QoS-Aware Load Balancing in Dense Cellular Networks with Dynamic User Traffic

Shuxin He, Tianyu Wang, and Shaowei Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
Email: MG1723062@smail.nju.edu.cn, {tianyu.alex.wang, wangsw}@nju.edu.cn

***Abstract*—Due to the dense deployment of small cells, the number of mobile users served by each access point is decreasing dramatically, which leads to an increasingly dynamic cell load distribution over time and space. In order to match dynamic traffic load with static infrastructure capacity, a variety of load balancing methods have been proposed. However, the existing load balancing approaches either run self-tuning algorithms to optimize local handover parameters, which may not be optimal for the network-wide performance, or formulate a static optimization problem for a "snapshot" network, which may require a huge amount of handover signaling in dynamic situations. In this paper, we consider the load balancing problem with dynamic traffic models, in which the average throughput over a certain time period is maximized, while the average packet delay is guaranteed to be below a certain threshold. Simulation results show that our proposed user association and resource allocation algorithm can highly increase the average throughput, compared with the baseline algorithm using the maximum SINR association and equal resource allocation.**

I. Introduction

The explosive growth of wireless traffic, especially enhanced mobile broadband services, requires dense deployment of heterogeneous infrastructures with small communication range. Unlike the traditional macro cells, small cells (e.g., microcells, picocells, and femtocells) usually cannot cover enough users to provide a statistical multiplexing effect, which leads to the highly dynamic traffic load in small access points over time and space [1], [2]. Thus, load balancing, i.e., matching the dynamic traffic demand of mobile users with the limited capacity supply of infrastructures, is recognized to be more important and more challenging in dense small cell networks as compared to macro cell networks [3]–[8].

Load balancing approaches can be roughly classified into handover optimization approaches based on self-tuning algorithms [9]–[13], and joint user association and resource allocation approaches based global optimization [14]–[19]. In [10] and [11], Markov decision process based algorithms are proposed for optimal vertical handover in heterogenous networks, in which load balancing performance is evaluated by the "connection reward" of each handover. In [12] and [13], the authors investigate the conflict between load balancing and robust handover, which intend to improve different performance indicators by tuning overlapping handover parameters, and propose distributed self-tuning algorithms to improve the overall handover performance, including the link failure ratio, the handover failure ratio, and the ping-pang handover ratio. However, handover optimization approaches are usually based on a heuristic self-tuning algorithm, which is lack of theoretic analysis and may not be optimal for the network-wide performance. Moreover, the variance of cell load is completely determined by the mobility of users as a constant data rate is always assumed, which is not suitable for small cell networks with highly dynamic user traffic.

Joint user association and resource allocation approaches formulate an optimization problem for the global network performance, in which various objectives and constraints are considered for different scenarios. In [14], the authors provide a low-complexity distributed algorithm with a theoretical performance guarantee to maximize the total utility of a heterogenous network. In [15], the objective function is defined to be related to the rate requirements of user traffic, and a classic gradient decent algorithm is utilized to give the optimal solution. In [17], the objective is to minimize the maximum load of access points in a millimeter-wave communication network, and a distributed algorithm based on Lagrangian duality theory is proposed to provide an asymptotically optimal solution. In [18], the authors consider the signaling load of the association between small base stations (SBSs) and device-to-device pairs, for which a cost function is defined and a heuristic algorithm is proposed. In [19], the authors introduce an efficient cell planning method from the viewpoint of workload balancing, where the service area is divided into multiple subregions with almost equal traffic demand. In [20], a heuristic QoS-Aware resource allocation algorithm is proposed to maximize the number of users and the overall Device-to-device (D2D) throughput gain while the data rate demands are satisfied. In [21], resource sharing scheme for D2D communication underlaying cellular network is investigated, where a subchannel can be shared by a CU and multiple D2D pairs to exploit the spatial reuse of D2D pairs to improve the spectrum efficiency of the considered cellular system. However, the existing user association and resource allocation approaches usually consider a "snapshot" network, in which the user traffic and network parameters are assumed to be unchanged. Therefore, their proposed optimization methods may not be optimal in dynamic situations, or lead to a huge amount of signaling overhead.

In this paper, we consider the QoS-aware load balancing

This work was partially supported by the National Natural Science Foundation of China (61671233), the Jiangsu Science Foundation (BK20151389, BK20170650), the Postdoctoral Science Foundation of China (BX201700118, 2017M621712), and the Jiangsu Postdoctoral Science Foundation(1701118B).

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

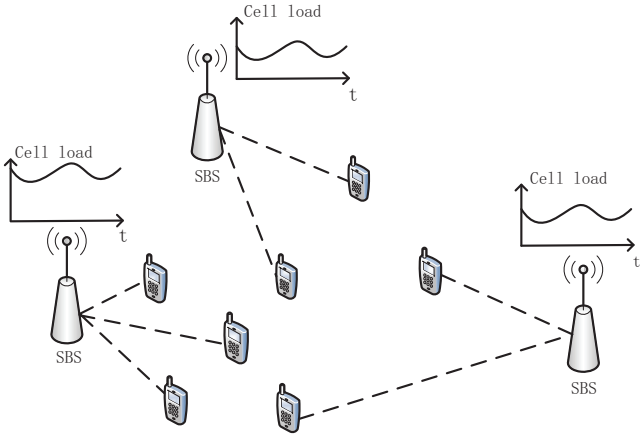

Fig. 1. Load balancing in dense small cell networks with dynamic user traffic.

problem in small cell networks with dynamic user traffic [23]. Specifically, the user traffic is modeled as a Poisson process instead of using a constant data rate, and we consider the average throughput during a certain time period, instead of an instantaneous throughput of a network snapshot. The QoS requirement is given by the average packet delay, which is constrained to be below a certain threshold. We formulate a mixed integer optimization problem and propose a suboptimal solution for the user association and resource allocation. Simulation results show that our proposed algorithm outperforms the baseline algorithm using maximum SINR association and equal resource allocation.

The rest of the paper is organized as follows. In Section II, we formulate the dynamic load balancing problem based on queuing theory. In Section III, we propose a lightweight user association and resource allocation algorithm to give suboptimal solutions. Simulation results are analyzed in Section IV, and conclusions are given in Section V.

II. System Model

Aa shown in Fig. 1, we consider a small cell network consisting of N SBSs, the set of which is denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$, and M mobile users, the set of which is denoted by $\mathcal{M} = \{1, 2, \ldots, M\}$. We denote by $x_{i,j}$ as the association indicator for SBS $i \in \mathcal{N}$ and user $j \in \mathcal{M}$, which is defined as

$$x_{i,j} = \begin{cases} 1 & \text{user } j \text{ is associated with SBS } i, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We assume that each mobile user be associated with at most one SBS and each SBS can serve unlimited number of mobile users, i.e.,

$$\sum_{i \in \mathcal{N}} x_{i,j} \leq 1, \forall j \in \mathcal{M}. \tag{2}$$

We assume that all SBSs serve mobile users with a constant transmit power and the total bandwidth W. We denote by $w_{i,j}$ as the bandwidth that SBS $i \in \mathcal{N}$ allocates to user $j \in \mathcal{M}$. Thus, the downlink transmission rate between SBS i and user j is given by

$$r_{i,j} = w_{i,j} \log \left[1 + \frac{P_0 (d_0/d_{i,j})^\alpha}{N_0 w_{i,j}} \right], \tag{3}$$

in which P_0 is the signal power received by a mobile user from a SBS at the reference distance d_0, $d_{i,j}$ is the distance between user j and SBS i, α is the pathloss exponent, and N_0 is the noise power spectral density. And we have the bandwidth constraint

$$\sum_{j \in \mathcal{M}} w_{i,j} \leq W, \forall i \in \mathcal{N}. \tag{4}$$

For any user $j \in \mathcal{M}$, we assume that the data packets arrive at a Poisson process with the arrival rate λ_j, and the packet size follows a negative exponential distribution with the mean B. If user j is associated with SBS i, i.e., $x_{i,j} = 1$, the transmission time of user j's packet follows a negative exponential distribution with mean $B/r_{i,j}$. We assume that a SBS can buffer at most K packets for each associated user. Therefore, the data traffic of user j can be formulated as an $M/M/1/K$ queueing system with the arrival rate λ_j and the service rate $\mu_{i,j} = r_{i,j}/B$. For any data packet arriving at user j, the probability that user j already has $n \in [0, K]$ packets buffered in SBS i is given by [22]

$$\pi_{i,j}(n) = \rho_{i,j}^n \frac{1 - \rho_{i,j}}{1 - \rho_{i,j}^{K+1}}, \tag{5}$$

in which $\rho_{i,j} = \lambda_j / \mu_{i,j}$ is defined as the occupation rate for user j in SBS i. Note that $\rho_{i,j} < 1$ must always be satisfied.

If the buffer is not full, i.e., $n < K$, the arriving packet can be served. And the average packet delay, i.e., the average time that the packet stays in the SBS, is equivalent to the average service time of total $n + 1$ packets, which is given by

$$\begin{aligned} \tau_{i,j} &= \sum_{n=0}^{K-1} \pi_{i,j}(n) \frac{(n+1)B}{\mu_{i,j}} \\ &= \frac{(1-\rho_{i,j})B}{(1-\rho_{i,j}^{K+1})\mu_{i,j}} \sum_{n=0}^{K-1} (n+1)\rho_{i,j}^n \\ &= \frac{(1-\rho_{i,j})B}{(1-\rho_{i,j}^{K+1})\mu_{i,j}} \sum_{k=0}^{K-1} \sum_{n=k}^{K-1} \rho_{i,j}^n \\ &= \frac{(1-\rho_{i,j})B}{(1-\rho_{i,j}^{K+1})\mu_{i,j}} \sum_{k=0}^{K-1} \frac{\rho_{i,j}^k - \rho_{i,j}^K}{1-\rho_{i,j}} \\ &= \frac{B}{(1-\rho_{i,j}^{K+1})\mu_{i,j}} \left[\frac{1-\rho_{i,j}^K}{1-\rho_{i,j}} - K\rho_{i,j}^K \right]. \end{aligned} \tag{6}$$

We assume that the average packet delay is constrained by a delay threshold D, i.e.,

$$\tau_{i,j} \leq D. \tag{7}$$

If the buffer is full, i.e., $n = K$, the arriving packet will be dropped by the SBS. And the packet blocking rate is equivalent

to the probability that K packets are buffered in the SBS, which is given by

$$p_{i,j} = \pi_{i,j}(K) = \rho_{i,j}^K \frac{1-\rho_{i,j}}{1-\rho_{i,j}^{K+1}}. \tag{8}$$

Thus, the throughput of user j is given by

$$R_{i,j} = \lambda_j(1-p_{i,j}). \tag{9}$$

We consider the optimal user association $\{x_{i,j}^*\}$ and resource allocation $\{w_{i,j}^*\}$ that maximizes the average system throughput over a certain time period, while ensuring that the average packet delay is below the delay threshold D. Therefore, the considered QoS-aware load balancing problem is formulated as follows:

$$\max_{\{x_{i,j}\},\{w_{i,j}\}} \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{M}} \lambda_j x_{i,j}(1-p_{i,j}) \tag{10a}$$

$$\text{s.t. } \sum\nolimits_{i\in\mathcal{N}} x_{i,j} \le 1, \forall j \in \mathcal{M}, \tag{10b}$$

$$x_{i,j} \in \{0,1\}, \forall i \in \mathcal{N}, j \in \mathcal{M}, \tag{10c}$$

$$\sum\nolimits_{j\in\mathcal{M}} w_{i,j} \le W, \forall i \in \mathcal{N}, \tag{10d}$$

$$x_{i,j}\tau_{i,j} \le \tau_0, \forall i \in \mathcal{N}, j \in \mathcal{M}, \tag{10e}$$

$$x_{i,j}\rho_{i,j} < 1, \forall i \in \mathcal{N}, j \in \mathcal{M}, \tag{10f}$$

in which the objective function (10a) is the average system throughput, (10b) and (10c) are the constraints of user association, (10d) is the constraint of total bandwidth, (10e) is the constraint of average packet delay, and (10f) is the constraint of the queueing system. Problem (10) is a mixed integer optimization problem, which is NP-hard in general, and we seek for suboptimal solutions.

III. Load Balancing Algorithms

In this section, we propose a suboptimal load balancing algorithm for problem (10). The proposed algorithm consists of a heuristic user association algorithm and a greedy resource allocation algorithm. Note that the proposed user association algorithm is based on the results of the proposed resource allocation algorithm.

A. Resource Allocation

We denote by $\mathcal{M}_i$ as the set of users associated with SBS $i \in \mathcal{N}$, i.e.,

$$\mathcal{M}_i = \{j | x_{i,j} = 1\}. \tag{11}$$

For each user $i \in \mathcal{M}_i$, the minimum required bandwidth can be calculated by jointly considering constraints (10e) and (10f), which is given by

$$w_{i,j}^{min} = \max\left\{w_1|_{\tau_{i,j}(w_1)=\tau_0}, w_2|_{\rho_{i,j}(w_2)=1}\right\}, \tag{12}$$

in which w_1 and w_2 are numerically calculated.

The proposed resource allocation algorithm is shown in Table I. The algorithm first decides weather the total bandwidth W is sufficient for the associated users in $\mathcal{M}_i$, i.e.,

$$\sum_{j\in\mathcal{M}_i} w_{i,j}^{min} \le W. \tag{13}$$

TABLE I
Resource Allocation Algorithm

Algorithm

$w_{i,j} = 0, j \in \mathcal{M}_i$;
$W_r = W$;
calculate $w_{i,j}^{min}, j \in \mathcal{M}_i$ as in (12);
if (13) is true
  $w_{i,j} = w_{i,j}^{min}, j \in \mathcal{M}_i$;
  calculate W_r as in (14);
  for $s = 1 : S$
    calculate j' as in (15);
    $w_{i,j'} = w_{i,j'} + W_r/S$;
  end for
end if
$w_{i,j}^* = w_{i,j}, j \in \mathcal{M}_i$;
return $\{w_{i,j}^*, j \in \mathcal{M}_i\}$;

If (13) is not satisfied, there is no feasible resource allocation of SBS i for the current user association. If (13) is satisfied, the algorithm first allocates the minimum bandwidth $w_{i,j}^{min}$ to each user $j \in \mathcal{M}_i$, and divide the rest bandwidth resources, which is given by

$$W_r = W - \sum_{j\in\mathcal{M}_i} w_{i,j}^{min}, \tag{14}$$

into S equal bandwidth units. Then, the bandwidth units are sequentially allocated in S rounds. In each round, one bandwidth unit W_r/S is allocated to the user with the maximal marginal performance improvement, which is defined as the difference between user throughput before and after the bandwidth unit is allocated, and the selected user is given by

$$j' = \underset{j\in\mathcal{M}_i}{\text{argmax}}\{\lambda_j[p_{i,j}(w_{i,j}+W_r/S) - p_{i,j}(w_{i,j})]\}. \tag{15}$$

The algorithm stops after S rounds and outputs the final bandwidth allocation $w_{i,j}^*, j \in \mathcal{M}_i$. We further denote by R_i^* as the maximal throughput of SBS i by using the proposed resource allocation algorithm, which is given by

$$R_i^* = \begin{cases} \sum\limits_{j\in\mathcal{M}_i} R_{i,j}(w_{i,j}^*) & \text{(13) is true,} \\ 0 & \text{(13) is false.} \end{cases} \tag{16}$$

B. User Association

For any association $\{\mathcal{M}_i, i \in \mathcal{N}\}$, we define a utility factor $u_{i,j}, j \notin \mathcal{M}_i$ for each potential association between SBS i and user j, which represents the marginal performance improvement if the association (i,j) is established, i.e.,

$$u_{i,j} = R_i^*(\mathcal{M}_i \cup \{j\}) - R_i^*(\mathcal{M}_i). \tag{17}$$

Association (i,j) is defined to be feasible if the utility factor is strictly positive, i.e., $u_{i,j} > 0$, and we define the set of feasible associations as

$$A = \{(i,j) | i \in \mathcal{N}, j \notin \mathcal{M}_i, u_{i,j} > 0\}. \tag{18}$$

The proposed user association algorithm is shown in Table II. It starts with the initial association in which no user and

TABLE II
USER ASSOCIATION ALGORITHM

Algorithm

1: $\mathcal{M}_i = \emptyset, i \in \mathcal{N}$;
2: calculate $u_{i,j}, i \in \mathcal{N}, j \notin \mathcal{M}_i$ as in (17);
3: calculate A as in (18);
4: **while** $A \neq \emptyset$
5: calculate (i^*, j^*) as in (19);
6: $\mathcal{M}_{i^*} = \mathcal{M}_{i^*} \cup \{j^*\}$;
7: calculate $u_{i,j}, i \in \mathcal{N}, j \notin \mathcal{M}_i$ as in (17);
8: calculate A as in (18);
9: **end while**
10: $\mathcal{M}_i^* = \mathcal{M}_i, i \in \mathcal{N}$;
11: **return** $\{\mathcal{M}_i^*, i \in \mathcal{N}\}$;

TABLE III
SIMULATION PARAMETERS

$P_0 = 25$dBW	Received power at the reference distance
$d_0 = 1$m	Reference distance
$\alpha = 3.76$	Path loss parameter
$N_0 = -184$dBm/Hz	Noise power spectral density
$W = 100$MHz	Total bandwidth
$B = 10^4$bit	Average packet size
$\lambda = 100$	Average packet arrival rate
$\delta^2 = 10$	Variance of packet arrival rate
$N = 20$	Number of SBSs

SBS are associated, i.e., $\{\mathcal{M}_i = \emptyset, i \in \mathcal{N}\}$, and sequentially builds up associations until there is no feasible association, i.e., $A = \emptyset$. In each round, the algorithm establish the feasible association with the highest utility factor, i.e.,

$$(i^*, j^*) = \underset{(i,j)\in A}{\operatorname{argmax}}\, u_{i,j}, \tag{19}$$

and the final user association established by the proposed algorithm is denoted by $\{\mathcal{M}_i^*, i \in \mathcal{N}\}$.

IV. SIMULATION RESULTS

In this section, we present and analyze the simulation results of the proposed load balancing algorithm, with comparison to a baseline algorithm using maximum SINR association and equal resource allocation. Specifically, we consider a square area with 2km side length, and the SBSs and mobile users are randomly distributed within the area. The packet arrival rate λ_j follows a Gaussian distribution with mean λ and variance δ^2. The simulation parameters are given in Table III.

In Fig. 2, we show the total throughput as a function of the total number of users M, where the delay threshold is given by $D = 0.02$s. We see that the total throughput increases with M in both algorithms, but our proposed algorithm outperforms the baseline algorithm by $20\% \sim 100\%$. The performance gain of our proposed algorithm is due to both the network-level optimization for user association and the SBS-level optimization for resource allocation.

In Fig. 3 and Fig. 4, we show the packet blocking rate and the packet delay of associated users as a function of the user number M, where the delay threshold is given by $D = 0.02$s. We see that the associated users in the proposed algorithm has a higher packet blocking rate and a higher packet delay, as compared to the baseline algorithm. The reason is that the bandwidth resources allocated to each user is minimized in our proposed algorithm, and the network-level performance improvement is achieved at the price of restricting the user-level performance. However, we see that the average packet delay is always below the delay threshold D, i.e., QoS support is always guaranteed for the associated users.

In Fig. 5, we show the total throughput as a function of the delay threshold D, where $M = 80$ users exist in the network. Note that the minimum bandwidth required to support the packet delay requirement in (10e) is a decreasing function of the delay threshold D. Therefore, more users can be served by the network as the delay threshold D goes up, and the total

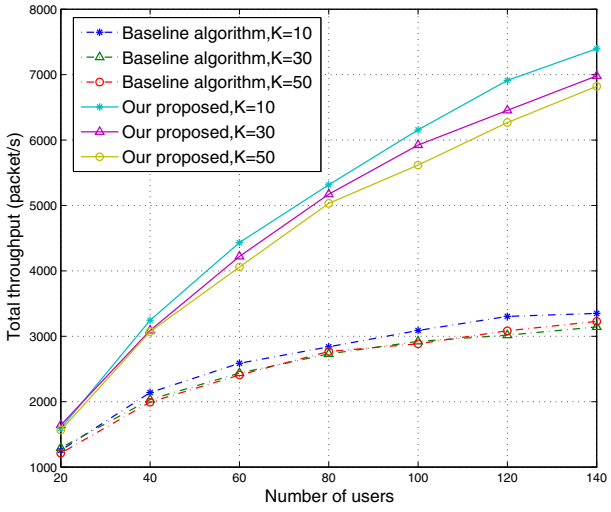

Fig. 2. Total throughput as a function of the number of users M.

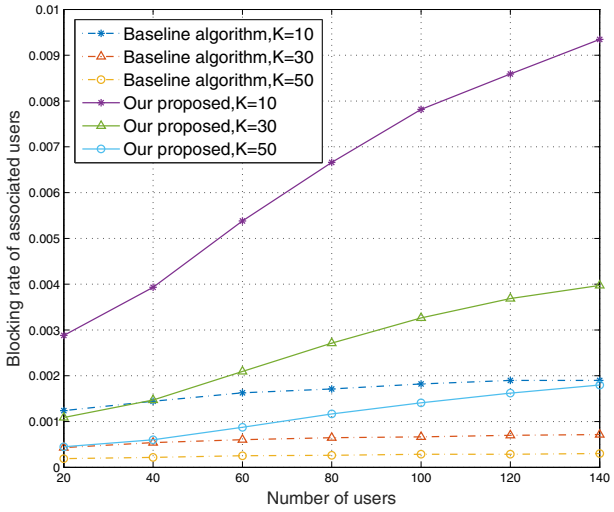

Fig. 3. Average packet blocking rate of associated users as a function of the number of users M.

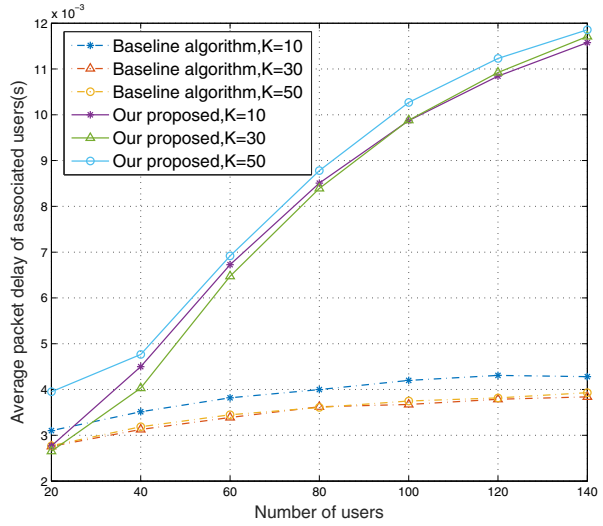

Fig. 4. Average packet delay of associated users as a function of the number of users M.

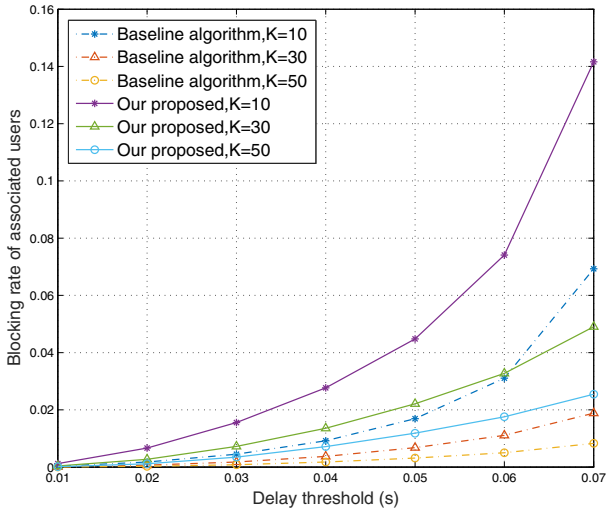

Fig. 6. Average packet blocking rate of associated users as a function of the delay threshold D.

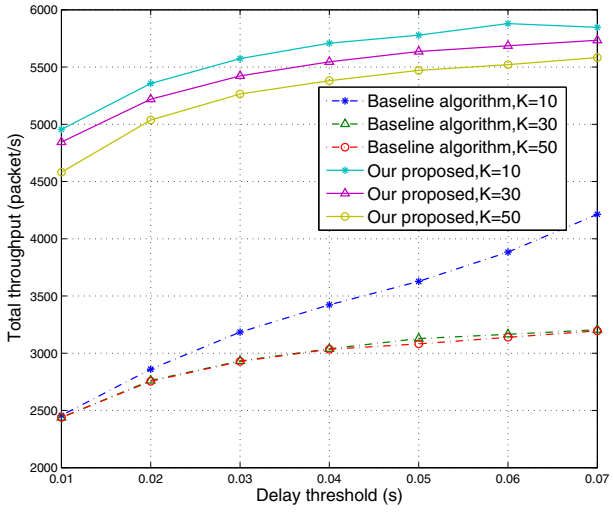

Fig. 5. Total throughput as a function of the delay threshold D.

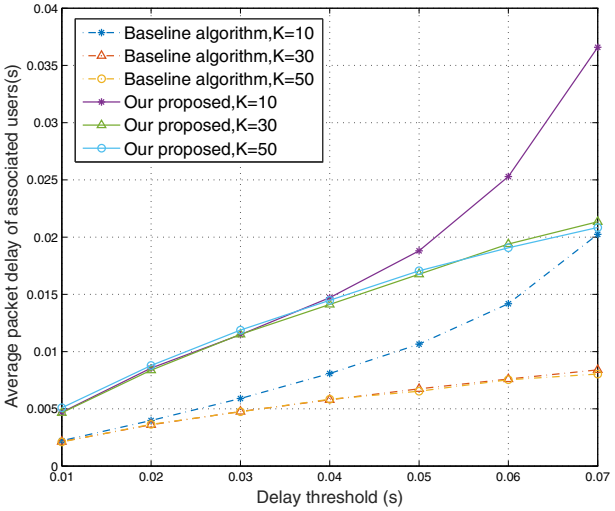

Fig. 7. Average packet delay of associated users as a function of the delay threshold D.

throughput increases, as shown in Fig. 5. Still, we see that the proposed algorithm outperforms the baseline algorithm by $20\% \sim 100\%$.

In Fig. 6 and Fig. 7, we show the packet blocking rate and the packet delay of the associated users as a function of the delay threshold D, where $M = 80$ users exist in the network. We see that the user-level performance decreases (i.e., the packet blocking rate or the packet delay increases) as the delay constraint is relaxed. The reason is that each user is allocated with less bandwidth resources and individual performance is constrained. Still, we see that the user-level performance of the proposed algorithm is sacrificed to achieve a network-level performance improvement.

Also, it can be seen that increasing the buffer size K can improve the user-level performance, i.e., the packet blocking rate and the average packet delay, while at the same time, decrease the network-level throughput. The reason is that the associated users with a large buffer size require more bandwidth to satisfy the delay constraint, which increases the individual user performance, but limits the total number of associated users and decreases the total throughput.

V. Conclusions

In this paper, we have considered the QoS-aware load balancing problem in dense cellular networks, in which the user traffic are modeled as Poisson processes instead of using a static or "snapshot" data rate assumption, and the average

packet delay is constrained to be below a delay threshold. We have formulated a mixed integer optimization problem, and proposed a lightweight algorithm to give suboptimal user association and resource allocation strategies. Simulation results have shown that the proposed algorithm can increase the number of associated users by restricting the individual user performance in terms of packet delay and packet blocking rate, and achieves a performance gain in terms of total throughput, compared to the baseline algorithm using the maximum SINR association and equal resource allocation.

REFERENCES

[1] M. Gong, S. Midkiff, and S. Mao, "On-demand routing and channel assignment in multi-channel mobile ad hoc networks,", *Ad Hoc Netw.*, vol. 7, no. 1, pp. 63-78, Jan. 2009.

[2] M. Gong, B. Hart, and S. Mao. "Advanced Wireless LAN Technologies: IEEE 802.11AC and Beyond," *GetMobile: Mobile Comp. and Comm.*, vol. 18, no. 4, pp. 48-52, Jan. 2015.

[3] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An Overview of Load Balancing in HetNets: Old Myths and Open Problems," *IEEE Wireless Commun.*, vol. 2, no. 21, pp. 18-25, Apr. 2014.

[4] C. Ran, S. Wang, and C. Wang, "Balancing Backhaul Load in Heterogeneous Cloud Radio Access Networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp.42-48, Jun. 2015.

[5] S. Wang and C. Ran, "Rethinking Cellular Network Planning and Optimization," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 118-125, Apr. 2016.

[6] W. Zhao, S. Wang, C. Wang, and X. Wu, "Approximation algorithms for cell planning in heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1561C1572, Feb. 2017.

[7] S. Wang, W. Zhao, and C. Wang, "Budgeted cell planning for cellular networks with small cells," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4797C4806, Oct. 2015.

[8] X. Lin and S. Wang, "Efficient Remote Radio Head Switching Scheme in Cloud Radio Access Network: A Load Balancing Perspective," in *Proc. IEEE INFOCOM*, Atlanta, GA, Apr./May 2017.

[9] A. Ahmed, L. M. Boulahia, and D. Gaiti, "Enabling Vertical Handover Decisions in Heterogeneous Wireless Networks: A State-of-the-Art and A Classification," *IEEE Commun. Surv. Tut.*, vol. 16, no. 2, pp. 776-811, Second Quarter 2014.

[10] E. Stevens-Navarro, Y. Lin, and V. Wong, "An MDP-based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 1243-1254, Mar. 2008.

[11] S. K. Lee, K. Sriram, K. Kim, Y. H. Kim, and N. Golmie, "Vertical Handoff Decision Algorithms for Providing Optimized Performance in Heterogeneous Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 865-881, Feb. 2009.

[12] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Coordinating Handover Parameter Optimization and Load Balancing in LTE Self-Optimizing Networks," in *Proc. IEEE VTC-Spring*, Budapest, Hungary, May 2011.

[13] P. Muñoz, R. Barco, and S. Fortes, "Conflict Resolution Between Load Balancing and Handover Optimization in LTE Networks," *IEEE Commun. Lett.*, vol. 18, no. 10, Aug. 2014.

[14] Q. Ye, B. Rong, Y. Chen, M. Al-Shlash, C. Caramanis, and J. G. Andrews,"User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706-2716, Jun. 2013.

[15] T. Zhou, Y. Huang, W. Huang, S. Li, Y. Sun, and L. Yang, "QoS-Aware User Association for Load Balancing in Heterogeneous Cellular Networks," in *Proc. IEEE VTC-Fall*, Vancouver, Canada, Sep. 2014.

[16] W. Zhao and S. Wang, "Traffic Density-Based RRH Selection for Power Saving in C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3157-3167, Dec. 2016.

[17] G. Athanasiou, P. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing Client Association for Load Balancing and Fairness in Millimeter-Wave Wireless Networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 3, pp. 836-850, Jun. 2015.

[18] C. Vlachos and V. Friderikos, "Optimal Device-to-Device Cell Association and Load Balancing," in *Proc. IEEE ICC*, London, United Kingdom, Jun. 2015.

[19] C. Ran, S. Wang, and C. Wang, "Cellular Networks Planning: A Workload Balancing Perspective," *Comp. Netw.*, vol. 84, no. 19, pp. 64-75, Jun. 2015.

[20] C. Liu and J. Zheng,"A QoS-Aware Resource Allocation Algorithm for Device-to-Device Communication Underlaying Cellular Networks," in *Proc. of IEEE VTC'17-Spring*, Sydney, Australia, Jun. 2017.

[21] W. Zhao and S. Wang, "Resource Sharing Scheme for Device-to-Device Communication Underlaying Cellular Networks," *IEEE Trans. Commun.*, vol. 63. no. 12, pp. 4838-4848, Dec. 2015.

[22] I. Adan and J. Resing, *Queueing Theory*, Eindhoven University of Technology, Feb. 2001.

[23] S. Mao, S. Panwar, and Y. Hou, "On minimizing end-to-end delay with optimal traffic partitioning," *IEEE Trans. Veh. Technol.*, vol. 55, no. 2, pp. 681-690, Mar. 2006.