Device-to-Device Communications Achieve Efficient Load Balancing in LTE-Advanced Networks

Citation:

Jiajia Liu, Yuichi Kawamoto, Hiroki Nishiyama, Nei Kato, and Naoto Kadowaki, "Device-to-Device Communications Achieve Efficient Load Balancing in LTE-Advanced Networks," IEEE Wireless Communications Magazine, vol. 21, no. 2, pp. 57-65, Apr. 2014.

URL:

http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6812292

Device-to-Device Communications Achieve Efficient Load Balancing in LTE-Advanced Networks

Jiajia Liu*, Yuichi Kawamoto*, Hiroki Nishiyama*, Nei Kato*, and Naoto Kadowaki†
*Tohoku University, Sendai, Japan
†National Institute of Information and Communications Technology (NICT), Tokyo, Japan

***Abstract*—In LTE-Advanced networks, besides the overall coverage provided by traditional macrocells, various classes of low-power nodes (LPNs, like pico eNBs, femto eNBs, and relays) can be distributed throughout the macrocells as a more targeted underlay to further enhance the area spectral efficiency, alleviate traffic hot zones, and thus improve the end user experience. Considering the limited backhaul connections within LPNs and the imbalanced traffic distribution among different cells, it is highly possible that some cells are severely congested while the adjacent cells are very lightly loaded. Therefore, it is of critical importance to achieve efficient load balancing among multi-tier cells in the LTE-Advanced networks. However, available techniques such as smart cell and biasing, although able to alleviate congestion or distribute traffic to some extent, cannot respond or adapt flexibly to the real-time traffic distributions among multi-tier cells. Towards this end, we propose in this article a device-to-device (D2D) communication based load balancing algorithm, which utilizes D2D communications as bridges to flexibly offload traffic among different tier cells and achieve efficient load balancing according to their real-time traffic distributions. Besides identifying the research issues which deserve further study, we also present numerical results to show the performance gains that can be achieved by the proposed algorithm.**

***Index Terms*—LTE-Advanced networks, device-to-device communication, load balancing, traffic offloading.**

I. Introduction

To cope with the exponential growth of mobile broadband data traffic and the unprecedented consumer demand for faster data connectivity, the Third Generation Partnership Project (3GPP) has been developing an enhanced Long-Term Evolution (LTE) radio interface called LTE-Advanced (LTE-A, also known as Release 10 and beyond), aiming to significantly enhance the current LTE and support much higher capacity and coverage, higher throughput and lower latency, higher peak rates and better user experience, etc.

In LTE-A networks, besides the overall coverage provided by traditional macrocells, various classes of low-power nodes (LPNs, like pico eNBs, femto eNBs, and relays) can be distributed throughout the macrocells as a more targeted underlay to further enhance the area spectral efficiency, alleviate traffic hot zones, and thus improve the end user experience. However, different from macro eNBs which are usually tower-mounted and equipped with high-speed backhaul connection, the underlaid LPNs may be subject to backhaul bottleneck. Specifically, as the femto eNBs are usually deployed in home and use backhaul connection such as digital subscriber line (DSL) or cable modem, clearly, a 10 MHz LTE femtocell is going to be limited by the backhual especially in the uplink [1]. Regarding the pico eNBs which are to be deployed at building corners or street poles, it may be financially prohibitive to install and maintain a high-quality backhaul connection at such locations.

Considering the limited backhaul connections within small cells and the imbalanced traffic distribution among different tiers, one can see that, for a HetNet consisting of multi-tier small cells, it is highly possible that some cells (areas) are severely congested while the adjacent cells are very lightly loaded. Consider a UE which is enjoying some online videos from the Internet (like Youtube) roams into a congested cell. How can the already congested eNB provide continuous and desirable Internet access for this new comer? What if a group of UEs? Therefore, it is of critical importance to achieve efficient load balancing among multi-tier cells in the LTE-A HetNets.

There has been a lot of researches on load balancing in different types of networks, such as Wide Area Network (WAN) [2], Wavelength Division Multiplexing (WDM) based packet networks [3], Wi-Fi networks [4], [5], backbone networks [6], mobile ad hoc networks [7], [8], satellite networks [9], mesh networks [10], etc. For cellular networks, most prior offloading techniques were based on borrowing channels from adjacent lightly-loaded cells, such as load balancing with selective borrowing [11], channel borrowing without locking [12], [13], etc. Other offloading techniques include direct retry [14], cell breathing [15], mobile-assisted call admission [16], and overlay ad hoc relays on top of cellular networks [17], [18]. It is noticed that the available schemes cannot be directly applied for traffic offloading in LTE-A HetNets, since the major target here is to balance traffic load among multi-tier cells which differ primarily in terms of physical size, maximum transmit power, etc. Furthermore, compared with previous works, D2D communication based offloading techniques have the following unique features: first, D2D communications are fully controlled by the operator including the transmit power of end users and D2D relays, transmit time slot, frequency resources, etc.; second, D2D communications use the same frequency resources as cellular transmissions for a better spectral efficiency and thus both inter- and intra-cell interference management are critical issues. While in the integrated network of cellular and ad hoc relays [17], [18],

there are only macrocells and the overlayed ad hoc relays are fixed and use dedicated ISM-band. Such scenarios are much simpler than D2D based offloading, because there is no need to address the challenging issues of relay mobility, inter- and intra-cell interference, etc.

Therefore, besides the common requirements for traffic offloading over multiple mobile terminals, D2D based offloading techniques should satisfy the following two basic requirements in LTE-A networks: on one hand, in order to avoid frequent change of offloading paths and provide satisfactory QoS for end users, it should be able to overcome various network dynamics (such as the movement of end users or D2D relays, the interference from surrounding terminals) which may easily deteriorate newly established D2D links; on the other hand, it should achieve efficient interference management via proper PRB assignment, transmit power control, etc., so as to effectively alleviate the impact of D2D communications on adjacent ongoing cellular transmissions.

Some techniques have been proposed for alleviating congestion and balancing traffic in LTE-A HetNets. One example is the idea of equipping small cells with big memory blocks and caching popular videos or other common downloads, which can be updated periodically in off-peak time period [19]. As all mobile user interactions will not have to traverse the backhaul and Internet, it can alleviate the backhaul congestion between small cells and core network to some extent, rather than the congestion over the air. That is, the air interface between UEs and eNBs can still be highly congested. Another approach is biasing which assigns small eNBs a bias value and pushes load onto small cells, by replacing the usual max-SINR association with the biased SINR [20]. Although the biasing technique with preconfigured bias values is simple yet effective, it is difficult to respond or adapt flexibly to the real-time traffic distributions among multi-tier cells. Furthermore, even the simple update of the bias value in a small cell, may affect the cell association of lots of already connected UEs in the cell and other adjacent cells, which is not favorable from the users' perspective. Therefore, how to efficiently offload traffic among different tier cells and achieve efficient load balancing according to their real-time traffic distributions remains a challenging problem.

Towards this end, in this article we show how to achieve efficient and real-time load balancing in LTE-A networks via device-to-device (D2D) communications. The rest of this article is organized as follows. In the next section, an overview of LTE-A networks and D2D communications is provided. This is followed by an application of D2D communication for direct traffic offloading. In the following section, we propose a D2D communication based algorithm for efficient loading balancing in LTE-A networks, and discuss some research issues which deserve further study. We present numerical results to show the performance gains that can be achieved by our D2D based algorithm, and finally conclude the article in the last section.

II. Overview of LTE-Advanced Networks and Device-to-Device Communications

A. LTE-Advanced Networks

As an evolution of LTE (Release 8 and 9), LTE-Advanced are proposed to meet or exceed the requirements of International Telecommunication Union (ITU) for the fourth generation (4G) cellular systems known as International Mobile Telecommunications-Advanced (IMT-A). LTE-A adopts orthogonal frequency-division multiple access (OFDMA) in the downlink (DL) and single-carrier FDMA (SC-FDMA) in the uplink (UL), along with spatial multiplexing using multi-layer multiple-input multiple output (MIMO) [21]. In LTE-A, carrier aggregation is employed to support flexible spectrum aggregation and maximum deployment bandwidths of 100 MHz. Via advanced MIMO techniques, LTE-A enables peak data rates of 1 Gbps in DL and 500 Mbps in UL, and performance gain of 1.4 to 1.6 from LTE in both capacity and cell-edge user throughput to be achieved. Other features include coordinated multipoint (CoMP) transmission in DL and CoMP reception in UL, enhanced intercell interference coordination (eICIC), self-optimizing networks (SON), multimedia broadcast/multicast service (MBMS), etc [22].

Among the techniques to achieve enhancements of capacity, coverage, and spectral efficiency, an important new development is the deployment of heterogeneous network, which enables various LPNs to be distributed across a macrocell network as an underlay. The LPNs includes pico eNBs (i.e., BS for Hotzone cells with typical transmit power of 33 dBm), home eNBs (i.e., BS for femtocells with typical transmit power of 20 dBm), relays, and remote radio heads (RRHs, also called distributed antenna systems (DAS)). Basically, pico eNBs, relays, and RRHs are placed indoors or outdoors and are open to all mobile UEs, and home eNBs are placed indoors and can be configured either as open access or for only closed subscriber group. Furthermore, pico eNBs have typically planned deployment by the operators, while home eNBs are usually randomly deployed by customers [23].

Besides the above HetNets, another technology component in LTE-A is D2D communication underlaying a cellular infrastructure, as to be introduced in the next section.

B. Device-to-Device Communications

Generally speaking, a UE pair moving within close proximity to each other in a LTE-A network, can establish a D2D link with or without the assistance of the serving eNB(s). Furthermore, the D2D communication can be operated in the unlicensed spectrum band, such as the industrial, scientific and medical (ISM) radio bands, or in the same licensed band as cellular UEs. For the case that mobile UEs conduct D2D communications utilizing the unlicensed band, like the WLANs band (i.e., 2.4 GHz and 5 GHz for IEEE 802.11/WiFi), and the IEEE 802.16 WiMAX band (2.5 GHz, 3.5 GHz, 5.8 GHz, etc.), it is very similar to that in classic ad hoc mobile networks if the cellular operators (i.e., the eNBs) are not involved in the communication process. In order to achieve efficient spatial reuse of the precious wireless band resources, in this article we restrict our interests to the case of operator assisted

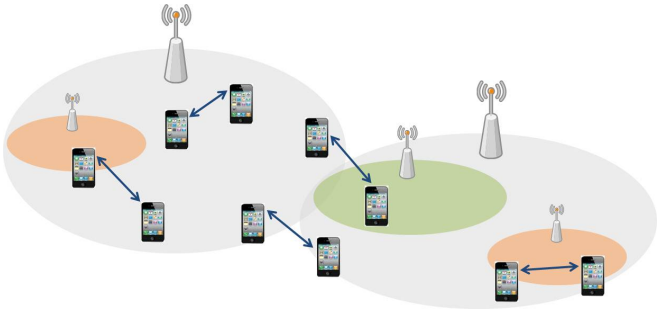

Fig. 1. Direct traffic offloading via D2D communications, where the transmitter and the receiver in each D2D link directly share the contents without routing the traffic via the serving eNBs or core network.

D2D communications and provide for D2D UE pairs access to the same licensed band as LTE-A cellular UEs. That is, the LTE-A eNBs will assist D2D pairs in the operations of peer discovery, link establishment, physical resource blocks (PRBs) assignment, transmit power control, intra-cell and inter-cell interference management, etc., and thus enable the D2D communications as a controlled or constrained underlay to the existing LTE-A networks. Note that it is different from the case of FlashLinQ [24], which works on dedicated licensed band for D2D communications, causing no interference to cellular connections.

Specifically, we focus on the application of D2D communication in traffic offloading, and illustrate its great potential to achieve efficient load balancing in future LTE-A networks, as to be detailed in ensuing sections. Note that in LTE-A networks, multi-hop D2D communications require much complicated procedures for resource allocation and interference management at the operator side, and also incur non-negligible signal overheads among D2D UEs and serving eNBs. Therefore, we focus on only the case of one-hop D2D communications in this article, and show how to apply it to balance the network traffic, increase system throughput, and enhance spectral efficiency in LTE-A networks.

III. Direct Traffic Offloading via D2D Communications

Direct traffic offloading via D2D communications corresponds to the case that two mobile UEs share the contents directly with each other without routing the data via the serving eNBs or core network. As shown in Fig. 1, the D2D links can be established between mobile UEs located within the same macrocell, picocell, and femtocell, or between mobile UEs served by different eNBs.

In such a scenario, the D2D session can be either initiated by the mobile UE (i.e., the transmitter), or initiated by the operator (i.e., the packet data network (PDN) gateway). In the former case, the transmitter explicitly requests to setup a D2D session by selecting a specific Session Initiation Protocol (SIP) uniform resource indicator (URI) extended with a special D2D keyword, such as .direct, .local, etc., which notifies the System Architecture Evolution (SAE) network the preference for a local D2D connection. While in the latter case, the PDN gateway which actually routes IP packets to the eNBs serving the destination UE, is able to detect potential D2D traffic after processing the IP headers of the data packets. If the data is to be destined to the same eNB or the eNBs serving neighboring cells, the eNB(s) notify mobile UEs the potential of D2D communication and request for measurement of the D2D link quality. One can find some suggested operation procedures in [25].

The direct D2D offloading has been intensively discussed in previous works [25]–[27]. One of its typical applications is the mobile P2P-style content sharing, where each mobile UE acts as a mobile P2P server, installs a big memory block and stores inside lots of popular contents, and registers its available contents to the operators. If the data requested by a mobile UE happens to be registered by a nearby UE, the operator can then forward the UE's request to the nearby UE and setup for them a D2D session, thus offloading the traffic from the serving eNB and core network.

The application of above direct D2D traffic offloading, although able to offload data from the serving eNB, is subject to the limitation that the receiver of the outgoing data (or the holder of the requested data) is just in close proximity to the transmitter. Furthermore, it cannot detour traffic from congested macro eNBs or pico eNBs to adjacent lightly loaded (uncongested) eNBs. Toward this end, in the next section, we propose a D2D communication based algorithm to achieve efficient load balancing among different tier cells in LTE-A networks.

IV. Load Balancing Among Multi-Tier Cells via D2D Communications

A. A D2D Communication Based Algorithm for Efficient Load Balancing in LTE-A Networks

We present a load balancing algorithm which is able to take advantage of D2D communications to efficiently detour traffic from congested macrocells, picocells or femtocells to adjacent uncongested cells. Without loss of generality, we focus on a three-tier HetNet consisting of macrocells, picocells and femtocells (with open access), and take a congested macrocell as an example to illustrate our D2D communication based load balancing algorithm. For the case of a congested picocell or femtocell in other multi-tier HetNets, the algorithm can be followed similarly. The algorithm has in total four steps, and it will proceed to the next step only after it fails in the current step.

D2D Communication Based Load Balancing Algorithm: suppose in the congested macrocell, an associated (or attached) mobile UE is requesting for access to Internet. As the serving macro eNB is already fully loaded, there is no available PRB for the requesting UE within the macrocell. The macro eNB operates as follows:

Step 1: the macro eNB tries to offload the requesting UE to an uncongested cell adjacent to the UE via a D2D relay. Specifically, the macro eNB first obtains the location of the requesting UE, and checks whether there exist any uncongested macro eNBs, pico eNBs, or femto eNBs adjacent to the requesting UE. If so, the macro eNB multicasts the location of the requesting UE to all uncongested candidate

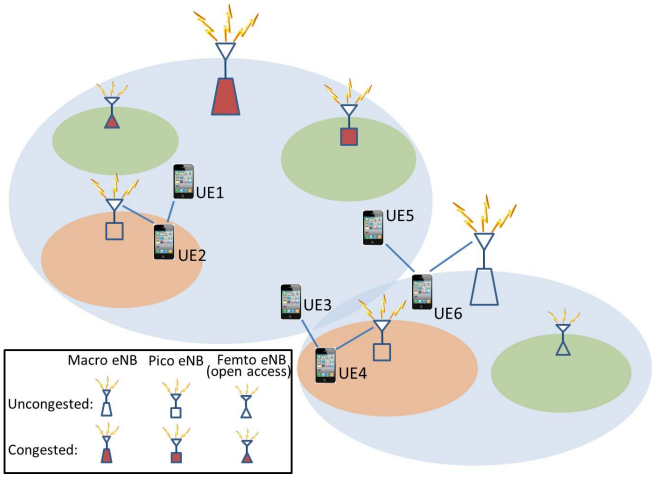

Fig. 2. Illustration of Step 1 in the D2D communication based load balancing algorithm. The congested macro eNB offloads the data traffic of the requesting UE 1, UE 3, and UE 5 to the adjacent uncongested picocells or macrocells via the D2D relays UE 2, UE 4, and UE 6, respectively.

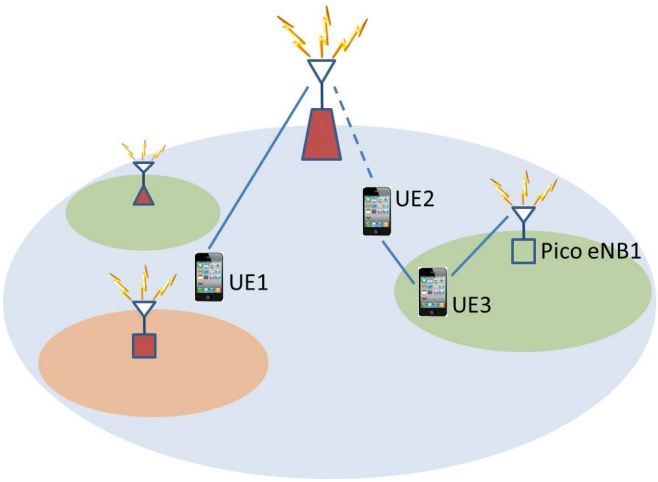

Fig. 3. Illustration of Step 2 in the D2D communication based load balancing algorithm. The congested macro eNB first offloads a currently being served UE, i.e., UE 2, and its ongoing traffic to the adjacent uncongested pico eNB 1 via D2D relay UE 3, then allocates the newly released PRBs (occupied previously by UE 2) to the requesting UE 1. The dashed line denotes a newly released link.

eNBs via X2 interface, then the uncongested eNBs reply with the information of their associated mobile UEs, which are in close proximity to the requesting UE and willing to relay the traffic for it. The macro eNB then instructs the requesting UE to measure and report the D2D link quality between itself and the potential relays. After selecting the D2D relay, the macro eNB collaborates with the eNB serving the relay to jointly assign PRBs and schedule transmissions for the D2D link between the requesting UE and the relay, and the cellular link between the relay and its serving eNB. Thus, the macro eNB manages to detour the traffic of the requesting UE to a neighboring lightly loaded cell via the D2D relay. The details of D2D relay selection, PRB assignment, and transmit power control are left to be discussed in the next section.

Fig. 2 shows an example of Step 1. The UE 1, UE 3, and UE 5 request the congested macro eNB to provide Internet access, and the macro eNB manages to offload their traffic to adjacent lightly loaded picocells or macrocells via the D2D relays UE 2, UE 4, and UE 6, respectively.

On the other hand, if there is no other eNBs around the requesting UE, or all neighboring eNBs are fully loaded, or the uncongested neighboring eNBs fail to find an eligible D2D relay, the macro eNB proceeds to Step 2.

Step 2: the macro eNB tries to release some occupied PRBs for the requesting UE, by offloading a currently being served macro-tier UE and its ongoing traffic to an adjacent uncongested cell via combined D2D link and cellular link. At the operator side, the procedures of offloading a connected UE are similar to that in Step 1. However, besides the basic requirements for setting up combined D2D link and cellular link in detouring traffic, the following condition should also be satisfied such that the macro-tier UE can be "seamlessly" offloaded: the newly established D2D link and cellular link should be able to provide equivalent (or at least comparable) QoS in terms of throughput and delay, and thus result in non-observable changes of QoE (quality of experience). According to such requirement, one therefore can see that only those connected macro-tier UEs each of which has good link quality with the D2D relay and can also get sufficient PRBs from the corresponding neighboring eNB, can be considered as candidates for traffic offloading. Otherwise, if the above requirements cannot be satisfied and the macro eNB fails to offload any being served macro-tier UE, i.e., no PRBs can be released from the macrocell for the requesting UE, the macro eNB proceeds to Step 3.

We use Fig. 3 to illustrate Step 2 of the D2D communication based load balancing algorithm. As shown in Fig. 3, UE 1 requests for Internet access in the congested macrocell. Since the macro eNB is already fully loaded and has no free PRBs, it first releases the PRBs occupied by UE 2 after offloading UE 2 to the uncongested neighboring pico eNB 1 via the D2D relay UE 3, then allocates the newly released PRBs to UE 1.

Step 3: the macro eNB tries to offload the requesting UE to a congested eNB which is close to the UE and able to release some PRBs by offloading a currently being served UE to its nearby uncongested cell. For the requesting UE, the QoS (like throughput and delay) of its traffic depends on the link quality between itself and the D2D relay and also the number of PRBs that can be released from the adjacent eNB. While for the eNB routing the traffic to (or from) the requesting UE, the requirements and procedures of offloading a being served UE are similar to that in Step 2. However, the complexity of operation in Step 3 is much higher than that in Step 2. In particular, one cellular link release and one D2D link setup are required in Step 2; in Step 3, it requires to release one cellular link and set up two D2D links. Furthermore, Step 2 involves the collaborations between two eNBs and one D2D relay, while Step 3 involves three eNBs and two D2D relays. As shown in Fig. 4, after the congested pico eNB 1 offloads its being served UE 3 to the uncongested pico eNB 2, the congested macro eNB offloads the requesting UE 1 to the pico eNB 1, where UE 2 and UE 4 are the D2D relays for offloading UE 1 and UE 3, respectively.

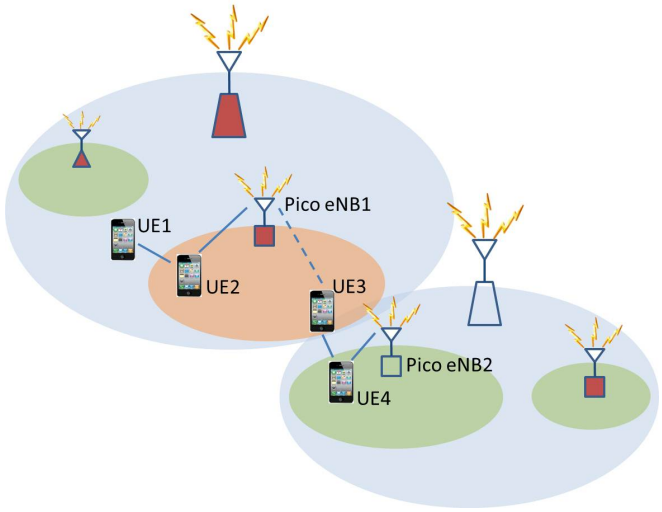

Fig. 4. Illustration of Step 3 in the D2D communication based load balancing algorithm. After UE 3 is offloaded from the congested pico eNB 1 to the uncongested pico eNB2 via D2D relay UE 4, the requesting UE 1 is offloaded from the congested macro eNB to the pico eNB 1 via UE 2. The dashed line denotes a newly released link.

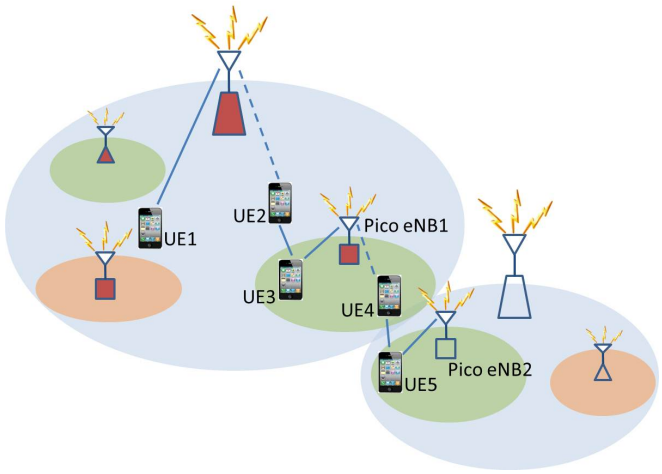

Fig. 5. Illustration of Step 4 in the D2D communication based load balancing algorithm. After pico eNB 1 offloads the being served UE 4 via UE 5 to the uncongested pico eNB 2, the congested macro eNB allocates to the requesting UE 1 the PRBs newly released by offloading the currently being served macro-tier UE 2 to the pico eNB 1 via UE 3. The dashed line denotes a newly released link.

If the macro eNB fails to offload the requesting UE to an adjacent congested eNB, due to lack of good channel quality, D2D relay(s), or sufficient PRBs, etc., it proceeds to the last step, i.e., Step 4.

Step 4: in the last step, the macro eNB tries to allocate to the requesting UE PRBs newly released by offloading a currently being served macro-tier UE to a nearby congested eNB, which is able to offload a being served UE to an adjacent uncongested cell. Similar to that in Step 3, it requires the collaborations between three eNBs and two D2D relays in UE offloading and PRB releasing. However, besides setting up two D2D links, Step 4 needs to release two cellular links which is one more than that in Step 3. As shown in Fig. 5, the macro-tier UE 2 can be offloaded via D2D relay UE 3 to the congested pico eNB 1, after UE 4 is offloaded to the uncongested pico eNB 2 via UE 5.

From the above algorithm, one can see that D2D communication is of significant potential in achieving efficient load balancing according to the real-time traffic distributions among different tier cells of LTE-A HetNets. For the case of LTE-A networks without D2D communications, it is possible to achieve a certain level of load balancing by assigning small eNBs different bias values and pushing load onto those less congested cells. However, since the preconfigured bias values are averaged over the statistical analysis of network traffic, such techniques cannot respond or adapt flexibly to the real-time network dynamics, in terms of eNBs powered on/off, mobile UEs coming/leaving, etc. Furthermore, by utilizing D2D communications, it could even directly route the traffic between a UE pair without occupying the air interface between eNB and UEs; while this is never the case for LTE-A networks without D2D, where every single bit has to be transmitted through the eNB.

It is noticed that in LTE-A networks, some areas may have severe interference among nodes depending on the node density there. Due to the limited channel resources at the eNB and the overly crowded nodes in the area, the wireless link (uplink or downlink) between a mobile UE and the eNB usually has very poor SINR, i.e., the air interface between UEs and the eNB is severely congested. For such scenarios, the proposed D2D based traffic offloading algorithm can also be utilized to alleviate the air interface congestion, increase system throughput and improve user experiences. Specifically, the serving eNB first communicates with the surrounding eNBs via the specific X2 interface, to obtain the list of neighboring eNBs which are relatively lightly loaded. Then, the serving eNB either directly offloads a requesting UE to an uncongested adjacent cell as defined in Step 1 or releases some occupied PRBs by offloading a currently being served UE as defined in Step 2, depending on the actual location of the UE. Note that after applying Step 1 and Step 2, the eNB only needs to allocate channel resources to a smaller number of UEs remaining in the area, i.e., the UEs that cannot be offloaded. Considering the much smaller transmit power and communication range adopted by D2D communications, the radio interference within the area can be effectively alleviated.

B. Discussions and Research Issues

In this section, we discuss on the requirements and difficulties in offloading traffic among multi-tier cells via D2D communications, and identify the challenging issues in terms of algorithm complexity, D2D relay selection, PRB assignment, transmission schedule, power control, interference management, network dynamics, QoS satisfaction, incentive stimulation, privacy and security, analytical modeling, etc.

Prerequisite: according to the SAE architecture in LTE systems, an eNB can obtain from the PDN gateway the distributions of adjacent eNBs and also the list of mobile UEs served by each eNB. Therefore, the only prerequisite of our algorithm is that each eNB is able to obtain the current locations of its associated mobile UEs. Clearly, this

TABLE I
SUMMARY OF NETWORK OPERATIONS IN THE PROPOSED D2D BASED TRAFFIC OFFLOADING

	Direct	Step 1	Step 2	Step 3	Step 4
Involved eNBs	1 or 2	2	2	3	3
Involved UEs	2	2	3	4	5
Established D2D links	1	1	1	2	2
Released cellular links	0	0	1	1	2

is not a technically challenging task for the operator, e.g., the popular function of locating one's iPhone. According to the triangulation, generally at least three eNBs are required to pinpoint a cell phone and its owner precisely. The only problem is the accuracy: in downtown or urban areas where there are lots of eNBs, it is easier for operators to accurately locate a UE; while in suburban or rural areas, it is less accurate but GPS can be a good alternative. On the other hand, as it is not so crowded or capacity demanding in rural areas, the desire to balance traffic among eNBs is not so obvious.

Supposed network applications and services: considering the mobility issues of mobile UEs, it is very difficult for an eNB to provide stable bandwidth for an end user through the selected D2D relay. Therefore, the proposed D2D communication based traffic offloading algorithm may be utilized for the network applications or services which are either delay tolerant or not bandwidth-hungry, such as browsing news, checking email, instant messaging, etc. Some popular exemplary applications include twitter, facebook, LINE, Viber, whatsapp, etc.

Complexity: the details of network operations in the proposed D2D based traffic offloading are summarized in Table I. Clearly, from "Direct offloading" to "Step 4", a monotonically non-decreasing varying trend can be observed for the number of involved eNBs, the number of involved UEs, the number of established D2D links, and the number of released cellular links. Due to the increasing complexity of operations, particularly the number of established D2D links and the number of released cellular links, we consider only four steps in the D2D based load balancing algorithm, i.e., we allow a congested eNB to try at most four different kinds of traffic offloading to provide Internet access for a requesting mobile UE. Generally speaking, one can further proceed to more complicated operations of traffic offloading, which unavoidably involves the collaborations between more eNBs and more mobile UEs, the setup of more D2D links, and the release of more cellular links.

Combined D2D link and cellular link: it is one of the common routes in Steps 1$\sim$4 for D2D based traffic offloading, which may largely affect the algorithm performance. Specifically, the following aspects deserve further study. First, how to select the best D2D relay. As two UEs closer in space may not necessarily own a better channel in practice, it is advisable to measure the actual D2D link quality before relay decision. In addition, the cellular link condition between the relay and its eNB should also be taken into account. Second, how to assign PRBs for D2D link and cellular link. Basically, the number of PRBs allocated for D2D link should be decided according to that allocated for the cellular link. Also, one should take into consideration the difference between the UL and DL for the combined D2D link and cellular link, and the difference between the frequency bands deployed in different tier eNBs to which the operator may allocate different bands according to their backhaul connections. Finally, how to schedule transmissions and manage interference. Besides the difference caused by TDD and FDD, another issue is that the D2D link extends over two adjacent cells, which means the D2D transmission will easily affect the cellular transmissions in two cells. Again, transmit power control can be an effective technique to manage the intra-cell and inter-cell interference.

Releasing cellular link: as a basic operation in the proposed algorithm, it is of vital importance that an eNB select proper connected UE to release and detour the ongoing traffic to neighboring cells. Besides the basic requirements in establishing for the released UE combined D2D link and cellular link, the newly combined route should provide for the UE at least comparable QoE. That means, it should be able to support QoS in terms of throughput, delay, and jitter, similar to that previously provided by the released cellular link. Furthermore, depending on the service contract customers signed with the operator, some customers may simply not allow to be offloaded.

Network dynamics and QoS satisfaction: the unpredictable network dynamics may have a nonnegligible impact on the performance of the proposed D2D based traffic offloading algorithm. Note that the communication range between a D2D pair is very limited. A mobile UE may easily fail to send/receive data through the newly established offloading path, due to the movement of itself or that of the D2D relay. Furthermore, the channel quality of the D2D link may become very poor due to the low battery power at relay node, or the interference from surrounding terminals especially in urban areas. Such network dynamics should be carefully accounted in D2D based traffic offloading, particularly in the selection of D2D relays and the selection of released cellular links. How to guarantee acceptable QoS at the end user and avoid frequent change of offloading path remains challenging and deserves further study. Also, it would be meaningful to conduct extensive simulations under the 3GPP recommended mobile scenarios, so as to further evaluate the performance of the proposed algorithm.

Incentive, privacy, and security: regarding incentive stimulation, it refers to how to encourage a mobile UE to participate in D2D communications as a D2D relay. According to the actual contributed traffic amount, the operator can appropriately reward the mobile UE which has delivered data to other users (for the case of direct offloading) or has forwarded data as a D2D relay (for the case of offloading among multi-tier cells). On the other hand, in the D2D based offloading, a mobile UE no longer receives (resp. sends) data directly from (resp. to) the eNB, but via another mobile UE instead, which may cause a lot of difficult threats. It is very difficult for operators to guarantee the secure transmission through multiple terminals, since the intermediate relays may attempt to do malicious attacks in lots of ways. Actually, in order for the proposed D2D based traffic offloading algorithm to address such issue, the eNBs and the end user have to take some actions before

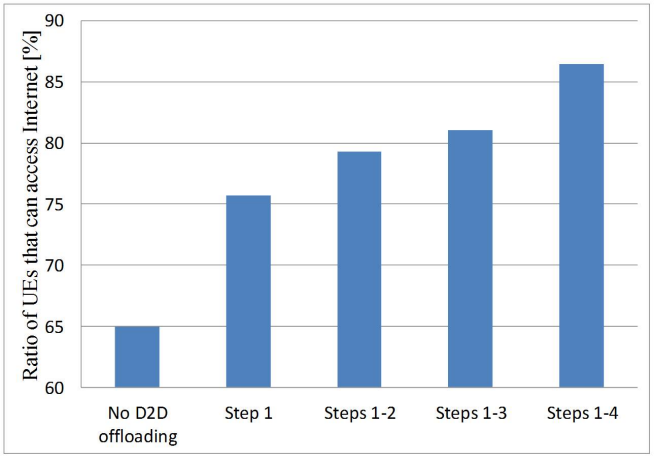

Fig. 6. The ratio of UEs that can access Internet to the UEs which request for Internet access.

starting to transmit data. Suppose a congested macro eNB has to offload a mobile user to an adjacent lightly loaded pico eNB. Using the route user-macro eNB-pico eNB, the user and the pico eNB can quickly establish a session key, which can be securely used for encrypting the data transmitted through the offloading path. Note that we utilize only one-hop operator assisted D2D communication for traffic offloading, i.e., we allow only one D2D relay in each offloading path. Therefore, the operator can effectively prevent the intermediate D2D relay from accessing, tampering, or falsifying the data, because the relay has no idea of the session key and it is selected by the operator (That is, the operator has the physical information of the relay, such as IMEI, etc.). However, for the case of offloading through more than one D2D relays, a malicious D2D relay may tamper the packets, inject falsified data, or just simply drop some incoming packets. How to identify and avoid such behaviors remains challenging and deserves further study.

Performance modeling and analysis: it is never too much to emphasize the importance of developing models for performance analysis. Due to the introduction of small cells, it appears to be nontrivial to accurately simulate a multi-tier LTE-A HetNet. Traditional hexagonal grid based model seems inapplicable, since the small cells (like femtocells) are usually irregularly scattered or clustered within the existing macrocell area. On the other hand, recently, the Poisson point process (PPP) based models has received wide attention, which is analytically tractable and appears to capture the main trend of HetNet performances. It is reported that besides giving similar shapes of SINR distributions, the grid model and the PPP model differ mostly in absolute SINR: with the former being optimistic and the latter being pessimistic [28].

V. Numerical Results

In this section, we present some numerical results to illustrate the performance gains that can be achieved by our D2D based load balancing algorithm. We consider a simple scenario which consists of a single marcocell underlaied by two picocells and two femtocells (configured as open access). The coverage area of macrocell, picocell and femtocell are assumed to be circles of radius 250 m, 100 m, and 50 m, respectively. Marco eNB, pico eNB, and femto eNB are assumed to have the same frequency resources of 50 orthogonal channels. That is, each eNB of macrocell, picocell, and femtocell can provide Internet access for 50 users at the same time. In total 600 UEs are uniformly distributed in the coverage area of macrocell, and each UE requests for Internet access with a probability of 33.3%. Moreover, the maximum distance allowed for D2D communication between a UE pair is set to 20 m.

With the above parameter settings, we obtain the ratio of UEs that can access Internet (to the UEs that request for Internet access) under our D2D based load balancing algorithm, and compare it with the scenario where our algorithm is not applied. The numerical results are summarized in Fig. 6. From the figure, one can clearly see that the ratio of UEs that can access Internet increases after applying more steps of our algorithm. Specifically, after applying Steps 1, 2, 3 and 4 of the proposed algorithm, it enables around 86% of the requesting UEs to access Internet simultaneously; while only 65% of UEs can be supported when not applying the algorithm. This is because via D2D communications, the requesting users can be offloaded from the congested macro eNB to the pico eNBs and femto eNBs which are relatively lightly loaded, thus enabling more users to be served and achieving higher spectral efficiency.

VI. Conclusions

When we expect the LTE-A networks to significantly enhance the current LTE and support much higher capacity and coverage, higher throughput and lower latency, higher peak rates and better user experience, etc., it is necessary to consider effective technology to address the congestion there caused by imbalanced traffic distributions among multi-tier cells. Different from available techniques, in this article, we highlighted a D2D communication based technique for load balancing, which is able to efficiently offload traffic among multi-tier cells according to their real-time traffic distributions. In addition, we presented numerical results to show the great promise of applying the proposed algorithm. Here we would like to emphasize also the challenges in applying D2D communication for traffic offloading, such as PRB allocation and transmission schedule for combined D2D link and cellular link, interference management, cellular link releasing, mobility, incentive stimulation, security issues, etc. Also, it is meaningful to further evaluate the performance of D2D based traffic offloading, when applied together with other techniques like cell biasing.

References

[1] J. G. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, March 2013.

[2] J. He, "An architecture for wide area network load balancing," in *ICC*, 2000.

[3] A. Narula-Tam and E. Modiano, "Dynamic load balancing for wdm-based packet networks," in *INFOCOM*, 2000.

[4] P.-H. Hsiao, A. Hwang, H. T. Kung, and D. Vlah, "Load-balancing routing for wireless access networks," in *INFOCOM*, 2001.

[5] H. Gong and J. Kim, "Dynamic load balancing through association control of mobile users in wifi networks," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 342–348, 2008.

[6] L. Georgiadis, P. Georgatsos, K. Floros, and S. Sartzetakis, "Lexicographically optimal balanced networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 6, pp. 818–829, December 2002.

[7] C.-F. Huang, H.-W. Lee, and Y.-C. Tseng, "A two-tier heterogeneous mobile ad hoc network architecture and its load-balance routing," in *IEEE VTC-Fall*, 2003.

[8] C. K. Toh, A.-N. Le, and Y.-Z. Cho, "Load balanced routing protocols for ad hoc mobile wireless networks," *IEEE Communications Magazine*, vol. 47, no. 8, pp. 78–84, 2009.

[9] T. Taleb, D. Mashimo, A. Jamalipour, N. Kato, and Y. Nemoto, "Explicit load balancing technique for ngeo satellite ip networks with on-board processing capabilities," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 281–293, 2009.

[10] S. Jung, M. Kserawi, D. Lee, and J.-K. K. Rhee, "Distributed potential field based routing and autonomous load balancing for wireless mesh networks," *IEEE Communications Letters*, vol. 13, no. 6, pp. 429–431, 2009.

[11] H. Jiang and S. Rappaport, "Cbwl: A new channel assignment and sharing method for cellular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 2, pp. 313–322, May 1994.

[12] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment," *Wireless Networks*, vol. 3, no. 2, pp. 333–347, 1997.

[13] ——, "A novel load balancing scheme for the tele-traffic hot spot problem in cellular networks," *Wireless Networks*, vol. 4, no. 2, pp. 325–340, 1998.

[14] B. Eklundh, "Channel utilization and blocking probability in a cellular mobile telephone system with directed retry," *IEEE Transactions on Communications*, vol. 34, no. 2, pp. 329–337, April 1986.

[15] X. Wu, B. Mukherjee, and S. H. G. Chan, "Maca-an efficient channel allocation scheme in cellular networks," in *IEEE Globecom*, 2000.

[16] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systemsg," in *INFOCOM*, 2003.

[17] H. Wu, C. Qiao, S. De, and O. K. Tonguz, "Integrated cellular and ad-hoc relay systems: icar," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2105–2115, October 2001.

[18] E. Yanmaz and O. Tonguz, "Dynamic load balancing and sharing performance of integrated wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 862–872, June 2004.

[19] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: wireless video content delivery through distributed caching helpers," in *INFOCOM*, 2012.

[20] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Communications Letters*, vol. 12, no. 6, pp. 2703–2716, 2013.

[21] "Physical layer aspects for evolved universal terrestrial radio access (utra)," 3GPP TR 25.814 V7.1.0, September 2006.

[22] A. GHOSH, R. RATASUK, B. MONDAL, N. MANGALVEDHE, and T. THOMAS, "Lte-advanced: Next-generation wireless broadband technology," *IEEE Wireless Communications Magazine*, vol. 17, no. 3, pp. 10–22, June 2010.

[23] "Evolved universal terrestrial radio access (e-utra); further advancements for e-utra physical layer aspects," 3GPP TR 36.814 V9.0.0, March 2010.

[24] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," in *Allerton*, 2010.

[25] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, December 2009.

[26] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Mikls, and Z. Turnyi, "Design aspects of network assisted device-to-device communications," *IEEE Communications Magazine*, vol. 50, no. 3, pp. 170–177, March 2012.

[27] L. Lei, Z. Zhong, C. Lin, and X. S. Shen, "Operator controlled device-to-device communications in lte-advanced networks," *IEEE Wireless Communications Magazine*, vol. 19, no. 3, pp. 96–104, June 2012.

[28] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, November 2011.