCI so that the eNB can transmit the C-RNTI of the new cluster head to the current cluster head, which can trigger the GO transfer procedure. The cluster head selection interval is implementation-specific and it is constrained by the delay of LTE network and group ownership transfer in WiFi Direct. This interval introduces a trade-off between signaling overhead and opportunistic gain. On one hand the opportunistic gain is maximized when the interval is set to the frame length (shortest possible interval). On the other hand, per-frame cluster head selection requires higher signaling overhead.

6.4.3 Scheduling

The existing LTE schedulers can be adapted to support our proposal with a minor modification. In LTE, the eNB selects the physical layer parameters based on the CSI of the scheduled UE. However, we require the eNB to select physical layer parameters according to the CSI of the cluster head so that the cluster head can decode the packets and forward them to the clients. Note that the eNB still uses the C-RNTI of the client in the DCI so that the cluster head is aware of its transceiving schedule in uplink and downlink. This also eliminates the need for an uplink intra-cluster scheduler in the cluster head.

6.4.4 Security

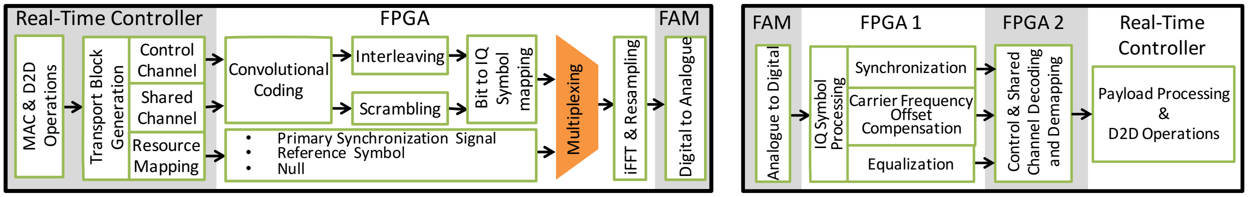
As mentioned, our proposal does not introduce any new security threats to the existing LTE architecture because the LTE packets are ciphered and integrity-protected before forwarding. We also propose to send *Security Mode Command* through the cluster head, so that the cluster head cannot exploit the resources of a UE that is not in the cluster. The only possible attack is a malicious cluster head that drops packets of its clients. The eNB can detect such behavior by tracking acknowledgements and act accordingly.

6.4.5 Policy control and billing

Since the cluster head is in charge of the LTE transmissions of its clients, it is important to ensure that the cluster head is not billed for the clients' traffic. The policy control and charging of LTE is done via Policy and Charging Enforcement Function (PCEF) which charges the UEs based on their IP address. Since each cluster member is given a separate IP address, our proposal does not cause any problem in the existing billing method. Interestingly, our scheme even allows the network to identify and offer rewards (e.g., discounts and extra data quotas) to relays for their contribution to the network welfare. It is also important to ensure that members do not *abuse* each other's resources. However, since the eNB schedules members individually, utilizing other cluster members' resources is not a concern. In case a malicious cluster head transmits its own packets on a slot allocated to another member, the eNB discards the cluster head data because it cannot be deciphered.

6.4.6 Protocol overhead

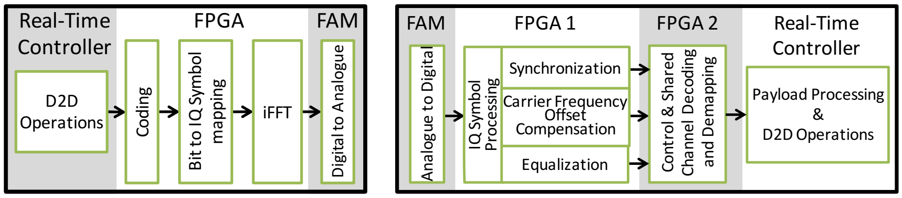
In a legacy LTE network, the users should send either low resolution CSI (i.e., wideband) or high resolution CSI (i.e., per sub-band) with very high interval to avoid flooding the control channel with feedback



(a) Architecture of the eNB

(b) Architecture of the UE

Fig. 5: The SDR architecture of the LTE interface.



(a) Architecture of the WiFi transmitter

(b) Architecture of the WiFi receiver

Fig. 6: The SDR architecture of the WiFi interface.

messages. As mentioned, our proposal can reduce the CSI-related signaling overhead. For example, the cluster clients can report the CSI over all sub-bands to the cluster head over WiFi. Then the cluster head filters these reports and sends the list of top candidates on each sub-band to the eNB. Alternatively, the cluster head reports the n highest Channel Quality Indicators (CQIs) to the eNB. The value of n imposes a trade-off between opportunistic gain and spectral efficiency. Cluster formation incurs signaling overhead but that does not directly translate to higher signaling overhead. These messages are sent only when a cluster is formed or a user joins/leaves the cluster. In this work, we assume that the overhead due to cluster formation is negligible in comparison to the time users spend in the cluster. Nonetheless, both neighbor discovery and D2D connection setup procedures consume time and energy for a few seconds. Thus, our proposal is not suitable for high speed mobile scenarios.

6.5 SDR implementation

We leveraged the National Instrument's LabVIEW SDR platform, and its RF equipments to prototype an eNB with an OFDMA transmitter and two UEs with an OFDMA receiver and a WiFi transceiver. Our SDR testbed consists of three Real-Time Controllers operating on Intel Core-i7-3610QE CPUs and eight FlexRIO FPGA modules (Kintex-7 and Virtex-5). The FPGA modules are attached to RF transceivers (i.e., Front Adaptor Modules (FAM)s) that convert baseband signal to bandpass and vice versa.

Fig. 5 illustrates the building blocks of the LTE-A transmitter interface at the eNB and the LTE-A receiver at the UE. The Real-Time controller runs LTE-A MAC layer and D2D operations with microsecond resolution. FPGAs are used to execute heavy operations such as Fast Fourier Transform (FFT), inverse FFT (iFFT) with nano-second resolution. Our SDR eNB prototype uses WRR—and Round Robin (RR) as a special case—and Proportional Fair (PF) schedulers at the eNB. Note that the current testbed only support downlink OFDMA transmission and uplink LTE-A packets are transmitted over ethernet to the eNB. Fig. 6 illustrates the architecture of WiFi transceiver at the UE. The majority of the WiFi blocks are implemented in FPGA. We run D2D-related services at the Real-Time controller because they are light operations that are executed at millisecond intervals.

7 Performance Evaluation

In this section, we perform numerical simulations, packet-level simulations, and experiments to benchmark our proposed D2D schemes (CL(WRR) and CL(MR)) against RR and PF schedulers in an FDD LTE-A system, whose capacity is 80.64 Mbps achieved by using a 20 MHz band, and neglecting LTE overheads (which would reduce the capacity to $\sim$ 75 Mbps). We consider a fully backlogged system. The numerical simulations are based on the results obtained using Mathematica software from the model presented in Sections 3 to 5. Each experiment is repeated 2000 times. The packet simulations are obtained from our home-grown Mathematica simulator that reproduces MAC (i.e., resource allocation and scheduling). The packet simulator allows us to measure performance figure which were not available in our numerical simulator such as delay and packet delivery ratios. The duration of packet simulations is 60 seconds, and simulations are repeated with 25 different seeds. The values of power related parameters are derived from [18] and [19]. The average packet size is $L_p = 1500$ B and average WiFi rate $R_{wifi} = 48$ Mbps. We also assume a Reyleigh fading channel with the mean value varying according to the channel quality of the user and a Poisson packet arrival rate. We use the random walk model for the

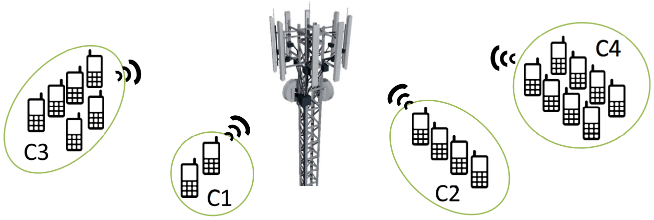


Fig. 7: Evaluation topology for static clusters.

mobile scenarios (speed between 0 and 5 km/h). The cluster formation is based on the merge and split algorithm defined in Section 5 and the maximum diameter of a cluster d_{m} is 100 m.

We assume that mobile users belong to one of three predefined SNR *classes*, which correspond to *poor*, *average*, and *good* mean SNR. The designated SNR for different classes are chosen so that the mean achievable rates for *poor*, *average*, and *good* users are 20%, 50%, and 80% of the maximum transmission rate achievable in the system, respectively. With the thresholds and MCS values adapted from [27], the designated SNR values are 7 dB, 16 dB, and 23 dB, respectively for *poor*, *average*, and *good* users. The use of non-homogeneous channel qualities enables evaluation of long-term system fairness under different (opportunistic) scheduling methods. The results include average, 25^{th} and 75^{th} percentiles of the achieved performance figures. The payoff allocation method is equal share (Eq. (22)) unless otherwise specified.

7.1 Performance of Static Clusters

This subsection provides a preliminary evaluation of the achievable throughput, fairness, and energy efficiency. For the sake of clarity, we consider a static scenario, formed by users with heterogeneous average SNR (see Fig. 7). In this scenario, clusters C1, C2, C3, and C4 have 2, 4, 6, and 8 users, respectively. In each experiment, the SNR class of each user is chosen as *poor*, *average*, or *good* with the same probability. Although the number of users is typically higher in a reality, this scenario is intended as a toy example that sheds light on potentials of the proposed schemes.

Fig. 8 illustrates the average user performance under different schedulers. Fig. 8(a) shows that users receive the lowest throughput under RR because they are scheduled irrespective of their channel quality. Instead, PF has remarkably better performance in terms of throughput, due to its opportunistic nature. Nevertheless, both RR and PF are significantly outperformed by D2D-clustering schemes in terms of throughput and energy efficiency. Interestingly, D2D-clustering schemes result in better energy efficiency than PF, although the users should maintain the WiFi interface active, in addition to the cellular interface. This stems from the higher throughput gain achieved by D2D-clusters and the insignificance of WiFi power consumption in comparison with LTE. Since in D2D cluster users with better channel quality are more active than those with poor channel quality, we illustrate the per-SNR class user throughput and user energy efficiency in Figs. 8(b) and 8(c). In terms of throughput, all classes of users enjoy higher throughput than RR and PF. D2D-clustering schemes also outperform RR and PF in terms of energy efficiency with the exception of CL(WRR) in which the *good* users can obtain higher energy efficiency under PF scheduler. Recall that in this scenario the clusters are fixed and users do not decide on the cluster formation. Thus, the *good* users may be forced to form a cluster with low throughput gain that leads to lower energy efficiency. This observation highlights the importance of cluster formation strategies in Section 5. Between D2D clustering schemes, CL(MR) has higher throughput and energy efficiency performance due to more aggressive opportunistic cluster selection scheme.

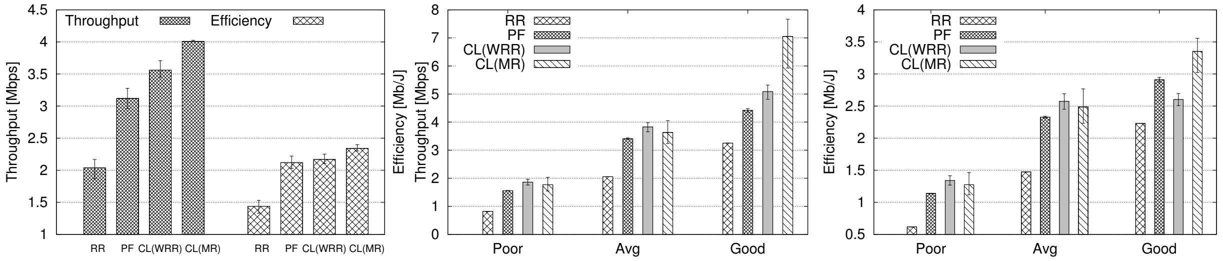
Fig. 9(a) shows the impact of size on the cluster throughput. For comparison, we also report the aggregate throughput achieved by cluster members if they were scheduled according to RR or PF. Hence, results of RR and PF scale linearly with the cluster size. Similarly, CL(WRR) shows linearity, while the high variability of results for CL(MR) does not allow us to confirm or reject the hypothesis that CL(MR) scales linearly. This behavior is due to the fact that CL(MR), differently from CL(WRR), does not guarantee a minimum airtime to any cluster, so that clusters not including *good* user receive little throughput.

Fig. 9(b) sheds light on the aggregate throughput performance. The figure reports results for three subscenarios with varying SNR class distribution. SC1 with the 60% *poor*, 30% *average* and 10% *good* users represents a cell with more low channel quality users. In SC2, we have equal distribution of different SNR classes (i.e., $33.\overline{3}\%$). Finally, SC3 represents a cell with more high channel quality users, i.e., with 10% *poor*, 30% *average* and 60% *good* users. The figure also reports the upper bound for the downlink throughput. We observe that RR and PF are outperformed by CL(WRR) and CL(MR). CL(MR) practically hits the upper bound, while the worst case for CL(WRR), i.e., when the number of *poor* users is predominant, outperforms RR and PF under their best performance.

So far, CL(MR) outperforms all other schedulers. However, considering fairness, CL(MR) is always the most unfair, especially when more *poor* users are present, while CL(WRR) performs like PF in terms of fairness, see Fig. 9(c).

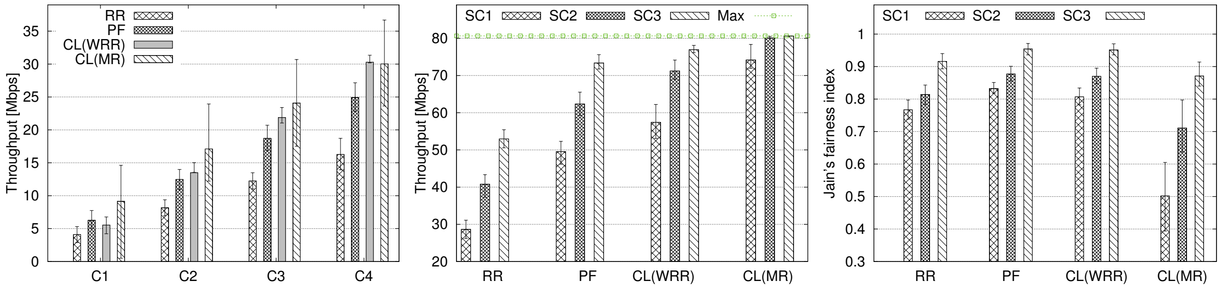
7.2 Packet simulation with static clusters

The previous scenario studied the network performance in saturation (i.e., fully backlogged). In order to better analyze the impact of clustering, we evaluate the same scenario (see Fig. 7) in a non-saturated network (i.e., 50 Mbps) using our home-grown LTE



(a) Throughput and energy efficiency (b) Throughput (c) Energy efficiency

Fig. 8: Average per-user and per-class throughput and energy efficiency performance.



(a) Per-cluster throughput (b) Aggregate cell throughput (c) Jain's fairness indices

Fig. 9: Per-cluster and aggregate performance.

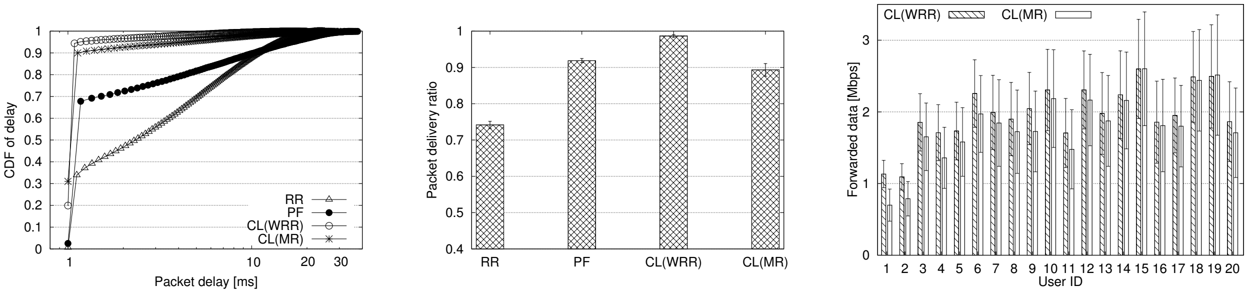
simulator written in Mathematica software. Here, we focus on the delay in the LTE cell and the load offered to the WiFi network, which provides us with better insight on the practicality of our scheme. The average SNR of users is selected randomly with a uniform distribution between 7 dB to 23 dB. The instantaneous channel quality of the users follows a Rayleigh distribution. Users have heterogenous Poisson packet arrivals with the total load of 50 Mbps that allows us to validate the benefits of our D2D-assisted scheme when the network load is below saturation.

Fig. 10(a) shows the delay CDF of the delivered packets. Here, we only account for the packet delivery time from the eNB to the UE, whereas the time to receive the ACK is not counted. The figure shows that our proposed schemes maintain a 1 ms delay with 90% probability while RR and PF require 10 ms to reach this threshold. The delay performance of a scheme is mainly affected by the achievable throughput and its prioritizing policy. If the achievable throughput is low, the packet waiting time increases which results in higher delays. On the other hand, a scheme that highly prioritizes certain class of users (e.g., MR) can potentially increase the queue size of the other classes of users. The latter is the reason why CL(MR) has lower delay performance than CL(WRR). D2D-clustering schemes can guarantee delays lower than 10 ms with 97% probability or higher, leaving at least 40 ms of delay budget for WiFi transmissions. Note that the WiFi delay budget is enough to support real time applications.

In Fig. 10(b), we can observe that CL(WRR) outperforms other schemes in terms of successful packet delivery ratio. The outstanding results of CL(WRR) are because of the throughput gain from D2D-clustering and fair resource allocations which avoids starvation of low priority users. On the other hand, CL(MR) and PF have comparable performance, although CL(MR) can potentially achieve higher throughput than PF. CL(MR) cannot outperform PF because of its greedy behavior in prioritizing high channel quality users.

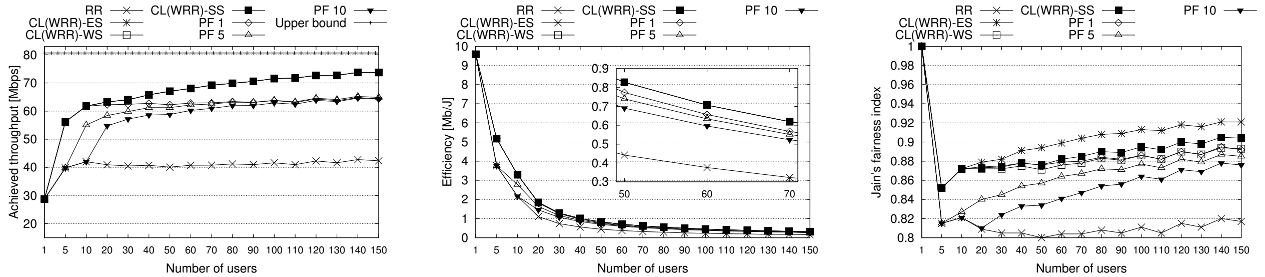
Fig. 10(c) illustrates the load offered to the WiFi network under CL(WRR) and CL(MR). This figure confirms that the WiFi Direct is not a bottleneck in our proposal. The figure also shows that the maximum load offered to C1 (users 1 and 2), C2 (users 3 to 6), C3 (users 7 to 12), and C4 (users 13 to 20) are less than 4 Mbps, 12 Mbps, 20 Mbps, and 31 Mbps, respectively. The load variation for different users depends on the channel quality. For instance, users 12 and 15 relay more traffic because they have higher average SNR w.r.t. the other users. In all cases, the traffic of each cluster is well below typical WiFi capacities.

To summarize the results for static scenarios, we have observed that the clustering proposal not only increases the throughput and the energy efficiency, but also increases the fairness. In particular, CL(WRR) achieves similar throughput and energy efficiency results as CL(MR), but it is much fairer. Thus, the advantage of using CL(WRR) is fourfold: (i) the possibility to gain higher throughput than legacy RR and PF; (ii) allowing each cluster to exploit the clustering gain proportionally to its size; (iii) near perfect fairness among users; (iv) higher energy efficiency than RR and PF; (v) the best delay performance compared with other schemes; (vi) high packet delivery ratio (almost 100%). Since numerical and packet simulations showed that CL(MR) may lead to poor fairness and



(a) Delay CDF of delivered packets.
(b) Packet delivery ratios.
(c) WiFi Direct load per user.

Fig. 10: Delay CDF, packet delivery ratio, and per-user WiFi direct loads.



(a) Aggregate throughput
(b) Average per-user efficiency
(c) Fairness

Fig. 11: Throughput, efficiency and fairness under different scheduling mechanisms.

packet delivery ratio, we will focus on CL(WRR) in the rest of the evaluation.

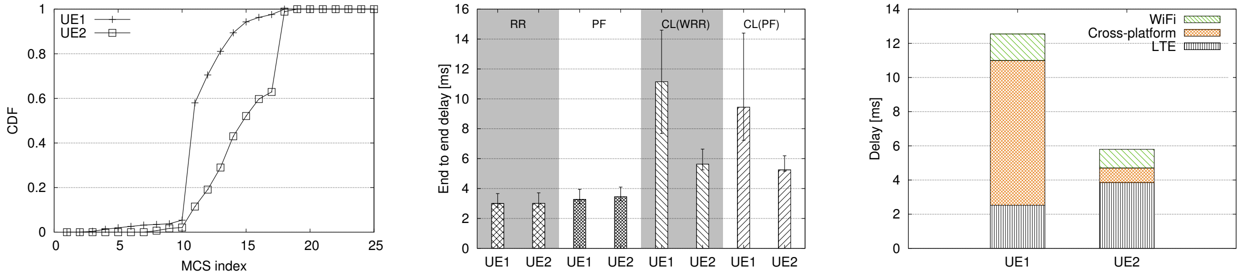
7.3 Performance of Dynamic Clusters

In order to evaluate our proposal in a more realistic setup, we simulate a network with variable number of users (from 1 to 100) with varying SNR, randomly placed in a circular-shaped cell with 500 m diameter. The SNR class of a user is selected at random with a probability distribution that changes according to the distance from the eNB. The users move with an average pedestrian speed between 0 and 5 km/h.

Fig. 11 illustrates the performance metrics for different user population sizes. The figure shows results achieved with RR, PF, CL(WRR) with equal share (i.e., CL(WRR)-ES), CL(WRR) with weighted share (i.e., CL(WRR)-WS), and CL(WRR) with Shapley share (i.e., CL(WRR)-SS). Moreover, we report results for PF when $m \geq 1$ users are scheduled per frame ("PF m" in the figure, $m \in \{1, 5, 10\}$). We report this comparison since user-based schedulers allocate multiple users per frame, and it is indeed common to schedule tens of users per scheduling interval, even with opportunistic schedulers. However, RR and CL(WRR) are not affected by the number of users scheduled per frame, due to the assumption that user's channels are independent and stationary.

In Fig. 11(a), we can observe that the clustering gain rises with the number of users in the system, and as soon as about 30 users are present, CL(WRR) achieves the highest aggregate network throughput, which approaches the upper bound with a reasonable cell population size of 100 users. Since CL(WRR) variations only redistribute the intra-cluster resources, they do not differ over the aggregate network throughput. The throughput of PF reduces significantly as the number of scheduled users per frame increases. However, all PF curves converge, for high number of users, to a value well below the throughput of CL(WRR). In Fig. 11(b), we can observe that the energy efficiency of CL(WRR) is the best. Overall, the energy efficiency decreases with the number of users, due to the fact that each additional user incurs a minimum cost due to activating the WiFi/LTE interfaces, while the cell capacity is upper bounded. However, e.g., with 70 users, the efficiency of CL(WRR) is higher than RR and PF 5 by $\sim 101\%$ and $\sim 13\%$, respectively. Recall that in Subsection 7.1 we observed that *good* users may obtain lower energy efficiency than PF. Here, the cluster formation is only allowed if all cluster members can achieve higher energy efficiency than what they can achieve under PF. This reduces the throughput gain of D2D schemes. As regards fairness, Fig. 11(c) shows that CL(WRR)-ES provides the highest fairness level followed by CL(WRR)-SS, while CL(WRR)-WS achieves results comparable to the best results achieved by PF. The ES method has better fairness due to equal resource distribution. SS method outperforms WS because SS distributes the resource based on the contribution of each user to the total revenue. The fairness improvement due to clustering with respect to RR and PF 5 or PF 10 is remarkable.

We also investigate the impact of payoff distribution methods using our simulator, the results are not shown here due to lack of space. Nonetheless, our investigation indicates that payoff distribution methods

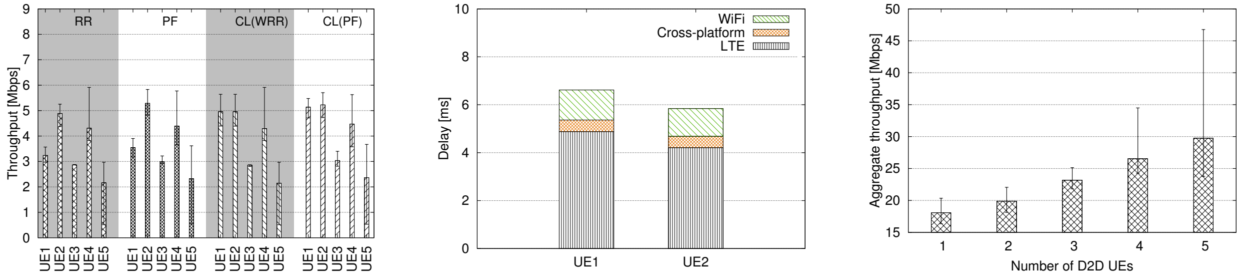


(a) CDF of MCS for each UE.

(b) Delay under different schemes.

(c) Delay at different parts of the link.

Fig. 12: The MCS distribution and delay results for two-user scenario.



(a) Per-UE throughput.

(b) Delay at different parts of the link.

(c) Aggregate throughput.

Fig. 13: Per user throughput, delay, and aggregate throughput.

act very similarly in non-saturated scenarios because users receive the required resources. When approaching saturation, the WS results in higher throughput and delay variations compared to ES and SS. Considering the minor impact of payoff distribution method in non-saturated networks, we can use simple payoff distribution methods such as ES instead of SS, that adds on to the practicality of D2D-clustering.

7.4 Experimental results

Although numerical and packet simulations provide good insights on the expected system performance, experimental evaluation is still the best method to verify the real performance. Thus, we prototyped our proposal in an SDR testbed as described in Section 6.5. Our eNB can support five UEs. However, we only have the available hardware for two fully operational UEs. Hence, we first show the result for a scenario with two UEs. Next we increase the number of UEs to five by using three *virtual* UEs that can send pre-recorded CQI, obtained from android smartphones, to the eNB. Although our virtual UEs cannot decode our eNB transmissions, we can still evaluate our proposal based on eNB transmission and an estimate of decoding performance at virtual UEs, in presence of actual cellular channel variation.

1. Two-user scenario. Here, we focus on the performance of RR and PF schedulers and D2D clustering schemes. To quantify channel quality variations in our experiments, we show the CDF of MCS for each UE in Fig. 12(a) . We can see that UE2 has better channel quality than UE1 on average. Fig. 12(b) shows the delay incurred by each UE. We observe that the additional delay overhead due to D2D clustering (i.e., in CL(WRR) and CL(PF)) is negligible for UE2 while it goes up to 7ms for UE1. Since the cluster head selection is done opportunistically based on the channel quality of the UEs, Most of UE1's packets are relayed through UE2 over the WiFi interface because UE1 has worse channel quality than UE2 on average. Thus, UE1 experiences higher delay with D2D clustering schemes. Fig. 12(c) provides a more detailed delay analysis. The figure reports three delay values: (i) LTE: the delay from eNB's MAC to the cluster head's LTE MAC, (ii) cross-platform: the delay from cluster head's LTE MAC to its WiFi MAC, and (iii) WiFi: delay caused by WiFi transmission. We observe that the cross-platform plays an important role in the additional delay overhead of UE1. The results show that the high cross-platform delay is due to packet processing time from LTE MAC to WiFi MAC of the cluster head. Since the relay traffic volume is very high (~ 20 Mbps) in this scenario, the cross-platform delay is more significant. As we see later, the cross-platform delay reduces in scenarios with more users.

2. Five-user scenario. In this scenario, we add the virtual UEs to the system. These UEs only send their CQI to the eNB. The eNB schedules them like the real UEs and transmits their packets over the air. In Figs. 13(a) and 13(b) only UE1 and UE2 are allowed to form a cluster and the virtual UEs (i.e., UE3 to UE5) do not join any cluster. We remove this limitation in Fig. 13(c) to illustrate the potential gain of clustering. Fig. 13(a) illustrates per-UE throughput under different schemes. The figure shows that the throughput of UE1 and UE2 increases (up to 2 Mbps) with the clustering schemes. D2D clustering schemes not only increase the throughput but also improve the fairness due to throughput equalization among D2D UEs. Fig.13(b) demonstrates different components of the delay experienced by D2D UEs. We can see that the cross-platform delay is significantly reduced in comparison to our observation in Fig. 12(c). Since per-UE throughput is lower in this scenario, the cross-platform delay also reduces accordingly. Finally, we show the impact of cluster size on the aggregate throughput in Fig. 13(c). The results confirm that the gain increases as the cluster size grows. We observe that the throughput gain increases from 13% to 76% when cluster size increases from 2 to 5.

7.5 Discussion

Our experimental observation is inline with our analytical and simulation results, which emphasizes on the importance of high user cooperation in D2D communication. Of course, due to hardware limitation and computational capacity of the FPGA design, we could not verify the simulation results with higher number of users with the SDR platform. However, the performance improvement observed in the simulation tallies with that of the experiments. For example, Fig.10(a) shows that the throughput gain is about 74% with a cluster of 4 which is in agreement with the experiment result for a cluster of 5 which is 76% (see Fig. 14(c)). This high throughput gain can certainly make up for the signaling overhead incurred during cluster formation procedure or the messages exchanged when cluster composition changes.

The reliability of outband D2D communication has always been an open challenge due to unregulated nature of ISM band. While service guarantee maybe not possible over such spectrums, we can guarantee quick adaptation and recovery. Indeed, we demonstrated the feasibility of quick recovery (switching between legacy and D2D modes) via our experimental evaluation. In our experiments, we also observed that the delay overhead of clustering increases with the relay load of the cluster head (up to 10 ms for 20 Mbps). This highlights the importance of on-chip solutions for cross-platform relaying in 5G networks.

8 Conclusions

In this manuscript, we have analyzed network-assisted opportunistic D2D clustering from a theoretical and practical aspect. The theoretical results illustrated our proposed architecture significantly outperforms legacy schedulers in terms of throughput, energy efficiency, and fairness by using simple schedulers and coalitional game theory tools. We also analyzed the practicality of implementing opportunistic D2D communication in 5G cellular networks using WiFi Direct and LTE-A. Our proposed protocol proved that not only D2D-assisted cellular communication is practical, but also that they can be achieved with minimal modifications to the current infrastructure. Finally, our experimental evaluation proved our theoretical finding and revealed the importance of cluster size and its impact on throughput and delay. Although throughput gain increases with cluster size, so does the cross-platform delay with the throughput achieved in the cluster. However, our results showed that networks with small clusters achieve large gains while experiencing little delay due to relay. Our analysis still holds if mm-Wave IEEE 802.11ad is used instead of WiFi Direct. In practice, using mm-Wave should results in better performance figures since it has much higher bandwidth compared to WiFi Direct. IEEE 802.11ad and alternative 5G D2D techniques will be evaluated in our future work.

9 Acknowledgements

The authors would like to thank National Instruments for the support in LabVIEW FPGA code development. The work has been partially supported by the Madrid Regional Government through the TIGRE5-CM program (S2013/ICE-2919) and the LOEWE initiative (Hessen, Germany) within the NICER Project. This research was supported in part by the Spanish Spanish Ministry of Economy and Competitiveness under grants TEC2014-55713-R and RYC-2014-01335.

References

[1] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," IEEE Communications Surveys Tutorials, no. 99, pp. 1–1, 2014.

[2] 3GPP, "3rd generation partnership project;technical specification group services and system aspects; policy and charging control architecture (release 13)," TR 23.203 V13.4.0, 2015.

[3] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," IEEE Communications Magazine, 2014.

[4] G. H. Sim, A. Loch, A. Asadi, V. Mancuso, and J. Widmer, "5G millimeter-wave and D2D symbiosis: 60 GHz for proximity-based services," IEEE Wireless Communications Magazine, 2016.

[5] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, "Intracluster device-to-device relay algorithm with optimal resource utilization," IEEE Transactions on Vehicular Technology, 2013.

[6] A. Asadi and V. Mancuso, "On the compound impact of opportunistic scheduling and D2D communications in cellular networks," in Proc. of ACM MSWIM, 2013.

[7] ——, "Wifi direct," in Proc. of IFIP Wireless Days.

[8] Wi-Fi Alliance, "Wi-Fi peer-to-peer (P2P) technical specification v1.1." [Online]. Available: www.wi-fi.org/wi-fi-peer-peer-p2p-specification-v11

[9] G. H. Sim, T. Nitsche, and J. Widmer, "Addressing MAC layer inefficiency and deafness of ieee802.11ad millimeter wave networks using a multi-band approach," in Proc. of IEEE PIMRC, 2016.

[10] G. H. Sim, R. Li, C. Cano, D. Malone, P. Patras, and J. Widmer, "Learning from experience: Efficient decentralized scheduling for 60GHz mesh networks," in IEEE WoWMoM, 2016.

[11] J. Seppala, T. Koskela, T. Chen, and S. Hakola, "Network controlled device-to-device (D2D) and cluster multicast concept for LTE and LTE-A networks," in Proc. of IEEE WCNC, 2011.

[12] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," IEEE JSAC, 2015.

[13] S. Andreev, J. Hosek, T. Olsson, K. Johnsson, A. Pyattaev, A. Ometov, E. Olshannikova, M. Gerasimenko, P. Masek, Y. Koucheryavy, and T. Mikkonen, "A unifying perspective on proximity-based cellular-assisted mobile social networking," IEEE Communications Magazine, 2016.

[14] A. Pyattaev, J. Hosek, K. Johnsson, R. Krkos, M. Gerasimenko, P. Masek, A. Ometov, S. Andreev, J. Sedy, V. Novotny, et al., "3GPP LTE-assisted Wi-Fi-Direct: Trial implementation of live D2D technology," ETRI Journal, 2015.

[15] J. G. Choi and S. Bahk, "Cell-throughput analysis of the proportional fair scheduler in the single-cell environment," IEEE Transactions on Vehicular Technology, 2007.

[16] E. Liu and K. K. Leung, "Proportional fair scheduling: Analytical insight under Rayleigh fading environment," in Proc. of IEEE WCNC, 2008.

[17] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in Proc. of IEEE ICC, 1995.

[18] J. Huang, F. Qian, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4G LTE networks," in Proc. of ACM MobiSys, 2012.

[19] P. Serrano, A. Garcia-Saavedra, G. Bianchi, A. Banchs, and A. Azcorra, "Per-frame energy consumption in 802.11 devices and its implication on modeling and design," IEEE/ACM Transactions on Networking, 2014.

[20] "3rd generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification (release 13)."

[21] A. Gupta and P. Mohapatra, "Energy consumption and conservation in WiFi based phones: a measurement-based study," in Proc. of IEEE SECON, 2007.

[22] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, "Coalitional game theory for communication networks," IEEE Signal Processing Magazine, 2009.

[23] L. Song, D. Niyato, Z. Han, and E. Hossain, "Game-theoretic resource allocation methods for device-to-device communication," IEEE Wireless Communications, 2014.

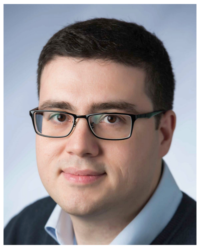
[24] W. Saad, Z. Han, T. Basar, M. Debbah, and A. Hjorungnes, "Coalition formation games for collaborative spectrum sensing," IEEE Transactions on Vehicular technology, vol. 60, no. 1, pp. 276–297, 2011.

[25] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, "Coalitional games for distributed collaborative spectrum sensing in cognitive radio networks," in IEEE INFOCOM, 2009.

[26] W. Saad, Z. Han, M. Debbah, and A. Hjorungnes, "A distributed coalition formation framework for fair user cooperation in wireless networks," IEEE Transactions on Wireless Communications, 2009.

[27] S. Sesia, I. Toufik, and M. Baker, LTE–the UMTS long term evolution: from theory to practice. Wiley, 2011.

Arash Asadi is a PostDoc researcher at the SEEMOO lab at the Technical University of Darmstadt as of March 2016. He received his PhD in Telematics Engineering from University Carlos III of Madrid (UC3M) with the highest distinction while he was affiliated with IMDEA Networks Institute. He also obtained a master's degree in Telematics Engineering from UC3M in October 2012, and another master's degree in Telecommunication Engineering from Multimedia University, Malaysia in January 2011.



Vincenzo Mancuso is Research Associate Professor at IMDEA Networks Institute. Previously, he was with University of Palermo, Rice University, and INRIA Sophia Antipolis. His research focuses on analysis, design, and experimental evaluation of protocols and architectures for wireless networks. He is currently working on analysis and optimization of power saving strategies for packet cellular networks.