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

Takehiro Ohashi 18M38015

B.S., University of California, Santa Barbara (2018)

Submitted to the Department of Mathematical and Computing Science

in partial fulfillment of the requirements for the degree of

Master of Science

at the

TOKYO INSTITUTE OF TECHNOLOGY

under supervision of

Naoto Miyoshi
Department of Mathematical and Computing Science
Professor

January 2021

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 restriction, 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.

Acknowledgements

I would like to thank all the people who have made contributions to my research work. First and foremost, my thesis advisor Professor Naoto Miyoshi for helping me develop a solid understanding of the field, and his discussion guides me in the right direction. I could never have gone so far without his valuable and insightful comments. Senior doctoral researcher Kiichi Tokuyama, for taking part in my research and helping me overcome the difficulties through the entire work. Without his help, this research was incomplete.

Thanks are also due to all the members in Miyoshi Lab for their support and advice through the process of researching and writing this thesis.

Last but not least, I would like to thank my whole family for always supporting me financially and mentally.

Contents

1 Introduction **7**

2 Literature Review **10**

2.1 Poisson Poisson Cluster Process 10

2.1.1 Probability Generating Functional of PPCP 11

2.2 Sum-product Functional of PPP 11

2.3 Discriminatory Processor Sharing Queue 12

3 System Model **14**

3.1 Network Architecture 15

3.2 Access and Cache Protocol 16

3.3 Tier Association Probability 18

4 Average Ergodic Rate **22**

4.1 Active BSs and D2D Links 22

4.2 Signal to Interference Plus Noise Ratio 22

4.3 The Average Ergodic Rate in Case 1 23

4.4 The Average Ergodic Rate in Case 2 26

4.5 The Average Ergodic Rate in Case 3 27

5 QoS of Clustered Deployment and Caching Aware Capacity Allocation **31**

5.1 User State and Class 31

5.2 User Request Arrival and Service Rate 32

5.3 DPS Queue and QoS Metric 33

6 Numerical Results **36**

7 Discussion and Conclusion **41**

List of Tables

3.1 LIST OF NOTATIONS. 14

List of Figures

2.1 Sample of Thomas cluster process with $\sigma = 100$, $\lambda_p = \frac{10}{\pi * 1000^2}$, $\bar{m} = 10$ on the window $2000^2[m^2]$ 11

3.1 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. 17

3.2 Illustration of contents requests in different cases. 18

6.1 The probability of a user active in different cases and associated in different tiers; $\alpha = 0.1$, $\{P_1, P_2, P_3\} = \{3, 13, 193\}$, $\beta = 4$, $\bar{m} = 10$, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\} = \{\frac{1000}{\pi * 1000^2}, \lambda_0 * \alpha, \frac{3 * \bar{m}}{\pi * 1000^2}, \frac{2}{\pi * 1000^2}\}$, and when the case of baseline, $\lambda_2 = \frac{30}{\pi * 1000^2}$. The variance of TCP is $\sigma = 250$. 36

6.2 The average ergodic rate in different cases with varying α; $\gamma = 0.8$, $\{P_1, P_2, P_3\} = \{73, 373, 1773\}$, $\beta = 4$, $\bar{m} = 10$, $A = 1000^2$, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\} = \{\frac{300 * A}{\pi * 500^2}, \lambda_0 * \alpha, \frac{3 * \bar{m} * A}{\pi * 500^2}, \frac{6 * A}{\pi * 500^2}\}$, and when baseline case is $\lambda_2 = \frac{30 * A}{\pi * 500^2}$. The variance of TPP $\sigma = 0.05$. 37

6.3 The mean number of requests and the mean throughput of SBS; the lines with pale color are PS queue results, and the lines with bright color are DPS queue results. The weights for baseline and clustered deployment:$\{w_{1,2}, w_{2,2}, w_{3,2}, w_{4,2}, w_{5,2}, w_{6,2}\} = \{1, 1, 1.1, 1.1, 1.5, 1.87\}$ with $\varsigma = 0.2$, $S = 100$[Mbits], $\omega = 70$MHz, $\varrho = 1$; parameters are the same as Figure 6.2. 38

6.4 The mean number of requests and the mean throughput of MBS; the lines with pale color are PS queue results, and the lines with bright color are DPS queue results. The weights for baseline:$\{w_{1,3}, w_{3,3}, w_{5,3}\} = \{1, 1, 1.8\}$ and for clustered deployment $\{w_{1,3}, w_{3,3}, w_{5,3}\} = \{1, 1, 1.5\}$ with $\varsigma = 0.2$, $S = 100$[Mbits], $\omega = 70$MHz, $\varrho = 1$; parameters are the same as Figure 6.2. 39

6.5 The mean throughput of D2D based on PS queue with $\varsigma = 0.2$, $S = 100$[Mbits], $\omega = 70$MHz, $\varrho = 1$; parameters are the same as Figure 6.2. . 40

Chapter 1

Introduction

In the recent advance of the cellular communication networks, the mathematical performance analysis of wireless networks has been sophisticated and succeeded; especially, the theory of stochastic geometry has been used widely and gathered attention from many researchers in the field of mathematics and engineering. Using stochastic geometry, we view the networks as a collection of nodes, located in a typical area, which can be adopted as transmitters or receivers and model the spatial dynamics of signals and noise configured by these nodes. Today's cellular networks are composed of various types of transmitters, such as macro base stations (is the large tower with many antennae, provides radio coverage serving high power signal), small base stations (is the smaller one, less radio coverage and low power signal), and device-to-device sharing links (is the device which can transmit the signal to the other device), and thus, it is called wireless heterogeneous networks (HetNets). The challenge in the current cellular networks given to researchers is to uncover such complex dynamics and innovate the solutions to support the ever-growing communication traffic and capacity demand in the networks. To this end, storing popular contents and reusing them in the networks was proposed and called "cache", which is a promising way to solve the intense traffic from devices. However, to carry out caching in the networks, the analysis of the fairness of quality of service (QoS) and the dynamics of the performance under the realistic deployment of base stations is limited.

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.

Chapter 2

Literature Review

2.1 Poisson Poisson Cluster Process

Let Φ be a cluster point process on $\mathbb{R}^d$ with parental process $\Psi_0 = \sum_{i=1}^{k_0} \delta_{X_i}$ with intensity measure $\Lambda_p(dx) = \lambda_p(x)dx$ where $x \in \mathbb{R}^d$ and daughter processes $\Psi_i = \sum_{j=1}^{k_i} \delta_{Y_{i,j}}$ where $i \in \mathbb{N}$ and δ_a is Dirac measure at a point $a \in \mathbb{R}^d$. Daughter processes are independent and identically distributed and finite point processes on $\mathbb{R}^d$, also independent of Ψ_0. When each Ψ_i is also Poisson point process (PPP) with intensity measure $\Lambda_d(dx) = \lambda_d(x)dx$ satisfying $\int_{\mathbb{R}^d} \lambda_d(x)dx = \bar{m}$, then Poisson Poisson cluster process (PPCP) is defined as

$$\Phi(A) = \sum_{i=1}^{k_0} \Psi_i(A - X_i) = \sum_{i=1}^{k_0} \sum_{j=1}^{k_i} \delta_{X_i + Y_{i,j}}(A), \quad A \in \mathcal{B}(\mathbb{R}^d), \tag{2.1}$$

where $A - x = \{y - x | y \in A\}$ for $x \in \mathbb{R}^d$ and $\mathcal{B}(\mathbb{R}^d)$ is Borel σ-field on a metric space $\mathbb{R}^d$. Let us define the radially symmetric daughter processes, so that $\lambda_d(x) = \bar{m} f_d(\|x\|)$. When $f_d(s) = \exp(\frac{-s^2}{2\sigma^2})/2\pi\sigma^2, \sigma > 0$, we call the process as Thomas cluster process (TCP), and when $f_d(s) = \mathbb{1}_{[0,r_d]}(s)/(\pi r_d^2), r_d > 0$, we call the process as Matérn cluster process (MCP). See Figure 2.1, which is a realization of Thomas cluster process.

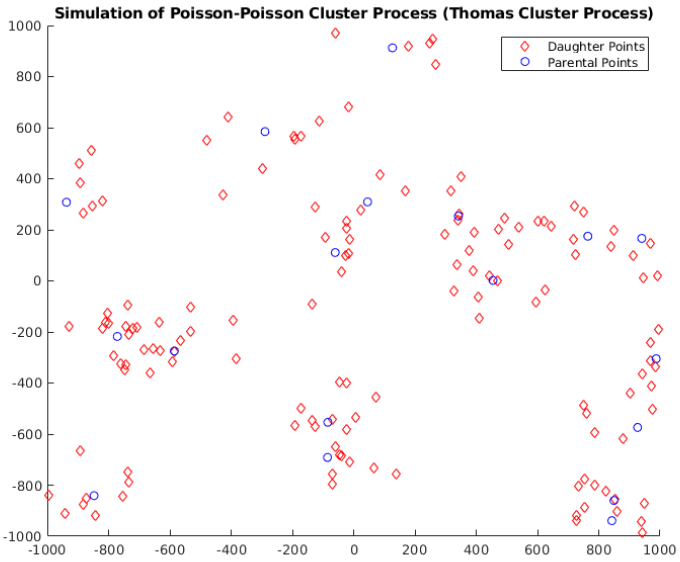

Figure 2.1: Sample of Thomas cluster process with $\sigma = 100$, $\lambda_p = \frac{10}{\pi * 1000^2}$, $\bar{m} = 10$ on the window $2000^2[m^2]$

2.1.1 Probability Generating Functional of PPCP

Let us define a measurable function such that $v : \mathbb{R}^d \mapsto [0, 1]$. Then, the probability generating functional (PGFL) of PPCP can be derived as

$$\mathbb{E}[\prod_{x_i \in \Psi_0} \prod_{y \in \Psi_i} v(y)] = \mathbb{E}[\prod_{x_i \in \Psi_0} \exp(-\int_{\mathbb{R}^d} (1 - v(x_i + y))\Lambda_d(dy))] \tag{2.2}$$

$$= \mathbb{E}[\prod_{x_i \in \Psi_0} \exp(-\bar{m} \int_{\mathbb{R}^d} (1 - v(x_i + y)) f_d(y)dy)] \tag{2.3}$$

$$= \exp(-\lambda_p \int_{\mathbb{R}^d} [1 - \exp(-\bar{m} \int_{\mathbb{R}^d} (1 - v(x_i + y)) f_d(y)dy)]dx_i), \tag{2.4}$$

where the equality (2.2) holds since each daughter process is a PPP, and (2.4) follows by PGFL of parental point process which is also PPP.

2.2 Sum-product Functional of PPP

Let Φ be PPP on $\mathbb{R}^d$ and $\mathcal{N}$ be the set of all counting measures of Φ on $\mathbb{R}^d$. Let $v(x) : \mathbb{R}^d \to \mathbb{R}^+$ and $\mu(y) : \mathbb{R}^d \to [0, 1]$ be measurable functions. For some realizations

such that $\varphi \in \mathcal{N}$, the sum product functional can be derived as

$$\mathbb{E}[\sum_{x\in\Phi} v(x) \prod_{y\in\Phi} \mu(y)] = \int_{\mathbb{R}^d} \int_{\mathcal{N}} \lambda(x)v(x) \prod_{y\in\varphi_x} \mu(y)\mathbf{P}_0(d\varphi)dx \tag{2.5}$$

$$= \int_{\mathbb{R}^d} \int_{\mathcal{N}} \lambda(x)v(x)\mu(x) \prod_{y\in\varphi} \mu(y)\mathbf{P}^!_x(d\varphi)dx \tag{2.6}$$

$$= \int_{\mathbb{R}^d} \lambda(x)v(x)\mu(x)dx \int_{\mathcal{N}} \prod_{y\in\varphi} \mu(y)\mathbf{P}^!_x(d\varphi) \tag{2.7}$$

$$= \int_{\mathbb{R}^d} \lambda(x)v(x)\mu(x)dx \int_{\mathcal{N}} \prod_{y\in\varphi} \mu(y)\mathbf{P}(d\varphi) \tag{2.8}$$

$$= \int_{\mathbb{R}^d} \lambda(x)v(x)\mu(x)dx \exp(-\int_{\mathbb{R}^d} \lambda(y)(1-\mu(y))dy), \tag{2.9}$$

where φ_x is a realization having a point at x and $\mathbf{P}_0$ is Palm measure at 0 and $\mathbf{P}^!_x$ is reduced Palm measure at x. The equality (2.5) follows by Campbell-Mecke theorem (see Section 8.2 [15]), and (2.6) holds since we remove a point x. The equality (2.8) holds by Slivnyak's theorem where it says $\mathbf{P}^!_x = \mathbf{P}$ when Φ is PPP (see Section 8.5 [15]).

2.3 Discriminatory Processor Sharing Queue

Let us assume that there are K classes of customers who arrive at a system according to an independent Poisson process with the rate $\lambda_k > 0$, and each class of customers has a service requirement denoted by B_k for $k = 1, ..., K$. All customers share a common resource and receive a portion of service from it. We define the overall arrival rate by $\lambda = \sum_{i=1}^K \lambda_i$ and total service requirement by $\frac{1}{\mu} = \sum_{i=1}^K \mathbb{E}[B_i]$. There are strictly positive wights $w_1, ..., w_K > 0$ associated to each class with number of customers $n_1, ..., n_K \in \mathbb{N}$. Then, let us define a system that allocates the available capacity of the resource to each class of customers according to the weights given, and the customers in that class share an equal portion of the allocated service. The service rate of class-k customers is then

$$\frac{w_k}{\sum_{i=1}^K n_i w_i}. \tag{2.10}$$

We denote $S_k(\lambda, b, w_1, ..., w_K)$ as the conditional sojourn time of a class-k customer with a given service requirement b and $\bar{S}_k(\lambda, b, w_1, ..., w_K) := \mathbb{E}[S_k(\lambda, b, w_1, ..., w_K)]$ as the mean of it. Then, the mean conditional sojourn times of class-k customer satisfy

the following system of integro-differential equations:

$$\frac{\partial \bar{S}_k(\lambda, b, w_1, ..., w_K)}{\partial \lambda} = 1 + \sum_{i=1}^{K} \int_0^{\infty} \lambda_i \frac{w_i}{w_k} \frac{\partial \bar{S}_i(\lambda, b, w_1, ..., w_K)}{\partial \lambda} [1 - F_i(y + \frac{w_i}{w_k} b)] dy + \int_0^b \frac{\partial \bar{S}_k(\lambda, b, w_1, ..., w_K)}{\partial \lambda} \sum_{i=1}^{K} \lambda_i \frac{w_i}{w_k} [1 - F_i(\frac{w_i}{w_k}(b - y))] dy, \quad (2.11)$$

where $F_k(b) = \mathbb{P}(B_k \leq b)$ is the distribution function of service requirement (see [11], [14]). Furthermore, the mean unconditional sojourn time with exponentially distributed service requirements of class-k customers denoted as $\bar{S}_k(\lambda, \mu, w_1, ..., w_K)$ is the unique solution of the following system of equations

$$\bar{S}_k(\lambda)(1 - \sum_{i=1}^{K} \frac{\lambda_i w_i}{\mu_i w_i + \mu_k w_k}) - \sum_{i=1}^{K} \frac{\lambda_i w_i \bar{S}_i(\lambda)}{\mu_i w_i + \mu_k w_k} = \frac{1}{\mu_k}, \quad (2.12)$$

which is given by [11], [14].

Chapter 3

System Model

We introduce our three-tier HetNets consisting of macro base stations (MBSs), small base stations (SBSs), and device-to-device sharing links (D2Ds). Note that our work follows the idea from [6] and regards their result as a baseline; here we consider SBSs in our system model that follows TCP while their analysis considers relays that follow independent PPP. The tier association probabilities in this network are also derived to characterize the request of contents from users to nearby base stations (BSs) or D2Ds. The list of notations is provided in Table 3.1 for the reader to review some of the notations used hereafter.

Table 3.1: LIST OF NOTATIONS.

Φ_i	The point process of the i-th BS/D2D tier
$f_{cd_i}, \quad \bar{F}_{cd_i}$	PDF and CCDF of contact distance of Φ_i
$f_d(s)$	The displacement kernel defined $\exp(\frac{-\|s\|^2}{2\sigma^2})/2\pi\sigma^2$
$\tau_i(r)$	$\begin{cases} \bar{m}\int_0^\infty 2\pi\lambda_{p_2}\frac{z}{\sigma}q(\frac{z}{\sigma},\frac{r}{\sigma})\times & \\ \exp(-\bar{m}(1-Q_1(\frac{z}{\sigma},\frac{\frac{P_2}{P_i}^{\frac{1}{\beta}}r}{\sigma})))dz, & \text{if } i=2 \\ 2\pi\lambda_i r, & \text{if } i=1,3 \end{cases}$
$f_{\mathcal{S}_i}, f_{\hat{\mathcal{S}}_i}, f_{\mathcal{S}_{1,j}}$	PDF of distance of a serving node given an event $\mathcal{S}_i, \hat{\mathcal{S}}_i, \mathcal{S}_{1,j}$
$\bar{F}_{\mathcal{S}_i}, \bar{F}_{\hat{\mathcal{S}}_i}, \bar{F}_{\mathcal{S}_{1,j}}$	CCDF of distance of a serving node given an event $\mathcal{S}_i, \hat{\mathcal{S}}_i, \mathcal{S}_{1,j}$
${}_2F_1[a,b;c;d]$	Gauss hyper-geometric function
α	Ratio of cache-enabled users ($\alpha \in [0,1]$)
β	Path loss exponent ($\beta = 4$)

3.1 Network Architecture

We consider three-tier HetNet consisting of MBSs, SBSs, and D2Ds. All SBSs are connected to the nearest MBS through backhauls, and then MBSs are connected to a core network (see Figure 3.1). We define that the nodes of users and MBSs follow independent homogenious PPPs in $\mathbb{R}^2$ denoted as Φ_i with intensity λ_i for $i = 0, 3$. The nodes of SBSs follow PPCP denoted as Φ_2 in $\mathbb{R}^2$ with conditional intensity function given the parental point process $\lambda_2(\mathbf{x}) = \bar{m} \sum_{\mathbf{z} \in \Phi_{p_2}} f_d(\mathbf{x}|\mathbf{z})$ where Φ_{p_2} is the parental point process following PPP with intensity λ_{p_2} and $\bar{m}$ is the mean number of daughter points distributed according to some distribution of $f_d(\mathbf{x}|\mathbf{z})$ around the center of each parental point $\mathbf{z}$. In this paper for the ease of our analysis, we assume that the daughter points are independently and normally distributed around a parental point $\mathbf{z}$, 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 PPCP is

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

$$\Phi_2 = \cup_{X_i \in \Phi_{p_2}} (X_i + \Psi_i), \tag{3.1}$$

where $\Phi_{p_2} = \{X_i\}_{i=1}^{\infty}$ *is the parental point process, homogeneous PPP with intensity* λ_{p_2} *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* $\lambda_d(\mathbf{y})$ *s.t.* $\int_{\mathbb{R}^2} \lambda_d(\mathbf{y}) d\mathbf{y} = \int_{\mathbb{R}^2} \bar{m} f_d(\mathbf{y}) d\mathbf{y} = \bar{m}$ *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 [16], [17]. Let $\mathcal{B}(\mathbb{R}^2)$ be the Borel σ-algebra on $\mathbb{R}^2$ and $A \subset \mathcal{B}(\mathbb{R}^2)$. Conditioned on Φ_{p_2}, the counting measure of Φ_2 is random measure which counts the number of points of Φ_2 falling in A and is given by (see [16], [17])

$$N_{\Phi_2}(A)|\Phi_{p_2} \sim Poisson(\bar{m} \sum_{\mathbf{x} \in \Phi_{p_2}} \int_A f_d(\mathbf{y} - \mathbf{x}) d\mathbf{y}). \tag{3.2}$$

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 ($\bar{m}\lambda_{p_2} = \lambda_2$).

For the user tier, we have the proportion of cache-enabled users who can transmit some contents cached 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 Φ_1 with intensity $\lambda_1 = \alpha\lambda_0$, where $\alpha \in [0, 1]$ and λ_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 is an 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 $M_1 \times S$ [bits]. It is a natural to consider $M_1 \ll M_2 \ll N$. See Figure 3.1, the blue disks are represented as caching of M_1 number of contents which all the cache enabled users to have and contain the same copies of the contents. SBSs have M_2 number of contents which includes M_1 number of contents as well. MBSs have all of the contents stored in their storage. The i-th popular content requested by the typical user follows Zipf distribution, and for arbitrary content popularity ranks a, b with $a < b$:

$$f_i = \frac{1/i^{\gamma}}{\sum_{j=1}^{N} \frac{1}{j^{\gamma}}}, \qquad \sum_{i=a}^{b} f_i \triangleq F_{pop}(a, b), \tag{3.3}$$

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 $F_{pop}(1, M_2)$.

3.2 Access and Cache Protocol

For the completeness of this paper, we summarize all the assumption from [6] and emphasis some of the important concepts here. We assume all users are following the max-power association rule. This means that all users request contents from a node which provides the highest power that is the closet node in a tier. For $i = 1, 2, 3$, the

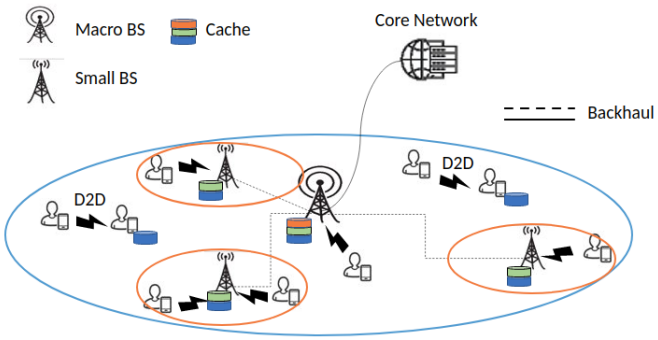

Figure 3.1: Illustration of three tier cache enabled heterogeneous networks. The blue circle is the coverage area of MBS. The orange circle is the coverage area of SBS.

maximum received power from the i-th tier is

$$C_i = \nu B_i P_i \|\mathbf{x}_i^*\|^{-\beta}, \tag{3.4}$$

where $\|\mathbf{x}_i^*\|$ is the distance from the typical user at $(0,0)$ to its closest BS in i-th tier, $\mathbf{x}_i^* \in \Phi_i$; that is $\mathbf{x}_i^* = \arg\max_{\mathbf{x}\in\Phi_i} P_i\|\mathbf{x}\|^{-\beta} = \operatorname{argmin}_{\mathbf{x}\in\Phi_i} \|\mathbf{x}\|$. P_i is a transmit power of the node in i-th tier, and ν, B_i, and β 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.

- 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.

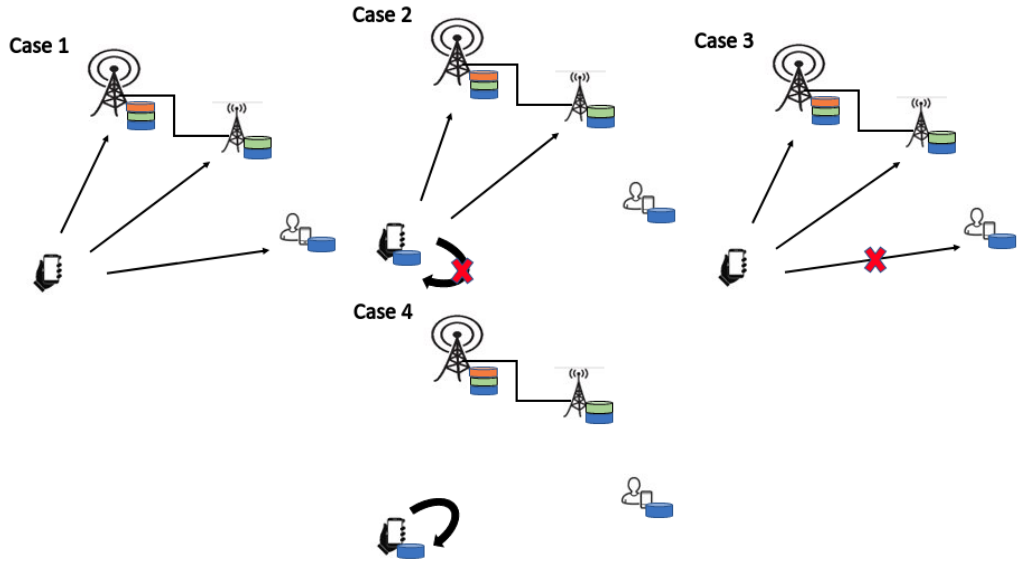

Figure 3.2: Illustration of contents requests in different cases.

Figure 3.2 is the illustration of all four cases. We assume that cache is placed according to proactive-caching: all contents are distributed to the storage of BSs/D2Ds during off-peak and ready to use. Additionally, when the typical user has requested contents to an SBS and the contents are not stored in the storage of the SBS, it requires wired backhauls to fetch requested contents from the storage of the nearest MBS. This is the case when the downlink data transmission is *backhaul-needed* (BH-needed), and we assume the typical user under BH-needed transmission experiences some delay.

3.3 Tier Association Probability

We derive the probability of ordering of maximum power received by the typical user from each tier Φ_i for $i = 1, 2, 3$, where Φ_1 and Φ_3 follow independent PPP and Φ_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. 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 Φ_i for $i \in \{1, 3\}$ denoted as f_{cd_i} and the complementary CDF denoted as $\bar{F}_{cd_i}$ are

$$f_{cd_i}(r_i) = 2\pi\lambda_i r_i e^{-\pi\lambda_i r_i^2}, \quad \bar{F}_{cd_i}(r_i) = e^{-\pi\lambda_i r_i^2}. \tag{3.5}$$

The conditional PDF of contact distance of Φ_2 given the parental point process Φ_{p_2} is derived (see, Lemma 1 [16], [17]) and given by

$$f_{cd_2}(r_2|\Phi_{p_2}) = \bar{m}\sum_{i=1}^{\infty}\frac{1}{\sigma}q(\frac{\|X_i\|}{\sigma},\frac{r_2}{\sigma})\prod_{i=1}^{\infty}\exp(-\bar{m}(1-Q_1(\frac{\|X_i\|}{\sigma},\frac{r_2}{\sigma}))), \tag{3.6}$$

where Q_1 is the first-order Marcum Q-function $Q_1(a,b) = \int_b^{\infty} z\exp(-\frac{z^2+a^2}{2})I_0(az)dz$ and $I_0(z) = \pi^{-1}\int_0^{\pi} e^{z\cos\phi}d\phi$ 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 [18], and the complementary CDF is given by

$$\bar{F}_{cd_2}(r_2) = \exp(-\int_0^{\infty} 2\pi\lambda_{p_2}z(1-\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma},\frac{r_2}{\sigma}))))dz). \tag{3.7}$$

Then, by using (3.6) we can derive the PDF of contanct distance distribution of TCP in the following Lemma 1.

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

$$\begin{aligned} f_{cd_2}(r_2) &= \mathbb{E}_{\Phi_{p_2}}[f_{cd_2}(r_2|\Phi_{p_2})] && (3.8)\\ &= \mathbb{E}_{\Phi_{p_2}}[\bar{m}\sum_{i=1}^{\infty}\frac{1}{\sigma}q(\frac{\|X_i\|}{\sigma},\frac{r_2}{\sigma})\prod_{i=1}^{\infty}\exp(-\bar{m}(1-Q_1(\frac{\|X_i\|}{\sigma},\frac{r_2}{\sigma})))] && (3.9)\\ &= \int_0^{\infty}\bar{m}\lambda_{p_2}2\pi\frac{z}{\sigma}q(\frac{z}{\sigma},\frac{r_2}{\sigma})\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma},\frac{r_2}{\sigma})))dz\times \\ &\exp(-\int_0^{\infty}2\pi\lambda_{p_2}z(1-\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma},\frac{r_2}{\sigma}))))dz). && (3.10)\end{aligned}$$

Proof. We can directly apply the sum-product functional with respect to parental PP which follows PPP (see Lemma 6 [17]). □

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*

$$\mathbb{P}(C_i > C_j > C_k) = \int_0^\infty \int_{(\frac{P_j}{P_i})^{\frac{1}{\beta}} r_i}^\infty \int_{(\frac{P_k}{P_j})^{\frac{1}{\beta}} r_j}^\infty \int_0^\infty \bar{m}\lambda_{p_2} 2\pi \frac{z}{\sigma} q(\frac{z}{\sigma}, \frac{r_2}{\sigma}) \times$$
$$\exp(-\bar{m}(1 - Q_1(\frac{z}{\sigma}, \frac{r_2}{\sigma})))dz \times$$
$$\exp(-\int_0^\infty 2\pi\lambda_{p_2} z(1 - \exp(-\bar{m}(1 - Q_1(\frac{z}{\sigma}, \frac{r_2}{\sigma}))))dz) \times$$
$$\prod_{l\in\{1,3\}} 2\pi\lambda_l r_l \exp(-\pi\lambda_l r_l^2) dr_k dr_j dr_i. \quad (3.11)$$

Proof. Let R_i, R_j and R_k be the random variable of contact distance to tier i, j and k, respectively. Then, we have the joint PDF of contact distance scaled by power from each tier as

$$\mathbb{P}(C_i > C_j > C_k) = \mathbb{P}((\frac{P_i}{P_k})^{-\frac{1}{\beta}} R_k > (\frac{P_i}{P_j})^{-\frac{1}{\beta}} R_j > R_i) \quad (3.12)$$
$$= \int_0^\infty \int_{(\frac{P_j}{P_i})^{\frac{1}{\beta}} r_i}^\infty \int_{(\frac{P_k}{P_j})^{\frac{1}{\beta}} r_j}^\infty \int_0^\infty f_{cd_k}(r_k) f_{cd_j}(r_j) f_{cd_i}(r_i) dr_k dr_j dr_i. \quad (3.13)$$

Since the PDF of contact distance distribution of each tier is given by (3.5) and (3.10), plugging in them into (3.13) gives the desired equation. □

The association probability conditional on parental PP is derived in Lemma 3 [17]. They assume an arbitrary many number of tiers in HetNets. Similarly, We can derive the unconditional tier association probability by the following.

Corollary 1. *The unconditional i-th tier association probability is*

$$\mathbb{P}(C_i > \max_{\forall n\neq i} C_n) = \int_0^\infty \tau_i(r) \exp(-\int_0^\infty 2\pi\lambda_{p_2} z$$
$$(1 - \exp(-\bar{m}(1 - Q_1(\frac{z}{\sigma}, \frac{(\frac{P_2}{P_i})^{\frac{1}{\beta}} r}{\sigma})))) dz) \exp(-\sum_{l\in\{1,3\}} \pi\lambda_l (\frac{P_l}{P_i})^{\frac{2}{\beta}} r^2) dr, \quad (3.14)$$

where $\tau_i(r)$ *is defined in Table 3.1, and* $n \in \{1, 2, 3\}$.

Proof. In the case $i = 2$,

$$\mathbb{P}(C_i > \max_{\forall n \neq i} C_n) = \mathbb{P}(\bigcap_{j \in \{1,3\}} P_2 R_2^{-\beta} > P_j R_j^{-\beta}) \tag{3.15}$$

$$= \mathbb{P}(\bigcap_{j \in \{1,3\}} R_j > (\frac{P_j}{P_2})^{\frac{1}{\beta}} R_2) \tag{3.16}$$

$$= \int_0^\infty \prod_{j \in \{1,3\}} \bar{F}_{cd_j}((\frac{P_j}{P_2})^{\frac{1}{\beta}} r) f_{cd_2}(r) dr. \tag{3.17}$$

In the case $i = 1, 3$, let $j = \{1, 3\} \setminus \{i\}$, then we have

$$\mathbb{P}(C_i > \max_{\forall n \neq i} C_n) = \mathbb{P}(R_2 > (\frac{P_2}{P_i})^{\frac{1}{\beta}} R_i) \mathbb{P}(R_j > (\frac{P_j}{P_i})^{\frac{1}{\beta}} R_i) \tag{3.18}$$

$$= \int_0^\infty \prod_{j \in \{1,3\}} \bar{F}_{cd_2}((\frac{P_2}{P_i})^{\frac{1}{\beta}} r) \bar{F}_{cd_j}((\frac{P_j}{P_i})^{\frac{1}{\beta}} r) f_{cd_i}(r) dr. \tag{3.19}$$

Plugging (3.5), (3.7) and (3.10) into (3.17) and (3.18) as well as rewriting the plugged equation by $\tau_i(r)$ gives our result in (11). □

We define the notations and use them for the convenience hereafter. $\mathbb{P}(C_i > \max_{\forall n \neq i} C_n) \triangleq \mathcal{G}_{3,i}$, and also, we denote $\mathbb{P}(C_i > C_j > C_k)$ and $\mathbb{P}(C_i > C_j)$ as $\mathbb{P}_{i,j,k}$ and $\mathbb{P}_{i,j}$, respectively. Note that $\mathbb{P}(C_i > C_j)$ is two tier case of Proposition 1, and the derivation is the same and therefore omitted.

Chapter 4

Average Ergodic Rate

In this chapter, the average ergodic rate of the downlink is analyzed under the three-tier HetNets described in the previous chapter. 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 Figure 3.2).

4.1 Active BSs and D2D Links

We define all active nodes which yield the interference. We denote λ_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 D2D nodes as $\lambda_1' \stackrel{\Delta}{=} \min\{\alpha\lambda_0, \lambda_0\mathcal{G}_{3,1}(1-\alpha)F_{pop}(1, M_1)\}$. For the other tiers, since all MBSs and SBSs are active over the time, we define $\lambda_i' \stackrel{\Delta}{=} \lambda_i$. We also define Φ_1' as the thinned PPP of active D2D links and $\Phi_2' \stackrel{\Delta}{=} \Phi_2$ and $\Phi_3' \stackrel{\Delta}{=} \Phi_3$, where active MBSs and SBSs are unchanged from original MBSs and SBSs.

4.2 Signal to Interference Plus Noise Ratio

The downlink power received from the serving BS/D2D at $\mathbf{x}_i \in \Phi_i'$ is $P_i h_{\mathbf{x}_i} \|\mathbf{x}_i\|^{-\beta}$ 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

$$SINR_i(\mathbf{x}) = \frac{P_i h_{\mathbf{x_i}} \|\mathbf{x}_i\|^{-\beta}}{\sum_{j=1}^{3} \sum_{\mathbf{y}\in\Phi'_j\setminus\{\mathbf{x}_i\}} P_j h_{\mathbf{y}} \|\mathbf{y}\|^{-\beta} + N_0}, \tag{4.1}$$

where $h_{\mathbf{x_i}}$ and $h_{\mathbf{y}}$ follow Rayleigh fading distributed as exponential with unit mean, and N_0 is thermal noise. In the denominator, $\|\mathbf{y}\|$ is the distance between reference user to its interfering active nodes in j-th tier. Let $\mathbf{x}$ 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]:

$$\mathcal{U}_i \triangleq \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{SINR_i}[\ln(1 + SINR_i(\mathbf{x}))]], \tag{4.2}$$

where the expectation of $SINR_i$ is with respect to all Φ'_j for $j = 1, 2, 3$ and the fading effect $h_{\mathbf{x_i}}$ and $h_{\mathbf{y}}$, and then the expectation of $\mathbf{x}$ 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.

4.3 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 $\mathcal{S}_i = \{P_i\|\mathbf{x}_i^*\|^{-\beta} > \max_{n\neq i} P_n\|\mathbf{x}_n^*\|^{-\beta} : n \in \{1, 2, 3\}\}$. The PDF of the distance distribution of the serving node conditional on parental PP and $\mathcal{S}_i$ is given in Lemma 4, [17]. Then, we can obtain the PDF of the distance between the typical user and the associated serving node unconditional on Φ_{p_2} and conditional on an event $\mathcal{S}_i$.

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*

$$f_{\mathcal{S}_i}(x) = \frac{\tau_i(x)}{\mathcal{G}_{3,i}} \exp(-\int_0^\infty 2\pi\lambda_{p_2} z(1 - \exp(-\bar{m}(1 - Q_1(\frac{z}{\sigma}, \frac{(\frac{P_2}{P_i})^{\frac{1}{\beta}} x}{\sigma})))) dz) \times \exp(-\pi \sum_{j\in\{1,3\}} \lambda_j (\frac{P_j}{P_i})^{\frac{2}{\beta}} x^2), \tag{4.3}$$

where $\tau_i(x)$ is given in Table 3.1.

Proof. The CCDF of $f_{\mathcal{S}_i}(x)$ can be expressed as

$$\bar{F}_{\mathcal{S}_i}(r|\mathcal{S}_i) = \mathbb{P}(\|\mathbf{x}_i^*\| > r|\mathcal{S}_i) \tag{4.4}$$

$$= \frac{\mathbb{P}(R_i > r, \mathcal{S}_i)}{\mathbb{P}(\mathcal{S}_i)} \tag{4.5}$$

$$= \frac{1}{\mathcal{G}_{3,i}} \mathbb{P}\Big(\bigcap_{j\in\{1,2,3\}\setminus\{i\}} P_i R_i^{-\beta} > P_j R_j^{-\beta}, R_i > r \Big). \tag{4.6}$$

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. □

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*

$$\mathcal{U}_{1,i} = \int_0^\infty \int_0^\infty \frac{\mathcal{M}_y(s,x)}{1+s} ds \frac{\tau_i(x)}{\mathcal{G}_{3,i}} \times \exp\Big(-\int_0^\infty 2\pi\lambda_{p_2} z \big(1 - \exp\big(-\bar{m}\big(1 - Q_1\big(\frac{z}{\sigma}, \frac{(\frac{P_2}{P_i})^{\frac{1}{\beta}} x}{\sigma}\big)\big)\big)\big) dz\Big) \times \exp\Big(-\pi \sum_{j\in\{1,3\}} \lambda_j (\frac{P_j}{P_i})^{\frac{2}{\beta}} x^2\Big) dx, \tag{4.7}$$

where

$$\mathcal{M}_y(s,x) = \prod_{j\in\{1,3\}} \exp\Big(-2\pi\lambda_j' x^2 \frac{s(\frac{P_j}{P_i})^{\frac{2}{\beta}}}{\beta(1-\frac{2}{\beta})} {}_2F_1[1, 1-\frac{2}{\beta}; 2-\frac{2}{\beta}; -s]\Big) \exp\Big(-2\pi\lambda_{p_2} \int_0^\infty \Big(1 - \exp\Big(-\bar{m} \int_{(\frac{P_2}{P_i})^{\frac{1}{\beta}} x}^\infty \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{y^2}{2\sigma^2})}{1 + (s\frac{P_2}{P_i})^{-1} (\frac{y+z}{x})^\beta} dy\Big)\Big) z dz\Big).$$

and ${}_2F_1[a,b;c;d]$ is Gauss hyper-geometric function.

Proof. Let x be the distance between the typical user and the serving node under the

max-power association law, where it is denoted as $\|\mathbf{x}_i^*\|$. We see that from (4.2)

$$\begin{aligned}\mathcal{U}_{1,i} &= \int_0^\infty \mathbb{E}_{SINR_i}[\ln(1+SINR_i(x))|\|\mathbf{x}_i^*\|=x]f_{\mathcal{S}_i}(x)dx \\ &= \int_0^\infty \int_0^\infty \frac{\mathcal{M}_y(s,x)-\mathcal{M}_{xy}(s,x)}{s}ds f_{\mathcal{S}_i}(x)dx, \end{aligned} \tag{4.8}$$

where

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

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

The second equality uses Lemma 1 given by Hamdi in 2010 (see, [19]). Let $\Psi_{[\mathbf{z}]}$ be the daughter points process around $\mathbf{z}\in\Phi_{p_2}$. Then, $\mathcal{M}_y(s)$ can be derived as

$$\begin{aligned}
&\mathcal{M}_y(s,x) = \mathbb{E}_{h_{\mathbf{y}},\Phi_j'}[\exp(-s(\sum_{j=1}^{3}\sum_{\mathbf{y}\in\Phi_j'\setminus\{\mathbf{x}_i^*\}}\frac{P_j}{P_i}h_{\mathbf{y}}\left(\frac{\|\mathbf{y}\|}{x}\right)^{-\beta}+\frac{N_0}{P_i x^{-\beta}}))] \\
&\overset{(a)}{=} \prod_{j\in\{1,3\}}\mathbb{E}_{\Phi_j'}[\prod_{\mathbf{y}\in\Phi_j'\setminus\{\mathbf{x}_i^*\}}\frac{1}{1+s\frac{P_j}{P_i}(\frac{\|\mathbf{y}\|}{x})^{-\beta}}]\mathbb{E}_{\Phi_{p_2},\Psi_i}[\prod_{\mathbf{z}\in\Phi_{p_2}}\prod_{\mathbf{y}\in\Psi_{[\mathbf{z}]}\setminus\{\mathbf{x}_i^*\}}\frac{1}{1+s\frac{P_2}{P_i}(\frac{\|\mathbf{y}\|}{x})^{-\beta}}]\exp(-s\frac{N_0}{P_i x^{-\beta}}) \\
&\overset{(b)}{=} \prod_{j\in\{1,3\}}\exp(-\lambda_j'\int_{\mathbb{R}^2\setminus b_0((\frac{P_j}{P_i})^{\frac{1}{\beta}}x)}\frac{1}{1+(s\frac{P_j}{P_i})^{-1}(\frac{\|\mathbf{y}\|}{x})^{\beta}}d\mathbf{y})\exp(-\lambda_{p_2}\int_{\mathbb{R}^2}(1-\exp(-\bar{m}\int_{\mathbb{R}^2\setminus b_0((\frac{P_2}{P_i})^{\frac{1}{\beta}}x)} \\
&\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y^2}{2\sigma^2})}{1+(s\frac{P_2}{P_i})^{-1}(\frac{\|\mathbf{y}+\mathbf{z}\|}{x})^{\beta}}d\mathbf{y}))d\mathbf{z})\exp(-s\frac{N_0}{P_i x^{-\beta}}) \\
&\overset{(c)}{=} \prod_{j\in\{1,3\}}\exp(-2\pi\lambda_j'\int_{(\frac{P_j}{P_i})^{\frac{1}{\beta}}x}^{\infty}\frac{1}{1+(s\frac{P_j}{P_i})^{-1}(\frac{y}{x})^{\beta}}ydy)\exp(-2\pi\lambda_{p_2}\int_0^\infty(1-\exp(-\bar{m}\int_{(\frac{P_2}{P_i})^{\frac{1}{\beta}}x}^{\infty} \\
&\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y^2}{2\sigma^2})}{1+(s\frac{P_2}{P_i})^{-1}(\frac{y+z}{x})^{\beta}}dy))zdz)\exp(-s\frac{N_0}{P_i x^{-\beta}}) \\
&\overset{(d)}{=} \prod_{j\in\{1,3\}}\exp(-2\pi\lambda_j' x^2\frac{s(\frac{P_j}{P_i})^{\frac{2}{\beta}}}{\beta(1-\frac{2}{\beta})}{}_2F_1[1,1-\frac{2}{\beta};2-\frac{2}{\beta};-s])\exp(-2\pi\lambda_{p_2}\int_0^\infty\exp(-\bar{m}\int_{(\frac{P_2}{P_i})^{\frac{1}{\beta}}x}^{\infty} \\
&\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y^2}{2\sigma^2})}{1+(s\frac{P_2}{P_i})^{-1}(\frac{y+z}{x})^{\beta}}dy))zdz)\exp(-s\frac{N_0}{P_i x^{-\beta}}),
\end{aligned} \tag{4.9}$$

where we move P_i and $\|\mathbf{x}_i^*\|$ to the denominator of SINR. The equation in (a) is valid since we apply the Laplace transform of $h_{\mathbf{y}}$, and all three tiers are independent. For (b), all interfering nodes are far from the serving node, and we use PGFL of PPP and PPCP [15], [20]. (c) holds by expressing Cartesian space to polar coordinates. (d) holds

by using change of variable $u = (\frac{y}{x})^{\beta}$ and integral expression from the table of integral (see 3.194 page 315, [21]). Then, for the case of $\mathcal{M}_{xy}(s)$,

$$\begin{aligned}\mathcal{M}_{xy}(s,x) &= \mathbb{E}_{h_{\mathbf{x}_i^*},h_{\mathbf{y}},\Phi_j'}[\exp(-s(h_{\mathbf{x}_i^*} + \sum_{j=1}^{3}\sum_{\mathbf{y}\in\Phi_j\setminus\{\mathbf{x}_i^*\}}\frac{P_j}{P_i}h_{\mathbf{y}}\left(\frac{\|\mathbf{y}\|}{x}\right)^{-\beta} + \frac{N_0}{P_i x^{-\beta}}))] \\ &\overset{(a)}{=} \mathcal{L}_{h_{\mathbf{x}_i^*}}(s)\mathcal{M}_y(s,x) \\ &= \frac{1}{1+s}\mathcal{M}_y(s,x), \end{aligned} \tag{4.10}$$

where (a) holds by independence of random variables and Laplace transform of $h_{\mathbf{x}_i}$. Then, we plug in (4.9) and (4.10) into (4.8) to get our result. Note that the background noise is dominated by the interference in HetNets, and therefore, we let $N_0 \to 0$. □

4.4 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 2, given an event $\hat{\mathcal{S}}_i = \{P_i\|\mathbf{x}_i^*\|^{-\beta} > \max_{j\neq i} P_j\|\mathbf{x}_j^*\|^{-\beta} : j \in \{2,3\}\}$ the unconditional PDF of the distance between the typical user and the serving node in the i-th is

$$\begin{aligned} f_{\hat{\mathcal{S}}_i}(x) = \frac{\tau_i(x)}{\mathbb{P}_{i,j}}\exp(-\int_0^{\infty} 2\pi\lambda_{p_2} z(1-\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma},\frac{(\frac{P_2}{P_i})^{\frac{1}{\beta}}x}{\sigma}))))dz)\times \\ \exp(-\pi\lambda_3(\frac{P_3}{P_i})^{\frac{2}{\beta}}x^2), \end{aligned} \tag{4.11}$$

where $i,j \in \{2,3\}$ s.t. $i \neq j$ and $\tau_i(x)$ is defined in Table 3.1. 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 Figure 3.2). Therefore, the possible distance between a requested user to the closest cache-enabled user denoted as a can be as close as 0; that is, the possible distance $a \to 0$. 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

$$\mathcal{U}_{2,i} = \int_0^\infty \int_0^\infty \frac{\mathcal{M}_y(s,x)}{1+s} ds \frac{\tau_i(x)}{\mathbb{P}_{i,j}} \times \exp(-\int_0^\infty 2\pi\lambda_{p_2} z(1-\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma}, \frac{(\frac{P_2}{P_i})^{\frac{1}{\beta}} x}{\sigma}))))dz) \times \exp(-\pi\lambda_3 (\frac{P_3}{P_i})^{\frac{2}{\beta}} x^2) dx, \quad (4.12)$$

where

$$\mathcal{M}_y(s,x) = \exp(-2\pi\lambda_1' x^2 \frac{s\frac{P_1}{P_i}(\frac{a}{x})^{2-\beta}}{\beta(1-\frac{2}{\beta})} {}_2F_1[1, 1-\frac{2}{\beta}; 2-\frac{2}{\beta}; -s\frac{\frac{P_1}{P_i}}{(\frac{a}{x})^\beta}]) \times \exp(-2\pi\lambda_3' x^2 \frac{s(\frac{P_3}{P_i})^{\frac{2}{\beta}}}{\beta(1-\frac{2}{\beta})} {}_2F_1[1, 1-\frac{2}{\beta}; 2-\frac{2}{\beta}; -s]) \times \exp(-2\pi\lambda_{p_2} \int_0^\infty (1-\exp(-\bar{m} \int_{(\frac{P_2}{P_i})^{\frac{1}{\beta}} x}^\infty \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{y^2}{2\sigma^2})}{1+(s\frac{P_2}{P_i})^{-1}(\frac{y+z}{x})^\beta} dy)) z dz).$$

Proof. 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 a can be small as 0. We derive the PGFL of interference from tier 1 as

$$\begin{aligned}
\mathbb{E}_{\Phi_1'}[\prod_{\mathbf{y}\in\Phi_1'\setminus\{\mathbf{x}_i^*\}} \frac{1}{1+s\frac{P_j}{P_i}(\frac{\|\mathbf{y}\|}{x})^{-\beta}}] &= \exp(-\lambda_1' \int_{\mathbb{R}^2\setminus b_0(a)} \frac{1}{1+(s\frac{P_1}{P_i})^{-1}(\frac{\|\mathbf{y}\|}{x})^\beta} d\mathbf{y}) \\
&= \exp(-2\pi\lambda_1' \int_a^\infty \frac{1}{1+(s\frac{P_1}{P_i})^{-1}(\frac{y}{x})^\beta} y dy) \\
&\overset{(a)}{=} \exp(-2\pi\lambda_1' x^2 \frac{s\frac{P_1}{P_i}(\frac{a}{x})^{2-\beta}}{\beta(1-\frac{2}{\beta})} {}_2F_1[1, 1-\frac{2}{\beta}; 2-\frac{2}{\beta}; -s\frac{\frac{P_1}{P_i}}{(\frac{a}{x})^\beta}]).
\end{aligned}$$

For the equality (a) we uses the change of variable $u = (\frac{y}{x})^\beta$ and the integral expression from the table of integral (see 3.194 page 315, [21]). Replacing this with the PGFL of interference from tier 1 derived in Theorem 1, we get our result. □

4.5 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 Figure 3.2). 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 $\mathcal{S}_{1,j} = \{P_1\|\mathbf{x}_1^*\|^{-\beta} > P_j\|\mathbf{x}_j^*\|^{-\beta} \cap P_j\|\mathbf{x}_j^*\|^{-\beta} > P_k\|\mathbf{x}_k^*\|^{-\beta} : j,k \in \{2,3\} \text{ s.t. } j \neq k\}$, the joint PDF of the distance between the typical user to the serving node in j-th tier is

$$f_{\mathcal{S}_{1,j}}(x,y) = \frac{\tau_1(x)\tau_j(y)}{\mathbb{P}_{1,j,k}} \exp(-\pi\lambda_1 x^2)\times \exp(-\int_0^\infty 2\pi\lambda_{p_2} z(1-\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma}, \frac{(\frac{P_2}{P_j})^{\frac{1}{\beta}} y}{\sigma}))))dz)\times \exp(-\pi\lambda_3(\frac{P_3}{P_j})^{\frac{2}{\beta}} y^2), \quad (4.13)$$

where $j,k \in \{2,3\}$ s.t. $j \neq k$ and $\tau_1(x)$ and $\tau_j(y)$ are defined in Table 3.1. 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*

$$\mathcal{U}_{3,j} = \int_0^\infty \int_0^{\frac{P_1}{P_j}^{\frac{1}{\beta}} y} \int_0^\infty \frac{\mathcal{M}_y(s,x,y)}{1+s} ds \frac{\tau_1(x)\tau_j(y)}{\mathbb{P}_{1,j,k}} \exp(-\pi\lambda_1 x^2)\times \exp(-\int_0^\infty 2\pi\lambda_{p_2} z(1-\exp(-\bar{m}(1-Q_1(\frac{z}{\sigma}, \frac{(\frac{P_2}{P_j})^{\frac{1}{\beta}} y}{\sigma}))))dz)\times \exp(-\pi\lambda_3(\frac{P_3}{P_j})^{\frac{2}{\beta}} y^2) dx dy, \quad (4.14)$$

where

$$\mathcal{M}_y(s,x,y) = \exp(-2\pi\lambda_1' y^2 \frac{s\frac{P_1}{P_j}(\frac{x}{y})^{2-\beta}}{\beta(1-\frac{2}{\beta})} {}_2F_1[1,1-\frac{2}{\beta};2-\frac{2}{\beta};-s\frac{\frac{P_1}{P_j}}{(\frac{x}{y})^{\beta}}])\times$$
$$\exp(-2\pi\lambda_3' y^2 \frac{s(\frac{P_3}{P_j})^{\frac{2}{\beta}}}{\beta(1-\frac{2}{\beta})} {}_2F_1[1,1-\frac{2}{\beta};2-\frac{2}{\beta};-s])\times$$
$$\exp(-2\pi\lambda_{p_2}\int_0^{\infty}(1-\exp(-\bar{m}\int_{(\frac{P_2}{P_j})^{\frac{1}{\beta}}y}^{\infty}\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y'^2}{2\sigma^2})}{1+(s\frac{P_2}{P_j})^{-1}(\frac{y'+z}{y})^{\beta}}dy'))zdz).$$

Proof. 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 Theorem 1, we only need to obtain $\mathcal{M}_y(s,x,y)$, where we see that

$$\begin{aligned}\mathcal{U}_{3,j} &= \int_0^{\infty}\mathbb{E}_{SINR_j}[\ln(1+SINR_j(x))|x,y]f_{\mathcal{S}_{1,j}}(x,y)dxdy \\ &= \int_0^{\infty}\int_0^{\infty}\frac{\mathcal{M}_y(s,x,y)-\mathcal{M}_{xy}(s,x,y)}{s}dsf_{\mathcal{S}_{1,j}}(x,y)dxdy \\ &= \int_0^{\infty}\int_0^{\infty}\frac{\mathcal{M}_y(s,x,y)}{1+s}dsf_{\mathcal{S}_{1,j}}(x,y)dxdy. \end{aligned} \tag{4.15}$$

Then, we derive $\mathcal{M}_y(s,x,y)$ as

$$\mathcal{M}_y(s,x,y) = \mathbb{E}_{h_{\mathbf{y}},\Phi'_n}[\exp(-s(\sum_{n=1}^{3}\sum_{\mathbf{y'}\in\Phi_n\setminus\{\mathbf{x}_j^*\}}\frac{P_n}{P_j}h_{\mathbf{y'}}\left(\frac{\|\mathbf{y'}\|}{y}\right)^{-\beta}+\frac{N_0}{P_j y^{-\beta}}))]$$

$$=\prod_{n\in\{1,3\}}\mathbb{E}_{\Phi'_n}[\prod_{\mathbf{y'}\in\Phi'_n\setminus\{\mathbf{x}_j^*\}}\frac{1}{1+s\frac{P_n}{P_j}(\frac{\|\mathbf{y'}\|}{y})^{-\beta}}]\mathbb{E}_{\Phi_{p_2},\Psi_i}[\prod_{\mathbf{z}\in\Phi_{p_2}}\prod_{\mathbf{y'}\in\Psi_{[\mathbf{z}]}\setminus\{\mathbf{x}_j^*\}}\frac{1}{1+s\frac{P_2}{P_j}(\frac{\|\mathbf{y'}\|}{y})^{-\beta}}]\exp(-s\frac{N_0}{P_j y^{-\beta}})$$

$$\overset{(a)}{=}\exp(-\lambda'_1\int_{\mathbb{R}^2\setminus b_0(x)}\frac{1}{1+(s\frac{P_1}{P_j})^{-1}(\frac{\|\mathbf{y'}\|}{y})^{\beta}}d\mathbf{y'})\exp(-\lambda_3\int_{\mathbb{R}^2\setminus b_0((\frac{P_3}{P_j})^{\frac{1}{\beta}}y)}\frac{1}{1+(s\frac{P_3}{P_j})^{-1}(\frac{\|\mathbf{y'}\|}{y})^{\beta}}d\mathbf{y'})\times$$

$$\exp(-\lambda_{p_2}\int_{\mathbb{R}^2}(1-\exp(-\bar{m}\int_{\mathbb{R}^2\setminus b_0((\frac{P_2}{P_j})^{\frac{1}{\beta}}y)}\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y^2}{2\sigma^2})}{1+(s\frac{P_2}{P_j})^{-1}(\frac{\|\mathbf{y'}+\mathbf{z}\|}{y})^{\beta}}d\mathbf{y'}))d\mathbf{z})\exp(-s\frac{N_0}{P_j y^{-\beta}})$$

$$\overset{(b)}{=}\exp(-2\pi\lambda'_1 y^2\frac{s\frac{P_1}{P_j}(\frac{x}{y})^{2-\beta}}{\beta(1-\frac{2}{\beta})}{}_2F_1[1,1-\frac{2}{\beta};2-\frac{2}{\beta};-s\frac{\frac{P_1}{P_j}}{(\frac{x}{y})^{\beta}}])\times$$

$$\exp(-2\pi\lambda_3 y^2\frac{s(\frac{P_3}{P_j})^{\frac{2}{\beta}}}{\beta(1-\frac{2}{\beta})}{}_2F_1[1,1-\frac{2}{\beta};2-\frac{2}{\beta};-s])\times$$

$$\exp(-2\pi\lambda_{p_2}\int_0^{\infty}(1-\exp(-\bar{m}\int_{(\frac{P_2}{P_j})^{\frac{1}{\beta}}y}^{\infty}\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y'^2}{2\sigma^2})}{1+(s\frac{P_2}{P_j})^{-1}(\frac{y'+z}{y})^{\beta}}dy'))zdz)\exp(-s\frac{N_0}{P_i y^{-\beta}}). \tag{4.16}$$

The equality of (a) holds by using PGFL of PPP and PPCP [15], [20]. 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 $j\in\{2,3\}$. Therefore, the distance of active D2D nodes in the case 3 is in the range of $0<x<(\frac{P_1}{P_j})^{\frac{1}{\beta}}y$. The rest of the proof is similar to Theorem 1. □

We denote $\mathcal{U}_4,i$ 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.

Chapter 5

QoS of Clustered Deployment and Caching Aware Capacity Allocation

In this chapter, 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]).

5.1 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 states 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 a matrix $\mathbf{D}_{8\times4}$, where each entry indicates the probability that the typical user requests contents in the corresponding states and the columns and the rows of the matrix indicate 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

$$\mathbf{D}_{8\times 4} = \begin{pmatrix} \mathcal{G}_{3,1}(1-\alpha)F(1,M_1) & \mathcal{G}_{3,2}(1-\alpha)F(1,M_2) & \mathcal{G}_{3,3}(1-\alpha) & 0 \\ 0 & \mathcal{G}_{3,2}(1-\alpha)F(M_2+1,N) & 0 & 0 \\ 0 & \mathbb{P}_{2,3}\alpha F(M_1+1,M_2) & \mathbb{P}_{3,2}\alpha F(M_1+1,N) & 0 \\ 0 & \mathbb{P}_{2,3}\alpha F(M_2+1,N) & 0 & 0 \\ 0 & \mathbb{P}_{1,2,3}(1-\alpha)F(M_1+1,M_2) & \mathbb{P}_{1,3,2}(1-\alpha)F(M_1+1,N) & 0 \\ 0 & \mathbb{P}_{1,2,3}(1-\alpha)F(M_2+1,N) & 0 & 0 \\ 0 & 0 & 0 & \alpha F(1,M_1) \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.1}$$

of user. The columns represent tiers from which the user in a certain state requests contents. For $i = 1, ..., 8$, and $j = 1, ..., 4$, we get the intensity of active users in each state by the thinning property of independent PPP; for example, $\lambda_0 \mathbf{D}_{i,j}$ is the intensity of users in the state of i-th row and j-th column. We define that the set of index of rows represents a class such that $\{\{1,2\},\{3,4\},\{5,6\},\{7,8\}\} \in \{\text{Class 1}, \text{Class 2}, \text{Class 3}, \text{Class 4}\}$, 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 $\{1,2,3,4\} \in \{\text{D2D}, \text{SBS}, \text{MBS}, \text{Local}\}$. Here, "Local" stands for the case when the users who request contents to their own local storage when the contents are cached.

5.2 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 $i \in \{1,2,...,8\}$ and $j \in \{1,2,3,4\}$. Each MBS cell has the average number of $\frac{\lambda_0}{\lambda_3}$ users, and the rate of a request by each user in that cell is modeled as homogeneous Poisson process with mean arrival rate of ς [requests/s]. Then, the requests of a single user is homogeneous Poisson process with parameter $\frac{\varsigma\lambda_3}{\lambda_0}$. Based on the matrix $\mathbf{D}_{8\times 4}$, we can obtain the average number of users who are active in the corresponding state in each state. For $i = 1, ..., 8$ and $j = 1, ..., 4$, the average number of users in each state is $\frac{\lambda_0 \mathbf{D}_{i,j}}{\lambda'_j}$, where $\lambda'_4 = \alpha\lambda_0$. The total request arrival rate in each BS/D2D is ready to be obtained as $\zeta_{i,j} = \frac{\lambda_0 \mathbf{D}_{i,j}}{\lambda'_j} \frac{\varsigma\lambda_3}{\lambda_0}$. 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 $\frac{1}{\varrho}$ [contents/request]. Next, we define

the average ergodic rate of the user in the state of (i, j) as follows:

$$\mathbf{A}_{8\times 4} = \begin{cases} A_{2m-1,j} = \eta\omega\mathcal{U}_{m,j}1\{\mathbf{D}_{2m-1,j} \neq 0\}, \\ A_{2m,j} = \eta\omega f(\mathcal{U}_{m,j})1\{\mathbf{D}_{2m,j} \neq 0\}, \end{cases} \tag{5.2}$$

where $m = 1, 2, 3, 4$ and $f(\cdot)$ is the effect of backhaul delay, is an arbitrary function. $\eta = 1.443$ is the conversion factor between [nats] and [bits], and ω 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.

5.3 DPS Queue and QoS Metric

Let $\mathbb{D} := \{1, 2, ..., 8\}$ 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 $\mathbf{x}_j = (x_{1,j}, x_{2,j}, ..., x_{8,j})$ which is counting the number of requests in each class. We introduce weights $w_{1,j}, w_{2,j}, ..., w_{8,j}$ 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

$$\begin{cases} q(\mathbf{x}_j, \mathbf{x}_j + \boldsymbol{\epsilon}_i) = \zeta_{i,j}, & \mathbf{x}_j \in \mathbb{N}^{\mathbb{D}}, \\ q(\mathbf{x}_j, \mathbf{x}_j - \boldsymbol{\epsilon}_i) = \frac{\mathbf{A}_{i,j}}{S/\varrho} \frac{x_{i,j}}{x_{\mathbb{D},j}} \frac{w_{i,j}}{w_{\mathbb{D},j}}, & \mathbf{x}_j \in \mathbb{N}^{\mathbb{D}}, \mathbf{x}_j > 0, \end{cases} \tag{5.3}$$

where $\boldsymbol{\epsilon}_i$ is the vector of $\mathbb{N}^{\mathbb{D}}$ with 1 in i-th element $(i = 1, ..., 8)$ and 0 elsewhere. $x_{\mathbb{D},j} = \sum_{i\in\mathbb{D}} x_{i,j}$ is the total number of requests in the queue, and $w_{\mathbb{D},j} = \sum_{i\in\mathbb{D}} w_{i,j}$ is the total weights of all classes. The class traffic demand can be obtained as $\rho_{i,j} = \frac{\zeta_{i,j}S}{\varrho}$, and the critical traffic value which the queue will be steady state is $\rho_{c,j} = \frac{\rho_{\mathbb{D},j}}{\sum_{i\in\mathbb{D}} \rho_{i,j}\mathbf{A}_{i,j}^{-1}}$ where $\rho_{\mathbb{D},j} = \sum_{i\in\mathbb{D}} \rho_{i,j}$ (see [6], [8]). We define a tier traffic intensity of DPS queue as $\rho'_{i,j} = \frac{\zeta_{i,j}}{S/(\varrho\mathbf{A}_{i,j})} \triangleq \frac{\lambda_{(i,j)}}{\mu_{(i,j)}}$, where $S/(\varrho\mathbf{A}_{i,j})$ 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* $\sum_{i\in\mathbb{D}} \rho^{'}_{i,j} < \rho_{c,j} < 1$, *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* $i \in \mathbb{D}$ *and* $j \in \{1,2,3,4\}$ *can be obtained as,*

$$\bar{N}^{INT}_{i,j} = \lambda_{(i,j)} \bar{S}^{INT}_{(i,j)}(\lambda_{(i,j)}, \mu_{(i,j)}, w_{1,j}, w_{2,j}, ..., w_{8,j}), \tag{5.4}$$

$$\bar{D}^{INT}_{i,j} = \bar{S}^{INT}_{(i,j)}(\lambda_{(i,j)}, \mu_{(i,j)}, w_{1,j}, w_{2,j}, ..., w_{8,j}), \tag{5.5}$$

$$\bar{T}^{INT}_{i,j} = \frac{\rho^{'}_{i,j}}{\bar{N}^{INT}_{i,j}}, \tag{5.6}$$

where the approximated sojourn time of DPS queue is

$$\bar{S}^{INT}_{(i,j)}(\lambda_{(i,j)}, \mu_{(i,j)}, w_{1,j}, w_{2,j}, ..., w_{8,j}) = \frac{1}{\mu_{(i,j)}} + \frac{1}{\mu_{(i,j)}} \sum_{k=1}^{8} (\frac{\lambda_{(k,j)}}{\mu_{(k,j)}}) + \sum_{k=1}^{8} (\frac{\lambda_{(k,j)}}{\mu_{(k,j)}} \frac{w_{k,j} - w_{i,j}}{w_{k,j}\mu_{(k,j)} - w_{i,j}\mu_{(i,j)}}) + \frac{(\sum_{k=1}^{8} (\frac{\lambda_{(k,j)}}{\mu_{(k,j)}}))^2}{1 - \sum_{k=1}^{8} (\frac{\lambda_{(k,j)}}{\mu_{(k,j)}})} \frac{1}{w_{i,j}\mu_{(i,j)}} \frac{\sum_{k=1}^{8} (\frac{\lambda_{(k,j)}/\lambda_{(\mathbb{D},j)}}{\mu^2_{(k,j)}})}{\sum_{k=1}^{8} (\frac{\lambda_{(k,j)}/\lambda_{(\mathbb{D},j)}}{\mu^2_{(k,j)} w_{k,j}})}. \tag{5.7}$$

Proof. Since $\sum_{i\in\mathbb{D}} \rho^{'}_{i,j} < \rho_{c,j} < 1$ 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 (5.4) follows from Little's law, and (5.5) and (5.6) are by the definition (see Proposition 1 [8]). □

This study aims to see the effectiveness and behavior of caching aware capacity allocation system, where its simulation is infeasible. The accuracy of the approximation

of mean unconditional sojourn time of DPS queue has already studied extensively by [14]. Therefore, we explore the results of numerical calculation only.

Chapter 6

Numerical Results

In this chapter, 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 compering with the case of non-clustered deployment of SBSs which is our baseline examined by the previous work done by [6].

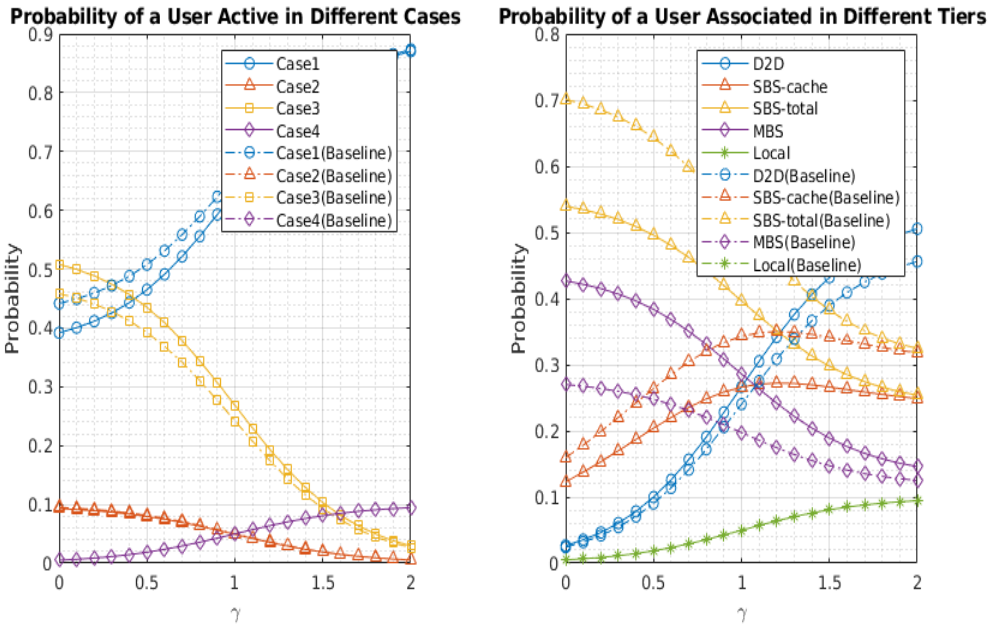

Figure 6.1: The probability of a user active in different cases and associated in different tiers; $\alpha = 0.1$, $\{P_1, P_2, P_3\} = \{3, 13, 193\}$, $\beta = 4$, $\bar{m} = 10$, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\} = \{\frac{1000}{\pi * 1000^2}, \lambda_0 * \alpha, \frac{3*\bar{m}}{\pi*1000^2}, \frac{2}{\pi*1000^2}\}$, and when the case of baseline, $\lambda_2 = \frac{30}{\pi*1000^2}$. The variance of TCP is $\sigma = 250$.

Figure 6.1 shows the probability of a user active in different cases (left) and different tiers (right) with changing the parameter of the distribution of contents. Recall that γ is

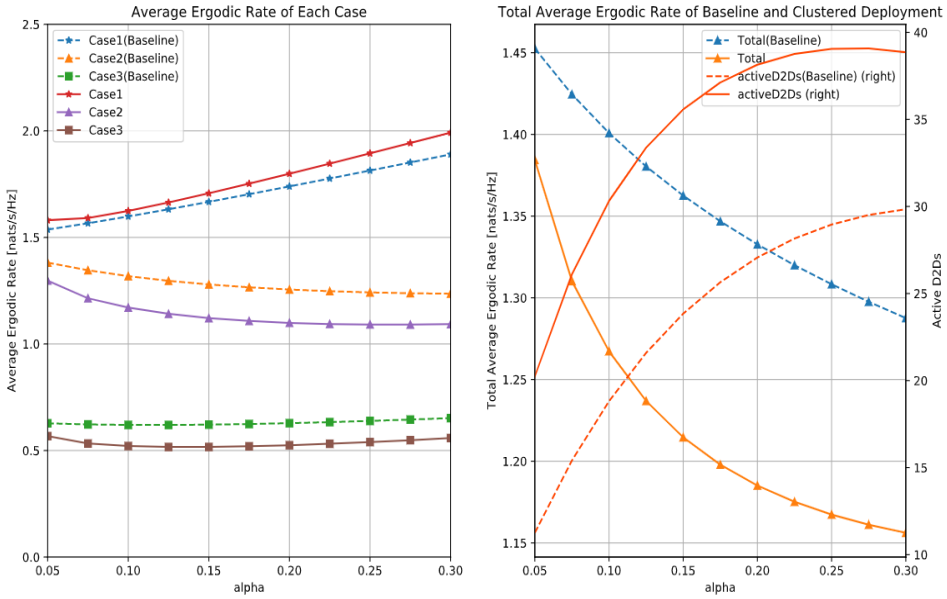

Figure 6.2: The average ergodic rate in different cases with varying α; $\gamma = 0.8$, $\{P_1, P_2, P_3\} = \{73, 373, 1773\}$, $\beta = 4$, $\bar{m} = 10$, $A = 1000^2$, $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\} = \{\frac{300*A}{\pi*500^2}, \lambda_0 * \alpha, \frac{3*\bar{m}*A}{\pi*500^2}, \frac{6*A}{\pi*500^2}\}$, and when baseline case is $\lambda_2 = \frac{30*A}{\pi*500^2}$. The variance of TPP $\sigma = 0.05$.

the parameter of content distribution. The larger value of γ implies that the distribution is steeper and dominated by popular contents. From the plot on the left, we see that only case 1 and case 3 are sensitive to the distribution of contents. When the popularity of the content is even, the clustered deployment of SBSs reduces the probability of user active in case 1 and increase the probability in case 3. The plot on the right coincides with our intuition; the clustered deployment of SBSs reduces the probability of association to itself from users whereas it increases the probability of association to MBSs. As γ increases, the probability of user associated with D2D transmission increases, and when the deployment of SBSs are clustered, the increase is accelerated.

Figure 6.2 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.

Figure 6.3 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

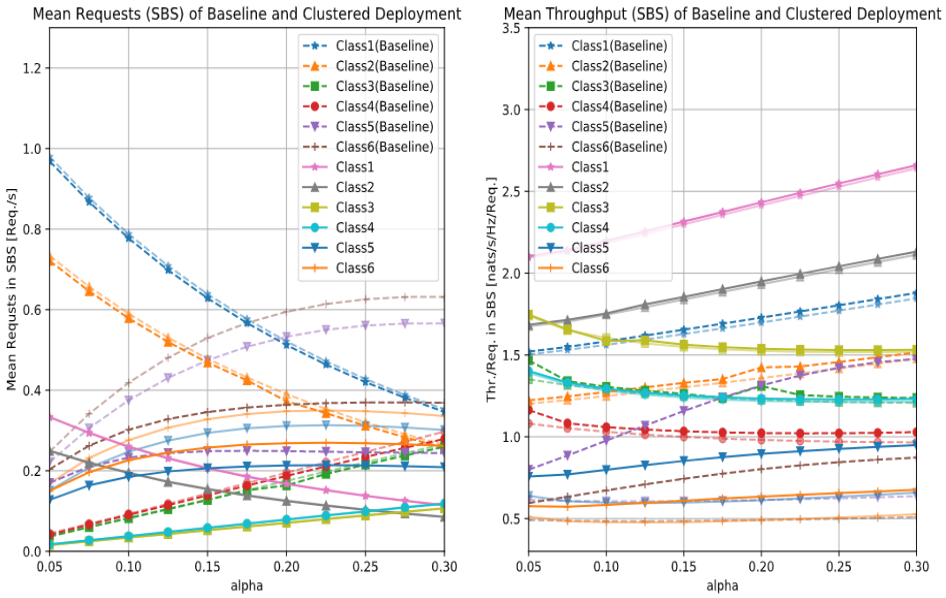

Figure 6.3: The mean number of requests and the mean throughput of SBS; the lines with pale color are PS queue results, and the lines with bright color are DPS queue results. The weights for baseline and clustered deployment:$\{w_{1,2}, w_{2,2}, w_{3,2}, w_{4,2}, w_{5,2}, w_{6,2}\} = \{1, 1, 1.1, 1.1, 1.5, 1.87\}$ with $\varsigma = 0.2$, $S = 100$[Mbits], $\omega = 70$MHz, $\varrho = 1$; parameters are the same as Figure 6.2.

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 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. This result confirms the result from Figure 6.1 where the probability of a user associated with SBS is smaller than that of the baseline. Next, while α (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 case 2 and 3 users belonging to 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.

Figure 6.4 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 Figure 6.3. Next, we found that the traffic load of MBSs increases under the clustered deployment of SBSs and this

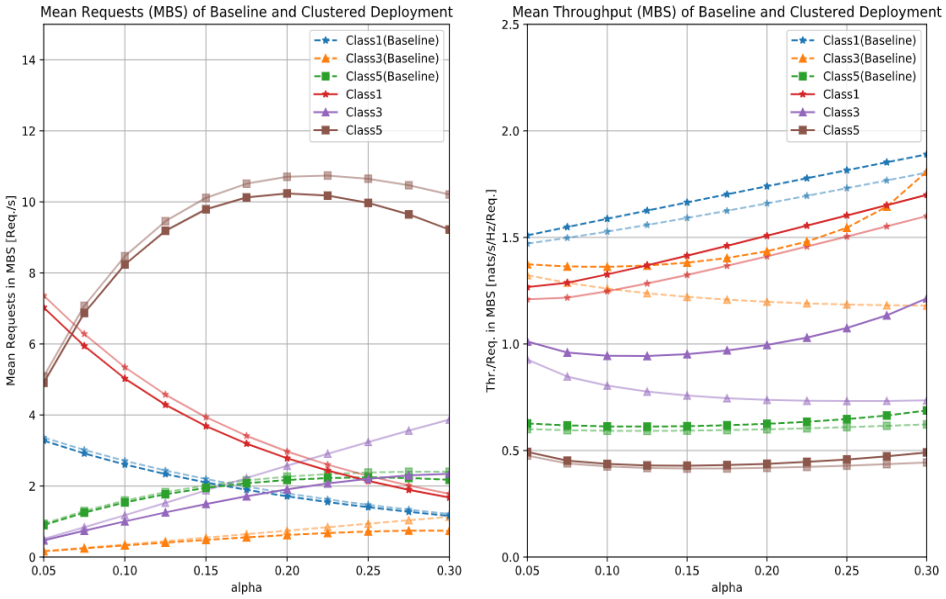

Figure 6.4: The mean number of requests and the mean throughput of MBS; the lines with pale color are PS queue results, and the lines with bright color are DPS queue results. The weights for baseline:$\{w_{1,3}, w_{3,3}, w_{5,3}\} = \{1, 1, 1.8\}$ and for clustered deployment $\{w_{1,3}, w_{3,3}, w_{5,3}\} = \{1, 1, 1.5\}$ with $\varsigma = 0.2$, $S = 100$[Mbits], $\omega = 70$MHz, $\varrho = 1$; parameters are the same as Figure 6.2.

increase of traffic load is conspicuous from class 5 users. Also, this result confirms the result from Figure 6.1 where the probability of a user associated with MBSs is larger than that of baseline.

Figure 6.5 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.

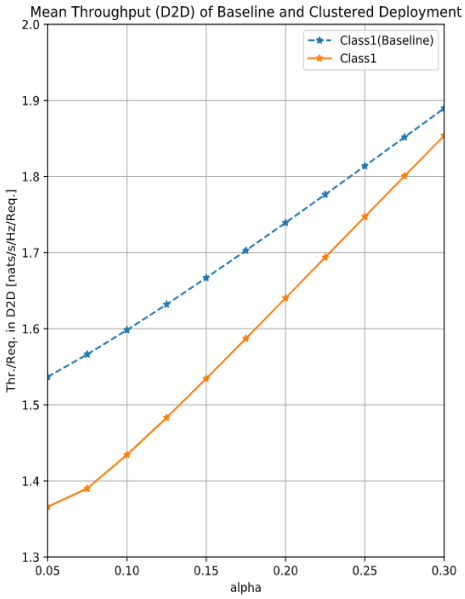

Figure 6.5: The mean throughput of D2D based on PS queue with $\varsigma = 0.2$, $S = 100$[Mbits], $\omega = 70$MHz, $\varrho = 1$; parameters are the same as Figure 6.2.

Chapter 7

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 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 are 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 [22] into the system might be an interesting point to start.

Bibliography

[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. Baştuğ, "Distributed caching methods in small cell networks," *CentraleSupélec Paris-Saclay University*, 2015.

[5] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "On the delay of geographical caching methods in two-tiered heterogeneous networks," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2016, pp. 1–5.

[6] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 131–145, 2016.

[7] Y. Wu, C. Williamson, and J. Luo, "On processor sharing and its applications to cellular data network provisioning," *Performance Evaluation*, vol. 64, no. 9, pp. 892–908, 2007, Performance 2007, ISSN: 0166-5316. DOI: https://doi.org/10.1016/j.peva.2007.06.011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166531607000612.

[8] M. K. Karray and M. Jovanovic, "A queueing theoretic approach to the dimensioning of wireless cellular networks serving variable-bit-rate calls," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 6, pp. 2713–2723, 2013.

[9] B. Błaszczyszyn, M. Jovanovic, and M. K. Karray, "Performance laws of large heterogeneous cellular networks," in *2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2015, pp. 597–604.

[10] B. Błaszczyszyn and M. K. Karray, "Performance analysis of cellular networks with opportunistic scheduling using queueing theory and stochastic geometry," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5952–5966, 2019.

[11] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a processor among many job classes," *J. ACM*, vol. 27, no. 3, pp. 519–532, Jul. 1980, ISSN: 0004-5411. DOI: 10.1145/322203.322212. [Online]. Available: https://doi.org/10.1145/322203.322212.

[12] K. Avrachenkov, U. Ayesta, P. Brown, and R. Nunez-Queija, "Discriminatory processor sharing revisited," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 2, 2005, 784–795 vol. 2.

[13] E. Altman, K. Avrachenkov, and U. Ayesta, "A survey on discriminatory processor sharing," *Queueing Systems*, vol. 53, no. 1, pp. 53–63, Jun. 2006, ISSN: 1572-9443. DOI: 10.1007/s11134-006-7586-8. [Online]. Available: https://doi.org/10.1007/s11134-006-7586-8.

[14] A. Izagirre, U. Ayesta, and I. M. Verloop, "Sojourn time approximations for a discriminatory-processor-sharing queue," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, vol. 1, no. 1, art n°5, Mar. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01089833.

[15] M. Haenggi, *Stochastic Geometry for Wireless Networks*, 1st. USA: Cambridge University Press, 2012, ISBN: 1107014697.

[16] N. Miyoshi, "Downlink coverage probability in cellular networks with Poisson-Poisson cluster deployed base stations," *IEEE Wireless Commun. Letters*, vol. 8, no. 1, pp. 5–8, 2019. DOI: 10.1109/LWC.2018.2845377. [Online]. Available: https://doi.org/10.1109/LWC.2018.2845377.

[17] C. Saha, H. S. Dhillon, N. Miyoshi, and J. G. Andrews, "Unified analysis of hetnets using Poisson cluster processes under max-power association," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3797–3812, 2019.

[18] M. Afshang, C. Saha, and H. S. Dhillon, "Nearest-neighbor and contact distance distributions for Thomas cluster process," *IEEE Wireless Communications Letters*, vol. 6, no. 1, pp. 130–133, 2017. DOI: 10.1109/LWC.2016.2641935.

[19] K. A. Hamdi, "A useful lemma for capacity analysis of fading interference channels," *IEEE Transactions on Communications*, vol. 58, no. 2, pp. 411–416, 2010.

[20] C. Saha, M. Afshang, and H. S. Dhillon, "3GPP-inspired hetnet model using Poisson cluster process: Sum-product functionals and downlink coverage," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2219–2234, 2018. DOI: 10.1109/TCOMM.2017.2782741.

[21] "3–4 - definite integrals of elementary functions," in *Table of Integrals, Series, and Products (Seventh Edition)*, A. Jeffrey, D. Zwillinger, I. Gradshteyn, and I. Ryzhik, Eds., Seventh Edition, Boston: Academic Press, 2007, pp. 247–617, ISBN: 978-0-12-373637-6. DOI: https://doi.org/10.1016/B978-0-08-047111-2.50013-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780080471112500133.

[22] B. Btaszczyszyn, A. Brochard, and H. P. Keeler, "Coverage probability in wireless networks with determinantal scheduling," in *2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2020, pp. 1–8.