Timely Throughput of Heterogeneous Cellular Networks

Gongzheng Zhang∗†, Aiping Huang∗†, Tony Q. S. Quek‡, and Hangguan Shan∗†

Abstract—Network densification via deploying dense small cells is one of the dominant evolutions towards future cellular network to increase spectrum efficiency. Packet transmission delay and reliability in the resultant interference-limited heterogeneous cellular network (HCN) are essential performance metrics for system design. By modeling the locations of base stations (BSs) in HCN as superimposed of independent Poisson point processes, we propose an analytical framework to derive the timely throughput of HCN, which captures both the delay and reliability performance. In the analysis, the BS activity and temporal correlation of transmissions are taken into consideration, both of which have significant effect on network performance. The effect of mobility, BS density, and association bias factor is investigated through numerical results, which shows that network performance derived ignoring the temporal correlation of transmissions is optimistic.

I. INTRODUCTION

Densely deploying different types of base stations (BSs), e.g., pico and femto cell BSs, is conceived to be one of the dominant evolutions in future generation of cellular networks (5G) to increase network capacity towards the next decade [1]–[3]. The unplanned locations of different types of BSs make the network more irregular towards a random heterogeneous cellular network (HCN). Although dense deployment with aggressive frequency reuse can potentially improve the area spectrum efficiency and coverage performance, the complex and random interference due to unplanned BSs may deteriorate packet transmission reliability and increase transmission delay due to retransmissions. Due to a wider range of services in future cellular networks, especially the ones with delay or reliability requirements, the need to understand the effect of key network parameters on packet transmission delay and reliability becomes more essential [4].

As the basic element of queueing and end-to-end delays, packet transmission delay measures the time that a user spends to successfully receive a packet from a BS. It is determined by network circumstances and reliability requirements. Specifically, more retransmissions or longer delay are needed under higher reliability requirement or worse channel conditions. The relationship between delay and reliability is not well understood, especially in the context of HCN. There exists many literature that models the HCN using Poisson point process (PPP) and analyzes network capacity and coverage performance [5]–[8]. However, delay and reliability cannot be directly obtained from these results because temporal correlation of transmissions has not been included in their models. Furthermore, BSs serving no users can be in inactive mode and do not contribute to any network interference. This effect becomes more obvious with continuously increasing of BS density [9]. As a result, both the temporal correlation and BS activity have a significant impact on network performance of HCN, and these should be taken into consideration in performance analysis and system design.

In this paper, we present an analytical framework to study the packet transmission delay and reliability performance of HCN based on the theory of timely throughput [10]. Differing from the work in [11] where they cover the details of the lower layer by assuming known success probability of each link, we take the main essential features of HCN into consideration, including scheduling and BSs being in inactive mode. The HCN is modeled as a multi-tier cellular network, where BSs of each tier are modeled using PPP of different spatial densities and transmit powers [7]. Specifically, the main contributions of our work are summarized as follows:

• We derive the timely throughput of HCN analytically, which handles both delay and reliability performance. Temporal correlation of transmissions and BS activity are considered in our analysis.
• The effect of mobility, BS density, and association bias factor on the performance is evaluated through numerical results. The results show that via increasing the BS density, the packet transmission reliability under certain delay constraint can be steadily improved for both static and high mobility scenarios. Furthermore, there exists an optimum bias factor, with which the maximal timely throughput can be achieved.

II. HETEROGENEOUS CELLULAR NETWORK MODEL

We consider a downlink HCN consisting of $K$ tiers of BSs. For notational simplicity, we denote the set $\{1, 2, \cdots, K\}$ by $\mathcal{K}$. The BSs across tiers may differ in terms of spatial densities, transmit powers, and path loss exponents. The locations of BSs in the $j$th tier ($j \in \mathcal{K}$) are modeled as a homogeneous PPP, $\Phi_j$, of density $\lambda_j$ on the infinite two-dimensional plane. The locations of users are modeled as another PPP, $\Phi_u$, of density $\lambda_u$ and independent of $\Phi_j$. Without loss of generality, we analyze the performance of a typical user located at the...
origin \( o \) and name its associated BS as the tagged BS. As a consequence of Slivnyak’s Theorem, the distributions of users and BSs are not affected by conditioning on a user being at the origin (typical user) [7].

In the following, the spectrum is reused by all BSs of all tiers. A BS associated with no user is inactive and does not transmit. Thus, except the tagged BS, all other active BSs become interfering BSs to the typical user. The active BSs of the \( j \)-th tier transmit with the same power \( P_j \). The downlink desired and interference signals experience standard distance dependent path loss with path loss exponent \( \alpha_j > 2 \) for the \( j \)-th tier. The random channel gain is modeled by Rayleigh fading with unit average power. Since HCNs are primarily interference-limited, we ignore thermal noise for simplicity.

### A. User Association and BS Activity

We assume open access for the HCN, which means that the typical user is allowed to access any tier’s BS. A maximum biased-average-received-power association rule is used, where a user associates with a BS of the \( k \)-th tier if

\[
k = \arg \max_{j \in K} B_j P_j R_j^{-\alpha_j}.
\]

Here, \( R_j \) is the distance from the typical user to its nearest BS in the \( j \)-th tier, \( B_j \) is the bias factor that is set identical for all BSs in the \( j \)-th tier, and a larger \( B_j \) implies that more user are pushed to the \( j \)-th tier. We call the typical user associate with a BS in the \( k \)-th tier or with the \( k \)-th tier equally hereafter. Note from (1) that the association metric depends on the path loss but not on instantaneous fading gain, which means that it only depends on the long-term average received power.

With this association rule, the probability that a typical user is associated with the \( k \)-th tier is given by [8]

\[
A_k = 2\pi \lambda_k \int_0^\infty r \exp \left\{ -\pi \sum_{j \in K} \lambda_j \left( \frac{B_j P_j}{B_k P_k} \right)^{\delta_j} \frac{2z_j}{r^{\delta_j}} \right\} dr
\]

where \( \delta_j = \frac{2}{\alpha_j} \). If \( \{ \alpha_j \} = \alpha \), then \( \delta = 2/\alpha \), and the expression of association probability is simplified to

\[
A_k = \frac{\lambda_k (B_k P_k)^{\delta}}{\sum_{j \in K} \lambda_j (B_j P_j)^{\delta}}.
\]

The association model described above leads to a result that the collection of association region of BSs is a multiplicatively weighted Voronoi tessellation. Let \( C_k \) denote the cell coverage area of a randomly chosen BS in the \( k \)-th tier. We adopt the approximation method proposed in [8] to model the distribution of \( C_k \), which is given by

\[
f_{C_k}(c) = \frac{3.5^{3.5} \lambda_k}{\Gamma(3.5) A_k} \left( \frac{\lambda_k}{A_k} c \right)^{2.5} \exp \left( -3.5 \frac{\lambda_k}{A_k} c \right)
\]

where \( \Gamma(\cdot) \) is the gamma function. Furthermore, the coverage area of the tagged BS is denoted as \( C_k \), with the distribution given by

\[
f_{C_k}(c) = \frac{3.5^{3.5} \lambda_k}{\Gamma(3.5) A_k} \left( \frac{\lambda_k}{A_k} c \right)^{3.5} \exp \left( -3.5 \frac{\lambda_k}{A_k} c \right)
\].

Now, the BS Activity Probability is defined as the probability that a randomly chosen BS is in active mode. The activity probability of BSs in the \( k \)-th tier is given in the following Lemma.

**Lemma 1.** The probability that a randomly chosen BS in the \( k \)-th tier is active is given by

\[
P_{a,k} = 1 - (1 + 3.5^{-1} \rho_k)^{-3.5}
\]

where \( \rho_k = \frac{\lambda_k A_k}{\lambda_k} \) is the mean load of the \( k \)-th tier.

**Proof:** The result can be proved by applying (4) to Proposition 1 of [9].

Note that the activity probability of BSs in the \( k \)-th tier is only determined by and a decreasing function of its mean load \( \rho_k \). In the following, we model the set of active BSs in the \( k \)-th tier as an independently thinning of \( \Phi_k \) of density \( \lambda_k P_{a,k} \).

### B. Scheduling and Transmissions

In the following, we consider that one packet for the typical user is transmitted within one time slot. The transmission is successful if the received signal to interference ratio (SIR) with respect to the tagged BS is above a given threshold \( \theta \). Retransmissions are required in case of unsuccessful transmissions, which introduces a delay. A simple retransmission protocol is employed, in which the packet is repeatedly transmitted until it is successfully received or a delay deadline is violated. The undelivered packets beyond the delay deadline are dropped, which incurs unreliability. Only the packet of the selected (scheduled) user has the opportunity to be transmitted. To guarantee the fairness among users, a random scheduling scheme is adopted, in which the BS randomly selects one of its associated users with equal probability in each time slot. Denote the number of other users, except the typical user, associating with the tagged BS as \( N_o \), then the typical user is scheduled with a probability of \( \frac{1}{1+o} \).

Obviously, both delay and reliability performance of the typical user are influenced by two factors. One is the probability of being selected/scheduled per time slot, which is determined by the number of users associated with the tagged BS. The other is the received SIR determined by the signal strength from the tagged BS and the interference strength from the other active BSs. Regarding these two factors, we make the following assumption throughout this paper.

**Assumption 1.** The number of total users associated with the tagged BS and the received SIR of the typical user from the tagged BS are approximated to be independent.

Assumption 1 implies that scheduling and transmission in one time slot for the typical user are independent given the tagged BS, which is shown to be accurate in [9].

To investigate the effect of mobility on network performance, we consider two mobility scenarios in our analysis.

- **Static scenario:** The locations of the users are modeled as a static PPP. Therefore, the users have fixed locations and associated BSs over different time slots. The typical user associates with a fixed BS in a fixed tier over time.
- **High mobility scenario:** In each time slot, the locations of users are modeled as an independent PPP, i.e., a new
realization of the PPP $\Phi_u$ for users’ locations is drawn in each time slot. Therefore, the associated BS of each user may change over different time slots. The typical user may associate with a new BS, possibly in different tiers, over different time slots.

### III. TIMELY THROUGHPUT OF HETEROGENEOUS CELLULAR NETWORK

In this section, we adopt the timely throughput framework to study the reliability performance of HCN under delay constraints [10]. In this framework, the time slots are grouped to intervals, each consisting of $\tau$ consecutive time slots, as illustrated in Fig. 1. At the beginning of each interval, the typical user requests a new packet from the tagged BS with the end of the interval as the transmission deadline. Then, the BS transmits the requested packet to the user according to the aforementioned scheduling and retransmission schemes. If the packet is successfully received (named as delivered for short hereafter), the BS removes the packet from its buffer. The packet that is not delivered by the end of the interval is dropped from the BS. The timely throughput measures the average number of packets delivered before their deadline, which handles the delay requirement and reliability performance jointly [10].

#### A. Timely Throughput

First, we define the user timely throughput as the average number of delivered packets per interval per user, and denote it as $T_o$. It actually measures the mean probability that a user’s packet is delivered before its transmission deadline, which evaluates the reliability performance under a delay constraint of $\tau$ time slots. According to the definition, the user timely throughput is given by [10]

$$T_o \equiv \lim_{L \to \infty} \frac{\sum_{l=1}^{L} \eta_o (l)}{L}$$

(7)

where $l$ is the index of the interval, and $\eta_o (l)$ is an indicator defined as

$$\eta_o (l) = \begin{cases} 1, & \text{the packet is delivered in the } l\text{th interval} \\ 0, & \text{otherwise} \end{cases}$$

When the system is in stationary state, the user timely throughput can be obtained as

$$T_o = \mathbb{E} \{ \eta_o (l) \} = \mathbb{E} \{ \eta_o \}$$

(8)

where $\mathbb{E} \{ \cdot \}$ is the expectation operation, the packet transmission indicator $\eta_o (l)$ does not depend on the interval index $l$ and can be abbreviated as $\eta_o$.

In addition, we define the area timely throughput as the average number of delivered packets per interval per unit area, and denote it as $T$. Then, the area timely throughput can be expressed by the user timely throughput as

$$T = \lambda_u T_o.$$  

(9)

#### B. Main Results

The timely throughput of HCN under general case is stated in the following theorem.

**Theorem 1.** The area timely throughputs of HCN in the static and high mobility scenarios are given, respectively, by (10) and (11) (at the top of next page), where $B_k (t, \theta, \delta_j, B_j) = \sum_{\ell \geq 1} (\ell - 1)^{\ell + 1} \ell! \frac{\theta^{\ell-\delta_j}}{\ell - \delta_j} \left(\frac{\theta}{\ell}\right)^{\ell} F_1 (\ell, \ell - \delta_j; \ell - \delta_j + 1; -\theta \frac{\delta_j}{\ell})$ with $F_1 (\cdot, \cdot; \cdot; \cdot)$ being the Gauss hypergeometric function and $\theta$ being the transmission threshold, and $B (x, y) = \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)}$ is the beta function.

Dividing $T_{st}$ and $T_{hm}$ by $\lambda_u$ gives the user timely throughput in the static and high mobility scenarios, respectively.

**Proof:** The event that the packet is delivered within one interval conditioned on the number of users associated with the tagged BS and the locations of BSs, follows a Bernoulli distribution with

$$\mathbb{P} \{ \eta_o | k, N_o, \phi_K = 1 \} = 1 - \left(1 - p_{sel} | k, N_o, \phi_K \right)^T$$

(12)

where $\phi_K$ denotes all BSs of all $K$ tiers, i.e., $\{ \phi_k \}_{k \in K}$. Here, $p_{sel} | k, N_o$ is the selection probability in one time slot conditioned that there are $N_o$ users associated with the tagged BS belonging to the $k$th tier except the typical user. $p_{suc} | k, \phi_K$ is the success probability in one time slot conditioned on the locations of all BSs and that the tagged BS belongs to the $k$th tier. Using (8), the user timely throughput can be derived as

$$T_o = \mathbb{E} \{ \eta_o | k, N_o, \phi_K \} = \sum_{t=1}^{\tau} \left(\frac{T_o}{t}\right) \left(1 - p_{suc} | k, \phi_K \right)^T = \sum_{t=1}^{\tau} \left(\frac{T_o}{t}\right) \left(1 - p_{suc} | k, \phi_K \right)^T \mathbb{E} \{ p_{suc} | k, N_o, \phi_K \}$$

(13)

where $(a)$ follows from the binomial expansion of (12) and the expectation is taken over $N_o, \phi_K, \text{and } k$.

**In the static scenario**, the typical user always associates with a fixed BS belonging to a fixed tier, and $N_o$ and $\phi_K$ are also fixed over the $\tau$ time slots. Based on Assumption 1, $p_{sel} | k, N_o$ and $p_{suc} | k, \phi_K$ are independent conditioned on $k$. Therefore, we have

$$\mathbb{E} \{ p_{sel} | k, N_o, \phi_K \} = \sum_{k \in K} \mathbb{E} \{ p_{sel} | k, N_o \} \cdot \mathbb{E} \{ p_{suc} | k, \phi_K \} \cdot \mathbb{E} \{ p_{suc} | k, \phi_K \}$$

(14)

where $p_{sel}^{(t)} | k$ is the conditional joint selection probability in $t$ time slots, and $p_{suc}^{(t)} | k$ is the conditional joint success probability in $t$ transmissions, both conditioned that the typical user associates with the $k$th tier.
\[
T_{st} = \lambda_u \sum_{t=1}^{\tau} \left( t \right) (-1)^{t+1} \sum_{n=0}^{\infty} \frac{1}{(n+1)!} B(n+1, 3.5) \frac{\lambda_k}{3.5 (1 + 3.5^{-1} \rho_k)^{n+1.5}} \\
\times \int_0^\infty 2\pi r \exp \left( -\pi \sum_{j=1}^K \lambda_j \left( \frac{B_j P_j}{B_k P_k} \right)^{\delta_j} \right) \left( 1 + P_{a,j} B_k(t, \theta, \delta_j, B_j) \right)^{2\delta_j} dr \\
\]

Here, applying (5) to Proposition 2 of [9], \( p_{\text{sel}|k}^{(1)} \) can be simplified as

\[
p_{\text{sel}|k}^{(1)} = \frac{P_{a,k}}{\rho_k}. \tag{20}
\]

Substituting (17) with \( t = 1 \) and (20) into (19) yields the area timely throughput in the high mobility scenario given by (11).

The expression of timely throughput can be simplified if the path loss exponents are equal (\( \{\alpha_j\} = \alpha \)), given by the following corollary.

**Corollary 1.** When \( \{\alpha_j\} = \alpha \), the area timely throughput of HCN in the static and high mobility scenarios are given, respectively, by

\[
T_{st,\alpha} = \lambda_u \sum_{t=1}^{\tau} \left( t \right) (-1)^{t+1} \sum_{n=0}^{\infty} \frac{1}{(n+1)!} B(n+1, 3.5) \\
\times \frac{\lambda_k}{3.5 (1 + 3.5^{-1} \rho_k)^{n+1.5}} \frac{1}{1 + 3.5 (1 + 3.5^{-1} \rho_k)^{n+1.5}} \\
\times \int_0^\infty 2\pi r \exp \left( -\pi \sum_{j=1}^K \lambda_j \left( \frac{B_j P_j}{B_k P_k} \right)^{\delta_j} \right) \left( 1 + P_{a,j} B_k(t, \theta, \delta_j, B_j) \right)^{2\delta_j} dr \\
\]

\[
T_{hm,\alpha} = \lambda_u \left[ 1 - \left( 1 - \sum_{k=1}^K \frac{\lambda_k P_{a,k}}{\rho_k} \right)^T \right] \\
\times \frac{1}{3.5 (1 + 3.5^{-1} \rho_k)^{n+1.5}} \left( 1 + P_{a,k} B_k(t, \theta, \delta_j, B_j) \right)^{2\delta_j} \\
\]

**Proof:** If \( \{\alpha_j\} = \alpha \), then \( \{\delta_j\} = \delta \). Since

\[
\int_0^\infty 2\pi r \exp \left( -\pi S r^2 \right) dr = \frac{1}{2}, \tag{23}
\]

the integration in (10) can be simplified as

\[
\frac{1}{\sum_{j=1}^K \lambda_j \left( \frac{B_j P_j}{B_k P_k} \right)^{\delta_j} \left( 1 + P_{a,j} B_k(t, \theta, \delta_j, B_j) \right)^{2\delta_j}}.
\]

And replacing \( \delta_j \) by \( \delta \) in (10) gives the result in (21). Similarly, (11) leads to the result in (22).

As shown in (21) and (22), the effect of bias factor is complicated. The bias factor not only directly influences the mean load and BS activity probability of each tier, but also has an effect on the transmission embedded in \( B_k(t, \theta, \delta, B_j) \). This effect can be found from the numerical results in Section IV.
IV. NUMERICAL RESULTS

In this section, we consider a two-tier HCN, consisting of macrocell network as the 1st tier overlaid with a picocell network as the 2nd tier. The effect of mobility, key network parameters, BS density, and bias factor on timely throughput is evaluated. We apply the same path loss exponent for both tiers, and the results are obtained using Corollary 1. The densities of users and macrocell BSs are fixed as $1 \times 10^{-3}/\text{m}^2$ and $5 \times 10^{-5}/\text{m}^2$, respectively.

Fig. 2 shows the variation of user timely throughput with the density of picocell BSs. From the figure, one can observe a higher user timely throughput in the high mobility scenario as compared with the static scenario. This implies that mobility has a positive effect on delay-related network performance, due to lower temporal correlation of transmission (or scheduling) induced by mobility. A higher path loss exponent leads to a slightly higher user timely throughput in both scenarios, due to lower interference. Furthermore, the user timely throughput increases with the density of picocell BSs in both scenarios. This means that increasing picocell BS density can steadily improve the network performance, e.g., the user timely throughput, that depends on network load. As the area timely throughput is the product of the user timely throughput and the user density, we can conclude that the area timely throughput also increases with the BS density.

User timely throughput with different interval lengths is presented in Fig. 3, which reveals the tradeoff between delay and reliability. As shown, the user timely throughput increases with the interval length with a diminishing rate. Thus, we can trade off delay for reliability in low user timely throughput regime. It can also be seen that the user timely throughput experiences a much lower increasing rate via increasing the interval length in static scenario as compared with the high mobility scenario. It reminds us that by relaxing the delay deadline, the reliability can be improved more significantly in high mobility scenario than in the static scenario. This again validates the significant effect of temporal correlation of transmission on network performance. It can be found that densifying the network (as depicted in Fig. 2) is more effective on improving the reliability than relaxing the delay deadline.

Fig. 4 presents variation of user timely throughput with the bias factor for picocells under different ratios of picocell BS density to macrocell BS density. As observed from Fig. 4, bias factor has a significant impact on network performance, especially in the static scenario. Furthermore, there exists an optimal bias factor that maximizes the user timely throughput for a given network parameter setting. On one hand, increasing the bias factor of picocells, $B_2$, can push users to picocells with lower load, targeting at load balancing. On the other hand, if $B_2$ is increased, the coverage area of picocell BS is enlarged, which increases the picocells’ load and BS activity probability, and in turn increases interference to the network. And, the former/later effect dominates when $B_2$ is below/above the optimal value.

V. CONCLUSIONS

In this paper, we have presented an analytical framework to investigate the delay/reliability performance of HCN in terms of timely throughput. As observed from the numerical results, densifying the network can improve network performance in terms of timely throughput, and mobility has a positive effect
on timely throughput. Furthermore, an optimal bias factor exists for achieving a maximal timely throughput for a two-tier HCN, while behaves differently in static and high mobility scenarios. In summary, this work provides an important step to fundamentally understand HCN, especially the interaction of delay and reliability.

**APPENDIX**

The conditional joint success probability is defined as

$$p_{\text{suc}}^{(t)}[k, R_k] = \mathbb{P}(\text{SIR}_k > \theta, j = 1, \ldots, t | R_k)$$

where $S_l$ denotes the event that the $i$th transmission succeeds. Conditioned on that the distance between the typical user and the tagged BS is $R_k$, the set of active BSs in the $j$th tier outside the area $B(o, z_j)$ form the interfering BSs, denoted as $\Phi_j$. Here, $B(o, z_j)$ is the ball centered at the origin with radius $z_j$, and $z_j = (\frac{B_o P_k}{\alpha R_k})^{\frac{1}{\alpha R_k - 1}}$, which is due to the association rule. Then, the conditional joint success probability conditioned on $R_k$ is obtained as

$$p_{\text{suc}}^{(t)}[k, R_k] = \mathbb{P}(\text{SIR}_k > \theta, j = 1, \ldots, t | R_k)$$

where $h_{j,i}$ is the Rayleigh fading coefficient. (a) holds due to the independence of $h_{k,i}, i = 1, \ldots, t$ and the independence of $\Phi_j, j = 1, \ldots, K$; (b) holds due to the independence of $h_{j,i}, i = 1, \ldots, t$ and the expectation of the exponential distributed random variable; (c) follows from the probability generating functional of PPP. Employing the change of variables $u = \nu^{\alpha_j}$, the integration $F_j$ can be calculated as

$$F_j = 2\pi \int_{\nu_{\alpha_j}}^{\infty} \left( \frac{1}{1 + \theta R_k^{\alpha_j} P_j / P_k} \right)^{\frac{1}{\alpha_j}} \frac{1}{\nu^{\alpha_j}} d\nu$$

where (d) follows from the binomial expansion, (e) employs the change of variables $u = \theta R_k^{\alpha_j} P_j / P_k$, and $F_j(\ell, \theta, \delta_j) = \frac{\theta^{1-\delta_j} \lambda R_k^{\alpha_j} P_j / P_k}{1 + \theta R_k^{\alpha_j} P_j / P_k}$, with $F_j(\ell, \theta, \delta_j)$ being the Gauss hypergeometric function $[12, \text{Eqn. 3.194.2}]$. Therefore, the conditional joint success probability in (25) is obtained as

$$p_{\text{suc}}^{(t)}[k, R_k] = \prod_{j=1}^{K} \mathbb{E}_{\Phi_j} \left\{ \prod_{b \in \Phi_j} \mathbb{P}[h_{j,b} \geq \lambda R_k^{\alpha_j} P_j / P_k] \right\}$$

Furthermore, the distance $R_k$ is a random variable with the probability density function given by [8]

$$f_{R_k}(r) = \frac{2\pi \lambda_k}{\theta R_k^2} \exp \left( -\pi \sum_{j=1}^{K} \lambda_j \left( \frac{B_j}{B_k} \right)^{\delta_j} \frac{2\pi R_k^{\alpha_j}}{\nu^{\alpha_j}} \right).$$

Then, (17) is obtained by averaging over $R_k$.

**REFERENCES**


