Title

Delay-Aware LTE WLAN Aggregation for 5G Unlicensed Spectrum Usage

Permalink

https://escholarship.org/uc/item/6fs503cg

Authors

Liu, Bin
Zhu, Qi
Zhu, Hongbo

Publication Date

2017-11-16

Peer reviewed

Delay-Aware LTE WLAN Aggregation for 5G Unlicensed Spectrum Usage

Bin Liu Qi Zhu and Hongbo Zhu
Jiangsu Key Laboratory of Wireless Communications,
Nanjing University of Posts and Telecommunications, Nanjing, China
Email: mliubin@hotmail.com, zhuqi@njupt.edu.cn, zhuhb@njupt.edu.cn,

***Abstract*—In 5G heterogeneous evolution, the unlicensed band has captured much attention. Specified by 3GPP Release 13, LTE WLAN aggregation (LWA) is deemed as an effective approach for spectrum integration of 5G heterogeneous networks (HetNet). However, most of previous works about LWA lie in the architecture design, and rarely investigate LWA algorithm analytically. In this paper, we formulate the network access and aggregation problem for delay-tolerant application in multiple slots, and further develop a delay-aware LTE WLAN aggregation algorithm (DLWA) based on dynamic programming, which is aimed to minimize the user payment with QoS requirement. To reduce the complexity, we prove optimal decision policy and simplify the searching space of scheme sets. Simulation results show that, comparing with the current WLAN interworking solutions, the algorithm could lower the payment and achieve high completion probability under specified transmission deadline. The framework presented can support WLAN offloading scheme as well, which enables the best use of unlicensed resource.**

Keywords—heterogeneous network; LTE-WLAN aggregation; WLAN offloading; delay-tolerant; dynamic programming

I. INTRODUCTION

Credited to the popularity of various multimedia services, mobile data has seen unprecedented growth, almost 4000-fold over the past few years [1]. As Cisco forecasts, the trend is continuously soaring, and expected to surpass 24.3 exabytes per month by 2019. Dramatic escalation has imposed great challenges on 5G systems design. Even though advances in cellular technology, such as massive MIMO, millimeter wave (mmWave), could improve the network performance and capacity, these alone are not sufficient to meet the data explosion in 5G [2]. Motivated by the exponentially increasing mobile traffic and limited licensed spectrum, 5G wireless system will inevitably evolve towards heterogeneous architecture.

Recently, much attention has been drawn to the use of unlicensed spectrum, since it could be used to leverage the over-loaded cellular network with its free and abundant band [3]. For the best use of unlicensed band, the Release 13 LWA solution was formally approved by the 3GPP RAN Plenary in March 2016[4]. LWA can be deployed without challenging the existing WLAN infrastructure. With LWA, the two radio access technologies are integrated to offer a compelling service experience [5]. The tactical features make it more prospective for the industry deployment. On 19 August 2016, Singapore's M1 with Nokia announced Singapore's first commercial LWA HetNet rolled out. They expect the LWA could promote the peak rate to more than 1Gbps by 2017[6]. It seems that LWA has bright future for evolution towards 5G heterogeneous integration.

Studies for LWA are illustrated in [7]-[11]. Zhu *et al* verifies the feasibility for carrier aggregation with 256-QAM WLAN and 40 MHz LTE in digital polar transmitter [7]. LWA layer 2 structure proposals illustrated in [8] could be well compatible to existing LTE and WLAN specifications. Moreover, LWA system design in [9] exhibits the aggregation mechanism and infrastructure coordination between the LTE and WLAN infrastructure. Several approaches for LWA system deployments are briefly compared in [10]. Peng *et al* [11] envisions the prospects for LWA, and further extends the mmWave small cell discovery mechanism and beamforming (BF) technology into aggregation. Most of these prior works, however, mainly concentrate on the architecture and viability examination, and barely involve in resource allocation for various applications. Singh *et al* in [12] develops a solution for the optimal proportional fairness aggregation with the backhaul. In addition, it's worth noting that the user payment increases in folds while multi-link boosting transmission rate.

In fact, most applications are delay-tolerant, such as video download, software update, and minutes delay for these applications will not have significant negative impacts on users' satisfactions [13].According to a survey in [14], over half of the respondents are willing to wait for 10 minutes to download videos with monetary incentives. The delay tolerant enables to trade off between data demand and payment.

In this paper, we consider LTE WLAN aggregation for delay-tolerant application, and formulate it as a finite-horizon sequential decision problem. Furthermore, the delay-aware LTE WLAN aggregation algorithm (DLWA) is proposed based on dynamic programming, aiming to minimize data usage payment under deadline constraint. By proving non-decreasing properties of optimal cost equation, we verify the optimal policy and reduce the decision scheme sets space, which could reduce the computational complexity in backward induction. The algorithm allows dynamically to perform LWA and WLAN offloading with consideration of transmission emergency for each slot. This compatible DLWA framework enables the best use of unlicensed resource and supports a smooth transition from deployed systems to the 5G heterogeneous integration. To the best of our knowledge, this is the first paper that studies aggregation algorithm for LWA analytically, which balances the user payment and QoS requirement.

978-1-5090-5932-4/17/$31.00 ©2017 IEEE

The rest of the paper is organized as follows. In Section II, we describe the system model and analyze the LTE WLAN aggregation problem for delay-tolerant application. Then, based on optimal policy, the DLWA algorithm is proposed in Section III. Simulation results are given in Section IV, followed by the conclusion in Section V.

II. System Model and Problem Formulation

A. System Model

The anchor-booster downlink framework for LWA scenario is illustrated as Fig.1, where the LTE eNB is the anchor while the WLAN APs act as the boosters. It should also be noted that alternative architectures to LWA are still under discussion in 3GPP.

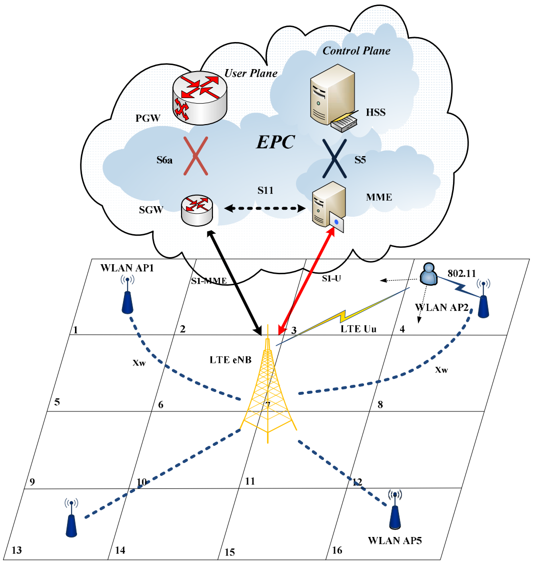

Fig.1 LWA downlink deployment scenario

In the LWA scheme, eNB bears data and control planes and connects to the Core Network (CN) via regular S1 interfaces (S1-C and S1-U). On the downlink, the eNB scheduler decides to send PDCP PDU on LTE or WLAN. The eNB collects UE feedback such as user enquiry information (including data demand and transmission deadline), Channel Quality Indication (CQI), user mobility, and other RRM functionalities for deciding which cells to use and how to schedule data packets. The interface between eNB and WLAN is standardized by 3GPP as X_W. The eNB has backhaul to a WLAN logical entity. Both ideal and non-ideal backhaul situations are considered in this paper. For WLAN offloading scheme in DLWA, the data will be directly transmitted by WLAN connecting to CN by it interfaces, specified for offloading services standardized before Rel-13 (such as ePDG/TWAG).

We consider the large scale fading in downlink channel. The channel from eNB to the n-th user is modeled as $h_n = \kappa d_n^{-\alpha_0}$, where d_n is the distance between n-th user and eNB, and α_0 is the path loss exponent. The achievable rate of n-th user is given by

$$R(n) = B\log_2(1+\frac{P|h_n|^2}{N_0}) \qquad (1)$$

where B is the bandwidth of LTE subcarrier, P denotes the transmitting power of eNB, and N_0 is the power of additive noise.

B. Delay-aware Aggregation Problem Formulation

In this subsection, we investigate LWA problem analytically. Rather than opportunistic aggregation that performs LTE WLAN aggregation whenever possible, delay-aware LTE WLAN aggregation is to lower the user payment by deciding the reasonable access and aggregation strategies.

The UE initiates an application and demand S bits data within T time slots. These cases often occur in our daily life, for example the users want to download a video with size of 750Mbytes in the 10 minutes on their commuting, or share a series of photos with friends by uploading about 50Mbytes data in 2 minutes on their journey.

We try to find the optimal policy to dynamically perform LTE WLAN aggregation to minimize the total cost and guarantee the completion probability. Without loss of generality, we normalize the length of the time slot Δt to be one. The access action is made at each decision epoch $t \in \mathcal{T} = \{1,\ldots,T\}$.

The user state is described as $\boldsymbol{e} = (s,\alpha)$, where $s \in \mathcal{S} \subseteq [0,S]$ denotes the data sequence (in bits) to be transferred. The state element $\alpha \in \mathcal{A} \subseteq [0,A]$ is the location index. The possible areas the user may arrive contain in the set $\mathcal{A} = \{1,\ldots,A\}$. The whole locations set $\mathcal{A}$ is divided into two subsets $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(0)}$ according to whether WLAN is available or not. Based on the past mobility pattern of the UE, a Markovian mobility model can be derived, which has been used in [15].

The action u represents the transmission decision taken by UE at each decision epoch. To be specific, we have action set as $\mathcal{U} = \{0,1,2,3\}$, where u=0 means that the UE will remain idle without data transmission, u=1 means that transmits through LTE only, u=2 indicates that the WLAN offloading will be performed, and u=3 denotes that the transmission will be in LTE WLAN aggregation. Apparently, actions u=0 and u=1 are always available in whole scenario, whereas the action u=2 and u=3, however, is only chosen a location $\alpha \in \mathcal{A}^{(1)}$. Thus, the action u depends on the location α, $u \in \mathcal{U}^{(\alpha)} \subseteq \mathcal{U}$, where $\mathcal{U}^{(\alpha)}$ is the set of valid actions set at location α:

$$\mathcal{U}^{(\alpha)} = \begin{cases} \{0,1\}, & \text{if } \alpha \in \mathcal{A}^{(0)} \\ \{0,1,2,3\}, & \text{if } \alpha \in \mathcal{A}^{(1)} \end{cases} \qquad (2)$$

Usage-base charging policy is adopted here, which means the uses' data payment is proportional to its data usage quantity. Let $\gamma(\alpha,u)$ be the price per unit of use for choosing action $u \in \mathcal{U}^{(\alpha)}$ at location α .Let $R(\alpha,u)$ be the estimated throughput of the user at location α with action $u \in \mathcal{U}^{(\alpha)}$, especially $R(\alpha,0)=0$, $\forall \alpha \in \mathcal{A}$ when u=0 and the UE in idle state. Thus, with backhaul latency time $t_b(\alpha,u)$, the payment with action $u \in \mathcal{U}^{(\alpha)}$ at time slot $t \in \mathcal{T}$ is

$$m_t(\boldsymbol{e},u) = m_t(s,\alpha,u) = \min\left\{s, R(\alpha,u)\left(\Delta t - t_b(\alpha,u)\right)\right\} \cdot \gamma(\alpha,u) \tag{3}$$

For the QoS requirement, we define the penalty for not being able to finish the file transfer before deadline. The penalty at state $\boldsymbol{e}$ is

$$\overline{m}_{T+1}(\boldsymbol{e}) = \overline{m}_{T+1}(s,\alpha) = \chi(s) \tag{4}$$

where $\chi(s)$ is a non-negative function decreasing with s and $\chi(0)$=0. The penalty holds upon the deadline arrival (from $T+1$ time slot) when the QoS can't be guaranteed. The chosen of $\chi(s)$ is based on the application's sensitivity to deadline.

The user state transition probability is described as $p(\boldsymbol{e}' \mid \boldsymbol{e},u) = p\left((s',\alpha'),u \mid (s,\alpha),u\right)$,which denotes the probability of entering $\boldsymbol{e}'$=(s',α') in the next time slot if action u is taken at state $\boldsymbol{e}$=(s,α). The mobility of user is independent with the remaining sequence s and action u, we have

$$p(\boldsymbol{e}' \mid \boldsymbol{e},u) = p\left((s',\alpha'),u \mid (s,\alpha),u\right) = p\left(\alpha' \mid \alpha\right) p\left(s' \mid (s,\alpha),u\right) \tag{5}$$

where $p\left(s' \mid (s,\alpha),u\right)$=$\begin{cases} 1 & \text{if } s' = \left[s - R(\alpha,u)\left(\Delta t - t_b(\alpha,u)\right)\right]^+ \\ 0 & \text{otherwise} \end{cases}$

and $[a]^+$=$\max\{0,a\}$. The probability for UE to move from location α to location α' is $p\left(\alpha' \mid \alpha\right)$, and it is obtained from the estimation of the user's past mobility pattern.

Since the process involves multiple time slots, the delay-aware aggregation is a finite-horizon Markov sequence decision problem [16]. We formulate transmission decision function at state $\boldsymbol{e}$=(χ,α) and slot t as $\xi_t : \mathcal{S} \times \mathcal{A} \to \mathcal{U}$. The decision policy set $\pi = \{\xi_t(s,\alpha), \forall s \in \mathcal{S},\ \forall \alpha \in \mathcal{A},\ \forall t \in \mathcal{T}\}$ is defined for action in each state and slot.

With the consideration of payment and deadline, the objective function is established as:

$$\min_{\pi \in \Pi} E_{e_1}^{\pi}\left[\sum_{t=1}^{T} m_t(\boldsymbol{e}_t^{\pi}, \xi_t(s_t^{\pi},\alpha_t^{\pi})\ ) + \overline{m}_{T+1}(\boldsymbol{e}_{T+1}^{\pi})\right] \tag{7}$$

$$\text{Subject to} \qquad \xi_t = \{0,1,2,3\}$$

$$\boldsymbol{e}_1 = (S,\alpha_1)$$

where, $\boldsymbol{e}_t^{\pi}$=$(s_t^{\pi},\alpha_t^{\pi})$ represents the state when the policy π is chosen at time slot t, The feasible π contains in the set Π. $E_{e_1}^{\pi}$ indicates an the expectation function , which is related to probability distribution in the user mobility model. The objective function starts with initial location α_1 in at $t=1$ in state $\boldsymbol{e}_1$=(S,α_1).

For simplicity, we assume that the total use payment and penalty have equal weights. If we put a larger weight on the penalty, then the completion probability of file transmission would increase with more payments.

III. Delay-aware LTE and WLAN Aggregation Algorithm

In this section, we solve finite-horizon Markov sequence aggregation problem in Section II based on dynamic programming (DP). With balance in payment and QoS, the delay-aware LTE WLAN aggregation algorithm is developed for the optimal transmission and aggregation policy. Further, to simplify the searching space and reduce the induction complexity, we prove that optimal cost equation is non-decreasing both in dimension of remaining data sequence and transmission time and present the optimal policy with these properties.

The minimum expected cost (including the penalty) from decision epoch t to $T+1$ is $w_t(e)$ .The optimal equation [16] can well describe the minimum expected cost at each decision step at $t \in \mathcal{T}$. Thus, the optimal cost equation in state $\boldsymbol{e}$ is

$$w_t(\boldsymbol{e}) = w_t(s,\alpha) = \min_{u \in \mathcal{U}^{(\alpha)}} \{\mu_t(s,\alpha,u)\} \tag{8}$$

where for $s \in \mathcal{S}$, $\alpha \in \mathcal{A}$ and $u \in \mathcal{U}^{(\alpha)}$

$$\begin{aligned} \mu_t(s,\alpha,u) &= m_t(s,\alpha,u) + \sum_{\alpha' \in \mathcal{A}} \sum_{s' \in \mathcal{S}} p\left((s',\alpha') \mid (s,\alpha),u\right) w_{t+1}(s',\alpha') \\ &= \min\left\{s, R(\alpha,u)\left(\Delta t - t_b(\alpha,u)\right)\right\} \cdot \gamma(\alpha,u) \\ &\quad + \sum_{\alpha' \in \mathcal{A}} p\left(\alpha' \mid \alpha\right) w_{t+1}\left(\left[s - R(\alpha,u)\left(\Delta t - t_b(\alpha,u)\right)\right]^+, \alpha'\right) \end{aligned} \tag{9}$$

In the equation, the total cost from t to $T+1$ is divided into two parts: the current payment for the action u taken in the epoch t and the expected cost for the remaining time slots by choosing action u . The second equation is derived by substituting (3)-(6) into the former one. Immediately after the deadline $t=T+1$, the equation only contains the penalty function. It is can be seen as a boundary constraint

$$w_{T+1}(\boldsymbol{e}) = \overline{m}_{T+1}(s,\alpha) = \chi(s), \quad \forall s \in \mathcal{S}\ ,\ \forall \alpha \in \mathcal{A} \tag{10}$$

The following assumptions are presented for proof of optimal policy:

Assumption 1: (a) The penalty function $\chi(s)$ is convex and non-decreasing in remaining sequence s; (b) Free WiFi can be deemed as the incentive to delay tolerance (i.e., $\gamma(\alpha,2)$=0, $\forall \alpha \in \mathcal{A}^{(1)}$), and it is also applicable since free WiFi is not hard to find in resident blocks, offices, or shopping centers. (c) The LTE use price per unit is location independent (i.e., $\gamma(\alpha,1) = \gamma(\alpha',1)$, $\forall \alpha,\alpha' \in \mathcal{A}^{(1)}$, $\alpha \neq \alpha'$); (d) WiFi data rates are location-independent. Thus, the throughput for $\forall \alpha \in \mathcal{A}^{(1)}$ is H_2=$R(\alpha,2)\Delta t$. (e) With the large scale fading in transmission in LTE, the throughput in the same block area is identical. Furthermore, for the block set at same distance to eNB A_1=$\{2,3,5,8,9,12,14,15\}$,A_2=$\{1,4,13,16\}$,A_3=$\{6,7,10,11\}$, UE could have same throughput. Thus, the throughput of LTE can be quantized as several values in the set $H_1 = H(\alpha,1)$=$R(\alpha,1)\Delta t$. (e) The latency for non-idea backhaul is location-independent. (i.e. t_b=$t_b(\alpha,u)$). Throughput by aggregation is H_3=$H(\alpha,3)$=$\left[R(\alpha,1)+R(\alpha,2)\right](\Delta t-t_b)$ (f).We approximate $\min\{s,R(\alpha,1)\Delta t\}$ in (3) by the throughput H_1 for action $u=1$ and $\min\{s,R(\alpha,3)(\Delta t-t_b)\}$ in (3) by H_3 for action $u=3$. Specially, H_3=H_1+H_2 is for ideal backhaul case in aggregation. For the same expression, the idle action $u=0$ has H_0=$H(\alpha,0)=R(\alpha,0)\Delta t=0$.

With these assumptions, the payment is only produced with LTE usage.

Let l to be an indication for LTE network usage.

$$l=l(u)=\begin{cases}1 & u=1 \text{ for } \alpha\in\mathcal{A}^{(0)} \text{ or } u=3 \text{ for } \alpha\in\mathcal{A}^{(1)}\\ 0 & u=0 \text{ for } \alpha\in\mathcal{A}^{(0)} \text{ or } u=2 \text{ for } \alpha\in\mathcal{A}^{(1)}\end{cases} \tag{12}$$

Namely, l=1 when the action for cellular alone $u=1$ or in $u=3$ LTE WLAN aggregation. Thus, the payment function at state e with action u at time slot t could be modified from (3) as

$$m_t(\boldsymbol{e},u)=\delta(l=1)\cdot\gamma_u=\begin{cases}\gamma_u & \text{if } l=1\\ 0 & \text{otherwise,}\end{cases} \quad u=1,3 \tag{13}$$

where $\forall\alpha\in\mathcal{A}^{(\alpha)}$, $\delta(\cdot)$ is the indicator function, and γ_1=$\gamma(\alpha,1)R(\alpha,1)\Delta t$, the payment for aggregation only contains the use of LTE γ_3=$\gamma(\alpha,1)R(\alpha,1)(\Delta t-t_b)$. Then, $\mu_t(s,\alpha,u)$ in (9) can be rewritten as

$$\mu_t(s,\alpha,u)\text{=}\delta(l=1)\cdot\gamma_u+\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-R(\alpha,u)]^+,\alpha') \tag{14}$$

A. Properties of the Optimal Policy

First, we present related properties in optimal cost equation for verifying the optimal policy under assumptions above.

Proposition 1: the properties of optimal cost equation $w_t(s,\alpha)$:

(a) $w_t(s,\alpha)$ is a non-decreasing with s ,$\forall\alpha\in\mathcal{A}$, $\forall t\in\mathcal{T}$.

(b) $w_t(s,\alpha)$ is a non-decreasing with t , $\forall s\in\mathcal{S}$, $\forall\alpha\in\mathcal{A}$

The proof of *Proposition* 1 is given in Appendix A.

Since the unlicensed band has monetary incentives, the WiFi is preferred to reduce the payment for $\alpha\in\mathcal{A}^{(1)}$. The optimal policy can be intuitive concluded in *Proposition* 2, which is proved in Appendix B.

Proposition 2: For location $\alpha\in\mathcal{A}^{(1)}$, we have:

(a) If $H(\alpha,2)\le H(\alpha,3)$, for $\forall s\in\mathcal{S}$,$\forall t\in\mathcal{T}$

$$\mu_t(s,\alpha,0)\ge\mu_t(s,\alpha,2)\text{, }\mu_t(s,\alpha,1)\ge\mu_t(s,\alpha,3)\text{ ,}$$

(b)If $H(\alpha,3)\le H(\alpha,2)$ for non-ideal backhaul, then $\xi_t^*(s,\alpha)$=2

Proposition 2 (a) indicates that WiFi always has a higher priority if free charged. (b) If aggregation is not ideal and satisfied, for example the latency or the aggregation deterioration is heavy, the optimal policy should WLAN offload and UE directly connect to the Internet with AP interfaces.

Notice, even if the WiFi is not free, the property still holds with the incentive to use unlicensed resources. With theses properties, the optimal action sets exclude the action u=0 and u=1 for lower action cost. The set of optimal policy could be reduced as:

$$\tilde{\mathcal{U}}^{(\alpha)}=\begin{cases}\{2,3\}, & \text{if } \alpha\in\mathcal{A}^{(1)}\\ \{0,1\}, & \text{if } \alpha\in\mathcal{A}^{(0)}\end{cases} \tag{15}$$

The optimality equation in (8) thus simplified as

$$w_t(s,\alpha)=\min_{u\in\mathcal{U}^{(\alpha)}}\{\mu_t(s,\alpha,u)\}\text{=}\min_{u\in\tilde{\mathcal{U}}^{(\alpha)}}\{\mu_t(s,\alpha,u)\} \tag{16}$$

B. The optimality of Delay-Aware LTE WLAN Aggregation (DLWA) Algorithm

Theorem: If the optimal equation holds

$$\xi_t^*(s,\alpha):=\arg\min_{u\in\tilde{\mathcal{U}}^{(\alpha)}}\{\mu_t(s,\alpha,u)\} \tag{17}$$

Thus, the policy $\pi^*=\{\xi_t^*(s,\alpha),\forall s\in\mathcal{S}, \forall\alpha\in\mathcal{A}, \forall t\in\mathcal{T}\}$ is the optimal solution of problem (7).

Proof: The optimal equation (11) is the subproblems for the entire optimization problem in (7). The optimal policy π^* is a contingency set that contains information about the optimal decision at all the possible states (s,α) in any time slot $t\in\mathcal{T}$, and the system computes it before transmission. With Principle of Optimality [17]: A problem is said to satisfy the Principle of Optimality if the subsolutions of an optimal solution of the problem are themselves optimal solutions for their subproblems, we thus derive that π^* is the optimal solution of problem (7) only if each element in π^* is optimal subsolutions for (11).

TABLE I. ALGORITHM DESCRIPTION

Algorithm 1 *Delay-Aware LTE WLAN Aggregation (DLWA) Algorithm*

```
Plan Phase (for eNB):
Set $m_{T+1}(s,\alpha)$ , $\forall s\in\mathcal{S}$ , $\forall\alpha\in\mathcal{A}$ using (10)
Set $t:=T$ and begin plan in recursive backward
while $t>1$
    for $\alpha\in\mathcal{A}$
        Set $s:=0$
        while $s\le S$
            Calculate $\mu_t(s,\alpha,u)$ , $\forall u\in\mathcal{U}^{(\alpha)}$ using (9)
            Set $\xi_t^*(s,\alpha):=\arg\min_{u\in\mathcal{U}^{(\alpha)}}\{\mu_t(s,\alpha,u)\}$
            Set $w_t(s,\alpha)=\mu_t(s,\alpha,\xi_t^*(s,\alpha))$
            Set $s:=s+\sigma$
        end while
    end for
    Set $t:=t-1$
end while
Output the optimal decision policy $\pi^*$ and announce UE
    for receiving preparation according to the policy, and
    schedule the AP via Xw for possible packet transmission
```

The solution to the optimal equation at epoch t is optimal value from the epoch t to the $T+1$ [16]. With optimal policy, the DLWA algorithm is illustrated in the Table I. The granularity of the discrete state element s is σ and obviously $\sigma > 0$ (such as 1 Mbits). The optimal policy π^* is obtained by backward induction in solving the objective function in (7) with the optimality equation in (8) and the boundary constraint in (10). The proposition in [17 pp.86] clarifies the optimal policy can be derived from the solution to the optimal equation from slot $t=1$ and backward to $t=T$. Then, the recursive backward update from time slot $t=T$ to time slot $t=1$ gives the optimal decision $\xi_t^*(s,\alpha)$ and payment value $m_t(s,\alpha)$. The latter is to schedule eNB and AP with UE for transmission.

IV. Performance Evaluation

In this section, the DLWA algorithm is evaluated with the following benchmarks (1) WLAN Preferred (WP), (2) 3GPP Release 12 WLAN interworking solution (Rel-12), (3)Always Aggregation. In WLAN Preferred scheme, the UE always connects to the WLAN whenever available. In Rel-12, the UE associates to the WLAN AP only when the rate in LTE below a certain threshold. The optimal value for this threshold is empirically found for the comparisons. To maximum the transmission rate, Always Aggregation is also investigated. Always Aggregation is a kind of opportunistic LTE WLAN aggregation scheme, which uses all the resource available, including LTE and WLAN, and performs aggregation as much as possible.

The WLAN APs locate randomly in the $A=16$ possible blocks as shown in Fig.1, whose random distribution is in deployment density ρ=0.4 .The density of WLAN reflects the sum capacity of unlicensed band usage. The block is the ideal coverage of AP and can only deploy a single AP. We assume no overlap and interference between the neighboring APs. We run simulations 1000 times with randomized AP locations and the user mobility trajectories, and give the average results. The evaluation is compared in term of total cost, completion probability, and payment.

The parameter setting is: $N_0=-80$ dBm , κ=0.1 , α_0=2 , $p(\alpha'|\alpha)$=0.6 , the latency time t_b=20ms for non-ideal backhaul, the granularity σ=10Mbits the block is a square with the length of 100m. The length of time slot Δt is one second. The rate for WLAN is 80Mbps .Due to the contention mechanism in 802.11, the probability of WiFi connection at the available location is normalized as 0.5. The price for LTE data usage is .We take the convex function as a penalty for not finishing a task as

$$\chi(s)=10s^2, \quad \forall s\in \mathcal{S} \tag{18}$$

TABLE II. SIMULATION ASSUMPTIONS

LTE	
Notation	Description
LTE Carrier Frequency	2 GHz
Channel/UE speed	[IMT] UMa Macro, UE speed= 3 km/hr
LTE mode	Downlink FDD; 20 MHz for DL
Transmission power	25dBm
No. antennas (LTE macro,UE);	(2, 2)
Antenna configuration	small cell &UE: co-polarized,
Max rank per UE	2 (SU-MIMO)
Feedback/control channel errors ;UE channel estimation	No Error ; Ideal
PHY Abstraction	Mutual Information
Scheduler ; Scheduling granularity	Proportional-Fair Scheduler ; 5 PRBs
Receiver type	Interference unaware MMSE
Feedback periodicity	10ms
CQI & PMI feedback granularity in frequency ;	5 PRBs
PMI feedback	3GPP Rel-10 LTE codebook (per sub-band)
Outer loop for target FER control	10% PER for 1st transmission
Link adaptation	MCSs based on LTE transport Format
HARQ scheme	Chase Combining (CC)
WLAN	
WiFi deployment	IEEE 802.11n based APs, no overlap
WiFi Frequency, Channelization	2.4 GHz band, 3 frequency bands, 20 MHz channels; least power based channel selection
AP Transmit power	20dBm outdoor
WiFi mode	Downlink only.
Scheduler	Proportional Fair and Round Robin
TX-OP;PHY Abstraction ; MPDU Size	1ms;RBIR ; 1500 Bytes

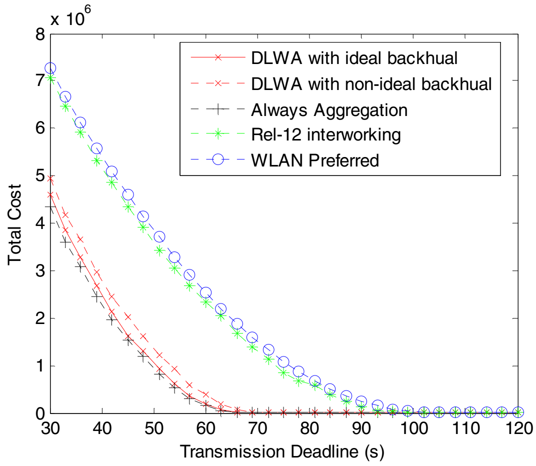

Fig.2 Total cost versus deadline

Firstly, we evaluate the performance of four schemes in short deadline requirement. Total cost and completion probability for transferring data sequence $S=1.2$Gbytes are illustrated in Fig.2 and Fig.3 respectively.

In, Fig.2 and Fig.3, the Always Aggregation has the lowest total cost and high completion probability when deadline is short, since it always uses all the resource it could get. Notice that the low cost advantages result from not being penalized rather than fee reduction. Actually, while high throughput it may get, it is expensive to use LTE all the time as shown in Fig.4.and can scarcely leverage the cellular load. Both the WP and Rel-12 solutions enable to UEs associating with either LTE or WLAN, and no traffic aggregation is done. They can alleviate traffic with unlicensed band in some way, but they often incurs a penalty for violating the deadline since not aware of the deadline. The QoS can't be guaranteed by poor completion probability especially when the deadline is tight.

In contrast, the DLWA dynamically arranges the access with both consideration of payment and the QoS requirement. When challenged by deadline, it has a high completion approximately to the Always Aggregation with much lower payment. With the best use of unlicensed band, DLWA enables cellular network to interwork with WLAN in a more reasonable way.

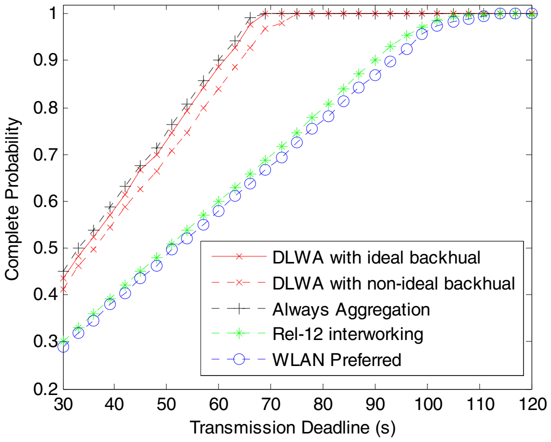

Fig.3 Probability of completion transfer versus deadline

Furthermore, we compare the performance with low deadline requirement for the same task $S = 1.2$Gbytes in Fig.4 and Fig.5. As shown in Fig.4, payment in DLWA is lowest and reduces with the larger time-tolerance, because the low QoS requirement in time sensitivity allows more slots for WLAN access. On the other side, larger time tolerance benefits little for payment reduction, since the WP and Rel-12 are heuristic schemes without awareness of deadline.

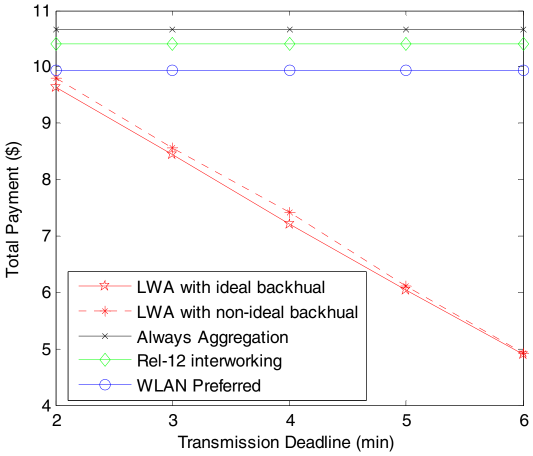

Fig.4 Total data usage payment versus deadline

In Fig.5, we give the performance of LWA with different WLAN deployment density. With the increase deployment density, the unlicensed spectrum can be utilized in more efficiency way, and leverage more data from cellular network at cost of more APs deployment.

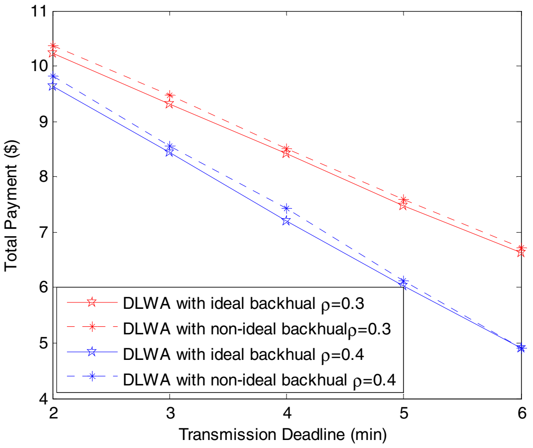

Fig.5 Payment for LWA under different WLAN deployment density

V. Conclusion

LWA is prospective for the smooth evolution toward 5G heterogeneous integration, since it compatible with current wireless network systems. In this paper, we investigate the unlicensed spectrum usage for 5G HetNet, and present a delay-ware LTE WLAN aggregation algorithm, which could lower the user payment in aggregation with QoS guarantee. DLWA can be also well compatible to WLAN offloading strategy, which enables to make the best use of unlicensed resource. To the authors' best knowledge, this is the first work to propose and demonstrate LWA algorithm analytically for delay-tolerant application. Future work could emphasize congestion and fairness in LTE and WLAN interworking and aggregation.

Acknowledgment

The work is supported by Natural Science Foundation of China (No.61571234, 61631020), National Program on Key Basic Research Project of China (No. 2013CB329005), and the Graduate Student Innovation Plan of Jiangsu Province (KYLX16_0659).

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update: 2015-2020," February, 2016.

[2] M. R. Palattella M. Dohler, A. Grieco *et al.*, "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 510-527, March 2016.

[3] O. Galinina, A. Pyattaev, S. Andreev, M. Dohler and Y. Koucheryavy, "5G Multi-RAT LTE-WiFi Ultra-Dense Small Cells: Performance Dynamics, Architecture, and Trends," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1224-1240, June, 2015.

[4] 3GPP RAN RP-1606003 "New Work Item on Enhanced LWA, Intel, Qualcomm, China Telecom,", Sweden, March. 2016.

[5] 5G Americas, "LTE Aggregation & Unlicensed Spectrum" White Paper, September. 2015.

[6] M1, "M1, Nokia announce Singapore’s first commercial nationwide HetNet rollout". August 2016.

[7] Q. Zhu, S. Yu, S. Wang, L. Huang, *et al*. "A digital polar transmitter with DC-DC converter supporting 256-QAM WLAN and 40 MHz LTE-A carrier aggregation," *IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, San Francisco, CA, 2016, pp. 198-201.

[8] Y. Ohta, N. Michiharu, S. Aikawa and T. Ode, "Link layer structure for LTE-WLAN aggregation in LTE-Advanced and 5G network," *IEEE Conference on Standards for Communications and Networking (CSCN)*, Tokyo, 2015, pp. 83-88.

[9] A. Rasha, H. Artail, and M. G. David. "LTE-WiFi carrier aggregation for future 5G systems: A feasibility study and research challenges." *Procedia Computer Science* , 2014, no.34, pp.133-140.

[10] P. Nuggehalli, "LTE-WLAN aggregation," *IEEE Wireless Communications*, vol. 23, no. 4, pp. 4-6, August 2016.

[11] H. Peng, K. Moriwaki and Y. Suegara, "Macro-Controlled Beam Database-Based Beamforming Protocol for LTE-WiGig Aggregation in Millimeter-Wave Heterogeneous Networks," *IEEE Vehicular Technology Conference (VTC Spring)*, Nanjing, 2016, pp. 1-6.

[12] S. Singh, M. Geraseminko, S. p. Yeh, N. Himayat and S. Talwar, "Proportional Fair Traffic Splitting and Aggregation in Heterogeneous Wireless Networks," *IEEE Communications Letters*, vol. 20, no. 5, pp. 1010-1013, May 2016.

[13] Y. Im, C. Joe-Wong, S. Ha, S. Sen, T. Taekyoung Kwon and M. Chiang, "AMUSE: Empowering Users for Cost-Aware Offloading with Throughput-Delay Tradeoffs," *IEEE Transactions on Mobile Computing*, vol. 15, no. 5, pp. 1062-1076, May 2016.

[14] S. Sen, C. Joe-Wong, S. Ha and M. Chiang, "Incentivizing time-shifting of data: a survey of time-dependent pricing for internet access," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 91-99, November 2012.

[15] A. Nadembega, A. Hafid and T. Taleb, "A Destination and Mobility Path Prediction Scheme for Mobile Networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, pp. 2577-2590, June 2015.

[16] M. L. Puterman, “Markov Decision Processes: Discrete Stochastic Dynamic Programming,” New York, NY: John Wiley and Sons, 2005.

[17] D. P. Bertsekas, “Dynamic Programming and Optimal Control” vol. 1, 3rd ed. Athena Scientific, 2005.

Appendix

A. Proof of Proposition 1

(a) The proof *Proposition* 1 is in the induction method.

At the end of transmission ($T+1$ slot), the total cost only contains the penalty $w_{T+1}(s,\alpha)=\chi(s)$. Thus, $w_{T+1}(s,\alpha)$ is non-decreasing in s . Further, we suppose that $w_{t+1}(s,\alpha)$ is a non-decreasing function in s for $\forall\alpha\in\mathcal{A}$. Then, the cost function in (9) is rewritten as

$$\mu_t(s,\alpha,u)=\delta(l=1)\gamma_u+\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,u)]^+,\alpha') \quad (19)$$

if $p(\alpha'|\alpha)\geq 0$, $\forall\alpha,\alpha'\in\mathcal{A}$ and $\delta(l=1)\gamma_u$ is independent of s , then $\mu_t(s,\alpha,u)$ is non-decreasing function in s .From (8), the minimum of $\mu_t(s,\alpha,u)$ is $w_t(s,\alpha)$, and the minimum property will not change the monotonicity about s . Thus, $w_t(s,\alpha)$ is also the non-decreasing function in s .

(b) Prove it by induction. Firstly, we give the proof for propriety in penalty boundary is

$$\begin{aligned} w_T(s,\alpha)&=\min_{u\in\mathcal{U}^{(\alpha)}}\{\mu_T(s,\alpha,u)\}\leq\mu_T(s,\alpha,0)\\ &=\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{T+1}(s,\alpha')=\chi(s)=w_{T+1}(s,\alpha)\end{aligned} \quad (20)$$

The first equation holds by the definition of cost function $w_t(s,\alpha)$ in (8). From the expansion of (14), we get the third one.

Secondly, suppose that $\forall s\in\mathcal{S}$, $\forall\alpha\in\mathcal{A}$, $w_{t+1}(s,\alpha)$ is a non-decreasing with t .

In (19), the $p(\alpha'|\alpha)\geq 0$, $\forall\alpha,\alpha'\in\mathcal{A}$ and $\delta(l=1)\gamma_u$ is independent of t . Therefore, $\mu_t(s,\alpha,u)$ is non-decreasing in t . From (8), the minimum of $\mu_t(s,\alpha,u)$ is $w_t(s,\alpha)$, and the minimum property will not change its monotonicity about t . Thus, $w_t(s,\alpha)$ is also the non-decreasing function in t .

B. Proof of Proposition 2

For any $s\in\mathcal{S}$ and $\alpha\in\mathcal{A}$ be given.

(a) With *Proposition 1*, we firstly to prove

$\mu_t(s,\alpha,1)\geq\mu_t(s,\alpha,3)$ when $H(\alpha,2)\leq H(\alpha,3)$

$$\begin{aligned} &\mu_t(s,\alpha,1)=\gamma_1+\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,1)]^+,\alpha)\\ &\geq\gamma_1+\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,1)-H(\alpha,2)]^+,\alpha')\\ &\geq\gamma_3+\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,3)]^+,\alpha')\\ &=\mu_t(s,\alpha,3)\end{aligned} \quad (21)$$

where the first equality is by (14) and the inequality is based on the properties that the $w_{t+1}(s,\alpha)$ is non-decreasing function in s .

Similarly, the proof for the second is:

$$\begin{aligned} \mu_t(s,\alpha,0)&=\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}(s,\alpha)\\ &\geq\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,2)]^+,\alpha')=\mu_t(s,\alpha,2)\end{aligned} \quad (22)$$

(b) If $R(\alpha,3)(\Delta t-t_b)\leq R(\alpha,2)\Delta t$, we have $H(\alpha,3)\leq H(\alpha,2)$

$$\begin{aligned} \mu_t(s,\alpha,3)&=\gamma_3+\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,3)]^+,\alpha')\\ &\geq\sum_{\alpha'\in\mathcal{A}}p(\alpha'|\alpha)w_{t+1}([s-H(\alpha,2)]^+,\alpha')=\mu_t(s,\alpha,2)\end{aligned} \quad (23)$$

Combined with the *Proposition 2* (a), the optimal action can be obtained by $\xi_t^*(s,\alpha)=2$, $\forall s\in\mathcal{S}$, $\forall t\in\mathcal{T}$.