User Association in Backhaul Constrained Small Cell Networks

© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.

Citation:

Tao Han and Nirwan Ansari, “User Association in Backhaul Constrained Small Cell Networks”, accepted to Proc. IEEE WCNC 2015, New Orleans, Louisiana, USA.

URL:
http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7127713
User Association in Backhaul Constrained Small Cell Networks

Tao Han, Student Member, IEEE, and Nirwan Ansari, Fellow, IEEE
Advanced Networking Laboratory
Department of Electrical and Computer Engineering
New Jersey Institute of Technology, Newark, NJ, 07102, USA
Email: {th36, nirwan.ansari}@njit.edu

Abstract—Explosive data traffic growth has led to a continuous surge in capacity demands in mobile networks. In order to provision high network capacity, small cell base stations (SCBSs) are widely deployed. Owing to the close proximity to mobile users, SCBSs can effectively enhance the network capacity and offloading traffic load from macro BSs (MBSs). However, the cost-effective backhaul may not be readily available for SCBSs that leads to backhaul constraints in small cell networks. In this paper, we investigate the traffic offloading in backhaul constrained small cell networks. We have proposed a network latency aware user association scheme that balances traffic loads among base stations (BSs) to minimize the average traffic delivery latency of the mobile network. The proposed network latency aware user association scheme considers the traffic delivery latency in both BSs and their backhaul during the process of establishing user associations. We have proved that the proposed user association scheme converges to the optimal solution that minimizes the average traffic delivery latency of the network. The simulation results show that the proposed scheme reduces the average traffic delivery latency by 63% and 34% as compared to the user association scheme that considers only the traffic delivery latency in BSs and a two-tier data rate bias user association scheme, respectively.

I. INTRODUCTION

Owing to the proliferation of mobile devices and applications, mobile data traffic grows exponentially, thus leading to a continuous surge in network capacity demands [1]. In order to provision high network capacity, small cell base stations (SCBSs) are widely deployed [2]. SCBSs, with a small coverage area, can significantly improve the spectrum utilization in mobile networks. Moreover, by capitalizing on their close proximity to mobile users, SCBSs are able to provide high network capacity to mobile users.

In small cell networks, traffic load balancing is an essential task for optimizing the network performance and has been extensively studied [3]. In mobile networks, traffic load balancing is achieved by executing a user association process in which mobile users are assigned to base stations (BSs) for services. Various user association algorithms have been proposed. Jo et al. [4] proposed cell biasing algorithms to balance traffic loads among macro BSs (MBSs) and SCBSs. The cell biasing algorithms perform user association according to the biased measured pilot signal strength. Ye et al. [5] modeled the traffic load balancing problem as a utility maximization problem and developed distributed user association algorithms based on the primal-dual decomposition. Kim et al. [6] proposed an optimal user association algorithm to achieve flow level load balancing under spatially heterogeneous traffic distribution. The proposed algorithm may maximize different network utilities, e.g., the traffic latency and the network throughput, by properly setting the value of $\alpha$. In addition, game theory has been exploited to model and solve the traffic load balancing problem. Aryafar et al. [7] modeled the traffic load balancing problem as a congestion game in which users are the players and user association decisions are the actions. Han and Ansari [8], [9] have proposed a user association scheme that jointly optimizes the average traffic delivery latency and the green energy utilization.

Most of the existing user association solutions optimize the traffic load balancing in a mobile network with the implication that the air interface between BSs and mobile users is the bottleneck of the network. This implication is generally correct for MBSs whose deployments are well planned. However, in small cell networks, considering the potentially dense deployment of SCBSs, various backhaul solutions, e.g., xDSL, non-line-of-sight (NLOS) microwave, wireless mesh networks, rather than ideal backhaul such as optical fiber and LOS microwave, are adopted [2]. As a result, the backhaul, instead of BSs, may become the bottleneck of small cell networks. Therefore, it is desired to optimize the user association with the consideration of backhaul constraints.

In this paper, we propose a network latency aware user association scheme to minimize the average traffic delivery latency for backhaul constrained small cell networks. We model the backhaul constrained small cell network as a tandem queue system in which mobile users experience delays at queues in both BSs and backhaul. In establishing user associations, the proposed network latency aware user association scheme considers the delays in both BSs and their backhaul. The proposed user association scheme is a distributed scheme consisting of a BS side algorithm and a user side algorithm. The BS side algorithm measures traffic loads in BSs and their backhaul while the user side algorithm selects the serving BS based on users’ data rates, BSs’ backhaul capacity, and the traffic load in both BSs and their backhaul. The distributed scheme is proved to converge to the optimal user association that minimizes the average traffic delivery latency of the
network.

The rest of the paper is organized as follows. In Section II, we define the system model and formulate the user association problem in backhaul constrained small cell networks. Section III presents the proposed user association scheme and analyzes its properties. Section IV shows the simulation results, and concluding remarks are presented in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

![Diagram](image)

Fig. 1. The backhaul constrained small cell network.

A. Traffic and QoS Model

Denote $B$ as the set of BSs including both MBSs and SCBSs and $A$ as the coverage area of all BSs. Here, a BS refers to either a MBS or a SCBS. We assume that the traffic arrives according to a Poisson process with the arrival rate per unit area to location $x$ equaling to $\lambda(x)$, and the traffic load follows an exponential distribution with the average traffic load of $\nu(x)$. Assuming a mobile user at location $x$ is associated with BS $j$, the data traffic reaches the user at location $x$ through BS $j$ and its backhaul. Therefore, we model the traffic delivery process as a tandem queue system as shown in Fig. 2.

![Diagram](image)

Fig. 2. The traffic delivery process as a tandem queue system.

We assume that the users associated with BS $j$ are uniformly distributed in its coverage area, and the traffic arrival processes are independent. Since the traffic arrival at a location follows a Poisson process, the traffic arrival in BS $j$’s backhaul, which is the sum of the traffic arrivals from its coverage area, is also a Poisson process. Although BSs may adopt different access technologies as their backhaul, it is reasonable to assume the expected data rates of the backhaul are constant in the time duration of one user association process. Since the traffic load follows an exponential distribution, the traffic delivery time (service time) of the backhaul is also an exponential distribution. Therefore, the traffic delivery in the backhaul realizes an M/M/1 queuing system.

A mobile user, who is located at $x$ and associated with BS $j$, is assumed to have the traffic load $\nu(x)$. Denoting $R_j$ as the average data rate of BS $j$’s backhaul. To fulfill the user’s traffic demand, the required service time in the backhaul is

$$\tilde{g}(x) = \frac{\nu(x)}{R_j}. \quad (1)$$

The average traffic load density of a user at location $x$ in BS $j$’s backhaul is

$$\tilde{\theta}_j(x) = \frac{\lambda(x)\nu(x)\eta_j(x)}{R_j} \quad (2)$$

Here, $\eta_j(x) = \{0, 1\}$ is an indicator function. If $\eta_j(x) = 1$, the user at location $x$ is associated with BS $j$; otherwise, the user is not associated with BS $j$. Assuming mobile users are uniformly distributed in the area and denoting $A$ as the coverage area of all the BSs, the traffic load in BS $j$’s backhaul can be expressed as

$$\tilde{\rho}_j = \int_{x \in A} \tilde{\theta}_j(x) dx. \quad (3)$$

According to [10], the average waiting time for traffic load $\nu(x)$ in BS $j$’s backhaul is

$$\bar{W}_j(x) = \frac{\tilde{\rho}_j \nu(x)}{R_j (1 - \tilde{\rho}_j)} \quad (4)$$

Denote $\tilde{\mu}_j(x)$ as the latency ratio that measures how much time a user at location $x$ must be sacrificed in waiting for per unit service time in BS $j$’s backhaul.

$$\tilde{\mu}_j(x) = \frac{\bar{W}_j(x)}{\gamma(x)} = \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j} \quad (5)$$

According to Eq. (5), $\tilde{\mu}_j(x)$ only depends on the traffic load in BS $j$’s backhaul. Therefore, all the users associated with BS $j$ have the same latency ratio. Thus, we define

$$\tilde{\mu}_j = \frac{\tilde{\rho}_j}{1 - \tilde{\rho}_j} \quad (6)$$

as the latency ratio of BS $j$’s backhaul. A smaller $\tilde{\mu}_j$ indicates that BS $j$’s backhaul introduces less latency to its associated users.

According to Burke’s Theorem, the traffic arrival to location $x$ in BS $j$ follows a Poisson distribution with average arrival rate $\lambda(x)$ [10]. In BS $j$, users at different locations may have different data rates depending on channel conditions. If a user at location $x$ associates with BS $j$, the user’s data rate, $r_j(x)$, can be generally expressed as a logarithmic function of the perceived signal to interference plus noise ratio, $SINR_j(x)$, according to the Shannon Hartley Theorem [6],

$$r_j(x) = \log_2(1 + SINR_j(x)). \quad (7)$$

Here,

$$SINR_j(x) = \frac{P_jg_j(x)}{\sigma^2 + \sum_{k \in B, k \neq j} P_kg_k(x)}, \quad (8)$$

where $P_j$ is the transmission power of BS $j$ and $\sigma^2$ denotes the noise power level. Since the user data rate is generally distributed, the service time in BS $j$ follows a general distribution. Therefore, a BS’s downlink transmission process realizes a M/G/1 processor sharing (PS) queue, in which multiple users share the BS’s downlink radio resource [10].
In mobile networks, various downlink scheduling algorithms have been proposed to enable proper sharing of the limited radio resource in a BS. According to the scheduling algorithm, users may be assigned different priorities on sharing the radio resource. For analytical simplicity, we assume that mobile users are served based on the round robin fashion. Then, the average traffic load density at location $x$ in BS $j$ is

$$
\varrho_j(x) = \frac{\lambda(x)\nu(x)\eta_j(x)}{r_j(x)}.
$$

(9)

The traffic load in BS $j$ can be expressed as

$$
\rho_j = \int_{x \in A} \varrho_j(x)dx.
$$

(10)

This value of $\rho_j$ indicates the fraction of time during which BS $j$ is busy. To fulfill the traffic demand of a user located at $x$, the required service time in BS $j$ is

$$
\gamma(x) = \frac{\nu(x)}{r_j(x)}.
$$

(11)

Since the traffic delivery process in a BS realizes a M/G/1-PS queue, the average traffic delivery time for the user in BS $j$ [10] is

$$
T_j(x) = \frac{\nu(x)}{r_j(x)(1-\rho_j)}.
$$

(12)

In BS $j$, the average waiting time for traffic load $\nu(x)$ is

$$
W_j(x) = T_j(x) - \gamma(x) = \frac{\rho_j \nu(x)}{r_j(x)(1-\rho_j)}.
$$

(13)

Denote $\mu_j(x)$ as the latency ratio of BS $j$ for a user at location $x$,

$$
\mu_j(x) = \frac{W_j(x)}{\gamma(x)} = \frac{\rho_j}{1-\rho_j}.
$$

(14)

According to Eq. (14), $\mu_j(x)$ only depends on the traffic load in BS $j$. Therefore, all the users associated with BS $j$ have the same latency ratio. Thus, we define

$$
\mu_j = \frac{\rho_j}{1-\rho_j}
$$

(15)

as the latency ratio of BS $j$. A smaller $\mu_j$ indicates that BS $j$ introduces less latency to its associated users. Aiming to minimize the traffic delivery latency, we adopt $\tilde{\mu}_j + \mu_j$ as the QoS model that indicates the average latency of delivering traffic through BS $j$.

B. Problem Formulation

In determining the user associations, the network aims to enhance the network QoS by reducing the traffic delivery latency in the network. Considering the traffic delivery latency in both BSs and their backhaul, the user association problem is

$$
\begin{align*}
\text{min}_{(\eta_j(x), x \in A, j \in B)} & \quad \sum_{j \in B} \frac{\tilde{p}_j}{1-\tilde{p}_j} + \frac{\rho_j}{1-\rho_j} \\
\text{subject to :} & \quad 0 \leq \tilde{p}_j \leq 1 - \epsilon, \\
& \quad 0 \leq \rho_j \leq 1 - \epsilon.
\end{align*}
$$

(16)

Here, $\epsilon$ is an arbitrary small positive constant to guarantee the queuing system is stable.

III. NETWORK LATENCY AWARE USER ASSOCIATION SCHEME

In this section, we present the network latency aware user association scheme and prove the properties of the proposed scheme. The network latency aware user association scheme considers the traffic delivery latency in both the BSs and their backhaul in determining user associations. The proposed scheme consists of a user side algorithm and a BS side algorithm. The BS side algorithm measures the traffic loads in both the BSs and their backhaul, and updates the advertised traffic load to mobile users. Based on the advertised traffic load, the user data rates, and the data rate of BSs’ backhaul, the user side algorithm selects the optimal BS for individual users to minimize $\psi(\eta) = \sum_{j \in B} \frac{\rho_j}{1-\rho_j} + \frac{\rho_j}{1-\rho_j}$. Here, $\eta = \{\eta_j(x) | j \in B, x \in A\}$.

In order to guarantee convergence of the distributed user-BS association scheme, we assume that the time scale of the arrival and departure process of user data traffic is faster relative to that of BSs in advertising their traffic loads. In other words, BSs broadcast their traffic loads after the system exhibits the stationary performance. We assume that all the BSs are synchronized and advertise their traffic loads simultaneously. We define the time interval between two consecutive traffic load advertisements as a time slot. We assume that the data rate of a BS’s backhaul is a constant during the time period of establishing a stable user-BS association.

A. The User Association Scheme

1) The User Side Algorithm: At the beginning of the $k$th time slot, BSs broadcast their traffic load $\rho_j(k)$, the traffic load in backhaul $\tilde{p}_j(k)$, and backhaul data rates to mobile users. Let

$$
\phi_j(k) = \frac{(1-\tilde{p}_j(k))^2(1-\rho_j(k))^2}{R_j(1-\rho_j(k)) + r_j(x)(1-\rho_j(k))};
$$

(18)

the BS selection rule for a user at location $x$ can be expressed as

$$
b^k(x) = \arg\max_{j \in B} R_j r_j(x) \phi_j(k).
$$

(19)

Here, $b^k(x)$ is the index of the BS selected by the user.

2) The BS Side Algorithm: After mobile users select their associating BSs, the coverage area of BS $j$ is updated and the user association in BS $j$ in the $k$th time slot is

$$
\eta^k_j(x) = \begin{cases} 1, & \text{for } j = b^k(x), \forall x \in A \\ 0, & \text{for } j \neq b^k(x), \forall x \in A. \end{cases}
$$

(20)

Given the user association, the perceived traffic load in BS $j$ and its backhaul can be calculated based on Eq. (3) and Eq. (10), respectively. On updating the traffic load advertisement, BS $j$ calculates an intermediate user association $\tilde{\eta}^k_j(x)$ as

$$
\tilde{\eta}^k_j(x) = (1-\delta)\eta^k_j(x) + \delta \eta^{k-1}_j(x).
$$

(21)

Here, $0 < \delta < 1$ is an exponential averaging parameter. With the intermediate user association, the advertised traffic load in BS $j$’s backhaul is

$$
\tilde{p}_j(k+1) = \int_{x \in A} \frac{\lambda(x)\nu(x)\tilde{\eta}^k_j(x)}{R_j} dx,
$$

(22)
and the traffic load in BS $j$ is advertised as
\begin{equation}
\rho_j(k+1) = \int_{x \in A} \frac{\lambda(x)\nu(x)\eta^k_j(x)}{r_j(x)} dx.
\end{equation}

**B. The Properties**

In this subsection, we show the convergence and the optimality of the user association scheme. The feasible set for the user association problem is
\begin{equation}
\mathcal{F} = \{\eta| 0 \leq \rho_j \leq 1 - \epsilon, 0 \leq \rho_j \leq 1 - \epsilon, \sum_{j \in B} \eta_j(x) = 1, \eta_j(x) = \{0, 1\}, \forall j \in B, \forall x \in A\}
\end{equation}

Since $\eta_j(x) = \{0, 1\}$, $\mathcal{F}$ is not a convex set. Thus, the traffic updates in the BS side algorithm cannot guarantee the advertised traffic load is in the feasible set. In order to show the convergence, we first relax the constraint to let $0 \leq \eta_j(x) \leq 1$ and then prove the traffic load vector converges to the traffic load vector that is in the feasible set. Define $\tilde{\mathcal{F}}$ as the relaxed feasible set.

**Lemma 1.** The relaxed feasible set $\tilde{\mathcal{F}}$ is a convex set.

*Proof:* The lemma is proved by showing that the set $\tilde{\mathcal{F}}$ contains any convex combination of the user association vector $\eta$.

**Lemma 2.** $\psi(\eta)$ is a convex function of $\eta$ when $\eta$ is defined in $\tilde{\mathcal{F}}$.

*Proof:* The lemma can be proved by showing $\nabla^2 \psi(\eta) > 0$ when $\eta$ is defined in $\tilde{\mathcal{F}}$.

**Lemma 3.** Let $\eta^k = \{\eta^k_j(x)|j \in B, x \in A\}$. When $\eta^k \neq \eta^{k-1}$, $\eta^k$ provides a descent direction of $\psi(\eta)$ at $\eta^k$.

*Proof:* Since $\psi(\eta)$ is a convex function over $\eta$, the lemma can be proved by showing $\langle \nabla \psi(\eta)|_{\eta = \eta^k}, \eta^k - \eta^{k-1} \rangle < 0$.

\begin{equation}
\langle \nabla \psi(\eta)|_{\eta = \eta^k}, \eta^k - \eta^{k-1} \rangle < 0
\end{equation}

Thus, $\langle \nabla \psi(\eta)|_{\eta = \eta^k}, \eta^k - \eta^{k-1} \rangle < 0$.

Denote $\rho = \{\rho_j| j \in B\}$ and $\rho = \{\rho_j| j \in B\}$ as the traffic load vectors in BSs and their backhaul, respectively. When traffic loads are converged, the user association is also determined.

**Theorem 1.** The traffic load vector $\rho$ and $\rho$ converge to the optimal traffic load vectors $\rho^*$ and $\rho^*$, respectively, such that $\psi(\eta)$ is minimized.

*Proof:* Since both traffic load vector $\rho$ and $\rho$ are updated based on $\eta$, when one of the traffic load vectors converges, the other traffic load vector will also converge. Therefore, we only have to prove one of the traffic load vectors, e.g., $\rho$, converges.

\begin{align}
\rho_j(k+1) &= \int_{x \in A} \frac{\lambda(x)\nu(x)\eta^k_j(x)}{r_j(x)} dx \\
&= \int_{x \in A} \lambda(x)\nu(x)(1 - \delta)\eta^k_j(x)dx \\
&+ \int_{x \in A} \lambda(x)\nu(x)\delta\eta^{k-1}_j(x)dx \\
&= (1 - \delta)\rho_j(k) + \delta\rho_j(k - 1).
\end{align}

Since $\rho_j(k+1), \rho_j(k)$ and $\rho_j(k-1)$ belong to $[0, 1 - \epsilon]$, according to the fixed point theorem, there exit $\delta$ and $\rho_j^* \in [0, 1 - \epsilon]$ such that mapping, $\rho_j(k+1) = (1 - \delta)\rho_j(k) + \delta\rho_j(k - 1)$, can converge to $\rho_j(k+1) = \rho_j^*$.

Since
\begin{align}
\rho_j(k) - \rho_j(k - 1) &= \int_{x \in A} \lambda(x)\nu(x)\eta^k_j(x)dx - \rho_j(k - 1) \\
&= (1 - \delta)\int_{x \in A} \lambda(x)\nu(x)(\eta^k_j(x) - \eta^{k-1}_j(x))dx
\end{align}

and $\eta^k$ provides a descent direction of $\psi(\eta)$ at $\eta^k$, $\psi(\eta)$ gradually reduces until $\rho_j(k) = \rho_j(k - 1)$. When the traffic load converges, $\psi(\eta)$ is minimized. Since the optimal traffic load is determined based on the user side algorithm, $\eta^k_j(x) = 0, 1$. Thus, when the traffic load converges, the user association $\eta^*$ is in the feasible set $\tilde{\mathcal{F}}$.

**IV. SIMULATION RESULTS**

We set up system level simulations to investigate the performance of the proposed network latency aware user association scheme for the downlink traffic load balancing in backhaul constrained small cell networks. In the simulation, three MBSs and seven SCBSs are randomly deployed in a $2000m \times 2000m$ area. The total bandwidth is $10$ MHz and the frequency reuse factor is one. The channel propagation model is based on COST 231 Wallis-ikegami [11]. The model and parameters are summarized in Table I. Here, $PL_{MBS}$ and $PL_{SCBS}$ are the path loss between the users and MBSs and SCBSs, respectively. $d$ is the distance between users and BSs. The transmit power of a MBS and a SCBS are 43 dBm and 33 dBm, respectively. In the simulations, the average data rate of SCBSs’ backhaul is 5 Mbps.
TABLE I
Channel Model and Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$PL_{MBS}$ (dB)</td>
<td>$PL_{MBS} = 128.1 + 37.6 \log_{10}(d)$</td>
</tr>
<tr>
<td>$PL_{SCBS}$ (dB)</td>
<td>$PL_{SCBS} = 38 + 10 \log_{10}(d)$</td>
</tr>
<tr>
<td>Rayleigh fading</td>
<td>9 dB</td>
</tr>
<tr>
<td>Shadowing fading</td>
<td>5 dB</td>
</tr>
<tr>
<td>Antenna gain</td>
<td>15 dB</td>
</tr>
<tr>
<td>Noise power level</td>
<td>-174 dBm</td>
</tr>
<tr>
<td>Receiver sensitivity</td>
<td>-123 dBm</td>
</tr>
</tbody>
</table>

We compare the performance of the proposed network latency aware user association scheme in terms of the average traffic delivery latency with the BS latency aware user association scheme and a two-tier data rate bias algorithm. The BS latency aware user association scheme minimizes the average traffic delivery latency in BSs. The formulation of the BS latency aware user association problem is

$$
\min_{\rho} \sum_{y \in B} \frac{\rho_j}{1 - \rho_j} \quad (33)
$$

subject to:

$$
0 \leq \rho_j \leq 1 - \epsilon. \quad (34)
$$

For the two-tier data rate bias scheme, we assume that BSs in the same tier have the data rate bias. In the simulation, MBSs are in the first tier while SCBSs are in the second tier. The data rate bias of a MBS is set to one. We vary the data rate bias of a SCBS to investigate the performance of the scheme. In the data rate bias scheme, a user selects the serving BS to maximize the biased data rate.

Fig. 3 compares the network’s average traffic delivery latency under the network latency aware user association scheme and the BS latency aware user association scheme. The network latency aware user association converges in about 20 iterations and reduces around 63% of the average traffic delivery latency as compared to the BS latency aware user association. The performance enhancement comes from the consideration of the traffic load in the data rate constrained backhaul during the user association process. When the backhaul is congested, the network latency aware user association scheme is able to offload data traffic from the BS that experiences congestion in its backhaul to BSs with lightly loaded backhaul.

The coverage area of the network latency aware user association and the BS latency aware user association are shown in Figs. 5 and 6, respectively. In these figures, different colors indicate the coverage areas of different BSs. With the awareness of latency in the backhaul, the network latency aware user association scheme reduces the coverage area of some backhaul constrained SCBSs, e.g., SCBS 8, by offloading data traffic to their neighboring BSs. If we further reduce a SCBS’s, e.g., SCBS 5’s, backhaul data rate, the network latency aware user association scheme will update BSs’ coverage areas and the updated BS coverage areas are shown in Fig. 7. Either the BS latency aware user association or the two-tier data rate bias scheme are unable to adapt the BS coverage areas according to the data rates of backhaul. This leads to excessive traffic delivery latency. As shown in Fig. 8, the average traffic delivery latency of the BS latency aware user association is 7.56 times more than that of the network latency aware user association.
Fig. 9 shows the impact of backhaul data rates on the average traffic delivery latency. When the data rates of backhaul are low (the backhaul is congested), e.g., 5 Mbps in the simulation, the network aware user association significantly outperforms the BS latency aware user association. As the data rates of the backhaul increase, the backhaul congestion is gradually mitigated. As a result, the network latency aware user association and the BS latency aware user association achieves almost the same traffic delivery latency when the backhaul data rate is high, e.g., 8 Mbps in the simulation.

Fig. 9. The average traffic delivery latency versus backhaul data rates.

V. CONCLUSION

In this paper, we have proposed a network latency aware user association scheme for backhaul constrained small cell networks. During the procedure of establishing user associations, the network latency aware scheme considers the traffic delivery latency in both BSs and their backhaul. By optimizing the user associations, the network latency aware user association scheme significantly reduces the average traffic delivery latency in backhaul constrained small cell networks. As compared with the user association scheme that only considers the traffic delivery latency in BSs and the two-tier data rate bias scheme, the proposed network latency aware user association scheme reduces about 63% and 34% of the average traffic delivery latency, respectively.