Analysis of QoS in Heterogeneous Networks with Clustered Deployment and Caching Aware Capacity Allocation

Takehiro Ohashi*
Department of Mathematical and Computing Science, Tokyo Institute of Technology
Email: *ohashi.t.ag@m.titech.ac.jp

Abstract—In cellular networks, the densification of connected devices and base stations engender the ever-growing traffic intensity, and caching popular contents with a smart management is a promising way to alleviate such consequences. Our research extends the previously proposed analysis of three-tier cache enabled Heterogeneous Networks (HetNets). The main contributions are threefold. We consider the more realistic assumption; that is, the distribution of small base stations is following Poisson-Poisson cluster processes, which reflects the real situations of geographic restrictions, user dense areas, and coverage-holes. We propose the allocation of downlink data transmission capacity according to the cases of requested contents which are either cached or non-cached in nearby nodes and elucidate the traffic efficiency of the allocation under the effect of clustered deployment of small base stations. The throughput and delay of the allocation system are derived based on the approximated sojourn time of Discriminatory Processor Sharing (DPS) queue. We present the results of achievable efficiency and such a system’s performance for a better caching solution to the challenges of future cellular networks.

Keywords: Heterogeneous cellular networks, Spatial stochastic models, Queuing Theory, Stochastic Geometry, Poisson-Poisson cluster processes, proactive caching, D2D transmission, Discriminatory Processor Sharing queue, Quality of Service (QoS), capacity allocation, service differentiation.

I. INTRODUCTION

The ever-growing communication traffic and capacity demand in cellular networks with the rise of complex networks composed of heterogeneity of cells necessitate smart traffic management. To expand the fifth-generation cellular systems (5G) network or even proceed beyond sixth-generation cellular systems (6G) in reality, the challenge is to support hundreds of gigabits of traffic from user’s devices to the core network through the backhauls as well as the high throughput of user-to-node data transmission while satisfying the extreme requirements such as availability, latency, energy, and cost-efficiency [1]–[3]. To this end, caching popular contents in the networks was proposed and analysed extensively for it’s potentials [1]–[5]. In this paper, we extend the previous analysis of the cache-enabled heterogeneous network conducted by [6]. The three-tier network, consisting of base stations (BSs), relays, and device-to-device sharing links (D2Ds) was considered, and their result showed that the global throughput of the proposed caching system can be increased significantly. To carry out such a system, in reality, more analysis is needed because of the complexity of the spatial dynamics of cellular networks. We also tackle the challenge of allocating network resources for better efficiency under this complexity.

The contributions of this work are three-fold. To reinforce the result of the previous analysis, instead of considering all three tiers following independent Poisson point process (PPP), we consider that the deployment of small cells (either relay or small base stations) is distributed as Thomas cluster process (TCP), a class of Poisson-Poisson cluster processes (PPCP), which exhibit attractive point patterns and reflect the real situations of small cell deployment influenced by geographic restrictions, user dense areas and coverage-holes. This assumption of the attractive pattern of small cell deployment is in-negligible when we analyze the performance of the caching system since the performance of downlink data transmission is largely influenced by whether the requested contents are cached nearby cells or not. Next, we propose a new system for more efficient caching solutions, that is, allocating the downlink data transmission capacity according to the different circumstances of contents requests from users. This capacity allocation system is modeled by DPS queue, a variant of multi-class processor sharing queue, which assigns various weights on each service rate of different user classes and enables the service differentiation. The previous work in [6] used egalitarian processor sharing (EPS), a simple queuing model that shares service rate equally. The property of flow-level performance of data transmission in a cellular network was well studied using PS queue as well as DPS queue [7], [8] but without taking an account of complex spatial stochastic dynamics of users and base stations. The study of cellular networks is very limited without considering spatial stochastic dynamics since the signal to interference noise ratio (SINR) perceived by a user depends on complicated interaction among the spatial distribution of base stations. Recently, successful analyses are using both stochastic geometry approach and queuing theory to overcome this complex spatial dynamics such as [6], [9], [10]. Thus far, EPS queue and generalized processor sharing (GPS) queue were used while the DPS queue was not the case. Mathematically, DPS queue is complex and some tractability is lost, and many analyses are limited under special circumstances [11]–[13]. Instead of using the exact analysis of the DPS queue, we use the best approximation derived by [14] to show the possible improvement of a cellular
network traffic efficiency by the allocation system. To the best of our knowledge, this is the first application DPS queue while taking an account of the effect of spatial stochastic patterns of users and base stations to analyze the capacity allocation performance of HetNets. Lastly, we compare numerically the results obtained by [6] with clustered small cells and our capacity allocated system. We show that the clustered deployment of small cells differs the traffic of entire networks from the uniformly distributed one and by prioritizing downlink data transmission of a particular class of users the throughput of entire networks can be improved further.

A. Paper Organization

The rest of the paper is organized as follows. In section II, we introduce the system model of our three-tier HetNets and protocols of caching as well as tier association probability among users in different circumstances. Note that the tier association probabilities derived in this section are needed to express the average downlink data transmission rate of a typical user as well as the probability of a user active in different circumstances. In section III, we derive the average downlink data transmission rate called average ergodic rate. In section IV, we derive QoS metrics such as the mean number of daughter points among users in different circumstances. Note that the tier protocols of caching as well as tier association probability of contents and have a caching storage with the fixed size of contents and have a caching storage with the size $M_t$ [bits]. All SBSs can cache $M_t$ number of contents and have a caching storage with the same size $M_t$ [bits]. Similarly, the cache enabled users can cache $M_t$ number of contents and have a caching storage with the fixed size

A. Network Architecture

We consider three-tier HetNet consisting of MBSs, SBSs, D2Ds. All SBSs are connected to the nearest MBS through backhauls, and then MBSs are connected to a core network (see Fig. 1). We define that the nodes of users and MBSs follow independent homogenous PPPs in $\mathbb{R}^2$ denoted as $\Phi_i$ with intensity $\lambda_i$ for $i = 0, 3$. The nodes of SBSs follow PPPC denoted as $\Phi_2$ in $\mathbb{R}^2$ with conditional intensity function given the parental point process $\lambda_2(x) = \bar{m}\sum_{x \in \Phi_2} f_d(x|x)$ where $f_d$ is the parental point process following PPP with intensity $\lambda_p$, and $\bar{m}$ is the mean number of daughter points distributed according to some distribution of $f_d(x|x)$ around the center of each parental point $z$. In this paper for the ease of our analysis, we assume that the daughter points are independently and normally distributed around a parental point $z$, i.e. $f_d(x) = \exp(-||x||^2/(2\sigma^2))/(2\pi\sigma^2)$ which is called the Thomas cluster point process (TCP). The formal definition of PPPC is

**Definition 1. Stationary PPCP on $\mathbb{R}^2$**

$$\Phi_2 = \cup_{X_i \in \Phi_{MBS}} (X_i + \Psi_i),$$  \hspace{1cm} (1)

where $\Phi_{MBS} = \{X_i\}_{i=1}^\infty$ is the parental point process, homogenous PPP with intensity $\lambda_p$, and a mark $\Psi_i = \{Y_{i,j}\}_{j=1}^N_i$ of the point $X_i$ is in-homogeneous PPP with the intensity function $\lambda_d(y)$ s.t. $\int_{\mathbb{R}^2} \lambda_d(y)dy = \int_{\mathbb{R}^2} \bar{m} f_d(y)dy = \bar{m}$ is the mean number of daughter points. $f_d$ is the density function of each daughter point around its parent.

When a PPCP is conditioned on the parental point process, it is an in-homogeneous PPP. The more extensive description of PPCP can be found in [15], [16]. Let $\mathcal{B}(\mathbb{R}^2)$ be the Borel $\sigma$-algebra on $\mathbb{R}^2$ and $\mathcal{A} \subset \mathcal{B}(\mathbb{R}^2)$. Conditioned on $\Phi_{MBS}$, the counting measure of $\Phi_2$ is random measure which counts the number of points of $\Phi_2$ falling in $\mathcal{A}$ and is given by (see [15], [16])

$$N_{\Phi_2}(A)|\Phi_{MBS} \sim \text{Poisson}(\bar{m} \sum_{x \in \Phi_{MBS}} \int_A f_d(y-x)dy).$$  \hspace{1cm} (2)

We consider [6] as a baseline in this paper, and in order to compare the results we let the intensity of SBSs in baseline case (which is assumed to follow independent homogeneous PPP) be the same as our scenario ($\bar{m}\lambda_p = \lambda_2$).

For the user tier, we have the proportion of cache-enabled users who can transmit some contentscached in their local storage to some users who are not cache-enabled. Those cache enabled users are distributed as a thinning of the homogeneous PPP of $\Phi_1$ with intensity $\lambda_1 = \alpha \lambda_0$, where $\alpha \in [0, 1]$ and $\lambda_0$ is the intensity of the nodes of users. Considering the real circumstances, we set the intensity as $\lambda_0 \gg \lambda_2 > \lambda_3$.

There are $N$ total number of contents, and all of them are stored in MBSs. We assume all the contents are the same size denoted as $S$ [bits]. All SBSs can cache $M_2$ number of contents and have a caching storage with the size $M_2 \times S$ [bits]. Similarly, the cache enabled users can cache $M_1$ number of contents and have a caching storage with the fixed size

| $\Phi_i$ | The point process of the $i$-th BS/D2D tier |
| $f_{cd,i}$, $F_{cd,i}$ | PDF and CCDF of contact distance of $\Phi_i$ |
| $f_d(s)$ | The displacement kernel defined $\exp(-\frac{||x||^2}{2\sigma^2})/2\pi\sigma^2$ |
| $\tau_i(r)$ | $\left\{ \begin{array}{ll} m \int_{0}^{\infty} 2\pi \lambda_2 \frac{1}{2}(\frac{r}{\sigma})^{\frac{3}{2}} \exp(-m(1-Q_1(\frac{r}{\sigma}, \frac{\sigma^2}{\sigma^2})))dz, & \text{if } i = 2 \\ 2\pi \lambda_i r, & \text{if } i = 1, 3 \end{array} \right.$ |
| $f_{S,i}, f_{S_{1,i}}, f_{S_{1,1,i}}$ | PDF of distance of a serving node given an event $S_i$, $S_1, S_{1,1}$ |
| $F_{S,i}, F_{S_{1,i}}, F_{S_{1,1,i}}$ | CCDF of distance of a serving node given an event $S_i$, $S_1, S_{1,1}$ |
| $2F_1[a, b, c; d]$ | Gaus hyper-geometric function |
| $\alpha$ | Ratio of cache-enabled users ($\alpha \in [0, 1]$) |
| $\beta$ | Path loss exponent ($\beta = 4$) |
\[ M_1 \times S \text{ [bits]. It is a natural to consider } M_1 \ll M_2 \ll N. \text{ See Fig. 1, the blue disks are represented as caching of } M_1 \text{ number of contents which all the cache enabled users to have and contain the same copies of the contents. SBSs have } M_2 \text{ number of contents which includes } M_1 \text{ number of contents as well. MBSs have all of the contents stored in their storage. The } i\text{-th popular content requested by the typical user follows the Zipf distribution, and for arbitrary content popularity ranks } a, b \text{ with } a < b:\]

\[ f_i = \frac{1}{\sum_{j=1}^{N} \frac{1}{j^\gamma}}, \quad \sum_{i=a}^{b} f_i \triangleq F_{pop}(a, b), \quad (3) \]

where \( \gamma \geq 0 \) is the parameter of skewness of the content popularity distribution. Then, we assume all the cache-enabled users have cached 1st to \( M_1 \)-th popular contents in their storage. This implies that the probability of a content requested by the typical user that is stored in cache-enabled user is \( F_{pop}(1, M_1) \). Similarly, for all SBSs, 1st to \( M_2 \)-th popular contents are cached in the storage. Therefore, the probability of a content requested by the typical user that is stored in SBSs is \( F_{pop}(1, M_2) \).}

![Fig. 1: Illustration of three tier cache enabled heterogeneous networks. The blue circle is the coverage area of MBS. The orange circle is the coverage area of SBS.](image)

**B. Access and Cache Protocol**

For the completeness of this paper, we summarize all the assumptions from [6] and emphasis some of the important concepts here. We assume all users are following the max-power association rule explained in the previous subsection. This means that all users request contents from a node which provides the highest power that is the closest node in a tier. For \( i = 1, 2, 3 \), the maximum received power from the \( i\)-th tier is

\[ C_i = \nu B_i P_i ||x_i^*||^{-\beta}, \quad (4) \]

where \( ||x_i^*|| \) is the distance from the typical user at \((0,0)\) to its closest BS in \( i\)-th tier, \( x_i^* \in \Phi_i \); that is \( x_i^* = \arg \max_{\Phi_i} P_i \|x\|^{-\beta} = \arg \min_{\Phi_i} \|x\| \). \( P_i \) is a transmit power of the node in \( i\)-th tier, and \( \nu, B_i, \) and \( \beta \) are the propagation constant, bias, and path-loss exponent, respectively. For the ease of analysis, we let \( \nu = 1 \) and \( B_i = 1 \). In this paper we consider the path-loss exponent \( \beta = 4 \). Under this max-power association rule there are four different cases that the typical user connects to corresponding BSs/D2Ds.

![Fig. 2: Illustration of contents requests in different cases.](image)

- Case 1: when the typical user is not cache-enabled, the user requests contents to a nearby node, either MBS, SBS, or cache-enabled user (D2D).
- Case 2: when the typical user is cache-enabled and the requested contents by the user are not cached in the storage, the only choice is to request the contents to either MBS or SBS.
- Case 3: when the typical user is not cache-enabled and a cache-enabled user (D2D) provides the highest power but the requested contents by the user were not cached in the storage of the cache-enabled user, the user requests the contents to either MBS or SBS.
- Case 4: when the typical user is cache-enabled and the requested contents by the user were cached in his/her local storage, the contents were retrieved and reused immediately.

![Fig. 2](image)

**C. Tier Association Probability**

We derive the probability of ordering of maximum power received by the typical user from each tier \( \Phi_i \) for \( i = 1, 2, 3 \), where \( \Phi_1, \Phi_2, \) and \( \Phi_3 \) follow independent PPP and \( \Phi_2 \) follows independent TCP. This probability is used to characterize the requests of contents from the typical user to BSs/D2Ds under the max-power association rule explained in the previous subsection. It is commonly called the tier association probability which is the probability that the typical user communicates to a tier.

The PDF of contact distance of \( \Phi_i \) for \( i \in \{1, 3\} \) denoted as \( f_{cd_i} \) and the complementary CDF denoted as \( F_{cd_i} \) are

\[ f_{cd_i}(r_i) = 2\pi \lambda_i r_i e^{-\pi \lambda_i r_i^2}, \quad F_{cd_i}(r_i) = e^{-\pi \lambda_i r_i^2}. \quad (5) \]
The conditional PDF of contact distance of $\Phi_2$ given the parental point process $\Phi_{p2}$ is derived (see, Lemma 1 [15], [16]) and given by

$$f_{cd2}(r_2|\Phi_{p2}) = \bar{m} \sum_{i=1}^{\infty} \frac{1}{\sigma} q(\frac{\|X_i\|}{\sigma}, \frac{r_2}{\sigma}) \prod_{i=1}^{\infty} \exp(-\bar{m}(1-Q_1(\frac{\|X_i\|}{\sigma}, \frac{r_2}{\sigma}))),$$

where $Q_1$ is the first-order Marcum Q-function $Q_1(a, b) = \int_{\frac{a^2}{2}}^{\infty} z \exp(-\frac{z^2 + a^2}{2})I_0(az)dz$ and $I_0(z) = \pi^{-1} \int_{0}^{\pi} e^{\cos \varphi} d\varphi$ is the modified Bessel function of the first kind with order zero. $q(a, b)$ is PDF of Rician distribution, and its cumulative distribution function is $1 - Q_1(a, b)$. Note that the CDF of contact distance distribution of TCP was derived in [17], and the complementary CDF is given by

$$\bar{F}_{cd2}(r_2) = \exp(-\int_{0}^{\infty} 2\pi \lambda_p z(1 - \exp(-\bar{m})(1 - Q_1(\frac{z}{\sigma}, \frac{r_2}{\sigma}))))dz).$$

Then, by using (6) we can derive the PDF of contact distance distribution of TCP in the following Lemma 1.

**Lemma 1.** The PDF of contact distance distribution of TCP is

$$f_{cd2}(r_2) = \mathbb{E}_{\Phi_{p2}}[f_{cd2}(r_2|\Phi_{p2})] = \mathbb{E}_{\Phi_{p2}}[\bar{m} \sum_{i=1}^{\infty} \frac{1}{\sigma} q(\frac{\|X_i\|}{\sigma}, \frac{r_2}{\sigma}) \prod_{i=1}^{\infty} \exp(-\bar{m}(1-Q_1(\frac{\|X_i\|}{\sigma}, \frac{r_2}{\sigma})))].$$

**Proof.** We can directly apply the sum-product functional with respect to parental PP which follows PPP (see Lemma 6 [16]).

Next, we derive the probability of ordering the maximum power received by the typical user from each tier.

**Proposition 1.** The ordered tier association probability of three tier $i, j, k \in \{1, 2, 3\}$, $i \neq j \neq k$ is

$$P(C_i > C_j > C_k) = \int_{0}^{\infty} \int_{0}^{\infty} \frac{1}{\sigma} \bar{m} \prod_{i=1}^{\infty} \exp(-\bar{m}(1-Q_1(\frac{\|X_i\|}{\sigma}, \frac{r_2}{\sigma})))dz \times \exp(-\int_{0}^{\infty} 2\pi \lambda_p z(1 - \exp(-\bar{m}(1 - Q_1(\frac{z}{\sigma}, \frac{r_2}{\sigma}))))dz) \times \prod_{i=1}^{\infty} 2\pi \lambda_p r_i \exp(-\pi \lambda_p r_i^2)dr_kdr_jdr_i.$$
III. AVERAGE ERGODIC RATE

In this section, the average ergodic rate of the downlink is analyzed under the three-tier HetNet described in the previous section. We assume that the communication links between BSs/D2Ds to users share the same frequency bandwidth, and this yields the interference among the links. Recall that there are four cases of the typical user’s requests of contents, and the rate of average downlink data transmission differs in those different cases (see Fig. 2).

A. Active BSs and D2D Links

We define all active nodes which yield the interference. We denote \( \lambda_i \) by the intensity of active nodes in tier \( i \) for \( i = 1, 2, 3 \). For the tier 1, not all cache-enabled users need to be active when the number of requests from non-cache-enabled users through the D2D transmission is less than the cache-enabled users. Thus, we define the intensity of active BSs in tier \( i \) as \( \lambda_i = \min \{ \lambda_0, \lambda_0 \Phi_i, (1 - \alpha) P_{\text{pop}}(1, M_i) \} \). For the other tiers, since all MBs and SBSs are active over the time, we define \( \lambda_i = \lambda_i \). We also define \( \Phi_i \) as the thinned PPP of active BSs in tier \( i \), and \( \Phi_2 \) and \( \Phi_3 \) are unchanged from original MBs and SBSs.

B. Signal to Interference Plus Noise Ratio

The downlink power received from the serving BS/D2D at \( x_i \in \Phi_i \) is \( P_i h_{xi} ||x_i||^{-\beta} \) for \( i = 1, 2, 3 \), and the signal-to-interference-plus-noise ratio (SINR) of the typical user located at the origin connecting to that node is

\[
\text{SINR}_i(||x_i||) = \frac{P_i h_{xi} ||x_i||^{-\beta}}{\sum_{j=1}^{3} \sum_{y \in \Phi'_j \setminus \{x_i\}} P_j h_y ||y||^{-\beta} + N_0},
\]

where \( h_{xi} \) and \( h_y \) follow Rayleigh fading distributed as exponential with unit mean. In the denominator, all \( ||y|| \) is the distance between reference user to its interfering active nodes in \( j \)-th tier. \( N_0 \) is thermal noise. Let \( x \) be the distance of a requesting user to the serving node. The average ergodic rate of the typical user when it communicates with the \( i \)-th tier is given by [6]:

\[
U_{i,t} = \mathbb{E}[\mathbb{E}_{\text{SINR}_i}[\ln(1 + \text{SINR}_i(x))||x|| = x]],
\]

where the expectation of \( \text{SINR}_i \) is with respect to all \( \Phi'_j \) for \( j = 1, 2, 3 \) and the fading effect \( h_{xi} \) and \( h_y \), and then the expectation of \( ||x|| \) is over the distance between the typical user and its serving node. The equation above indicates the average ergodic rate, which is the mean rate of data transmitted, of a randomly chosen user associated with the \( i \)-th tier in a cell.

C. The Average Ergodic Rate in Case 1

Let us define the association event such that the typical user requests contents to the \( i \)-th tier as \( S_i = \{ ||x_i||^{-\beta} > \max_{n \neq i} P_n ||x_n||^{-\beta} : n \in \{1, 2, 3\} \} \). The PDF of the distance distribution of the serving node conditional on parental PP and \( S_i \) is given in Lemma 4, [6]. Then, we can obtain the PDF of the distance between the typical user and the associated serving node unconditional on \( \Phi_{p2} \) and conditional on an event \( S_i \).

**Lemma 2.** The PDF of the distance from the typical user associated to a node in the \( i \)-th tier for \( i = 1, 2, 3 \) is

\[
f_{S_i}(x) = \frac{\tau_i(x)}{G_{3,i}} \exp(-\int_0^x 2\pi \lambda_{p2} z \times \left(1 - \exp(-m(1 - Q_1(\frac{z}{\sigma}, \frac{(P_{p2}/P_i)^{\frac{\beta}{\beta-1}})\sigma)))dz\right) \times \exp(-\pi \sum_{j \in \{1,3\}} \lambda_{p2} \frac{(P_{p2}/P_i)^{\frac{\beta}{\beta-1}} x^2})
\]

where \( \tau_i(x) \) is given in the TABLE 1.

**Proof.** The CCDF of \( f_{S_i}(x) \) can be expressed as

\[
\bar{F}_{S_i}(r|S_i) = \mathbb{P}([x_i] > r|S_i)
\]

\[
= \frac{\mathbb{P}(R_i > r, S_i)}{\mathbb{P}(S_i)}
\]

\[
= \frac{1}{\mathbb{G}_{3,i}} \mathbb{P}(\bigcap_{j \in \{1,2,3\}\setminus \{i\}} P_i R_i^{\beta} > P_j R_j^{\beta}, R_i > r).
\]

Then, the last equation is similar to what we derived in Corollary 1, and by taking the derivative with respect to \( r \) we get the desired result. \( \square \)

Since we know that the distribution of the distance of the typical user and the associated serving node, we are ready to obtain the average ergodic rate in case 1 as follows.

**Theorem 1.** The average ergodic rate of the typical user connecting to a node in the \( i \)-th tier in case 1 is

\[
U_{i,1} = \int_0^\infty \int_0^\infty \mathcal{M}_{y}(x,s) \frac{\mathbb{G}_{3,i}}{1+s} \frac{d \tau_i(x)}{d s} \times \exp(-\int_0^x 2\pi \lambda_{p2} z (1 - \exp(-m(1 - Q_1(\frac{z}{\sigma}, \frac{(P_{p2}/P_i)^{\frac{\beta}{\beta-1}})\sigma)))dz) \times \exp(-\pi \sum_{j \in \{1,3\}} \lambda_{p2} \frac{(P_{p2}/P_i)^{\frac{\beta}{\beta-1}} x^2})dx,
\]

where

\[
\mathcal{M}_{y}(x,s) = \prod_{j \in \{1,3\}} \exp(-2\pi \lambda_{j,x} x^2 \frac{s(P_{p2}/P_i)^{\frac{\beta}{\beta-1}}}{\beta(1 - \frac{\beta}{s})} 2F_1[1, 1 - \frac{2}{\beta}; 2 - \frac{2}{\beta}; -s])
\]

\[
\exp(-2\pi \lambda_{p2} \int_0^\infty (1 - \exp(-m \int_0^\infty \frac{1}{\sqrt{2\pi \sigma^2}} \exp(-\frac{y^2}{2\sigma^2})dy))dz).
\]

and \( 2F_1[a; b; c; d] \) is Gauss hyper-geometric function.

**Proof.** See Appendix A. \( \square \)
D. The Average Ergodic Rate in Case 2

In case 2, the typical user requests contents to either MBSs or SBSs. Therefore, by following the similar derivation as Lemma 1, given an event \( S_i = \{ P_1 \| x_i^* \|^{- \beta} > \max_{j \neq i} P_j \| x_j^* \|^{- \beta} : j \in \{2,3\}\} \) the unconditional PDF of the distance between the typical user and the serving node in the \( i \)-th is

\[
f_{S_i}(x) = \frac{\tau_i(x)}{F_{i,i}} \exp(- \int_0^\infty 2\pi\lambda_{P_2} z x \times (1-\exp(-\tilde{m}(1-\frac{P_2}{\pi\lambda_{P_2}})^\frac{z}{\sigma})))dz) \times \exp(-\pi\lambda_3(\frac{P_3}{F_i})^{\frac{1}{2}}x^2),
\]

(21)

where \( i,j \in \{2,3\} \) s.t. \( i \neq j \) and \( \tau_i(x) \) is defined in TABLE 1. The derivation is similar to Lemma 2 and therefore omitted. For the average ergodic rate of case 2, the interfering active D2D nodes can be closer than a serving node, which is either MBS or SBS. This is because the typical user is cache-enabled and can not request the contents from the nearest cache-enabled user through the D2D link (see Fig. 2). Therefore, the possible distance between a requested user to the closest cache-enabled user denoted as \( \alpha \) can be as close as 0; that is, the possible distance \( \alpha \rightarrow 0 \). Then, by considering this we have the following theorem.

Theorem 2. The average ergodic rate of the typical user connecting to a node in the \( i \)-th tier in case 2 is

\[
U_{2,i} = \int_0^\infty \int_0^\infty \frac{\mathcal{M}_y(s,x)}{1+s} ds \frac{\tau_i(x)}{F_{i,i}} \times \exp(- \int_0^\infty 2\pi\lambda_{P_2} z (1-\exp(-\tilde{m}(1-\frac{P_2}{\pi\lambda_{P_2}})^\frac{z}{\sigma})))dz) \times \exp(-\pi\lambda_3(\frac{P_3}{F_i})^{\frac{1}{2}}x^2)dx,
\]

(22)

where

\[
\mathcal{M}_y(s,x) = \exp(-2\pi\lambda_1 x^2 \frac{s \frac{P_3}{F_i} (\frac{P_2}{\pi\lambda_{P_2}})^{\frac{1}{2}}}{\beta(1-\frac{2}{\beta})} F_1[1,1-\frac{2}{\beta};\frac{2}{\beta};-s \frac{P_3}{F_i} (\frac{P_2}{\pi\lambda_{P_2}})^{\frac{1}{2}}]) \times \\
\exp(-2\pi\lambda_3 x^2 \frac{s \frac{P_3}{F_i} (\frac{P_2}{\pi\lambda_{P_2}})^{\frac{1}{2}}}{\beta(1-\frac{2}{\beta})} F_1[1,1-\frac{2}{\beta};\frac{2}{\beta};-s]) \times \\
\exp(-2\pi\lambda_{P_2} \int_0^\infty (1-\exp(-\tilde{m}) \int_0^\infty \frac{1}{(s \frac{P_3}{F_i})^{\frac{1}{2}} x + (s \frac{P_3}{F_i})^{\frac{1}{2}} (\frac{y+\frac{1}{x}}{2})^2} dy)dz).
\]

Proof. See Appendix B.

E. The Average Ergodic Rate in Case 3

Recall that in case 3, the typical user who is not cache-enabled and the highest power provider is a D2D node but the requested contents are not cached in its local storage; therefore, the only choice for the user is to request the contents to either MBS or SBS (see Fig. 2). In this case, the PDF of the distance between the typical user and the associated serving node (MBS or SBS) is a joint PDF under the event such that a cache-enabled user is the node providing the highest power while either MBS or SBS is the next highest one. Let \( x \) be the distance between the typical user and the cache-enabled user providing the highest power, and let \( y \) be the distance between the user and the associated serving node in \( j \)-th tier (MBS or SBS). Given an event \( S_{ij} = \{ P_1 \| x_i^* \|^{- \beta} > P_j \| x_j^* \|^{- \beta} \cap P_j \| x_j^* \|^{- \beta} > P_k \| x_k^* \|^{- \beta} : j,k \in \{2,3\} \) s.t. \( j \neq k \)}, the joint PDF of the distance between the typical user to the serving node in \( j \)-th tier is

\[
f_{S_{ij}}(x,y) = \frac{\tau_i(x)\tau_j(y)}{F_{i,j,k}} \exp(-\pi\lambda_1 x^2) \times \\
\exp(- \int_0^\infty 2\pi\lambda_{P_2} z (1-\exp(-\tilde{m}(1-\frac{P_2}{\pi\lambda_{P_2}})^\frac{z}{\sigma})))dz) \times \\
\exp(-\pi\lambda_3(\frac{P_3}{F_i})^{\frac{1}{2}}x^2 y^2),
\]

(23)

where \( j,k \in \{2,3\} \) s.t. \( j \neq k \) and \( \tau_i(x) \) and \( \tau_j(y) \) are defined in TABLE 1. The derivation is similar to the Lemma 2 using the result of Proposition 1. Therefore, we skip the derivation. Then, following a similar argument as case 1 and case 2, we obtain the next theorem.

Theorem 3. The average ergodic rate of the typical user connecting to a node in the \( j \)-th tier in case 3 is

\[
U_{3,j} = \int_0^\infty \int_0^\infty \frac{\mathcal{M}_y(s,x,y)}{1+s} \frac{\tau_i(x)\tau_j(y)}{F_{i,j,k}} \exp(-\pi\lambda_1 x^2) \times \\
\exp(- \int_0^\infty 2\pi\lambda_{P_2} z (1-\exp(-\tilde{m}(1-\frac{P_2}{\pi\lambda_{P_2}})^\frac{z}{\sigma})))dz) \times \\
\exp(-\pi\lambda_3(\frac{P_3}{F_i})^{\frac{1}{2}}x^2 y^2) dx dy,
\]

(24)

where

\[
\mathcal{M}_y(s,x,y) = \exp(-2\pi\lambda_1 y^2 \frac{s \frac{P_3}{F_i} (\frac{P_2}{\pi\lambda_{P_2}})^{\frac{1}{2}}}{\beta(1-\frac{2}{\beta})} F_1[1,1-\frac{2}{\beta};\frac{2}{\beta};-s \frac{P_3}{F_i} (\frac{P_2}{\pi\lambda_{P_2}})^{\frac{1}{2}}]) \times \\
\exp(-2\pi\lambda_3 y^2 \frac{s \frac{P_3}{F_i} (\frac{P_2}{\pi\lambda_{P_2}})^{\frac{1}{2}}}{\beta(1-\frac{2}{\beta})} F_1[1,1-\frac{2}{\beta};\frac{2}{\beta};-s]) \times \\
\exp(-2\pi\lambda_{P_2} \int_0^\infty (1-\exp(-\tilde{m}) \int_0^\infty \frac{1}{(s \frac{P_3}{F_i})^{\frac{1}{2}} y + (s \frac{P_3}{F_i})^{\frac{1}{2}} (\frac{y+\frac{1}{x}}{2})^2} dy)dz).
\]

Proof. See Appendix C.

We denote \( U_{d,i} \) as the average ergodic rate of the typical user in case 4. The rate is extremely fast since the requested contents can be fetched from their local device immediately.
We are interested in the QoS of the downlink data rate among the cases and therefore ignore case 4 for our study.

IV. QoS of Clustered Deployment and Caching Aware Capacity Allocation

In this section, we analyze the QoS of the cluster deployment and the caching aware capacity allocation system. We define all essential settings and parameters for modeling the traffic flow of caching aware capacity allocation. We follow the same scenarios of the system and the configurations of the parameters as the baseline work (see Section V, [6]).

A. User State and Class

First, we define the classes of users in all four cases to characterize the presence of users who requests contents under different circumstances. We consider that the typical user is in a certain state depending on which tier to request and whether the request is BH-needed or not. The probability that the typical user is active in different states is expressed in the typical user is active in different states is expressed in the state of the active user. This matrix is essentially identical to what is given in the baseline, where they provide it as the comprehensive tabular format (see TABLE I [6]). The rows of the matrix indicate all four cases with BH-needed or not, and there are eight different states in total, where we regard those eight states as classes of user. The columns represent tiers from which the user in a certain state requests contents. For \( i = 1, \ldots, 8 \), and \( j = 1, \ldots, 4 \), we get the intensity of active users in each single BS/D2D by the thinning property of independent PPP; for example, \( \lambda_{0}D_{i, j} \) is the intensity of users in the state of \( i \)-th column and \( j \)-th column. We define that the set of index of rows represents a class such that \( \{1, 2, 3, 4\} \) \( \times \{5, 6, 7, 8\} \) \( \in \{\text{Class 1, Class 2, Class 3, Class 4}\} \), and the index with even number represents BH-needed while the odd number represents backhaul free (BH-free). The set of index of columns represents a tier where we see that \( \{1, 2, 3, 4\} \in \{\text{D2D, SBS, MBS, Local}\} \). Here, "Local" stands for the case when the users who request contents to their own local storage when the contents are cached.

B. User Request Arrival and Service Rate

Next, we define the total request arrival rate of class of users in each state represented as \( (i, j) \) for \( i \in \{1, 2, \ldots, 8\} \) and \( j \in \{1, 2, 3, 4\} \). Each MBS cell has the average number of \( \lambda_{0}D_{i, j} \) users, and the rate of a request by each user in that cell is modeled as homogeneous Poisson process with mean arrival rate of \( \lambda \) [requests/s]. Then, the requests of a single user is homogeneous Poisson process with parameter \( \lambda \). Based on the matrix \( D_{8 \times 4} \), we can obtain the average number of users who are active in the corresponding state in each state. For \( i = 1, \ldots, 8 \) and \( j = 1, \ldots, 4 \), the average number of users in each state is \( \lambda_{i, j}D_{i, j} \), where \( \lambda_{i, j} = \alpha_{i, j} \lambda_{0} \). The total request arrival rate in each BS/D2D is ready to be obtained as \( \lambda_{i, j} = \sum_{i, j} \frac{\lambda_{i, j}D_{i, j}}{\lambda_{0}} \). Each single request by a user consist of a set of contents, and the volume of a set of contents per request is a random variable denoted as \( B \) and follows the exponential distribution with mean \( \frac{1}{\eta} \) [contents/request]. Next, we define the average ergodic rate of the user in the state of \( (i, j) \) as follows:

\[
A_{8 \times 4} = \begin{cases} 
A_{2m-1, j} = \eta_0 \mathcal{U}_{m, j} \mathbb{1}_{\{D_{2m-1, j} \neq 0\}}, \\
A_{2m, j} = \eta_0 f(U_{m, j}) \mathbb{1}_{\{D_{2m, j} \neq 0\}}, 
\end{cases}
\]

where \( m = 1, 2, 3, 4 \) and \( f() \) is the effect of backhaul delay, is an arbitrary function. \( \eta_0 = 1.443 \) is the conversion factor between [nats] and [bits], and \( \omega \) is bandwidth [Hz] shared among different tiers. Consider each serving node (MBS, SBS, or D2D) as a server which processes the arrivals of requests from the different class of users, and each user receives the service of downlink transmission at a rate configured by the corresponding cases with BH-needed or BH-free. Then, according to the class of users, a serving node allocates the downlink transmission capacity to communicating users. This system can be viewed as the DPS queue.

C. DPS Queue and QoS Metric

Let \( \mathbb{D} := \{1, 2, \ldots, 8\} \) be the set of class, and let \( \{X_j(t) : t \geq 0\} \) be the process of the number of users who request contents to a tier for \( j \in \{1, 2, 3, 4\} \) with a vector \( x_j = (x_{1, j}, x_{2, j}, \ldots, x_{8, j}) \) which is counting the number of requests in each class. We introduce weights \( w_{1, j}, w_{2, j}, \ldots, w_{8, j} \) to the class of users to differentiate the priority of processing their requests. Then, we claim that \( \{X_j(t) : t \geq 0\} \) has discrete state space \( \mathbb{N}^\mathbb{D} \) with a continuous-time Markov process generated by

\[
q(x_j, x_j + \epsilon_i) = \zeta_{i, j}, \quad x_j \in \mathbb{N}^\mathbb{D},
q(x_j, x_j - \epsilon_i) = \frac{\rho i,j}{\sum_i \rho i,j} w_{i, j}, \quad x_j \in \mathbb{N}^\mathbb{D}, x_j > 0,
\]

where \( \epsilon_i \) is the vector of \( \mathbb{N}^\mathbb{D} \) with 1 in \( i \)-th element \( (i = 1, 2, \ldots, 8) \) and 0 otherwise. \( \rho_{D, j} = \sum_{i \in \mathbb{D}} x_{i, j} \) is the total number of requests in the queue, and \( \rho_{W, j} = \sum_{i \in \mathbb{D}} w_{i, j} \) is the total weights of all classes. The class traffic demand can be obtained as \( \rho_{i,j} = \frac{\zeta_{i,j} S}{\rho} \), and the critical traffic value which the queue will be steady state is \( \rho_{c,j} = \frac{\rho_{i,j}}{\sum_i \rho_{i,j} A_{i,j}} \) where \( \rho_{D, j} = \sum_{i \in \mathbb{D}} \rho_{i,j} \) (see [6], [8]). We define a tier traffic intensity of DPS queue as \( \rho_{i,j} = \frac{\zeta_{i,j} S}{S/(\rho A_{i,j})} \Delta \frac{\lambda_{i,j}}{\rho_{c,j}} \), where \( S/(\rho A_{i,j}) \) is the rate of completion of transmission of requests [requests/s]. Recall that \( S \) is the size of each content. We need to obtain mean sojourn time of DPS queue described thus far; however, the total class of users is large, and this makes deriving the exact mean sojourn time intractable. To this end, we use the best approximated mean sojourn time of DPS queue (see [14]). We assume that the traffic load of the tier is less than the critical traffic value, which is less than 1, and all weights are strictly positive, which implies the queue is stable. We also assume that arrivals of requests do not change over time and finite. This yields the condition that the queue is stable and ergodic; Little’s law is applicable.
Proposition 2. For $\sum_{i \in \mathcal{D}} \rho_{i,j}' < \rho_{c,j} < 1$, all $w_{i,j} > 0$, and $\mu_{(i,j)} < \infty$, approximated mean number of requests, delay, and throughput per user request of the $i$-th class at $j$-th tier for $i \in \mathcal{D}$ and $j \in \{1, 2, 3, 4\}$ can be obtained as,

\[
\tilde{N}^\text{INT}_{i,j} = \lambda_{(i,j)} \tilde{S}^\text{INT}_{(i,j)}(\lambda_{(i,j)}, \mu_{(i,j)}, w_{1,j}, w_{2,j}, \ldots, w_{8,j}), \quad \text{(28)}
\]

\[
\tilde{D}^\text{INT}_{i,j} = \tilde{S}^\text{INT}_{(i,j)}(\lambda_{(i,j)}, \mu_{(i,j)}, w_{1,j}, w_{2,j}, \ldots, w_{8,j}), \quad \text{(29)}
\]

\[
\tilde{T}^\text{INT}_{i,j} = \frac{\rho_{i,j}}{\tilde{N}^\text{INT}_{i,j}}, \quad \text{(30)}
\]

where the approximated sojourn time of DPS queue is

\[
\tilde{S}^\text{INT}_{(i,j)}(\lambda_{(i,j)}, \mu_{(i,j)}, w_{1,j}, w_{2,j}, \ldots, w_{8,j}) = \frac{1}{\mu_{(i,j)}} + \frac{1}{\mu_{(i,j)}} \sum_{k=1}^{8} \left( \lambda_{(k,j)} \frac{w_{k,j} - w_{i,j}}{w_{k,j} \mu_{(k,j)} - w_{i,j} \mu_{(i,j)}} \right) + \frac{(\sum_{k=1}^{8} \lambda_{(k,j)} \mu_{(i,j)})^2}{1 - \sum_{k=1}^{8} \lambda_{(k,j)} \mu_{(i,j)} w_{k,j} \mu_{(k,j)}} \sum_{k=1}^{8} \lambda_{(k,j)} \mu_{(i,j)} w_{k,j} \mu_{(k,j)} w_{k,j} \mu_{(k,j)}}, \quad \text{(31)}
\]

Proof. Since $\sum_{i \in \mathcal{D}} \rho_{i,j}' < \rho_{c,j} < 1$ and the weight of class $i$ is strictly positive and the service time distribution is finite, then this is sufficient that DPS queue is stable (see Theorem 1, [12]). We assumed that the arrival of the requests is ergodic. The approximated mean unconditional sojourn time when service time is exponentially distributed is given by using light and heavy traffic approximation order 2 (see [14]). The equation (28) follows from Little’s law, and (29) and (30) are by the definition (see Proposition 1 [6]).

This study aims to see the effectiveness and behavior of caching aware capacity allocation system, where its simulation is infeasible. The approximated mean unconditional sojourn time of DPS queue has already studied extensively by [14]. Therefore, we explore the results of numerical calculation only.

V. NUMERICAL RESULTS

In this section, we show the numerical results of the clustered deployment of SBSs and caching aware capacity allocation system. Also, the numerical results of the non-allocated system are presented for elucidating the effect of clustered deployment of SBSs comparing with the case of non-clustered deployment of SBSs which is our baseline examined by the previous work done by [6].

Fig. 3 shows the probability of a user active in different cases and associated in different tiers; $\alpha = 0.1$, $\{P_1, P_2, P_3\} = \{3, 13, 193\}$, $\beta = 4$, $\tilde{m} = 10$, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\} = \{\frac{1000}{\pi+1000\pi}, \alpha * \frac{3+\tilde{m}}{30}, \frac{2}{\pi+1000\pi}\}$, and when the case of baseline, $\lambda_2 = \frac{2}{\pi+1000\pi}$. The variance of TCP is $\sigma = 250$.

Fig. 4 compares the average ergodic rate of each case (left) and the total average ergodic rate (right) of both the baseline of [6] and clustered deployment of SBSs. The total average ergodic rate is the weighted sum by the probability of the user active in each case. From the plot on the right, we can confirm
that when the deployment of SBSs is clustered, more cache-enabled users are needed to be active to support the requests of the contents from users and the active D2D links yield more interference to degrade the total average data rate than that of the baseline.

Fig. 5 shows the mean number of requests and mean throughput in each SBS under the case of baseline or clustered deployment. When SBSs are clustered, the mean number of requests is less than the case when SBSs are scattered uniformly. This is because each SBS is proximal to the other SBSs and the requests from users who are associated to SBSs are decentralized by those SBSs in a cluster, and also, the reason is the smaller coverage area of SBSs compared to the baseline case. This result confirms the result from Fig. 3 where the probability of a user associated with SBS is smaller than that of the baseline. Next, while $\alpha$ (the ratio of cache-enabled users) increases, the amount of case 1 users belonging to class 1 and 2 decreases significantly; however, the amount of class 2 and 3 users which are class 3 to class 6 increases. The plot also shows the result of caching aware capacity allocation system. We found that assigning more weights on class 5 and 6 (users in case 3) improve the throughput and traffic load not only those users but the entire networks. This improvement is conspicuous as the number of cache-enabled users increases.

Fig. 6 compares the result of caching aware capacity allocated and non-allocated system in MBSs. First, we found that by weighting more on processing the requests from the users in class 5, the throughput can be increased, and the traffic load on MBSs can be alleviated significantly. This is similar to the result of Fig. 5. Next, we found that the traffic load of MBSs increases under the clustered deployment of SBSs and this increase of traffic load is conspicuous from class 5 users. Also, this result confirms the result from Fig. 3 where the probability of a user associated with MBSs is larger than that of baseline.

Fig. 7 shows the mean throughput of D2D transmission which has only class 1. The plot confirms that the throughput of D2D transmission is less efficient under clustered deployment of SBSs compared to the baseline case.

VI. DISCUSSION AND CONCLUSION

In this work, we have elucidated the performance of three tier cache-enabled HetNets consisting of macro base stations (MBSs), small base stations (SBSs), and device-to-device sharing links (D2Ds) under the clustered deployment of SBSs and proposed the capacity allocation system according to the caching circumstances. We found that when the deployment of SBSs is clustered the performance can be differed significantly compared to the case in which SBSs are uniformly
might be an interesting point to start. The determinantal scheduling proposed by [18] into the system scheduling of D2D transmissions with caching. Considering D2D caching in HetNets. These findings suggest that smart gain under the allocation system is significant. Our findings the number of cache-enabled users increases, the throughput interfered with active D2D links can improve not only the resources to serve the requests from a class of users that are using the approximated mean sojourn time. Allocating more and the other QoS metrics of the system are derived by capacity allocation according to the caching circumstances can efficiently. On the other hand, we also found that interference of those D2D links are essential to utilize available resources. The increase of traffic in MBSs is conspicuous reason is the smaller coverage area of SBSs compared to the baseline case. The increase of traffic from the non-cache-enabled users who are struggling from this yields more consumption of energy in user devices. We conclude that although employing caching in D2D sharing can active as D2D transmitters when SBSs are clustered. As a larger amount of cache-enabled users are needed to be interference in HetNets, and therefore, we let \( \mathcal{N}_0 \to 0 \).

A. proof of Theorem 1

Let \( x \) be the distance between the typical user and the serving node under the max-power association law denoted as \( \| \mathbf{x}_i \| \). We see that from (15)

\[
\mathcal{U}_{t,i} = \int_0^\infty \mathbb{E}_{SINR_{t}} \ln(1 + SINR_{t}(x)) \| \mathbf{x}_i \| = x \mathbb{E}_{S}(x) dx
\]

where

\[
\mathcal{M}_y(s, x) = \mathbb{E}_{h_y, \Phi'}(\exp(-s(\sum_{j=1}^{3} \sum_{y \in \Phi'}(x) \frac{P_j}{P_i} h_y \left( \frac{\| y \|}{x} \right)^{-\beta} + \frac{N_0}{P_i(x-x^\beta)}))]
\]

and

\[
\mathcal{M}_{xy}(s, x) = \mathbb{E}_{h_x, h_y, \Phi'}(\exp(-s(h_{x^*} + \sum_{j=1}^{3} \sum_{y \in \Phi'}(x) \frac{P_j}{P_i} h_y \left( \frac{\| y \|}{x} \right)^{-\beta} + \frac{N_0}{P_i(x-x^\beta)}))].
\]

The second equality uses the lemma 1 given by Hamdi in 2010 (see, [19]). Let \( \Psi_{[x]} \) be the daughter points process around \( x \in \Phi_{P_2} \). Then, \( \mathcal{M}_y(s) \) can be derived as (32), where we move \( P_i \) and \( \| x^* \| \) to the denominator of SINR. The equation in (a) is valid since we apply the Laplace transform of \( h_y \), and all three tiers are independent. For (b), all interfering nodes are far from the serving node, and we use PGFL of PPP and PCP (20), (21). (c) holds by expressing Cartesian space to polar coordinates. (d) holds by using change of variable \( u = \frac{\| y \|}{x} \) and integral expression from the table of integral (see 3.194 page 315, [22]). Then, for the case of \( \mathcal{M}_{xy}(s) \),

\[
\mathcal{M}_{xy}(s, x) = \mathbb{E}_{h_x, h_y, \Phi'}(\exp(-s(h_{x^*} + \sum_{j=1}^{3} \sum_{y \in \Phi'}(x) \frac{P_j}{P_i} h_y \left( \frac{\| y \|}{x} \right)^{-\beta} + \frac{N_0}{P_i(x-x^\beta)}))]
\]

where (a) holds by independence of random variables and Laplace transform of \( h_{x^*} \), plug in (32) to (35), we get our result. Note that the background noise is dominated by the interference in HetNets, and therefore, we let \( N_0 \to 0 \).

B. proof of Theorem 2

The derivation of Theorem 2 is similar to Theorem 1, except the PGFL of interfering nodes from tier 1 which is the Poisson point process of active D2D transmitter. Let denote \( a \) be the distance to the closest possible active D2D transmitter, and scattered. This is because each SBS is proximal to the other SBSs, and the requests from users associated with SBSs are decentralized by those SBSs in a cluster, and also, the reason is the smaller coverage area of SBSs compared to the baseline case. The increase of traffic in MBSs is conspicuous from the non-cache-enabled users who are struggling from the interference from active D2D links. We also found that a larger amount of cache-enabled users are needed to be active as D2D transmitters when SBSs are clustered. As a result, the interference from active D2D links deteriorates the efficiency of the overall data rate of the networks, and this yields more consumption of energy in user devices. We conclude that although employing caching in D2D sharing can improve the throughput of HetNets, the careful management of interference of those D2D links is essential to utilize available resources efficiently. On the other hand, we also found that capacity allocation according to the caching circumstances can improve the throughput of the entire network and alleviate the traffic load. Such an allocation system can be modeled as an elegant queuing theory of DPS queue, and the throughput and the other QoS metrics of the system are derived by using the approximated mean sojourn time. Allocating more resources to serve the requests from a class of users that are interfered with active D2D links can improve not only the throughput of that class of users but also the entire network. As the number of cache-enabled users increases, the throughput gain under the allocation system is significant. Our findings might be practically important when it comes to carrying out D2D caching in HetNets. These findings suggest that smart management of interference from D2D transmissions is essential, and it poses further research direction to incorporating scheduling of D2D transmissions with caching. Considering the determinantal scheduling proposed by [18] into the system might be an interesting point to start.

![Fig. 7: The mean throughput of D2D based on PS queue with ζ = 0.2, S = 100[Mbits], ω = 70MHz, ρ = 1; parameters are the same as Fig. 4.](Image 1)
\[ M_y(s,x) = \mathbb{E}_{h_y,\psi_j}[\exp\{-s\left(\sum_{j=1}^{3} \frac{P_j}{P_i} h_y \left(\frac{\|y\|}{x}\right)^{-\beta} + \frac{N_0}{P_i x^{-\beta}}\right)\}] \]

\[ \begin{aligned}
(a) & = \prod_{j \in \{1,3\}} \mathbb{E}_{\psi_j} \left[ \prod_{y \in \Phi_i \setminus \{x_1^*\}} \prod_{y \in \Phi_2} \prod_{y \in \Phi_3} \right] \frac{1}{1 + \frac{P_j}{P_i} \left(\frac{\|y\|}{x}\right)^{-\beta}} \exp\{-s\left[\frac{N_0}{P_i x^{-\beta}}\right]\}
\end{aligned} \]

\[ \begin{aligned}
(b) & = \prod_{j \in \{1,3\}} \exp\{-\lambda_j \int_{\mathbb{R}^2 \setminus b_0(a)} \frac{1}{1 + \left(\frac{P_j}{P_i}\right) \left(\frac{\|y\|}{x}\right)^{\beta}} \exp(-\lambda_j y)\} \exp\{-\lambda_{P_2} \int_{\mathbb{R}^2} \left(1 - \exp(-\bar{m} \int_{\mathbb{R}^2 \setminus b_0((\frac{P_j}{P_i})^{\frac{1}{\beta}})} y\right)\}
\end{aligned} \]

\[ \begin{aligned}
&= \prod_{j \in \{1,3\}} \exp\{-2\pi \lambda_j \int_{\mathbb{R}^2 \setminus b_0(a)} \frac{1}{1 + \left(\frac{P_j}{P_i}\right) \left(\frac{\|y\|}{x}\right)^{\beta}} \exp(-2\pi \lambda_{P_2} \int_{\mathbb{R}^2} \left(1 - \exp(-\bar{m} \int_{\mathbb{R}^2 \setminus b_0((\frac{P_j}{P_i})^{\frac{1}{\beta}})} y\right)\}
\end{aligned} \]

\[ \begin{aligned}
&= \prod_{j \in \{1,3\}} \exp\{-2\pi \lambda_j x^2 \left(\frac{P_j}{P_i}\right) \beta \int_{\mathbb{R}^2 \setminus b_0(a)} \exp(-2\pi \lambda_{P_2} \int_{\mathbb{R}^2} \left(1 - \exp(-\bar{m} \int_{\mathbb{R}^2 \setminus b_0((\frac{P_j}{P_i})^{\frac{1}{\beta}})} y\right)\}
\end{aligned} \]

\[ \begin{aligned}
&= \prod_{j \in \{1,3\}} \exp\{-2\pi \lambda_j x^2 \left(\frac{P_j}{P_i}\right) \beta \int_{\mathbb{R}^2 \setminus b_0(a)} \exp(-2\pi \lambda_{P_2} \int_{\mathbb{R}^2} \left(1 - \exp(-\bar{m} \int_{\mathbb{R}^2 \setminus b_0((\frac{P_j}{P_i})^{\frac{1}{\beta}})} y\right)\}
\end{aligned} \]

For the equality (a) we use the change of variable \( \bar{u} = \left(\frac{y}{x}\right)^{\beta} \) and the integral expression from the table of integral (see 3.194 page 315, [22]). Replacing this with the PGFL of interference from tier 1 derived in Theorem 1, we get our result.

C. proof of Theorem 3

Let \( x \) be the distance between the typical user and the closest active D2D node and \( y \) be the distance between the typical user and the closest serving node in \( j \)-th tier for \( j \in \{2,3\} \), and let \( y' \) be the distance of interfering nodes. By following the same argument as the proof of the Theorem 1 (Appendix A.), we only need to obtain \( M_y(s,x,y) \), where

we see that

\[ \begin{aligned}
U_{3,j} &= \int_0^\infty \mathbb{E}_{SINR_i} \ln(1 + SINR_i(x)) | x, y, f_{S_1,j}(x, y) | dx dy \\
&= \int_0^\infty \int_0^\infty M_y(s,x,y) - M_{xy}(s,x,y) \ ds f_{S_1,j}(x, y) | dx dy \\
&= \int_0^\infty \int_0^\infty M_y(s,x,y) \ ds f_{S_1,j}(x, y) | dx dy.
\end{aligned} \]

Then, we derive \( M_y(s,x,y) \) as (33). The equality of (a) holds by using PGFL of PPP and PPCP [20, 21]. The distance of the closest active D2D transmitters can be any range, and the distance of interfering nodes of MBSs and SBSs are far from the serving node in tier \( j = 2,3 \). Therefore, the distance of active D2D nodes in the case 3 is in the range of \( 0 < x < \left(\frac{P_j}{P_i}\right)^{\frac{1}{\beta}} y \). The rest of the proof is similar to Theorem 1.

REFERENCES

[1] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, “Toward 6G networks: Use cases and technologies,” IEEE Communications Magazine, vol. 58, no. 3, pp. 55–61, 2020.

[2] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, “5G backhaul challenges and emerging research directions: A survey,” IEEE Access, vol. 4, pp. 1743–1766, 2016.

[3] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, “A survey on low latency towards 5G: Ran, core network and caching solutions,” IEEE Communications Surveys Tutorials, vol. 20, no. 4, pp. 3098–3130, 2018.

[4] E. Baştığ, “Distributed caching methods in small cell networks,” CentraleSupélec Paris-Saclay University, 2015.
\[
\mathcal{M}_y(s, x, y) = \mathbb{E}_{h_y, \Phi^n}[^{1\over N_0} \mathbb{E}_{h_y, \Phi^n} \mathbb{E}_{h_y, \Phi^n} [\exp(-s (\sum_{n=1}^{3} \sum_{y' \in \Phi^n \setminus \{x\}} P_n P_j h_y' (|y'| - 1) + N_0 \mathbb{E}_{h_y, \Phi^n})] 
= \prod_{n \in \{1, 2\}} E_{\Phi^n} \prod_{y' \in \Phi^n \setminus \{x\}} \frac{1}{1 + s \frac{P_n}{P_j} (|y'| - 1)} E_{\Phi^n, \Phi^n} \prod_{y \in \Phi^n \setminus \{x\}} \frac{1}{1 + s \frac{P_n}{P_j} (|y'| - 1)} \exp(-s N_0) \]
\]

\[
= \exp(-\lambda_1 \int_{\mathbb{R}^2 \setminus \mathbb{B}(x)} \frac{1}{1 + (s \frac{P_n}{P_j} - 1)(|y'| - 1)} (dy') \exp(-\lambda_2 \int_{\mathbb{R}^2 \setminus \mathbb{B}(x)} \frac{1}{1 + (s \frac{P_n}{P_j} - 1)(|y'| - 1)} (dy') \exp(-s N_0) \]
\]

\[
= \exp(-2\pi \lambda_1 y \frac{8 \frac{P_n}{P_j}}{\beta (1 - 2 \beta)} s F_1[1, 1 - 2 \beta, 1 - 2 \beta; -s \frac{P_n}{P_j}]) \exp(-2\pi \lambda_2 y \frac{8 \frac{P_n}{P_j}}{\beta (1 - 2 \beta)} s F_1[1, 1 - 2 \beta, 1 - 2 \beta; -s]) \times \exp(-2\pi \lambda_2 y \int_0^\infty 1 - \exp(-m \int_0^\infty \frac{1}{1 + (s \frac{P_n}{P_j} - 1)(|y'| - 1)} (dy') \exp(-s N_0) \]
\]

(36)