Design of a Quality of Service-Based Load Balancing Relay Selection Mechanism for Long Term Evolution-Advanced Systems

San-Yuan Wang*, Guan-Hsiung Liaw[2], Chih-Hao Hsu[3], Lain-Chyr Hwang[4], Tain-Lieng Kao[5], Hsing-Yen Hsieh[6]
Dept. of Computer Science and Information Engineering,
**sywang@isu.edu.tw* (Corresponding author), [2]ghliaw@isu.edu.tw, [3]*isu9903021m@isu.edu.tw*
Department of Electrical Engineering, [4]*lain@isu.edu.tw*
Dept. of Communication Engineering, [5]*tlkao@isu.edu.tw*
Department of Healthcare Administration,[6]*hsiehhy@isu.edu.tw*
I-Shou University, Kaohsiung City, Taiwan,

Abstract**—**Serving as the fourth generation mobile communication standard, Long Term Evolution-Advanced provides various technical support to achieve high transmission speed. In particular, relays are an essential technology supported by the standard. Because a relay uses the resources within a communication system, user devices adopt the optimal relay method as the transmission pathway to optimize resource utilization. According to the quality of service required by various user applications, this paper fabricates a method for selecting the optimal load-balancing transmission pathway for user devices.

Keywords-*LTE-advanced; relay selection; load balancing; QoS*

*****

I. Introduction

International Mobile Telecommunications-Advanced (IMT-Advanced) is the fourth generation wireless bandwidth mobile communication system standard formulated by the International Telecommunication Union (ITU). According to the standards set by the ITU [1, 2], IMT-Advanced must support a peak rate of 100 Mb/s and 1 Gb/s in a high speed vehicular environment (up to 350 km/h) and a pedestrian environment (10 km/h), respectively. The transmission bandwidth of IMT-Advanced is expandable, ranging from 20 to 100 MHz. The bandwidth use rates in the uplink and downlink are [1.1, 15 b/s/Hz] and [0.7, 6.75 b/s/Hz], respectively.

To achieve the wireless access speed of IMT-Advanced and meet the related quality of service (QoS) requirements, the Third Generation Partnership Project (3GPP) initiated the standardization of Long Term Evolution (LTE) in late 2004, which was completed successfully in 2007. Subsequently, the 3GPP initiated the standardization of LTE-Advanced to meet the requirements of IMT-Advanced mobile communication systems through various types of communication technology [3, 4], including carrier aggregation, coordinated multiple point transmission and reception, and relays. These newly developed LTE-Advanced communication technology standards serve as the 3GPP’s candidate solutions for IMT-Advanced mobile communication systems.

The role of a relay station (RS) in LTE-Advanced is similar to that of a small low-power base station, or Evolved Node B (eNB), which enables wireless backhaul for linking to a core network. In other words, a user equipment(UE) treats an RS as an eNB, and an eNB treats an RS as a UE. An RS uses the resources of an eNB to serve as an intermediate transmitter for UEs; hence, it does not increase the overall system capacity. Nevertheless, the RS enhances the throughput of the system and expands the coverage of the eNB. RSs are preferred over eNBs because the former is more advantageous; in comparison, eNBs exert greater interference on LTE system that involves a frequency reuse factor of 1. Although an eNB exhibits a coverage range mostly identical to that of an RS and attains a greater throughput within the coverage range [5, 6], an RS is more advantageous than an eNB from a financial cost perspective [7].

An RS selection mechanism involves allocating an RS that is most suitable to the target UE. Because an RS uses the transmission resources of an eNB to transfer UE data, allocating an excessive amount of resources can decrease the throughput of the overall system; hence, a favorable RS selection mechanism should be formulated to assign each UE to an RS adequately. However, when only the overall system throughput is considered, the selection mechanism prioritizes assigning UEs with high transmission speed, thereby reducing the resources available for peripheral UEs. This causes the overall throughput to be lower than that before the selection mechanism is adopted. Conversely, when all UEs are considered to have the same transmission speed, peripheral UEs (i.e., UEs at the peripheries of transmission coverage or those with slow connections) may consume excessive transmission resources, thus reducing the overall system throughput.

When only an RS selection mechanism is adopted, an RS becomes a hotspot when it is near a high number of UEs, all of which transmit data through the RS. Therefore, the RS assigns the UEs according to a schedule. Several UEs might experience transmission delay as the wait time increases. In this scenario, other RSs may still have additional resources that have not been allocated to the UEs, thus wasting these resources. The purpose of load balancing is to reassign UEs in wait to another RS with subpar transmission speed. Accordingly, the UEs do not have to wait, and no resource is left unused. In this paper, the load interpretation method proposed by [8] is adopted. Through the QoS class system of LET, the method determines which UEs are reassigned to other RSs according to the application requirements of each UE.

Next, Section 2 discusses related literature review. Section 3 introduces various types of relay. Section 4 elucidates relay selection mechanisms, and Section 5 examines the design of QoS-based load balancing. The final section addresses the conclusion.

II. Literature Review

Fareed and Uysal [9] proposed a relay selection mechanism that does not require UEs to transmit channel state information (CSI). Specifically, the selection decisions are not made by the base station or relay. Instead, the UE determines the CSI between a RS and itself according to the signals broadcasted by the RS. Computing the CSI yields the signal-to-interference-plus-noise ratio (SINR) value, which is then used to determine the transmission speed achievable through the relay transmission. After the computation and comparison processes, the UE transmits a link request to the target relay, instructing the relay to join the transmission pathway.

Gkatzikis and Koutsopoulos [10] applied maximum weighted matching to a bipartite graph, and the Hungarian algorithm [11] was used to obtain the optimal solution. Their study first assumed that the CSI of three transmission pathways (namely eNB–RS, eNB–UE, and RS–UE) can be acquired completely. Through channel status estimation, the transmission speed of each pathway was then calculated. Next, a weight was assigned to each speed according to the corresponding pathway displayed in the bipartite graph. Finally, the Hungarian algorithm was employed to determine the optimal relay selection scheme.

Wang et al. [12] proposed integrating proportional fairness with void filling. This approach focuses on the long-term average transmission speed of UEs to ensure a favorable QoS without neglecting peripheral UEs. The void filling method enables fully utilizing unallocated resources.

Hu and Qiu [13] considered the mutual interferences of RSs in a multiple RS transmission system to propose a utility function approach according to the resources required by the RSs and UEs as well as the efficiency (transmission speed and fairness in resource allocation) achievable by the RSs. A greedy algorithm was used to select the most suitable RS for each UE, thereby enhancing the system transmission rate and sustaining fairness in resource allocation.

Wu et al. [14] proposed an approach to decrease the interferences between related systems and increase the usage rate of system resources. The selection decisions are formulated to maximize the connection and transmission rates of UEs and RSs.

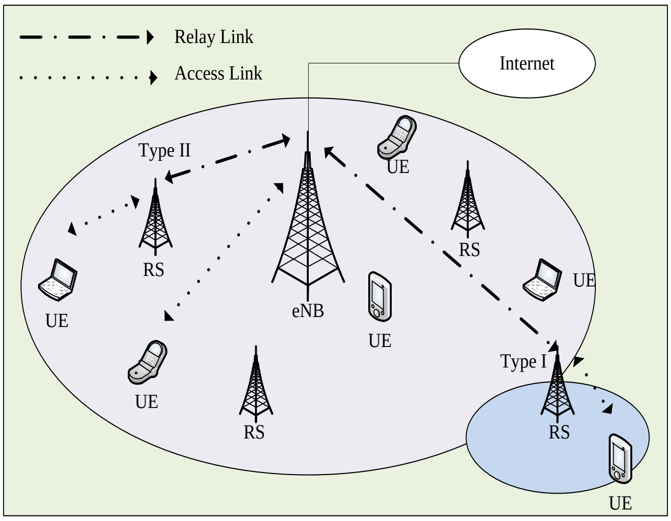
The aforementioned studies have aimed to maximize system transmission rates. Although several studies have considered resource allocation fairness, its contribution to the overall system load balancing is limited. The following studies have attempted to fully utilize system resources.

BouSaleh et al. [15] formulated a handover mechanism for UEs to select either an eNB or RS. Because UEs select connection methods according to the transmission power they received, decreasing the power threshold of which UEs select RSs enables the RSs to service additional numbers of UEs. However, UEs that are far away from the RSs can experience transmission speed lower that obtained through directly connecting to eNBs. Therefore, balance should be maintained between the number of serviced UEs and transmission speed.

Yu et al. [16] employed a heuristic algorithm to calculate the resources required by RSs and UEs for transmitting data. According to the calculation outcomes, an exhaustive method was used to determine the optimal UE and RS selection scheme, ensuring that fewest resources are left unused, thereby maximizing the system transmission speed. However, the heuristic approach faces a computation speed problem.

Jian and Wang [8] proposed a load-aware RS selection mechanism. Specifically, selection decisions are made according to the number of UEs serviced by each RS and the achievable transmission speed calculated through the CSI of each UE. If the optimal RS for a UE is already servicing many UEs, the mechanism assigns the UE to another RS or directly to an eNB according to the transmission speed. This enables preventing an extensive queue caused by assigning an excessively high number of UEs to an RS. Therefore, the system resources are fully utilized to increase the transmission speed effectively.

III. TYPES OF RELAY

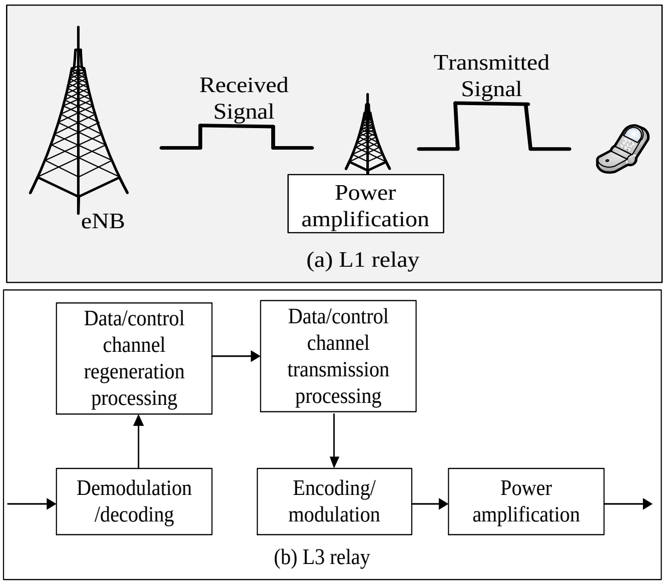


A network scenario with multiple RSs and multiple UEs.

Fig. 1 illustrates the data transmission of UEs through RSs. The transmission pathways of which the UEs connect to the RSs are called access links, and those of which the RSs connect to the eNB are called relay links, or backhaul links. The transmission pathways between the UEs and eNB are called direct links. RSs are divided into Type 1 and Type 2. Particularly, Type 1 RSs are deployed to service UEs that are not within the coverage range of the eNB by transmitting the common reference signal and control information of the UE to the eNB. In other words, the main purpose of Type 1 RSs is to extend the service range of the eNB, thereby increasing the overall system capacity. Type 2 RSs assist UEs within the coverage range of the eNB. Because the UEs might receive a low SINR value from the eNB due to environmental factors, these RSs facilitate the link between the UEs and eNB to achieve more favorable QoS and transmission speed. Transmitting the CRS and control information is unnecessary for this type of RS, which enhances the overall system throughput by increasing the UE transmission speed.

Relay links are categorized into L1 relay and L3 relay. In particular, an L1 relay employs an amplify and forward (AF) transmission method (Fig. 2a). When the relay receives the signal transmitted from the eNB, the signal is amplified and then sent to the UE. Similarly, when the UE transmits a signal to the eNB through an L1 relay, the signal is amplified by the relay before being sent to the eNB. Because the AF method only amplifies signals, the delay caused by signal processing is short. In addition, an L1 relay involves simple equipment, short installation time, and low costs. However, intercell interferences and background noises are amplified along with the transmission signal when processed by an L1 relay. This can decrease the SINR and thereby fail to increase the transmission speed.

An L3 relay employs a decode and forward (DF) transmission method (Fig. 2b). When a UE sends data to the eNB, the data first arrive at the relay and are demodulated and decoded. Next, the data are subjected to processes including encryption, integration, and splitting. Subsequently, the data are encoded and modulated again before they are sent back to the UE. The DF method reorganizes the received signals to prevent sending the wrong signals to the UE and eliminates inter-cell interferences and environmental noises. A high-level modulation process can be employed to increase the data transmission volume; however, such processes extend the overall processing time.



(a) L1 relay

(b) L3 relay

Figure 1. L1/L3 relay techniques

IV. RELAY SELECTION MECHANSIMS

A UE select relays according to various conditions, the final goal of which is to attain favorable transmission speed. Specifically, the following factors are considered when making selection decisions: the shortest path, minimum path loss, maximum receive-power transfer, and maximum SINR.

Most relay selection mechanisms are based on the aforementioned factors. Categorized into two types, the mechanisms are aimed to either 1) maximize the overall system transmission speed or 2) enable peripheral UEs to also receive satisfactory speed. In particular, the first type of mechanism attains the maximum speed at the cost of unfair resource allocation. Because UEs with favorable transmission conditions expend dissimilar amount of resources than do those with unfavorable transmission conditions, aiming for enhancing the overall system throughput cause UEs with unfavorable transmission conditions to attain extremely low transmission speed. By contrast, the second type of mechanism resolves this problem by allocating additional resources to UEs with unfavorable transmission conditions, thereby sustaining the transmission speed at the cost of decreasing the overall system throughput. Therefore, when designing a relay selection mechanism, the balance between the system transmission speed and fair resource allocation should be maintained. In this paper, a load balancing approach is adopted to facilitate fairness in resource allocation; hence, the first type of mechanism is employed to design the proposed mechanism.

The maximum SINR is chosen as the basis for relay selection because the maximum capacity can be calculated using the SINR value and Shannon Theorem. The data of SINR are obtained from the CSI transmitted from UEs to an eNB.

Specifically, k ∈ {0, 1, 2,…, K} indicates that the total number of transmission link is K + 1. K = 0 denotes a direct link with the eNB. k∈ {0, 1, 2,…, K} implies that information is transmitted through the kth RS. m∈ {0, 1, 2,…, M} represents the index of UEs. S denotes the eNB. If a direct transmission is employed between the eNB and UE, the attainable capacity can be calculated using (1):

$$C_{DT}(S,m) = W\log_2(1+SINR_{Sm}) \qquad □□□$$

where W represents the bandwidth of the transmission channel.

According to [17], the capacity of DF transmission (i.e., the information the eNB transmits to a UE through an RS) is calculated using (2):

$$C_{DTf}(S,k,m) = \frac{W}{2}\min\left\{\log_2(1+SINR_{Sm}),\log_2(1+SINR_{Sm}+SINR_{km})\right\} \qquad □□□$$

On the basis of (1) and (2), the maximum system capacity and the corresponding RS allocation scheme can be determined. The allocation scheme is represented by I = {1,2,…,M}. The corresponding position of UE is substituted with the calculated RS code. The allocation scheme is described by (3):

$$I = \arg\max\sum_{m=1}^{M}\sum_{k=1}^{K}\max\left\{C_{DT}(S,m),C_{DF}(S,k,m)\right\} \qquad □□□$$

According to the allocation outcome yielded by (3), RSs can be assigned to each UE to achieve the maximum system capacity.

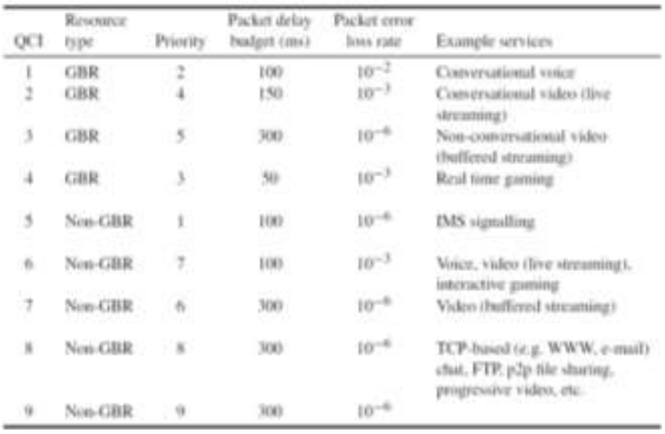
V. QoS-Based Load Balancing Relay Selection Mechanism

Within the coverage range of an eNB, UEs can access various types of application services, including Voice over Internet Protocol (VoIP) phone calls, website browsing, and data download through the File Transfer Protocol (FTP) or Transmission Control Protocol. These applications require distinctive classes of QoS. For example, VoIP demands stricter requirements on packet delay and jitter but more lenient requirements on packet error loss rate, whereas FTP transmission demands stricter requirements on packet error loss rate but more lenient requirements on packet delay. In response to various types of application, LTE designates nine types of QoS class identifier (QCI). The Evolved Packet System configures dissimilar bearers according to various classes of QoS. A bearer is the IP packet flow of a defined QoS class and can be categorized into two types according to the classes of QoS:

- Minimum guaranteed bit rate (GBR) bearer: This type of bearer is adopted by applications such as VoIP. When a GBR bearer is established, its specific transmission resource is allocated to the UE, thereby maintaining the QoS during the transmission process. Until the application is terminated, this resource is allocated to the UE permanently.
- Non-GBR bearer: This type of bearer, which neither guarantees the transmission speed nor allocatesresources to the UE permanently, is often applied to website browsing or FTP transmission.

Table I lists the nine classes of LTE QoS [18] according to various services. In a network access, the eNB is responsible for verifying the QoS required by the bearer. Each bearer has a corresponding QCI and allocation and retention priority (ARP). Each QCI is determined by QoS parameters such as priority level, packet delay budget, and acceptable packet loss rate, and a packet with high priority is scheduled first. The ARP is applied to admission control. For example, when a wireless network has high traffic, the priority of the ARP of a bearer is used to determine whether the link of the bearer should be established. After the link is established the ARP does not affect the subsequent packet transmission processes, including packet scheduling and speed control. Instead, these processes are determined by the QoS parameters such as the QCI and GBR.

TABLE I. STANDARDIZED QOS CLASS IDENTIFIER(QCIS) FOR LTE



QCI	Resource type	Priority	Packet delay budget (ms)	Packet error loss rate	Example services
1	GBR	2	100	10^{-2}	Conversational voice
2	GBR	4	150	10^{-3}	Conversational video (live streaming)
3	GBR	5	300	10^{-6}	Non-conversational video (buffered streaming)
4	GBR	3	50	10^{-3}	Real time gaming
5	Non-GBR	1	100	10^{-6}	IMS signalling
6	Non-GBR	7	100	10^{-3}	Voice, video (live streaming), interactive gaming
7	Non-GBR	6	300	10^{-6}	Video (buffered streaming)
8	Non-GBR	8	300	10^{-6}	TCP-based (e.g. WWW, e-mail) chat, FTP, p2p file sharing, progressive video, etc.
9	Non-GBR	9	300	10^{-6}	

In this paper, the QoS load balancing is based on the packet delay budgets and packet error loss rates required by the nine types of QoS class, among which four types are selected as the bases for prioritizing UEs to be reallocated to another RS. According to the descending order of priority, the selected classes are QCI-1 (priority 2), QCI-4 (priority 3), QCI-2 (priority 4), and QCI-6 (priority 7). These four classes are selected because, compared with other classes, they have higher requirements for packet delay budget but lower requirements for packet loss error rate. Therefore, prioritizing the reallocation of UEs with these QoS classes is advantageous for avoiding competition with other UEs, thus attaining low delay time. Although UEs with these QoS classes are processed by RSs with subpar speed and their SINR values are less favorable, their high tolerance for packet error loss rate enables decreasing the effect of unfavorable packet error loss rate on the QoS, thereby achieving the goal of load balancing.

The load sensing relay selection method proposed by [8] is used to assess the number of UEs processed by each RS. When an excessive load is being processed by the RS, additional UEs are transferred to other RSs with subpar speed. The QoS-based load balancing selection mechanism is detailed as follows:

First, the eNB periodically disseminates broadcast packets, through which the RSs and UEs estimate the CSI between themselves and the eNB. Similarly, the RSs also periodically disseminate broadcast packets, which enable the UEs to determine the CSI between themselves and the RSs. The packets disseminated by the RSs also inform on the number of UEs currently being serviced by the RSs. This number serves as a basis for the subsequent relay selection decisions.

When a UE establishes a new connection or is transferred to a new eNB, the UE calculates the achievable transmission speed according to the CSI acquired from the broadcast packets disseminated by the previous eNB and RSs. If the calculated transmission speed is faster than that obtained through connecting to another RS, the UE sends a transmission request to the eNB directly. Next, the eNB either accepts or rejects the request according to the decision of the admission control.

If following calculation, the UE decides to connect with an RS instead, the broadcast packets disseminated by nearby RSs are used determine the most suitable RS, the broadcast packet of which informs on the current number of UEs being serviced by the RS. According to the load sensing equation (4) proposed by [8] and (2), calculations are performed to determine which RS enables attaining the largest β value. The maximum β value denotes the maximum throughput. In (4), $E[C_{DF}(S,k,m)]$ represents the average speed attained when a UE (m) selects an RS (k).$E(M_k)$ denotes the number of UEs currently being serviced by the RS (k).

When a UE selects an RS that is servicing a high number of other UEs, the RS assesses the QoS class of the UE. If the priority level of the UE is 2, 3, 4, or 7, the UE repeats the RS selection process, in which the previously selected RS is omitted. By comparing the transmission speeds obtain through other connection means, the UE selects another subpar RS or directly connects to the eNB. By contrast, if the priority level of the UE is not 2, 3, 4, or 7, another UE that is currently being serviced by the RS and has the priority level of 2, 3, 4, or 7 is reallocated to another subpar RS. Nonetheless, the UEwill still be connected to the RS if none of the UEs serviced by the RS fulfills the reallocation requirement.

VI. CONCLUSION

Relays are an essential technique of LTE-Advanced and effectively enhance the coverage range and throughput of eNBs, in addition to having cost advantages. Therefore, increasing the efficiency of RSs is imperative. In this paper, a QoS-based load balancing relay selection method is proposed to not only improve the efficiency of relays but also maintain fairness in resource allocation for meeting the QoS levels of UEs. Furthermore, this mechanism enables fully utilizing system resources, ensuring that no resources are wasted or left unused.

REFERENCES

[1] ITU-R SG5, “Invitation for submission of proposals for candidate radio interface technologies for the terrestrial components of the radio interface(s) for IMT-advanced and invitation to participate in their subsequent evaluation,” Circular Lett. 5/LCCE/2, Mar. 2008.

[2] ITU-R Rep. M.2134, “Requirements related to technical performance for IMT-advanced radio interface(s),”International Telecommunications Union, 2008.

[3] 3GPP TR 36.913, “Requirements for further advancements for evolved terrestrial radio access (E-UTRA),” v. 8.0.1, March 2009.

[4] 3GPP TR 36.814 V1.2.1, “Further advancements for EUTRA: physical layer aspects,” Tech. Spec. Group Radio Access Network Rel. vol. 9, June 2009.

[5] 3GPP TSG RAN WG1 #58 R1-093315, “Comparing relays vs. pico eNBs deployments in coverage limited scenario,” Shenzhen, China, August 24- 28, 2009.

[6] A. Bou Saleh, S. Redana, B. Raaf, and J. Hamalainen, “Comparison of relay and pico eNB deployments in LTE-advanced,” In Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th, Sept. 2009, pp. 1-5.

[7] E. Lang, S. Redana, and B. Raaf, “Business impact of relay deployment for coverage extension in 3GPP LTE-advanced,” Proc. IEEE Int. Conf. Communications Workshops (ICC) 2009, June 2009, pp. 1-5.

[8] F. Jian and B. Wang, “A load balancing relay selection algorithm for relay based cellular networks,” Proc. 7th Int. Conf. Wireless Communications, Networking and Mobile Computing (WiCOM), Sept. 2011, pp. 1-5.

[9] M. M. Fareed and M. Uysal, “A novel relay selection method for decode and forward relaying,” Proc. Canadian Conf. Electrical and Computer Engineering, CCECE 2008, May 2008, pp. 000135-000140.

[10] L. Gkatzikis and I. Koutsopoulos, “Low complexity algorithms for relay selection and power control in interference-limited environments,” Proc. 8th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), June 2010, pp. 278-287.

[11] H. W. Kuhn, “The Hungarian method for the assignment problem,” Nav. Res. Log. Quart. vol. 2, pp. 83-97,March 1955.

[12] L. Wang, Y. Ji and F. Liu, “Joint optimization for proportional fairness in OFDMA relay-enchanced cellular networks,” Proc. Wireless Communications and Networking Conference (WCNC), April 2010, pp. 1-6.

[13] Y. Hu and L. Qiu, “A novel multiple relay selection strategy for LTE-advanced relay systems,” Proc. Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd, May 2011, pp. 1-5.

[14] D. Wu, G. Zhu, D. Zhao, and L. Liu, “Cross-layer design of joint relay selection and power control scheme in relay-based multi-cell networks,” Proc. Wireless Communications and Networking Conference (WCNC), March 2011, pp. 251-256.

[15] A. Bou Saleh, O. Bulakci, S. Redana, B. Raaf, and J. Hamalainen, “Enhancing LTE-advanced relay deployments via biasing in cell selection and handover decision,” Proc. 21st Annual IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC), Sept. 2010, pp. 2277-2281.

[16] Y. Yu, R. Q. Hu, C. S. Bontu, and Z. Cai, “Mobile association and load balancing in a cooperative relay cellular network,” IEEE Commun. Mag. vol. 49, pp. 83-89, May 2011.

[17] T. M. Cover and A. EL Gamal, “Capacity theorems for the relay channel,” IEEE Trans. Inform. Theory, vol. 25, pp. 572–584, Sept. 1979.

[18] S. Sesia, I. Toufik, and M. Baker, Eds., LTE: The UMTS Long Term Evolution. New York: John Wiley and Sons, 2009.