Department of Electronic and Electrical Engineering

University of Sheffield

Content Caching Policy and Performance Optimization in Fog Radio Access Networks

Haojiang Hua

This dissertation is submitted for the degrees of

Degree of Philosophy

November 2021

Acknowledgements

First of all, I am grateful to my first supervisor, Prof. Xiaoli Chu, for her patient guidance of my PhD research and great support in improving my research skills. I thank her for her selfless support and suggestions to help me work through my problems. Without her clear instruction, my research and thesis could not have been completed. As I was dealing with some personal issues, her patient guidance helped me get through my most difficult times. I will benefit from her advice all my life.

Second, I am grateful to my second supervisor, Dr. Wei Liu, for his kind suggestions and help. I am also indebted to Dr. Bleron Klaiqi, who helped me in the beginning of my PhD studies.

Last, my thanks go to my family and friends for their encouragement, consideration, and support.

Abstract

With the rapid growth of mobile device users and mobile communication applications, the explosive traffic demand has brought many new challenges the network system. For matching the increasing traffic demand, cloud radio access network (C-RAN) is proposed as a solution in optimizing the energy efficiency and spectral efficiency which is made up of a centralized baseband signal processing unit (BBU) in the cloud center with a set of remote radio heads (RRHs) densely deployed. C-RAN can centralize the functions of computing, signal processing, and content caching on the cloud platform through the computational capabilities of the BBU pool. However, the full-centralized architecture of C-RAN also brings heavy burdens on the fronthaul which is used for connecting the BBU pool and RRHs.

To solve these disadvantages and optimize system services, based on C-RAN architecture, the fog radio access network (F-RAN) is proposed. Different from the remote radio heads in C-RAN, which is only responsible for transmitting signals, fog access points (F-APs) are applied in fog radio access network with functions of signal processing, resource management and local content storage. Therefore, by distributing the contents close to the content requested users, fog radio access network has been proposed as a promising paradigm for the 5th generation of mobile networks to improve the quality of service. Also, fog radio access network has a broad application prospect in improving spectrum efficiency and energy efficiency by making full use of centralized cooperation and edge cache technology. When investigating the optimization problem of the network service in F-RAN systems, much attention has been paid to the multiple types of edge caching devices, fronthaul traffic and content caching policy. In F-RAN, the edge cache device that is assigned the signal processing function and the content caching function can be other devices besides the F-AP, such as ordinary users. Such users are referred as edge caching users (EUs). Since both the fog radio access point and edge-caching users can possess content caching and transmission functionalities, the case of providing content downloading services via device-to-device (D2D) communications and clustered cooperative D2D communications can be considered in the study of F-RAN. However, few existing works study the performance optimization of F-RAN system in considering the D2D communication and EUs. Therefore, in the case that multiple types of edge caching devices are enabled in F-RAN, the way of improving the content download service by optimizing the content transmission mode selection and content caching policy will become the main investigation direction in this thesis. This thesis studies the optimization of system performance which include energy efficiency, cache hit probability and average download delay by optimizing transmission mode selection, optimizing content caching policy for EUs, applying D2D communication and cooperative communication.

In Chapter 3, the optimization problem of average energy efficiency for content downloading service in F-RAN system with edge caching users assisted is investigated. To optimize the energy efficiency, the greedy algorithm and transmission mode selection method is considered as means of optimization. In this F-RAN system, there are totally two transmission modes for content requesting users, D2D mode, F-AP mode. We define content requested users as model option users (MOUs) when the D2D mode and F-AP can be both selected. The content downloading process in F-RAN system is summarized. In particular, the way of caching content in the edge caching devices by using caching policy during pre-fetching phase is studied. As a reference law for analyzing the probability distribution of things, Zipfs law is investigated and applied to represent the content popularity of this F-RAN system. Also, the energy efficiency in content downloading phase of different transmission modes is analyzed and summarized. For optimizing the average energy efficiency of D2D transmission mode, edge caching users allocation scheme based on greedy algorithm is proposed. Meanwhile, for improving the average energy efficiency of whole F-RAN system in content downloading service, the transmission mode selection method based on two transmission mode is proposed. The optimal transmission mode for each content requested users will be selected by comparing D2D transmission with F-AP transmission. Simulation results show that the proposed mode selection method can not only reduce the energy consumption in content transmission, but also significantly improve the energy efficiency per request in the F-RAN. In the simulation results, the average energy efficiency can be increased by up to 58% by using the proposed transmission mode selection method compared with not using mode selection when Zipf's exponent equals to 0.7.

Moreover, in Chapter 4, the average energy efficiency of the content downloading is studied when D2D clustered cooperative beamforming is applied as one additional content transmission method in F-RAN system. Cooperative beamforming typically provides a longer service distance than direct D2D transmission and consumes less energy than F-AP transmission. In order to optimize the cache hit probability of edge cache users and improve the average energy efficiency for content downloading service, this chapter not only enables D2D cluster cooperative beamforming as an additional transmission mode, but also proposes the new content caching policy called sub-popular caching policy in the pre-fetching process based on the tradi-

tional probability-based caching policy. There are totally three transmission modes for content requesting users, D2D mode, F-AP mode and D2D clustered cooperative beamforming mode. Next, based on the content popularity provided by Poisson's law and Zipf's law, we provide two different content sets for the pr-fetching of EUs. Based on these two content sets, two caching policies will be studied-sub-popularity content caching policy (SP caching policy) and probability based caching policy (PB caching policy). The energy consumption and energy efficiency of three transmission modes are studied, and the relay selection scheme based on timer is used in the analysis of D2D cluster cooperative beamforming. In addition, the expressions of cache hit probability in three transmission modes are summarized, and the selection probabilities of different transmission modes are given based on content caching policy. Simulation results show SP content caching strategy has lower Zipf exponent and higher DEU density than PB content caching policy. If the Zipf's exponent is less than 0.5, sub-popularity content caching policy can improve the average download energy efficiency compared with probability-based caching policy. In addition, increasing the service range of D2D cluster cooperative beamforming does not affect the selection of caching policy, but can optimize energy efficiency to a certain extent. When Zipf's exponent equals to 0.2, the density of D2D edge caching users equals to $3 \times 10^{-4} m^{-2}$ and the density of cooperative beamforming allowed edge caching users equals to $0.6 \times 10^{-4} m^{-2}$, using the sub-popularity content caching policy can provide approximately 10% percent more average energy efficiency than using probability-based caching policy. Meanwhile, compared with not using D2D cluster, using D2D cluster and applying sub-popularity caching policy can improve the average energy efficiency by 16.5 %. The improvement prove that using the D2D clustered cooperative beamforming and sub-popularity caching policy can assist the F-RAN system improve the average energy efficiency.

Chapter 5 mainly studies the impact of content caching policy in D2D assisted F-RAN system on cache hit probability and download delay. This Chapter assumes that the fog access point knows the spatial density of edge caching users and the popularities of contents. As for this, we propose the edge caching user classification content caching policy in F-RAN for reducing the impact of storage constraints of edge caching users. Also, combined with probability-based caching policy, an optimized content caching policy selection algorithm for maximizing cache hit probability at edge caching users and minimizing the average download delay for content requests. In the edge caching user classification content caching policy, edge caching users are divided into two groups, each with a different content set. We summarize and analyze the

process of pre-cache and file transfer in the system. Poisson Point Process provides a solution for calculating content popularity. M/D/1 queuing system is considered for dynamic requests generated by multiple users. By studying the learning queuing theory, the download delay is divided into waiting time in queue and service time. Based on the M/D/1 queuing system, the download delay expressions with different caching policies are studied. By studying the properties of the Poisson point process and M/D/1 queue system, We focus on maximizing the cache hit probability at edge caching users and analyzing the effect on average download delay. In the simulation, optimized content caching policy selection algorithm is compared with only selecting probability based content caching policy. Compared to only using the probability based caching policy, using optimized content caching policy selection algorithm can obtain 0 to 45 % for cache hit probability of edge caching users with Zipf's exponent ranging from 1.2 to 0.2. For average download delay, using caching policy selection can obtain the gain from 0 to 4.15%. The simulation results show that the proposed EU classification content caching policy can significantly increase the cache hit probability and reduce the average download delay when differences in popularity among the contents are small.

Contents

List of Tables

List of Figures

List of Abbreviations

BBU	Baseband processing unit
C-RAN	Cloud radio access network
CEU	Cooperative beamforming allowed edge caching user
CRRM	Cooperative radio resource management
CRSP	Collaboration radios signal processing
CSI	Channel stat information
D2D	Device-to-Device
DEU	D2D edge caching user
EUC	Edge caching user classification
EU	Edge caching user
F-AP	Fog radio access point
F-RAN	Fog radio access network
HPN	High-power node
MOU	Mode option user
PB	Probability based
PPP	Poisson point process

RRH	Remote radio head
RU	Content requested user
SNR	Signal to noise ratio
SP	Sub-popularity

Chapter 1

Introduction

1.1 Background and Motivation

Due to the explosive development of mobile devices and mobile applications, wireless traffic demands has been growing rapidly [1]. Cloud radio access network (C-RAN) is thought as a promising candidate to optimize the content download delay and energy efficiency. This network system dives the traditional base station into two parts: base band unit (BBU) pool and remote radio heads (RRHs). The high computational capabilities of cloud computing platforms have enabled C-RANs, where wireless signal processing and resource allocation functions are centralized in a baseband processing unit BBU pool. BBU pool is located at cloud with three functions: centralized control, centralized communication and centralized caching. RRHs located near user equipment only implement radio frequency functionality [2]. RRH can be regarded as a base station with only signal transmission function. As Fig. 1.1 shows, RRHs are connected to the BBU in the cloud by means of fronthaul links [2][4]. With a BBU pool and densely deployed RRHs, it can provide greater spectral efficiency and energy efficiency [3]. Besides, BBU pool and RRHs are linked using fronthaul which can provide high-speed transmission services [5]. However, heavy burden on and constrained capacity of fronthaul degrade the performance in C-RANs [6]. Fronthaul capacity limitations, processing latency in the BBU pool, and service delay are the major challenges facing C-RANs [7].

With the proposal of fog computing [9], the emerging technology of fog-radio access networks (F-

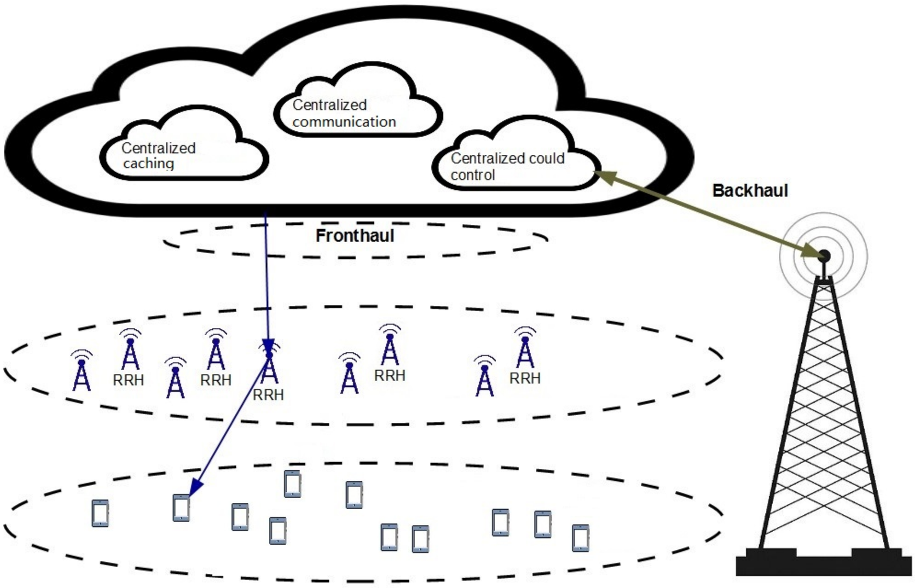

Figure 1.1: Cloud radio networks system

RANs) has been introduced. The F-RAN is a promising paradigm incorporating edge storage, cloud computing, and fog computing to improve the performance of the fifth generation of wireless networks [6][10]. In the F-RAN, fog means that some cloud functions are distributed to other devices to reduce traffic load at front haul and optimize content transmission. These devices are called fog nodes. As Fig. 1.2 shows, there are two types of fog nodes should be considered when we study F-RAN system: fog access point (F-AP) and edge caching user (EU). Different from C-RANs, in F-RANs, the content providing service cannot only be executed from the BBU pool, but also can be achieved by the fog nodes. It is well known that the constraints of front haul create significant bottlenecks in optimize system performance [7]. By moving storage and communication functionalities to the network edge, F-RAN can effectively decrease the service latency for cellular users, improve energy efficiency and reduce the fronthaul load. Consequently, in the F-RANs, switch content caching and data processing to fog nodes is the key to optimize the energy efficiency, content download delay and the burden at front haul. As a network edge communication technology that can support content delivery and sharing, D2D communication has attracted more and more attention in F-RAN study[8]. In F-RANs, EUs can be applied to reduce the traffic load at front haul and optimize the network performance. Using EUs to provide contents in F-AP means contents are cached at the EUs for content providing service via D2D transmission. In the traditional underlaid cellular networks, cache enabled D2D communication has already been studied in improving the system performance [11][12][13]. In addition, device-to-device (D2D) technology has been validated and studied in regard to the improvement of the content delivery performance of F-RANs [14][37]. EUs can provide content services to users who request content directly to RUs to reduce the content delivery latency. Thus, for optimizing the energy efficiency and content download latency, the D2D communication is considered in this thesis as one content transmission mode. Fig. 1.3 shows the structure of F-RAN. When the content download service is provided by BBU pool in the cloud center, the requested content can be sent from BBU pool to target user via RRH as the blue arrow shows in Fig. 1.3. Also, as the red arrow shows in Fig. 1.3, the F-AP with can transfer the requested content directly to the content requested user. Since EUs are enabled in F-RAN, D2D transmission can be used to provide content download services as shown in the black arrow. As a result, the content providing service can be executed by BBU pool, F-APs and EUs in F-RANs.

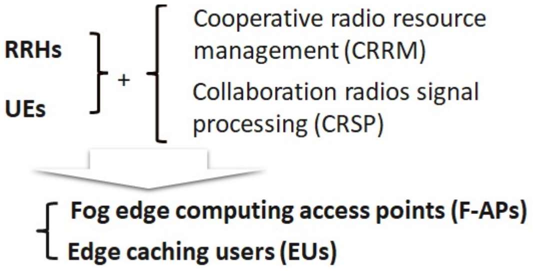
RRHs
UEs
+
Cooperative radio resource management (CRRM)
Collaboration radios signal processing (CRSP)
Fog edge computing access points (F-APs)
Edge caching users (EUs)

Figure 1.2: Fog nodes

Caching, computing, and signal transmission functionalities in F-APs and EUs help improve the performance of network systems. Optimization of energy efficiency, download delay, and cache hit probability are hot topics for F-RAN study. In the study of content caching policy, the popularity of requested contents generally follows Zipf distribution [16]. Zipf's law is a rank-size rule that captures the relationship between the frequency of a group of objects and their size. In the corpus of natural language, the frequency of a word's occurrence is inversely proportional to its rank in the frequency chart. So the most frequent word is about twice as common as the second most frequent word, and the second most frequent word is twice as common as the fourth most frequent word. This law is used as a reference for anything to do

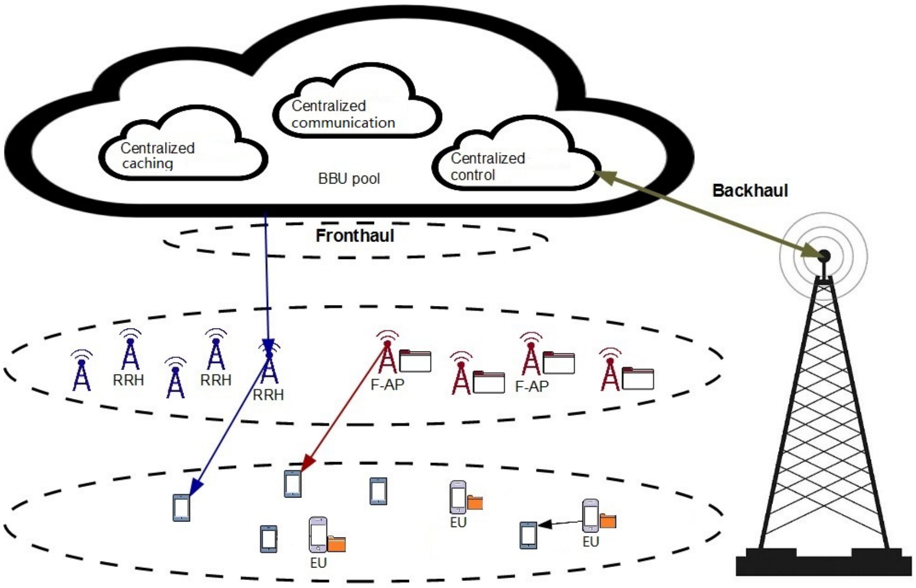

Figure 1.3: Fog radio networks system

with power law probability distributions [17]. For N contents, the content request probability of content f_x has a Zipf distribution: $p_x = \frac{1/x^{\beta}}{\sum_{x=1}^{N} 1/x^{\beta}}$, where $\beta \geq 0$ is the Zipf exponent that controls the difference of content requested frequency to different contents. Larger β means that the requirement to a small number of popular content accounts for the majority of total content requirement. As Zipf's exponent changes, the probability distribution of users' content requirement changes.

Much work has been done to improve content caching policies and mode selection in performance optermization. Reinforcement learning was used to learn user preferences, which was used to optimize the content placement strategy in the F-AP [19][18] . A time-varying popular content set was considered and the characterisation of the long-term normalised delivery time was analysed [20]. Optimized caching placement was also proposed [21], where a binary caching variable was defined to indicate whether cooperative content had been cached in a base station. The authors studied both cooperative and non-cooperative transmission to optimize the file delivery rate and transmission latency. Time-varying content requests have also been studied [22]. By investigating ultra-dense fog radio access networks, the authors used the dynamic caching state to represent the time-varying popularity for each item of content and optimized the request service delay. The service delay denotes the time delay in wireless transmission and fronthaul transmission. Neither study discusses the content caching policy of EUs.

Instead of discussing the download delay, the paper [23] studied the energy efficiency organisation by assigning different content sets to F-APs which have various active levels, without having to consider the edge caching user. Considering dynamic content-centric F-AP clustering under three content caching strategies, a popularity-aware content caching strategy was selected to provide better results, focusing on power cost and backhaul capacity [24]. The power consumption of edge caching and transitional radio resources were also studied [25]. The energy efficiency optimization of a downlink C-RAN was studied with the help of large-dimensional random matrix theory, and the optimal caching capacity was discussed while considering the fronthaul compression to improve energy efficiency and transmission power [26].

D2D communication enables the user to directly transmit a message to a nearby receiver [27]. Without transmission to the base station, it can improve energy and spectral efficiency [28]. In an F-RAN system, an EU can be considered as the D2D transmitter to provide the requested content to an RU through D2D communication. D2D technology has been validated and studied in regard to the improvement of the content delivery performance of F-RANs [14][37]. However, most work has focused on improving the caching placement of F-APs, and little has focused on optimizing the cached content placement for EUs, which can increase energy efficiency and cache hit probability, and decrease transmission latency.

User clustered cooperative beamforming has been investigated to improve the transmission rate of the user [29]. It is also a reliable wireless transmission technique to optimize energy efficiency [30]. An EU can transmit content with the help of idle users, which form a virtual antenna array to the RU. Joint cooperative beamforming in F-RAN was proposed to optimize total power consumption and backhaul capacity [31]. Cooperative beamforming transmission in this work is provided by radio units that are the same as F-APs. To minimize the fronthaul cost and total transmit power network, beamforming was considered, which was also provided by F-APs [32]. There has been little discussion of the performance of F-RANs with user clustered cooperative beamforming.

To optimize the average energy efficiency, cache hit probability, and content download delay for F-RAN, which allows D2D communication and D2D clustered cooperative beamforming, it is essential to analyse the transmission mode and caching policy. This thesis intends to devise mode selections for different transmission modes and content caching policies for edge caching devices to optimize the performance of an EU-assisted F-RAN.

1.2 Contributions

The main contributions of are as follows.

- For the EU assisted F-RAN system, a greedy mode selection method is proposed to optimize the average energy efficiency per request. RUs are defined as mode option users (MOUs) when their requested contents can be served by both EUs and F-AP. Based on the greedy algorithm and the distance between MOUs and the F-AP, the greedy mode selection method is proposed. In this method, the MOU with the furthest distance from the F-AP is first assigned best EU that can provide the highest energy efficiency of D2D content transmission among all the EUs. Then, comparing the energy efficiency between D2D transmission by the best EU and F-AP transmission, the MOU is assigned the optimal transmission mode that maximizes the average energy efficiency.

- Theoretical analysis of energy efficiency for providing requested content by cloud mode, F-AP mode and D2D mode considering the overhead of sending the request task to the F-AP, selecting transmission mode and providing the requested content. To optimize the average energy efficiency, the greedy mode selection for RUs is generated according to the analysis of energy efficiency and the D2D transmitter assignment method. The impacts of the value of the Zipf exponent and the number of D2D transmitters on the average energy efficiency are analysed through simulations. The optimal performance of the proposed greedy mode selection to improve energy efficiency is verified by the simulation results.

- This thesis investigate the D2D clustered cooperative beamforming in F-RAN for improving the energy efficiency, cache hit probability and mode selection probability. We define the cooperative beamforming allowed edge caching users (CEUs) as the source which can provide the requested content with the help of idle D2D users. Idle D2D users, which uniformly distribute around the CEU, act as relays. Timer based relay selection is applied in realizing the D2D clustered cooperative beamforming. RUs can obtain the required content through one of three different transmission modes: F-AP mode, D2D mode and D2D clustered cooperative beamforming mode. We provide EUs with a new set of content to choose from for caching.This content collection is less popular than the original provided content set called sub-popularity content set. For CEUs, a novel content caching policy called sub-popularity content caching policy is proposed . The proposed policy can make the EUs to cache a sub-popularity content set based on Zipf's law and the size of local caching space, which is different from the content set provided by the probability based caching policy.

- Theoretical analysis of energy efficiency, and mode selection probability for three transmission modes with constraints such as threshold data rate, fog node service range and symbol length of different messages. Power consumption with content provider selection, channel estimation, relays selection and content transmission is analyzed. Compared the mode selection probability for using PB caching policy and SP caching policy is carried out. Furthermore, the average energy efficiency per request with different caching policies is simulated and studied. The impacts of the value of Zipf exponent, threshold service range of CEUs and density of CEUs on the performance of PB caching policy and SP caching policy when applied to CEUs are studied. The suitable situation of choosing SP caching policy to improve energy efficiency is analyzed through simulations.

- An EU classification content caching policy is proposed for the EU-assisted F-RAN system. This divides EUs into two groups, each assigned a different set of content. The cache hit probability maximization problem is formulated and decomposed into two subproblems: (i) cache hit probability maximization when the EU classification-based caching policy is utilized; and (ii) construction of the content caching policy selection algorithm to maximize cache hit probability. The optimal ratio of the two EU groups that maximizes the cache hit probability is derived as a function of the Zipf exponent of contents, density of EUs, density of RUs, caching space size of EUs, and service range of EUs.

- A theoretical analysis is conducted of the cache hit probability of EUs for the probability-based and EU classification content caching policies, whose average download delays are studied utilizing the M/D/1 queuing model. The cache hit probability of EUs and average download delay per request are evaluated under the impact of the Zipf exponent and density of EUs for the following two cases: (i) with policy selection, each EU selects between the two content caching policies to maximize the EU cache hit probability; (ii) without policy selection, only a PB content caching policy is implemented for EUs. The reduction of average download delay by the proposed policy selection is verified by simulation.

1.3 Thesis Outline

The rest of the thesis is organized as follows.

Chapter 2 introduces and defines C-RAN and F-RAN. We review the literature of transmission modes, including D2D communication and cooperative beamforming. To study the F-RAN system, this chapter introduces edge caching in F-AP and the spatial distribution of user equipment. The literature of performance metrics for F-RAN is reviewed, including cache hit probability and content download delay.

In Chapter 3, the greedy algorithm and transmission mode selection method are proposed to maximize the average energy efficiency per request in an EU assisted F-RAN. Analysing the power consumption and energy efficiency of three transmission modes: F-AP mode, D2D mode, cloud mode. When the requested content of the RU is in the local caching space of the EUs and the F-AP, this RU is defined as the MOU. Besides, according to the location of RUs, energy efficiency expressions for different transmission modes and the greedy algorithm, the greedy algorithm and mode selection method are constructed . In the proposed mode selection method, the D2D user is first assigned to the MOU farthest from the F-AP. Energy consumption and energy efficiency will be considered in the selection of content provider. Average energy efficiency with mode selection and without mode selection is simulated and compared.

In chapter 4, D2D clustered cooperative beamforming is considered in F-RAN with timer-based relay selection. The average energy efficiency, cache hit probability, and mode selection probability are analysed and studied. Analysing the performance metrics for three transmission modes: F-AP, D2D, and D2D clustered cooperative beamforming. Location-based mode selection is considered in selecting the transmission mode for RUs. Sub-popularity-based content caching policy for CEUs is proposed and compared with probability-based content caching policy. The impact of the Zipf exponent, density of DEUs, and threshold service distance for cooperative beamforming are studied for caching policy selection to optimize the average energy efficiency.

In chapter 5, to mitigate the impact of limited storage at each EU, an EU classification-based (ECU) content caching policy is proposed for F-RANs, where EUs are divided into two groups, each caching a different content set. This cahapter investigate how to change the ratio between the two groups to optimize their cache hit probability. To maximize the cache hit probability at EUs, a caching policy selection algorithm is presented which allows the fog radio access point to choose between the proposed ECU content caching policy and the conventional probability-based caching policy. By modeling the content request queue at each EU as an independent M/D/1 queue model, the average download delay under the proposed caching policy selection algorithm is analysed. The simulation results show that the proposed algorithm can significantly increase the cache hit probability for EUs and reduce the average download delay for RUs.

In Chapter 6, the conclusion of this thesis is provide. Some possible researches in the future are discussed.

1.4 List of Publication

Chapter 5: H.Hua, and X.Chu, “Content Caching Policy with Edge Caching User Classification in Fog Radio Access Networks,” *2021 IEEE Wireless Communications and Networking Conference (WCNC).* pp.1392 - 1398, Mar. 2021

Chapter 2

Literature Review

2.1 C-RAN and F-RAN

2.1.1 Definition of Cloud Radio Access Networks

To achieve the desired higher system capacity and optimal energy efficiency, an emerging technology—cloud radio access networks (C-RANs) has been proposed [5] [3]. C-RAN is a promising network technology to optimize the energy efficiency, network capacity[35]. Using the computational capabilities of the cloud computing platform, C-RANs can realize the extension of network coverage and increased network capacity [6]. Different from common cellular network systems, for abundant use of the functions of centralized large-scale cloud computing, baseband processing in C-RANs—e.g., wireless signal processing and resource allocation—is centralized in the baseband processing unit (BBU) pool, as Fig. 2.1 shows. Remote radio heads (RRHs) located near user equipment only implement radio frequency functionality [2]. In the C-RAN, the RRHs are connected to the cloud center through fronthaul links. The fronthaul link is used to connect the remote radio heads and centralize baseband processing unit pool. Hence, C-RAN can apply joint signal processing and data management[36]. In other words, a C-RAN separates the capability of a traditional base station into a BBU pool and RRHs. By making use of large-scale coordinated signal processing, the interference between RRHs can be fixed. However, a C-RAN still has problems such as fronthaul capacity limitations, processing latency in the BBU pool, and separation between the control surface and service plane. Long time delays and heavy burdens on the BBU pool affect the properties of C-RANs [7].

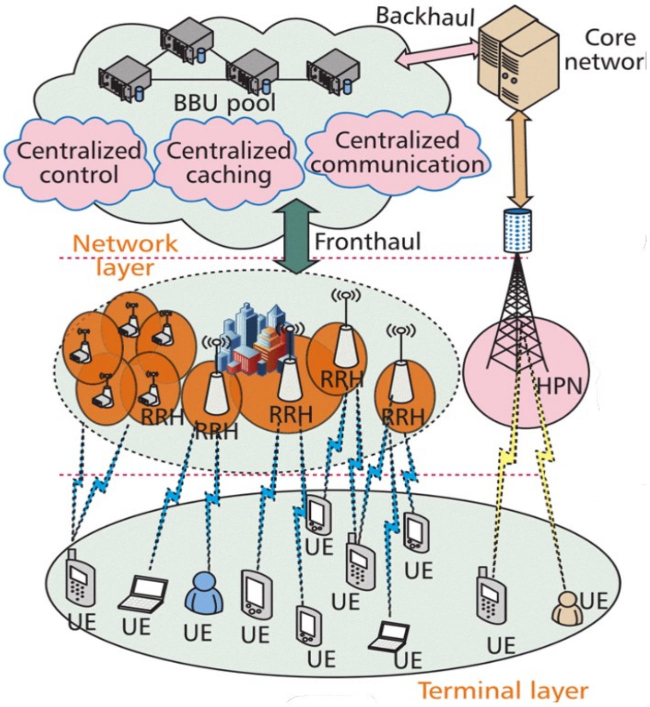

Figure 2.1: Cloud radio system networks [6]

2.1.2 Definition of Fog Radio Access Networks

To overcome the above disadvantages in a C-RAN system, fog radio access networks (F-RANs) have been proposed, which combine the emerging technologies of fog computing and edge caching [10]. In F-RAN, to alleviate the heavy burdens of cloud computing, while taking full advantage of a heterogeneous network (HetNet) and C-RAN, a centralized control function is segregated from the BBU pool to the existing high-power node (HPN). As Fig. 2.2 shows, an HPN connects with the BBU pool to provide the centralized cloud controller in the F-RAN. The key difference between F-RAN and C-RAN is that fog computing and edge caching are enabled in F-RAN. Fog computing was first proposed by Cisco [9] to achieve high performance at the network edge. It fully exploits the communication, control, computing, storage, measurement, management, and configuration of the network edge device, which is much closer to users than the BBU pool [45]. Edge caching devices are also discussed in the study of F-RANs. Different from C-RAN, in an F-RAN, RRHs and UEs are allowed to realize more functions, such as cooperative radio resource management (CRRM), distributed storage, and collaboration radio signal processing (CRSP). As a result of these new RRHs, which are equipped with storage, CRRM and CRSP, they are referred to as F-APs. Similarly, transitional user devices with a storage function are denoted as EUs. F-AP and EUs are called edge caching devices. Users in an F-RAN do not have to download data packets by connecting to the centralized storage in the BBU pool when the requested content can be provided from an adjacent F-AP or EU [46]. As edge caching devices are equipped, the edge caching process in F-RAN can be operated in two phases: the pre-fetching phase and delivery phases [14]citePengX[41]. Pre-fetching phase is executed during the off-peak traffic periods to storage the popular contents in edge caching devices. The delivery phase is characterized by transmitting the content requests to users [44]. Since pre-fetching the most popularity requested contents to the F-AP and EUs during off-peak network traffic periods is allowed to perform in F-RAN, the traffic load at fronthaul can be reduced by applying the F-AP and EUs during the peak traffic periods [37]. Also the requested content delivery latency can be obtained. Usage of an edge caching device causes an F-RAN to have different efficiency analysis compared with cloud radio access networks. In current research, different from traditional cellular networks and C-RAN, the investigation of fog radio access networks focus more on optimizing edge caching [47], waiting delay [48], latency ratio, cache hit probability, ergodic and content delivery rates by leveraging BBU pool and edge caching devices [39][41][42][43]. The consumption of fronthaul and the link between F-AP and EUs need to be taken into account when analyzing energy efficiency and spectrum efficiency in F-RAN. In conclusion, F-RAN takes advantage of fog computing and edge caching to reduce front haul traffic load and improve network performance compared to C-RAN.

2.1.3 Contents and Edge Caching in F-RAN

Currently, at present, the network data flow is mainly a request for multimedia content, such as video streaming on demand and media push [33][34]. The main feature of this type of data flow is asynchronous requests for pre-recorded contents, such as movies or user-generated content[40]. Besides, these contents are usually cached in advance, and most of the download requests for these contents are for contents with high popularity. According to these characteristics,

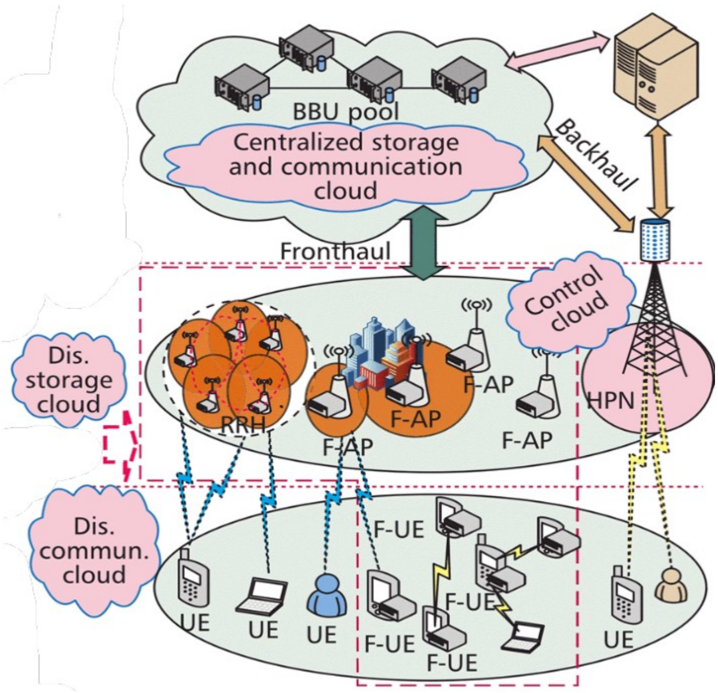

Figure 2.2: Fog radio system networks [6]

popular contents can be cached closer to the users to reduce the network transmission traffic and download delay of content delivery. Therefore, the research of pushing content caching to the edge of wireless network has attracted more and more attention in recent years[39][41][40]. During the off peak traffic periods, by caching popular contents at macro base stations, small base stations, or user equipments, the network performance can be optimized[40]. In the F-RAN, the popular contents are caching in edge caching devices like F-AP and EUs in general. Edge caching can effectively improve spectrum efficiency for the F-RAN, energy efficiency, and traffic latency. Functions such as front radio frequency, local distributed collaboration radio signal processing, and cooperative resource management are endowed in the F-AP. In the investigation of edge caching, the mode selection of content providing service and the caching strategy of pre-caching are the emphases of the research.

Since the edge caching process of F-RAN usually consists of a pre-cache phase and content transfer phase, the optimization of the F-RAN is mainly through optimizing the transmission mode selection of content requested users and the cache strategy of edge caching devices [24]. In the case that F-AP and edge caching users are used for content caching at the same time, the transmission mode selection of content transfer is an important optimization direction based on the different contents storage and transmission characteristic. On the other hand, in the cache-aided system, a pre-fetching phase must be completed before implementing the edge caching process [79]. The edge caching phase can be used to pre-fetch the proper files for an edge caching device through different strategies. Popularity-aware, probability, and random caching strategies have been considered [24]. Numerical results indicate that popularity-aware caching performs most effectively and uses less backhaul capacity. Social-awareness edge caching in fog radio access networks has also been discussed [37]. The scenario considered the impact of social relationships on the performance of an edge caching scheme, focusing on bandwidth consumption in content diffusion. The results showed that it was feasible to reduce bandwidth consumption. Pre-fetching for F-APs and edge caching users are limited by their local caching storage. Users are mostly interested in downloading the most popular content. It was presumed that a potential D2D transmitter and F-APs only cache the most popular content in their storage [24].

A standard Zipf law has been used to label the popularity of content. Zipf's law can be expressed such that in the natural language corpus, the frequencies of a word are inversely proportional to the frequency table ranking. So, the highest word frequency is about twice that of the second word, and the second frequency is four times the frequency of 2 times the word. This law is considered the reference point of all things related to probability distributions. The probability derived by Zipf's distribution can be used to explain the probability of requested content for a common user [46], and of content cached in storage [24].

In conclusion, this thesis will focus on the study of F-RAN system and edge caching. Besides, the algorithm of transmission mode selection and content caching policy in pre-fetching phase will be considered as the optimization proposal.

2.2 D2D Communication in F-RANs

2.2.1 D2D Communications

With the introduction of smart mobile devices and the rapid development of communication technology, increasing amounts of mobile equipment are limited by the spectrum resources of network systems. A promising technology, device-to-device (D2D) communication, addresses the problem of tremendous bandwidth demand. Within the cellular network system, D2D communications allow two close users to communicate directly without passing through a base station [49]. Generally, D2D communication takes place in a D2D pair, which consists of a D2D transmitter and a D2D receiver. Compared to common cellular communication, D2D is more stable and can provide higher data rates, lower power consumption, and lower time delay by reusing available cellular resources. In addition, sharing the signal directly by reusing the resource block for physically close users can lower overall interference, and the cellular network capacity and spectral efficiency can be optimized. While there is the effective utilization of the spectrum resources and extended network service coverage, this technological innovation generates new challenges. Due to the sharing of the spectrum with cellular users, interference between D2D pairs and the cellular system is inescapable [50].

2.2.2 D2D communication assisted F-RANs

Since D2D communications allow direct propagation between a D2D transmitter and receiver and can be implemented by exploiting edge caching user-assisted transmission, for alleviating the traffic load on the fronthaul and optimizing the transmit delay and make the best of edge caching. As a network edge communication technology that supports content delivery and sharing, D2D communication has been introduced to the C-RAN and F-RAN [8]. D2D communication is obviously a feasible scenario for the improvement of energy and spectrum efficiency. However, the social relationship is usually ignored and seldom considered in the optimal plan. Mobile users prefer to transmit to their friends, which is quite different from other content caching strategies. Compared to traditional D2D communication, social-awareness D2D communication could be grouped based on locations, interests, and backgrounds [5]. A social-awareness resource allocation optimal for D2D communication has been introduced [1], which shows that game theory is extensively used in discussing this topic. The social effect of F-RANs regarding energy and spectrum efficiency has not been much discussed previously. This thesis will optimize the content downloading service of F-RAN system in terms of cache policy based on content popularity.

In [51], the author aims at maximizing the total throughput in D2D communication assisted F-RAN by focusing on the stochastic optimization problem of resource allocation. The author further decomposes the optimization problem into three problem: mode selection, uplink beamforming design, and power control. Since the D2D communication is considered , the binary mode selection between C-RAN mode and D2D mode is studied in [51] where the C-RAN mode is a transmission mode that provides the request service to the user from the cloud via front haul.The proposed algorithm in [51] can success optimize the throughput and implies that F-RAN with D2D communication can provide low latency and high throughput. When studying the resource allocation problem in the F-RAN, [51] considers the dynamic traffic arrivals, the request delay of D2D communication and time-varying channel conditions. In the analysis of D2D transmission request delay, [51] only studies queuing delay in transmission process based on the relationship among queueing delay, the arrival rate and the transmission rate where the transmission rate is applied to the average throughput. According to the study of [51], D2D communication is a proofed technology for F-RAN. Also, the queueing delay is proposed to study in this thesis when D2D communication is applied in F-RAN. When discussing transmission delay, in the following research, is thesis not only consider the queuing delay between D2D pairs but also the time consumed by mode selection. Meanwhile, it can be seen from [51] that mode selection is an effective means to optimize F-RAN performance. As mentioned, in D2D communication assisted F-RAN, the request service can be applied by tree transmission modes: cloud transmission mode, F-A= mode and D2D mode. Since [51] dose not consider the F-AP transmission mode in mode selection, this thesis is proposed to continue study the transmission mode selection include F-AP transmission mode.

Presume that both the D2D transmitters (EUs) and the F-AP have cached the requested content of target UE. Different D2D pairs can use the same subchannel even in different cells, so interference from other cells' users should be considered [52] [51]. Consequently intra-interference from adjacent D2D transmitters in the same and different cells should be included in the calculation. Similarly, a D2D receiver will suffer from interference caused by cellular users from the same or different cells. While utilizing edge caching devices, a user requesting content can access EUs, RRHs, and F-APs based on different conditions and requirements. Various mode selection strategies lead to different results. If all users move at a low speed, D2D or relay-assisted mode can be used when D2D users are adjacent to users and the needed content is cached in the D2D transmitter. If no viable D2D user is around the UE, the nearest FAP mode can be used. If multiple F-APs are adjacent to the UE and can provide the desired content, then local distributed coordination mode is best. Otherwise, the traditional global C-RAN mode is feasible. For high energy efficiency, spectrum efficiency, and lower latency, the best mode choice is D2D or F-AP [37].

The mode selection for users in F-RAN is also investigated in [14]. In [14], researchers paid more attention to the cache hit probability and ergodic rate, which relate to the performance of edge caching. Different from [51], this paper considers both F-AP mode and D2D mode, and does not study the case of cloud transmission. A dynamic user access mode selection is proposed with evolutionary game for F-RAN. According to the evolutionary game, the active user is act as a player in the game. The payoff of a player denotes the performance satisfaction level. For choosing different transmission modes, the payoff of one player can b,e calculated based on the ergodic rate. In [14], author aims at optimization the average payoff in F-RAN system. Meanwhile, for investigate the cache hit probability in F-RAN, the coverage probability of D2D and F-AP are characterzed in [14]. Cache hit probability is one of the most important parameter in F-RAN study which means the probability that the request of one user can be served. As different transmission modes correspond to different cache hit probabilities, this parameter is commonly used in the study of transmission mode selection. In order to study the problem of mode selection, cache hit probability is continue investigated in this thesis.

When analyzing D2D communication, attention should also be paid to the trade-off between energy efficiency and spectral efficiency. [38] analyzed the trade-off between spectrum efficiency and energy efficiency by investigating optimal caching policy in Device-to-Device networks. Considering the influence of base station transmission, D2D-cache and self-cache, as well as the influence of cooperative distance, a clustering method with a specially designed power control and resource reuse strategy is proposed. In the analysis, [38] divides the situation into two categories: throughput-based design and energy efficiency-based design. By comparing throughput and energy efficiency evaluation, [38] shows that the optimal cooperation distance between them is different. These two categories conflict with each other. Therefore, there is a tradeoff between throughput and EE when choosing different cooperation distances.

In conclusion, utilizing edge caching user and D2D transmission in F-RAN can effectively achieve higher energy efficiency and lower latency. Also, mode selection is an effective means to optimize F-RAN performance when D2D transmission is considered. According to the literature, energy efficiency, content download delay and cache hit probability are important optimization parameters in F-RAN research. Therefore, this thesis will optimize the content downloading service in F-RAN from the direction of building the transfer mode selection algorithm. Moreover, the energy efficient, download delay and cache hit ratio will be studied and analyzed. The optimization direction of the literature rarely focus on the content caching policy at edge caching device, so we will focus on this to optimize the F-RAN system.

2.2.3 Spatial Distribution of User Equipment

To combine D2D communication and mobile edge caching in a fog radio access network system, the distribution of edge caching equipment should be considered. The performance difference between D2D caching and small cell caching was studied [46]. The distribution of small base stations and user equipment is modeled according to a Poisson point process (PPP) distribution in a two-dimensional phase. The Poisson distribution is a commonly used discrete probability distribution. Fog radio access networks were considered, in which fog access points and user equipment were also deployed based on a PPP distribution [24]. The Poisson process is used to describe the probability distribution of a number of random events occurring per unit of time or space [46][24]. User equipment categorization was based on the content request. A probability p $(0, 1]$ was set to explain whether a user makes a random request. To categorize user equipment, performance was analyzed based on the potential gain captured by the cache hit probability [46], i.e., the probability that a common user can find a requested content in the local cache. Then, based on the SNR and interference ratio, the success probability of cache-assisted transmission was discussed. Finally, power consumption was explained by D2D caching and small cell caching according to the average number of requests that can be transmitted successfully and handled simultaneously per unit area.

2.3 Cooperative Beamforming

When wireless nodes cooperate with each other, cooperative communication can be realized in a wireless network system to forward information and optimize total energy efficiency and overall throughput [53] [54]. This communication technology can be used to save energy costs in data transmission from relay nodes to destinations [55] [56] [57]. Therefore, under the proper conditions, it requires less overhead cost to broadcast information from a source to multi-relays, which forward it to its destination. A spectrum-sharing system based on co‘gnitive radio has been proposed [58] [59], using zero focusing beamforming, which is defined as spatial signal processing in wireless devices with multiple antennas. Two sources can share the spectrum with the primary users in this system, and interference will be generated during data transmission.

Due to the limited transmission power of each relay node, which is battery-powered, the optimal relays should be selected with the required channel stat information (CSI), and the overall energy consumption and energy efficiency are a focus of research. To estimate the energy consumption of obtaining the desired CSI to implement relay cooperative beamforming is a challenge in energy efficiency analysis. Previous work has ignored the cost of obtaining CSI. [60] discussed energy saving in cooperative beamforming with an amplifier and forward relays, and did not consider the overhead of acquiring CSI. It still be neglected in [56]. In [56], the ideal outage probability in relay is also be applied. In selection relaying, if the signal-to-noise ratio (SNR) of the relay’s source-relay channel exceeds a threshold, it will be seemed as the candidate selected relay [61]. Cooperative transmission established by relay nodes can be optimized by their proper selection [62]. Recently, best relay selection has been a subject of intense study, as it could be an effective approach for managing the overhead costs of cooperative transmission [63][66]. [67] and [68] proposed a distributed space-time coding method to perform cooperative beamforming. Authors in [69] considered optimized overall consumption in cooperative beamforming with simple decode-and-forward relay selection, and analyzed single optimal relay selection and successful decoding relay selection. For a single optimal relay selection scheme, the best relay will be selected from candidate relays that can correctly decode the information from the destination. A successful decoding relay selection scheme will select all of the candidate relays to execute relay cooperative beamforming. In the current literature, authors always select the most optimal relay, or all available relays, to compose cooperative beamforming transmission [67][70]. Authors in [71] proposed a time-based relay selection scheme with low power consumption. Another method used a location-aware cluster to obtain the required CSI [72]. This thesis considers a suitable number of relays to obtain higher energy efficiency.

2.3.1 D2D Clustering

As the number of mobile devices rises, the D2D cluster is allowed to be considered in implementing cooperative beamforming to improve the performance of the network system. A cluster in a D2D system is the aggregation of nodes into groups [73]. By forming virtual arrays, D2D clusters can cooperatively forward information from source to destination. D2D clusters have been discussed [74][75]. User-cooperation schemes focus on minimizing the average power consumption for user equipment [74]. In this thesis, a user located in a D2D cluster is denoted as a relay node. The uplink transmission strategy is considered in the frame structure of user cooperation to analyze D2D communication. To set a cluster, the user equipment should first complete the process of device discovery. After link setup [76], the cluster in D2D communication is built, with the base station noted as the destination. As usual, each cluster will select a head user. In this case, the head of the cluster has already been selected according to the power consumption calculation. A number of nodes from the same cluster are selected and assigned to help the source to propagate the information to a certain destination. As cellular channel reuse is encouraged in the D2D communication system, relay selection in D2D networks is quite different from that in cooperative networks [77][78]. Based on the definition of a D2D cluster, the D2D multi-hop cooperative network was realized in D2D networks [75]. A multi-hop mechanism is considered to forward the source message to the destination via cooperative transmission. Base station-assisted clustering techniques are discussed, where the base station groups successful decoding nodes into a cluster. In particular, the idea is given that if the destination is still inaccessible after using the first cluster, a second cluster will be built based on it [75]. Although the cluster made by the base station can transmit information forward, to improve network performance, two relay selection schemes, called random and SNR-based relay selection, were proposed. The base station's effect and interference were ignored in the subsequent calculation and analysis. In conclusion, D2D cluster can provide cooperative beamforming and can provide larger service range than D2D communication. Therefore, this thesis will analysis and study the D2D cluster assisted cooperative beamforming in content pre-caching and content transmission service. Also, optimize the average energy efficiency of content downloading service through transmission mode optimization will be analyzed.

2.4 Performance in F-RAN

2.4.1 Cache Hit Probability

The cache hit probability is one of the most common measures of content placement performance in edge caching analysis. This is the probability that a random content request can be fulfilled by an edge caching device that already has the content cached. A reasonable content caching policy for an edge caching device can effectively improve the cache hit probability. This is called the cache hit rate in some research. An unknown spatiotemporal content popularity and user preference were considered in an F-RAN system [18]. A Q-learning method is considered in this thesis to seek the optimal caching policy to improve the cache hit probability. Based on an ant colony algorithm, a centered user equipment selection scheme was proposed to find optimized user equipment to cache data and improve the delivery delay and cache hit rate [37].

2.4.2 Content Download Delay and Queue Model

Optimizing the content download delay is also a challenge in F-RANs. When the user equipment has an edge caching function, content download delay optimization is always discussed. The analysis of content caching was based on an F-RAN system that supports D2D communication, where both user equipment and F-APs can cache content [37]. Thus, two types of edge caching equipment are discussed in this thesis. To reduce the content download delay and optimize the content caching rate, the author [37] focused on the most appropriate user equipment selection for content caching, and proposed an ant colony optimization scheme to find the most reasonable user to cache the most requested content.

Queuing of requests at edge caching equipment must also be discussed when analyzing download delay. An M/G/1 queue system was considered regarding the request queue problem at a small base station to optimize the average download delay [80].

To identify the probability distribution assumed for service (and interarrival) times, a queueing model is conventionally labeled with the distribution of service times, number of servers, and distribution of interarrival times. The symbols used for the possible distributions are:

1) M = negative exponential distribution;

2) D = degenerate distribution (constant times);

3) G = general service-time distribution.

Depending on the case, different queue systems can be considered. A basic queue system consists of customers, a service facility, and a queue. In the F-AN system, customers can be seen as content requests, and the service facility is edge caching equipment. The time between consecutive costumer arrivals to a queueing system is called the interarrival time. The expected number of arrivals per unit time, i.e., the mean arrival rate λ, is one of the most important parameters in queue system analysis. When a customer enters the service, the elapsed time from beginning to end is referred to as the service time. Generally, the queue system has no control over when customers arrive. Random arrivals imply completely unpredictable interarrival times. The only probability distribution with this property is the negative exponential distribution, which describes the probability of a time interval for an event with a fixed value and the average number of occurrences per unit time. The latter has a Poisson distribution. Therefore, a Poisson process can be used to express the request arrival process [80]. In conclusion, to analyze a request queue system in an F-RAN system requires an examination of the types of queuing models. Then, based on the arrival rate and mean service time, the download delay can be further analyzed.

Chapter 3

Energy Efficiency Maximization by Greedy Algorithm and Mode Selection in F-RAN

F-RAN is implemented as an effective system to provide high energy efficiency. Edge caching is considered promising for the downloading of content to users through direct D2D communication. Edge caching user-assisted F-RAN can provide more transmission modes to download content.

We discuss the energy-efficiency optimization of edge caching user-assisted F-RAN with the greedy algorithm and transmission mode selection. F-AP can provide service to users requesting content according to the proposed mode selection method.

The chapter is organized as follows. In section 3.1, the network construction of F-RAN with three transmission modes is described, and the pre-fetching phase is introduced. Section 3.1.2 analyzes the power consumption and energy efficiency of the F-AP, D2D, and cloud modes. Greedy algorithm and mode selection are described in section 3.2. Section 3.3 presents simulation results.

3.1 F-RAN System Model

Fig. 3.1 shows the F-RAN system model, in which the cell tier includes one fog radio access point (F-AP) at the center of the system, and the user tier includes M_R content-requesting users (RUs) and M_D D2D transmitters. Letting $\mathcal{M}_r = \{1, 2, 3, ..., M_R\}$ and $\mathcal{M}_d = \{1, 2, 3, ..., M_D\}$ denote the sets of RUs and D2D transmitters, respectively, with edge-caching capacity, each D2D transmitter is allowed to provide the requested content directly. In addition, each user, whether a D2D transmitter or RU, is assumed to be a single antenna configuration, which means that it cannot send and receive information at the same time. In the proposed system, for effective and reliable content transmission, a user making a content request must first send it to a servicing base station, which is the F-AP. In Fig. 3.1, the blue route shows that the cloud center can provide the desired content to the corresponding user. If requested content can be found in local storage by edge-caching devices such as F-APs and adjacent D2D transmitters, the content will not have to be retrieved from the centralized content library in the cloud service center. It is considered that all D2D transmitters reuse the uplink spectrum resource along with other cellular users that are not requesting content. To simplify the system model, it is assumed that when D2D transmissions reuse the resource of one sub-channel, each sub-channel is assigned to at most one D2D transmission link. It is further assumed that the channel-state information (CSI) of both the uplink and downlink is the same.

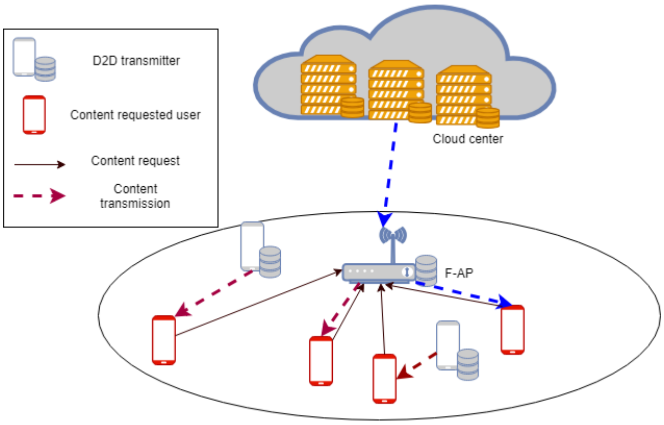

Figure 3.1: System model

The channel power gain g_{ij} between transmitter i and receiver j can be expressed as

$$g_{ij} = h_{ij} P_L(d_{ij}) = h_{ij} PL^{-1} d_{ij}^{-\alpha}, \tag{3.1}$$

where d_{ij} is the distance between the transmitter and receiver, and $P_L(d_{ij}) = PL^{-1} d_{ij}^{-\alpha}$ is the path loss, which consists of a path-loss exponent α and path-loss constant PL^{-1}.

3.1.1 Pre-fetching and Content Request Probability

Since network data streams are mainly requests for multimedia content, such as video streaming on demand, media push, etc., it can be pre-cached according to popular contents of this kind of data [33][34][40]. The requested content can be served by EUs via two transmission modes: D2D clustered cooperative beamforming transmission and direct D2D transmission. In the pre-fetching phase, the F-AP caches the content from the cloud center to its caching space through a fronthaul link. Setting D_c as the amount of content in the centralized cloud server, the content library of the cloud center is given as $F_C = \{f_1, f_2, f_3 ..., f_{D_c}\}$. The content request probability has a Zipf distribution, $p_x = \frac{1/x^{\beta}}{\sum_{x=1}^{N} 1/x^{\beta}}$, where p_x is the probability that content f_x is requested by an RU, and β is the Zipf exponent that controls the degree to which requests are focused on popular content. Therefore, a higher value of β leads to a larger probability of requesting content f_1, which is the most popular content in set F_C.

Since the RUs are interested in popular content [16], the probability-based content caching policy is applied in this study for F-APs and D2D transmitters. It is assumed that an F-AP can cache content with limited caching space D_f. Thus, the content set in the F-AP can be expressed as $F_F = \{f_1, f_2, f_3 ..., f_{D_f}\}$. Furthermore, according to the probability-based content caching policy and storage size D_d of EUs, $F_D = \{f_1, f_2, f_3 ..., f_{D_d}\}$ can be given, which denotes the local content caching space of edge caching users. In general, the probability that one random request can be from F_F is $p_{CC}^{F} = \sum_{x=1}^{D_f} p_x$. Similarly, the probability that one requested item of content can be provided from content set F_D is $p_{CC}^{T} = \sum_{x=1}^{D_d} p_x$. Note that the total transmission power consumption consists of transmission power and circuit power consumption [64]. Accordingly, the circuit power at user is assumed as the same and is denoted as P_{UE}^{C}. Also the F-AP circuit power consumption is P_{FAP}^{C}. Assumed that both P_{UE}^{C} and P_{FAP}^{C} are constant.

3.1.2 Transmission Modes and Energy Consumption

Since the distance between RU i and an F-AP is d_{iF}, according to the Shannon theory, the data rate can be expressed as: $R_T = log_2(1+\frac{P_{iF}^{T} g_{iF}}{P_N})$, where $\frac{P_{iF}^{T} g_{iF}}{P_N}$ denotes the signal to noise rati between RUi and F-AP. Further, the total transmission power can be given as

$$P_{iF}^{T} = \frac{(2^{R_T}-1)P_N}{g_{iF}} + P_{UE}^{C}, \tag{3.2}$$

where P_N is the noise power; R_T is the target data rate, in symbols per second; and g_{iF} is the channel power gain between the F-AP and RU_i. It is assumed that the cloud can allocate channels perfectly, and a user and the F-AP already know the CSI between them. The interference and outage probability will not be described in this study. With the symbol length of the request N_R, the energy overhead of sending a request is $E_R = P_{iF}^{T} N_R T_s$, where T_s is the duration time of one symbol. After the request has been sent, the transmission method can be selected by the associated F-AP from the F-AP, D2D, and cloud modes.

F-AP Mode

If the requested content is available at the F-AP, then it transmits it directly to the content requester RU_i with transmission power

$$P_{Fi}^{T} = \frac{(2^{R_T}-1)P_N}{g_{iF}} + P_{FAP}^{C}. \tag{3.3}$$

It is assumed that each item of content has a size expressed with N_C symbols. Total energy consumption for the F-AP mode can be expressed as

$$E_{FAP} = P_{Fi}^{T}(N_R + N_C)T_s. \tag{3.4}$$

D2D Mode

If the requested content is not available at the F-AP and is available at some edge-caching users, then the F-AP will select one of these EUs to send the content to the RU through a D2D link. To implement direct D2D communication, the selected D2D transmitter should receive the request, which is forwarded by the F-AP with total transmission power

$$P_{FD}^{T} = \frac{(2^{R_T} - 1)P_N}{g_{FD}} + P_{FAP}^{C}, \tag{3.5}$$

where g_{FD} is denotes the channel power gain between the F-AP and selected D2D transmitter.

After the target D2D transmitter has received the request from the F-AP, it sends N_T training symbols to the RU with total transmission power

$$P_{Di}^{Tr} = \frac{(1 - 2^{R_T})P_N}{\overline{g_{Di}}ln(1 - p_{out})} + P_{UE}^{C}, \tag{3.6}$$

where $\overline{g_{Di}}$ is the mean channel power gain, which is $1/PL_{d_{Di}}$; and p_{out} is the maximum allowed outage probability. Using the received pilot symbols, the content requester performs channel estimation and sends the estimated CSI using N_F symbols back to the D2D transmitter with total transmission power

$$P_{iD}^{F} = \frac{(2^{R_T} - 1)P_N}{g_{Di}} + P_{UE}^{C}. \tag{3.7}$$

In the content-transmission phase, with power $P_{iD}^{T} = P_{iD}^{F}$, the D2D transmitter sends the requested content to the RU with power $\dfrac{(2^{R_T} - 1)P_N}{g_{Di}}$. According to the previous calculation, the energy consumption to transmit the requested content can finally be obtained as

$$E_{D2D} = T_s(P_{FD}^{T}N_R + P_{Di}^{Tr}N_T + P_{iD}^{F}N_F + P_{iD}^{T}N_C). \tag{3.8}$$

Cloud Mode

It is assumed that any requested content is available at the cloud server. The power consumption at the cloud mainly consists of P_{Ct}. Since the F-AP already knows the CSI for the user requesting content, the energy consumption of the cloud mode is given as

$$E_{Cloud} = T_s(P_{Cf} + P_{Ct} + P_{FD}^T N_C). \tag{3.9}$$

To simplify the calculation, it is further assumed that the capacity of the fronthual is always sufficient.

3.2 Analysis of Energy Efficiency and Mode Selection Method

3.2.1 Greedy Algorithm with Mode Selection

According to the discussion in the section 3.1, a requested content can be served by the F-AP, D2D transmitter, or cloud center. In this section, there are two parts in the process of transmission mode selection. In the first part, the RUs who can only be served by the cloud center and the F-AP will be selected and assigned corresponding content providing devices. For each RUs, the request task should first be sent to the F-AP to categorize the transmission mode of the RU. Denote the content requirement of RU x as f_x where $x \in \mathcal{M}_r$. When the requirement of RU x meets the condition: $f_x \cup F_F = \emptyset$, this requested content can only be provided by the cloud center. Hence, the cloud mode is selected as the transmission mode. Therefore, the energy consumption can be calculated according to the equation 3.9. Since the content sets in F-AP and the D2D transmitter are F_F and F_D separately, if $f_x \cup F_D = \emptyset$ and $f_x \in F_F$, the F-AP is selected to send the requested content directly to the RU x with energy consumption $E_C = E_{FAP}$. In addition to RUs who can be sure that only the F-AP or cloud server can provide the request contents, there are RUs who need to make transmission mode selection. It is easily seen that if the requested content f_x satisfies $f_x \in F_D$, then both the F-AP and D2D transmitter can be considered to provide the requested content. Such content requested users are called mode optional users (MOUs). All users who cannot be served by F-AP and edge caching users will have their content provided by the cloud center. Therefore, according to the requested contents, the user can be divided into two parts: cloud service and F-AP or D2D transmission. Therefore, according to the requested contents, RUs can be divided into two parts: Rus only served by the cloud service, Rus can choose F-AP mode or D2D transmission mode .We focus on learning the transfer mode selection related to F-AP mode and D2D mode in this study

In second part, the transmission model selection will be discussed. Since each D2D transmitter can only serve one user requesting content, we investigate the problem of joint mode selection and D2D transmitter assignment, with the objective to increase the average energy efficiency for all users requesting content. Since only the MOUs can select the transmission mode, it is proposed to optimize the mode selection for MOUs, which can be ranked in descending order of distance to the F-AP. In this case, the greedy algorithm is considered to meet the energy efficiency optimization. The greedy algorithm refers to the optimal value obtained through several times of greedy selection from the initial state of the problem. In this study, the selection of transmission mode for each MOUs was regarded as a greedy selection. It is assumed that there are M_S MOUs in total, and the F-AP knows the location of each user. According to $E_{FAP} = P_{Fi}^{T} T_s(N_R + N_C)$, when using F-AP transmission mode to send contents, the farther away the MOU is from F-AP, the more transmission energy consumption is cost. This also shows that the farther away the MOU is from F-AP, the more it needs to optimize the transmission energy by selecting D2D transmission mode. Hence, MOUs are ranked in descending order of distance to the F-AP. The set of ranked MOUs can be given as $\mathcal{M}_s = \{1, 2, 3..., M_S\}$, where $d_{1F} > d_{2F} > ... > d_{iF} > ... > d_{M_S F}$, where d_{iF} is the distance between MOU i and the F-AP.

Accordingly, the mode selection based on the greedy algorithm starts with MOU 1, which is farthest MOU from the F-AP. For MOU 1, based on the above instantaneous energy efficiency expression, the energy efficiency attained by selecting D2D transmitter j ($j \in \mathcal{M}_d$) for MOU 1 can be given by

$$EE_{1j}^{D2D} = \frac{N_C R_T}{E_{R1} + E_{D2D1j}}, \tag{3.10}$$

where $E_{R1} = P_{1F}^{T} N_R T_s$ is the energy required to send the request to the F-AP. According to the D2D transmission mode, the energy consumption $E_{D2D1j} = T_s(P_{Fj}^{T} N_R + P_{1j}^{Tr} N_T + P_{1j}^{F} N_F + P_{1j}^{T} N_C)$. On the other hand, the energy efficiency when the F-AP is chosen to provide the requested content to MOU_1 is

$$EE_{1}^{FAP} = \frac{N_C R_T}{E_{R1} + E_{FAP1}}, \tag{3.11}$$

where $E_{FAP1} = P_{F1}^{T} T_s (N_R + N_C)$ is the energy required to transmit content from the F-AP to MOU 1. By definition, each MOU can be served by at only one D2D transmitter. It is assumed that EE_{1j}^{D2D} has a maximum value when $j = j^*$ ($j^* \in \mathcal{M}_d$). D2D transmitter j^* can be selected as the content provider for MOU_1, if the in-equation $EE_{1j^*}^{D2D} > EE_{1}^{FAP}$ is satisfied. In addition, the D2D transmitter j^* will not be considered in the mode selection for remaining MOUs. Also, a D2D transmitter set $\mathcal{D} = \{DU_1, DU_2, DU_3, ..., DU_i\}$ is defined as the set of selected D2D transmitters for MOUs. In this case, for MOU 1, the selected D2D transmitter can be expressed as: $DU_1 = j^*$. Since each D2D transmitter can only serve one RU at a time, when D2D transmitter j^* for MOU 1 is selected, it will not be considered a potential content provider for the remaining RUs. Otherwise, $DU_1 = \varnothing$, which means no D2D transmitter is selected and the requested content is provided directly by the F-AP. Based on the greedy algorithm, the MOU farthest from the F-AP is assigned as the first D2D transmitter. After a content provider has been selected for MOU 1, the above steps are repeated to select a content provider for MOU 2, and then for MOU 3, and so on, until each MOU is allocated an optimized content provider that increases the total energy efficiency. This mode selection process is presented as Algorithm 1.

Algorithm 1 Greedy Algorithm with Mode Selection

```
Input:
    Set of MOUs, $\mathcal{M}_s = \{1, 2, 3..., M_S\}$;
    Set of D2D transmitters, $\mathcal{M}_d = \{1, 2, 3, ..., M_D\}$;
    Set of selected D2D transmitters $\mathcal{D} = \{DU_1, DU_2, DU_3, ..., DU_i\}$;
    MOU $i$ where $i \in \mathcal{M}_s$;
    D2D transmitter $j$ where $j \in \mathcal{M}_d$;
Output:
    Set of selected D2D transmitters for MOUs, $\mathcal{D}$;
    Mode selection for MOU $i$;
Set $i = 1$;
while $i \leq M_S$ do
    $(i, j^*) = \text{arg max } EE_{i,j}^{D2D}$;
    if $EE_{i,j^*}^{D2D} < EE_i^{FAP}$ then
        For MOU $i$, the F-AP transmission mode is selected;
        $DU_i = \varnothing$;
    else
        For MOU $i$, the D2D transmission mode is selected;
        $DU_i = \{j^*\}$;
        $\mathcal{M}_j = \mathcal{M}_d - \{j^*\}$;
    end if
end while
return $\mathcal{D}$
```

3.2.2 Greedy Algorithm Without Mode Selection

Different from the section 3.2.1, this scheme only allocates D2D transmitters to MOUs through greedy algorithm, and does not consider providing F-AP transmission mode for these MOUs. According to the explanation above, for MOUs, the greedy selection for D2D transmitters starts from MOU 1 which is the MOU farthest from the F-AP. Basing on the equation EE_{1j}^{D2D} where $j \in \mathcal{M}_d$, the D2D transmitter that provides the highest energy efficiency for MOU 1 can be selected. The greedy selection will then continue to assign D2D transmitters to MOU2, MOU3. Hence, either each user has been assigned a D2D transmitter or no more D2D transmitter can be assigned to MOUs, the greedy selection of D2D transmitter for MOUs is ended. The F-AP transmission mode is no longer taken into account when assigning a content provider for MOUs, unless there are not enough D2D transmitters to serve all MOUs. In the case of an insufficient number of D2D transmitters, the remaining MOUs that are not assigned to the content provider are directly served by the F-AP. This process is presented as Algorithm 2.

Algorithm 2 Greedy Algorithm Without Mode Selection

```
Input:
    Set of MOUs, $\mathcal{M}_s = \{1, 2, 3..., M_S\}$;
    Set of D2D transmitters, $\mathcal{M}_d = \{1, 2, 3, ..., M_D\}$;
    Set of selected D2D transmitters $\mathcal{D} = \{DU_1, DU_2, DU_3, ..., DU_i\}$;
    MOU $i$ where $i \in \mathcal{M}_s$;
    D2D transmitter $j$ where $j \in \mathcal{M}_d$;
Output:
    Set of selected D2D transmitters for MOUs, $\mathcal{D}$;
    Mode selection for MOUs;
Set $i = 1$;
while $a \leq M_S$ & $\mathcal{M}_d \neq \varnothing$ do
    $(i, j^*) = \text{arg max } EE_{i,j}^{D2D}$;
    $DU_i = \{j^*\}$;
    $\mathcal{M}_d = \mathcal{M}_d - \{j^*\}$;
end while
MOUs that has not been assigned D2D transmitters download the requested contents by
    F-AP transmission mode;
return $\mathcal{D}$
```

3.3 Simulation and Results

Table 3.1: Simulation parameters for F-RAN

Parameters	Value	Description
D_c	2000	Number of items of content in BBU pool
D_f	500	umber of items of content in F-AP node
D_d	20	Number of items of content in D2D Tx
B	15×10^3 Hz [81]	Bandwidth
N_0	$-174 dBm/Hz$[81]	Noise power spectral density
P_{out}^t	0.1 [81]	Outage probability
α	4 [81]	Path loss exponent
N_C	$1 symbol$[72]	Length of request task
N_F	$1 symbols$[72]	Length of feedback
N_T	$300 symbols$[72]	Length of training symbols
N_R	$300 symbols$[72]	Length of content
P_{UE}^C	$100 mw$[65]	Circuit power of user
P_{FAP}^C	$100 mw$[65]	Circuit power of F-AP

We present simulation results to illustrate the performance of mode selection with a greedy algorithm. The 16-QAM modulation with $R_T = 4(bits/symbol)$ is employed. Main system parameters are listed in Table 6.1 [81]. In Fig. 3.2, the number of content requested users and D2D transmitters have the same value, where $M_R = M_D = 10$ and the number of contents in the D2D transmitter is equal to 20. It is obvious that, as the value of β increases, the average

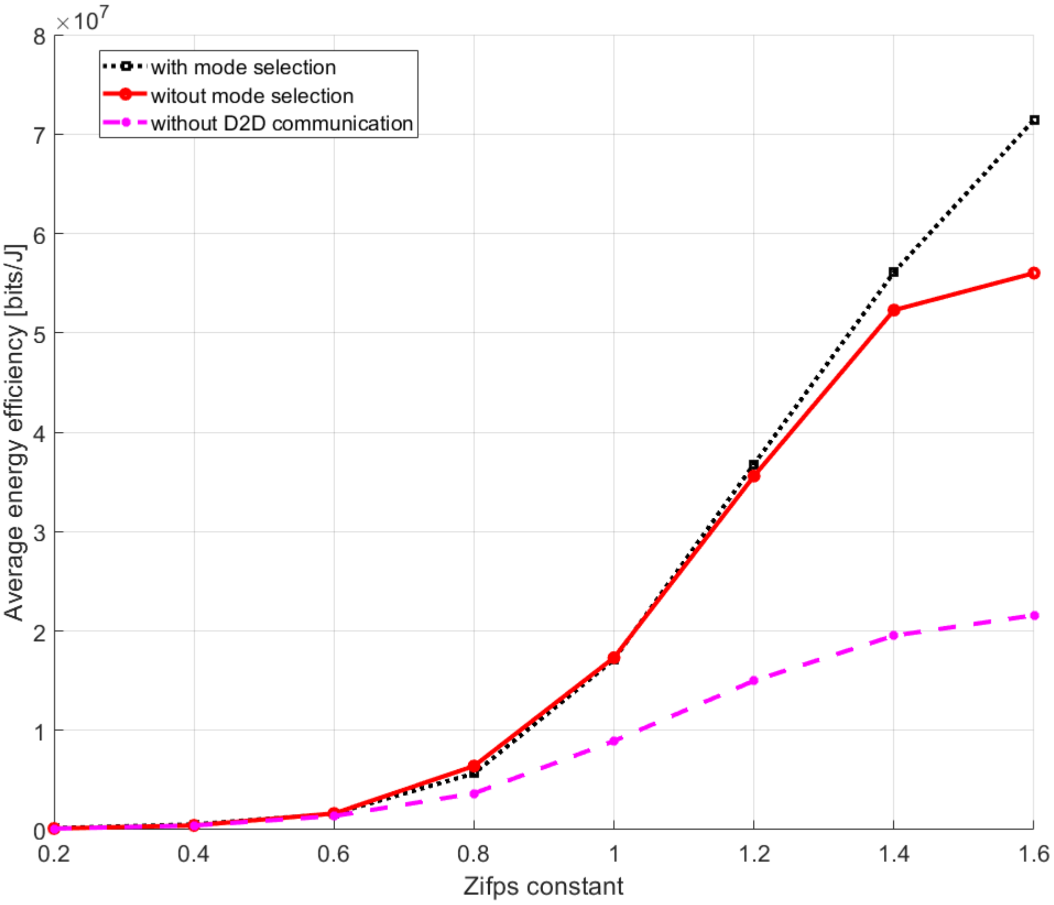

Figure 3.2: Energy efficiency of three different mode selection methods

energy efficiency using greedy mode selection is better optimized than without mode selection. Since a probability-based caching policy is used, to increase β increases the number of MOUs. As β increases, better average energy efficiency can be provided using mode selection. Using a mode selection method cannot only select RUs that are more suitable for the F-AP transmission mode, but can provide more D2D transmitter choices for the remaining RUs.

In Fig. 3.3, the average energy efficiency is plotted over different M_D values for $M_R = 20$ and β=0.7. With a varying number of D2D transmitters, i.e., from 5 to 30, the preponderance of greedy mode selection is increasingly clear. As mentioned previously, a lower β value provides a higher concentration of low-popularity contents. More different contents will be requested when β has a smaller value. Therefore, the higher β value creates more MOUs to be served by D2D transmitters. As mode selection can optimize the choice of content provider for each MOU, it

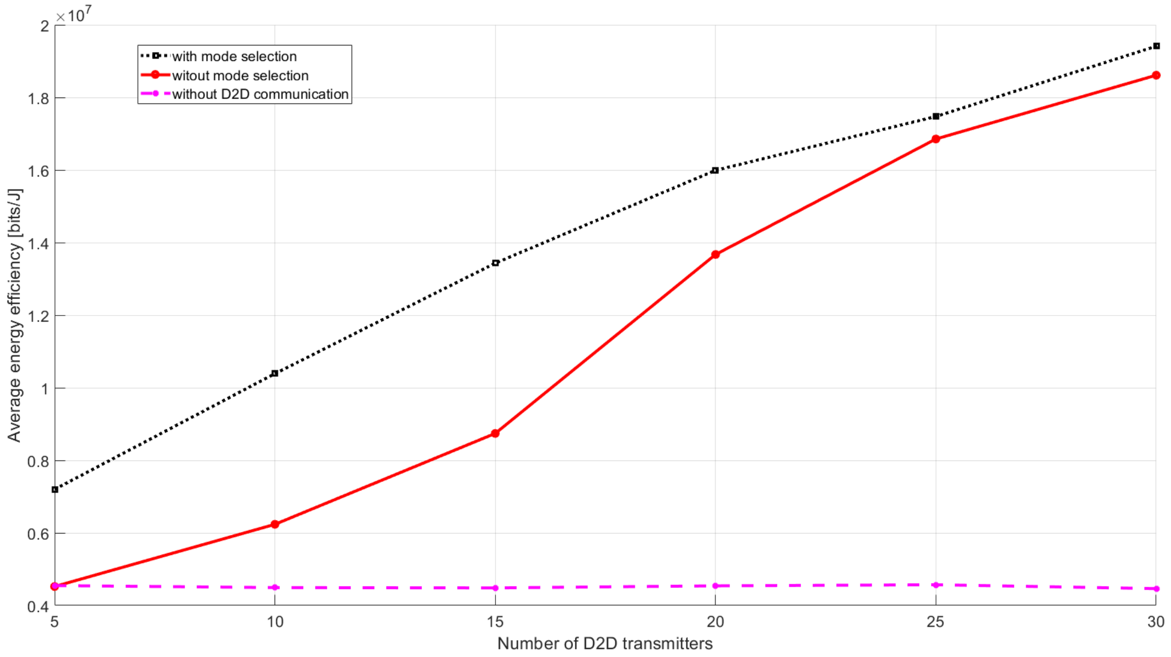

Figure 3.3: Energy efficiency with increasing number of D2D transmitters

gives better results than the two other methods. It can be seen that when the number of D2D transmitters is greater than or equal to 20, the gap between using a mode selection method and without using one becomes smaller. With increasing number of D2D transmitters, more D2D transmitters are able to provide direct transmission services to RUs in a short distance with higher energy efficiency than the F-AP. The distance between RUs and D2D transmitters are shortened and lower power consumption can be generated in D2D communication. Thus, for MOUs, there are fewer cases to for which the energy efficiency can optimized by selecting the F-AP transmission mode. Thus, fewer MOUs choose to use the F-AP transmission mode to optimize the energy efficiency, and the benefits of using mode selection is reduced.

In Fig. 3.3, the average energy efficiency is plotted over different M_D values for $M_R = 20$ and β=0.7. With the number of D2D transmitters varying from 5 to 30, the advancement of greedy algorithm and mode selection method is increasingly clear. As mentioned previously, a lower β provides a higher concentration of low-popularity content. More different contents will bez requested when β is smaller. Therefore, a higher β value results in more MOUs served by D2D transmitters. As mode selection can optimise the choice of content provider for each MOU, it gives better results than the two other methods. It can be seen that when the number of

D2D transmitters is greater than or equal to 20, the gap between using and not using a mode selection method decreases. With an increasing number of D2D transmitters, more of them can provide direct transmission service to RUs within a short distance with higher energy efficiency than the F-AP. The distances between RUs and D2D transmitters are decreased, and lower power consumption can be generated in D2D communication. Thus, for MOUs, there are fewer cases for which energy efficiency can be optimised by selecting the F-AP transmission mode, and so fewer MOUs choose that mode to optimise energy efficiency, and the benefits of using mode selection are reduced.

3.4 Conclusion

In this Chapter, we proposed a greedy algorithm and mode selection method for edge caching user assisted F-RAN in optimizing the average energy efficiency. The F-AP, D2D, and cloud transmission modes were discussed, and we analysed the power consumption and energy efficiency in fulfilling request tasks and transmitting content for these modes. According to the distance from users requesting content to the F-AP, the greedy algorithm and mode selection method were proposed to optimise energy efficiency. Simulation results showed that the proposed mode selection can improve energy efficiency. The impact of the Zipf exponent and number of D2D transmitters was also studied. With an increasing Zipf exponent value, the optimization effect of mode selection was more obvious. Moreover, limited by the number of users requesting content, the average energy efficiency of both using and not using the proposed mode selection with an increasing number of D2D transmitters gradually became small. In general,greedy algorithm, mode selection method and edge caching users can be considered to optimise the energy efficiency of F-RANs.

Chapter 4

Sub-popularity Caching Policy for D2D Cluster Assisted Fog-RAN

In this Chapter, we consider D2D clustered cooperative beamforming with timer-based relay selection in an F-RAN system. A sub-popularity content caching policy is proposed and compared with a probability-based caching policy. We propose to increase edge caching service coverage without increasing power consumption per request by exploiting D2D clustered cooperative beamforming, and discuss the edge caching performance of the cache hit probability and successful delivery probability.

In Section 4.1, We establish a F-RAN system model, where each type of devices is distributed following an independent homogeneous Poisson Point Process (PPP). The pre-fetching and content caching policy for different edge caching equipment is given. Besides, location based transmission mode selection method of F-AP and EUs is represented in this section. In Section 4.2, we defined three different transmission modes for cache requesting users which called D2D mode, D2D cluster mode and F-AP mode. Moreover, energy efficiency for these three transmission mode is analyzed. Section 4.3 analyzes the cache hit probability and mode selection probability for different transmission modes. The simulation results are represented in Section 4.4.

4.1 System Model

4.1.1 Network Model

In this network system, the traditional base station is divided into a base band unit (BBU) pool and remote radio heads (RRHs). The BBU pool is at the cloud center, and has the functions of centralised control, communication, and caching. An RRH can be regarded as a base station with a signal transmission function. In the Fog-RAN system, some functions can be distributed to several devices called fog nodes to the reduce traffic load and optimise content transmission. We consider two types of fog nodes: F-APs and EUs. An RRH with edge caching, signal processing, and resource manage functions is referred to as an F-AP. Similarly, cellular users (UEs) with these three functions are considered as EUs. In this paper, we assume that EUs are all idle users, so the power consumption caused by D2D transmission will have less impact on the battery usage of the corresponding users.

We consider an F-RAN system with a group of F-APs and several UEs in a disc plane ${\mathbb{D}_{\mathbb{R}}}^2$, as shown in Fig. 4.1. The cell tier includes RRHs and F-APs. The user tier includes RUs, EUs, and cellular users without requests. Content downloading can be categorised as either cloud- or fog node-assisted. As shown via the blue route in Fig. 4.1, using cloud-assisted content downloading service, the BBU pool transmits the requested content to the corresponding RU through one RRH. Otherwise, if the requested content can be found in local caching space by fog nodes such as F-APs, cooperative beamforming-allowed EUs, and common EUs, the requested content does not have to be received from the centralised cloud server. Fig. 4.3 shows the classification for fog nodes. Since we focus on analyzing the energy efficiency, cache hit probability, and mode selection probability of F-APs and EUs, the content downloading service provided by the cloud server through fronthaul and RRHs is not discussed in this study.

As shown in Fig. 4.4, EUs are separated into two types: cooperative beamforming allowed EUs (CEUs) and common EUs (DEU). We consider a D2D clustered cooperative beamforming transmission where the requested content is provided to the RU by forming a Multiple-Input Single-Output (MISO) virtual beamforming with the help of relays. In this research, the CUE not only has the function of edge caching, but it also has a sufficient density of idle D2D users

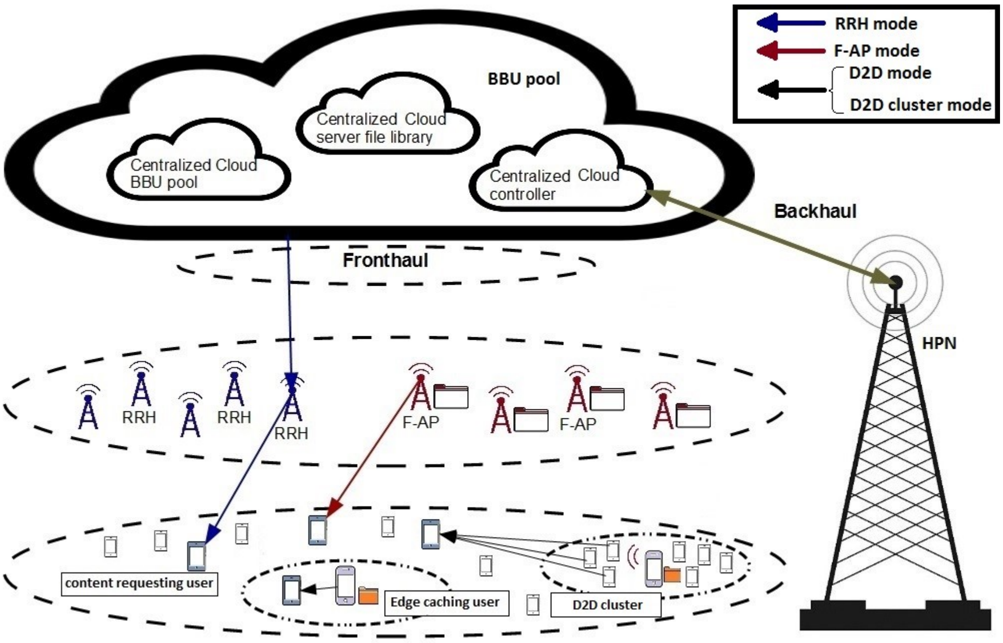

Figure 4.1: F-RAN system model with three different content downloading schemes

around it. These idle D2D users are treated as as half-duplex decode-and-forward relays in this network system and they do not have an edge caching function. As shown in Fig. 4.2, We assume that for each D2D cluster, N relays are uniformly distributed around the CEU within a definition distance of L_c [84]. Note that if the distance between the RU and the CEU is less than the threshold distance of direct D2D communication L_d, the requested content can be provided by the CEU directly without the relays in the D2D cluster. Therefore, CEUs can provide the content downloading service with different transmission modes according to the distance to the RU. Accordingly, the providing content for RUs can be realized in three different transmission modes:

- **F-AP**: Direct content downloading via F-AP;
- **D2D**: Direct D2D communication between the RU and EU;
- **D2D clustered cooperative beamforming**: CEU provides the requested content to the RU with the help of idle D2D UEs by forming a D2D clustered cooperative beamforming.

In this F-RAN system, the F-APs are deployed in a two-dimensional Euclidean plane $\mathbb{D}_{\mathbb{R}}{}^2$ according to a homogeneous Poisson point process (PPP) Φ_f with a density of λ_f. Let $\mathcal{K} =$

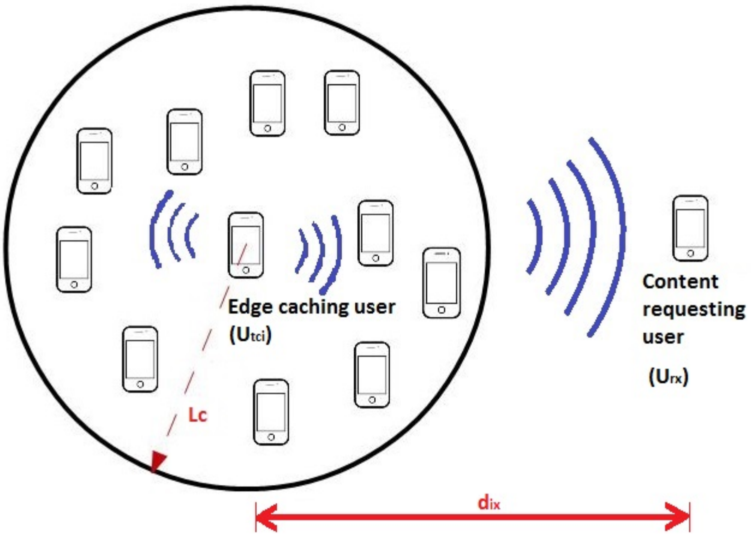

Figure 4.2: D2D clustered cooperative beamforming

$\{1, 2, 3, ..., M_f\}$ denote the set of F-APs. User equipment (UEs) is distributed while following an independent PPP Φ_u with a density of λ_u [14]. Since the UEs in an F-RAN makes a content request are called RUs, the location distribution of RUs is modeled as an independent PPP Φ_r with a density of $\lambda_r = p_r \lambda_u$ where $p_r \sim (0, 1]$. The set of RUs can be denoted as $\mathcal{R} = \{1, 2, 3, ..., M_r\}$. M_r means the number of RUs. Since EUs form a homogeneous PPP Φ_t with a density of $\lambda_t = p_t \lambda_u$ where $p_t \sim (0, 1]$, the set of EUs in the F-RAN can be explained as $\mathcal{T} = \{1, 2, 3..., M_t\}$. M_t is defined as the number of EUs in F-RAN. Remaining common UEs without content requesting and edge caching functions are distributed following PPP Φ_c with an intensity of $\lambda_c = p_c \lambda_u$, where $p_c = 1 - p_t - p_r$. The set of CEUs within the F-RAN can be indicated as $\mathcal{B} = \{1, 2, 3, ..., M_b\}$. We assume p_{tc} to be the probability that one EU acts as a CEU, which can implement cooperative beamforming in the F-RAN. The spatial distribution of CEU according to PPP is Φ_{tc} with an intensity of $\lambda_{tc} = p_{tc} \lambda_u$. Meanwhile, the remaining EUs form PPP Φ_{td} with a density of $\lambda_{td} = p_{td} \lambda_u$, where $p_t = p_{tc} + p_{td}$ and $\mathcal{D} = \{1, 2, 3, ..., M_d\}$. Furthermore, we can calculate the expected numbers for different devices as shown in Table 4.1 where $A({\mathbb{D}_{\mathbb{R}}}^2)$ is the area of the disc plane ${\mathbb{D}_{\mathbb{R}}}^2$.

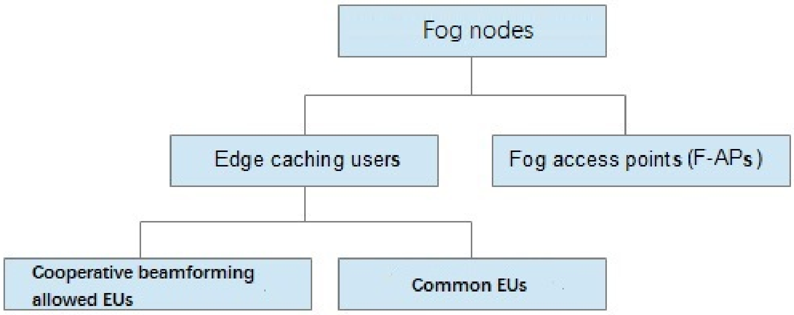
Fog nodes
Edge caching users
Fog access points (F-APs)
Cooperative beamforming allowed EUs
Common EUs

Figure 4.3: Classification of edge caching equipment

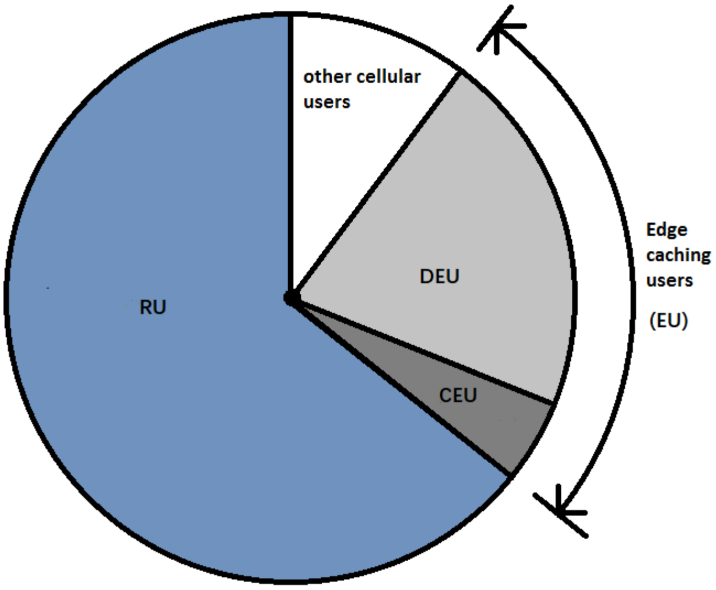

Figure 4.4: Classification of cellular users in F-RAN

4.1.2 Channel Model

The channel model consists of large-scale pathloss and small-scale Rayleigh fading. If we let $g(d_{ij})$ denote the channel power gain of the link between transmitter i and receiver j, then

$$g(d_{ij}) = h_{ij}P_L(d_{ij}) = h_{ij}PL^{-1}d_{ij}^{-\alpha}, \tag{4.1}$$

where $h_{ij} \sim exp(1)$ is the Rayleigh fading power coefficient, which is exponentially distributed with unit mean and variance, and $P_L(d_{ij}) = PL^{-1}d_{ij}^{-\alpha}$ obeys the standard path loss propagation with path loss exponent α. PL^{-1} is called the path loss constant. The variable d_{ij} denotes the distance between transmitter i and receiver j [85], and the additive white Gaussian noise

Table 4.1: Expected number and symbol notation for different equipment

Equipment	PPP	Intensity	Expected Number
F-APs	Φ_f	λ_f	$A(\mathbb{D}_{\mathbb{R}}{}^2)\lambda_f$
RU	Φ_r	λ_r	$A(\mathbb{D}_{\mathbb{R}}{}^2)\lambda_r$
EU	Φ_t	λ_t	$A(\mathbb{D}_{\mathbb{R}}{}^2)\lambda_t$
CEU	Φ_{tc}	λ_{tc}	$A(\mathbb{D}_{\mathbb{R}}{}^2)\lambda_{tc}$
DEU	Φ_{td}	λ_{td}	$A(\mathbb{D}_{\mathbb{R}}{}^2)\lambda_{td}$

(AWGN) has a power spectral density of $N_0(dBm/Hz)$. For simplicity, we assume that each link has a bandwidth of $B(Hz)$. Downlink control channels are used in mode selection when D2D mode and cooperative beamforming mode are selected. We assume that different F-APs share the same spectrum. To simplify the analysis, we assume that, within the service area of one F-AP, the cooperative radio resource management (CRRM) in the cloud center perfectly allocates all downlink control channels. Instances of uplink communication associated within one F-AP are assigned orthogonal channels. Therefore, intra-cell interference will not be discussed when using both uplink and downlink control channels. Inter-cell interference only exists when sub-channels are reused. As multiple relays are used to transmit the desired content to the same destination when D2D clustered cooperative beamforming is implemented, we assume that relays within identical clusters are allowed to reuse the same subchannel. The CEU and relays around it are predefined. Therefore, we ignore interference between a CEU and D2D cluster consisting of relays. In the remaining analysis, we assume perfect channel estimation at each node.

4.1.3 Location-Based Transmission Mode Selection

After an F-AP receives the request of one RU, the appropriate transmission mode is selected by analyzing the location and cache information of fog nodes. A content transfer terminal selection that can provide the required content to the RU satisfies two conditions: it can successfully transmit the signal to the target RU, and it has the requested content in the local caching space. We assume that each F-AP has the location information of UEs and can provide each RU with a node as a potential content-transfer terminal, which is defined as the nearest fog node to the target RU that can successfully transmit the signal. Therefore, once the requested content exists in the caching space of the potential content transfer terminal, the corresponding fog transmission nodes are selected directly by F-APs. Otherwise, the F-AP selects the nearest equipment that can successfully transmit content from the remaining fog nodes. In this study, as mentioned above, since the case for cloud service is not considered, we assume that all requested content can be served by F-APs.

Algorithm 3 Location-Based Transmission Mode Selection

```
Input:
    Set of F-AP, $\mathcal{K} = \{1, 2, 3, ..., M_f\}$;
    Set of RUs, $\mathcal{R} = \{1, 2, 3, ..., M_r\}$;
    Set of CEUs, $\mathcal{B} = \{1, 2, 3, ..., M_b\}$;
    Set of DEUs, $\mathcal{D} = \{1, 2, 3, ..., M_d\}$;
    Set of selected D2D transmitters $\mathcal{D} = \{DU_1, DU_2, DU_3, ..., DU_i\}$;
    $r$ where $i \in \mathcal{R}$;
    Requested content for each RUs;
    Set of contents in EUs;
    Set of contents in F-APs;
Output:
    Mode selection for RUs;
    Set $r = 1$;
    while $r \le M_r$ do
        while Content requirement of RU $r$ can be provided by the content storage in EUs do
            while The distance between the RU $r$ and its nearest DEU $\le L_d$ do
                D2D transmission is selected;
                while The distance between the RU $r$ and its nearest CEU $\le L_{cr}$ do
                    D2D Clustered cooperative beamforming transmission mode is selected;
                end while
            end while
        end while
        $r = r + 1$;
    end while RUs who has not been assigned a transmission mode will be provided with the
    F-AP transmission mode;
    return Mode selection for RUs
```

4.1.4 Pre-fetching Phase and Content Caching Policy

The pre-fetching phase should be carried out before implementing the content downloading service in the F-RAN. Before downloading content, each F-AP and EU must cache content based on popularity and the limited size of the edge caching space. In the pre-fetching phase, there exists a content library with content set $C_C = \{c_1, c_2, c_3..., c_{S_c}\}$ and size S_C. We assume that all content in the content library has the same size. Content stored in the cloud center is arranged in descending order according to popularity. This means that in content library C_C, c_i denotes the i-th most popular content. Following Zipf's law, for one RU, the probability of requesting content c_i can be expressed as

$$p_i = \frac{1/i^{\beta}}{\sum_{j=1}^{S_c} 1/j^{\beta}}, \tag{4.2}$$

where $\sum_{i=1}^{S_C} p_i = 1$, and β is the Zipf shape parameter that controls the correlation of content selection [86]. As β increases, more requests are generated from the RU for content with high popularity. Therefore, for c_a and c_b, if $a < b$, we have $p_a > p_b$, which means that c_1 has the largest probability of being requested by RUs. Each F-AP can cache contents from the cloud center to its caching space via fronthaul link with limited local storage size S_F. Moreover, some of the contents are cached from F-APs to EUs through wireless links. We presumed that each EU has same local storage size S_U where $S_C > S_F > S_U$. Since RUs are interested in the most popular content [16], probability-based content caching policy is defined which means that F-APs and EUs cache the content based on the content popularity. As the result, $C_F = \{c_1, c_2, c_3..., c_{S_F}\}$ and $C_U = \{c_1, c_2, c_3..., c_{S_U}\}$ are the corresponding local content cache libraries for F-APs and EUs respectively.

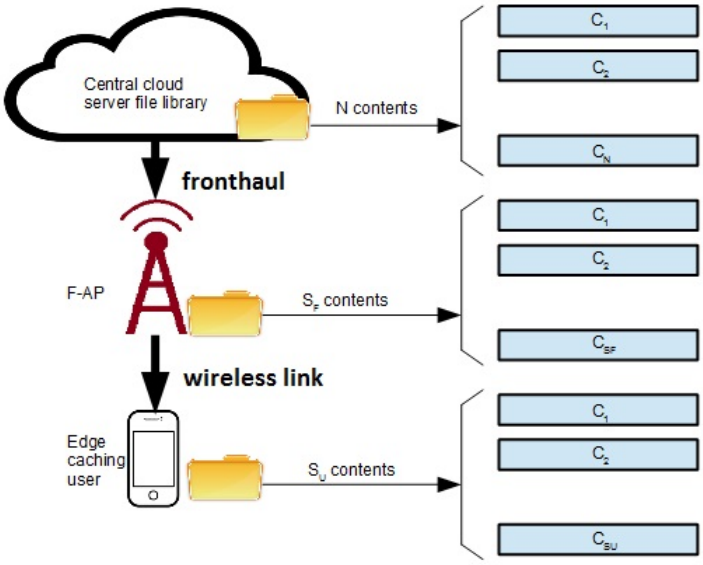

Figure 4.5: Illustration of the pre-fetching phase

In this study, we provide a new caching policy in which CEUs have two content sets that can be chosen. The first content set is generated with the probability based caching policy, and is same as content set $C_{B1} = C_U$. The second set of cache contents is $C_{B2} = \{c_{S_U+1}, c_{S_U+2}, c_{S_U+3} ..., c_{2S_U}\}$. The policy in which all of the CEUs choose to cache content set C_{B2} in the pre-fetching phase is called sub-popular content caching policy. The probability that the content requested by the RU has been cached by the assisted edge caching equipment is called to the content caching probability. Accordingly, the content caching probability for F-APs, DEU and CEU can be obtained respectively as follows:

$$p_{CC}^{F} = \sum_{i=1}^{S_F} p_i \qquad p_{CC}^{T} = \sum_{i=1}^{S_U} p_i \tag{4.3}$$

$$p_{CC}^{B1} = \sum_{i=1}^{S_U} p_i \qquad p_{CC}^{B2} = \sum_{i=S_{U+1}}^{2S_U} p_i \tag{4.4}$$

where p_{CC}^{B1} denotes the content caching probability for CEU when the probability based caching policy is selected. Meanwhile, when the sub-popular content caching policy is considered, the content caching probability for CEU is p_{CC}^{B2}. p_{CC}^{F} denotes the content caching probability for F-AP and p_{CC}^{T} is the probability for DEU.

Based on the location-based transmission mode selection method and and two content caching policies, the average energy efficiency per request and cache hit probability for edge caching equipment are analyzed in the remaining sections.

In section 4.2, we study the energy efficiency with acquiring CSI, selecting suitable relays and cooperative beamforming for different transmission methods with a fixed data rate R_T.

In section 4.3, we analyze the cache hit probability under two different content caching policies. In addition, the impact of UE density and content popularity on performance metrics are studied.

4.2 Energy Efficiency Analysis

Whenever one RU has a content requirement, for effective and reliable content transmission, it must first send its request to the nearest F-AP. Therefore, the transmission mode can be selected by the associated F-AP from D2D, F-AP, or D2D clustered cooperative beamforming mode. After selecting the transmission mode, the F-AP sends the selected results to the RU. Within the coverage of each F-AP, we assume that the location information of EUs has been collected, and the channel statement information (CSI) between EUs and the F-AP is known. To analyse the energy efficiency for three transmission modes, we use a fixed rate R_T (bits/symbol). To better discuss the differences of energy efficiency for three different transmission modes, we assume that the transmission power of the fog node is not limited by its maximum transmission power when calculating energy consumption. The outage probability is only considered in the channel estimation phase. The maximum transmission power of EU and F-AP is used only when calculating transmission interference.

4.2.1 F-AP Mode

Channel Estimation

In the F-AP transmission mode, the RU downloads the requested content from its nearest F-AP node directly. Defined the F-AP k as the content provider where $k \in \mathcal{K}$. In the channel estimation phase, N_T training symbols are sent from F-AP k to the RU r ($r \in \mathcal{R}$) under the transmission power P_T^F. As downlink communication analysis is considered in this transmission mode, the receiver RU r suffers the interference from other F-APs, which is denoted as I_f. Therefore, the corresponding interference I_f at RU r can be expressed as follows:

$$I_f(d_{fr}) = \sum_{f=1 f \neq k}^{M_f} P_{fr}^F g(d_{fr}), \tag{4.5}$$

where P_{fr}^F denotes the transmission power from F-AP f to RU r and d_{fr} denotes the distance between them. For simplify the interference calculation, the worst case will be discussed in the remaining interference analysis. As for this case, the value of each P_{fr}^F equals to the maximum transmission power of F-AP. Further, we can obtain the expression of inter-cell interference as below:

$$I_f(d_{fr}) = \sum_{i \neq k}^{M_f} P_f g(d_{fr}), \tag{4.6}$$

where P_f means the F-AP's maximum transmission power.

According to [82] and Shannon equation, the data rate R_T can be given as: $R_T = log_2(1 + SINR(-ln(1-p_{out})))$, where signal to noise radio can be expressed as:

$$SINR = \frac{P_T^F(d_{kr}, d_{fr})\bar{g}(d_{kr})}{P_N + I_f(d_{fr}))}. \tag{4.7}$$

Hence, the transmission power in broadcasting N_T training symbols can be expressed as follows:

$$P_T^F(d_{kr}, d_{fr}) = \frac{(1 - 2^{R_T})(P_N + I_f(d_{fr}))}{\bar{g}(d_{kr})ln(1 - p_{out})}, \tag{4.8}$$

where $\bar{g}(d_{kr})$ is the mean channel power gain, which equals $1/(PLd_{kr}^{\alpha})$; p_{out} is the outage probability; P_N is the noise power; and R_T is the fixed data rate of the user requesting content. Once the channel state information between F-AP k and RU r has been estimated at the RU, it is sent to the F-AP as feedback. Since no inter-cell interference is considered in F-AP k, we assume that one EU exists for each remaining F-AP on the same uplink channel. These EUs can generate interference at the target F-AP on the same sub-channel. For the worst case, the interference at F-AP k can be expressed as $(P_N + I_d)$, where

$$I_d(d_{dir}) = \sum_{i \neq k}^{M_f} P_d g(d_{dir}), \tag{4.9}$$

and P_d is the maximum transmission power of the EU. $P_d g(d_{dir})$ expresses the inter-cell interference from the EU in F-AP i on the same uplink channel. We assume that there is only one EU in each remaining F-AP coverage area that multiplexes the same uplink channel and generates interference. Since N_{FB} symbols are used to feed back the CSI to F-AP k with transmission power P_{FB}^F, the transmission power is obtained as

$$P_{FB}^F(d_{kr}, d_{dir}) = \frac{(2^{R_T} - 1)(P_N + I_d(d_{dir}))}{g(d_{kr})}. \tag{4.10}$$

Content Transmission

Once F-AP k has received the target CSI, the desired content can be able to sent with minimum transmission power P_{CT}^{F} under the fixed value of the data rate R_T.

$$P_{CT}^{F}(d_{kr}, d_{fr}) = \frac{(2^{R_T} - 1)(P_N + I_f(d_{fr}))}{g(d_{kr})}. \tag{4.11}$$

Energy Efficiency

In the F-AP transmission mode, the calculation of energy consumption can be classified in two parts: channel estimation and content transmission. Note that the circuit energy consumption is not discussed in this phase. In the first part, the energy consumption for the channel selection in F-AP mode is

$$E_{CE}^{F}(d_{kr}, d_{fr}, d_{dir}) = P_{T}^{F}(d_{kr}, d_{fr})N_T T_S + P_{FB}^{F}(d_{kr}, d_{dir})N_{FB} T_S, \tag{4.12}$$

where $T_S = 1/B$ denotes the time duration of symbol.

Suppose that N_C is the number of the symbols per content. The energy consumption for transmitting content from F-AP k to the target user RU r is given as:

$$E_{CT}^{F}(d_{kr}, d_{fr}) = P_{CT}^{F}(d_{kr}, d_{fr})N_C T_S. \tag{4.13}$$

According to the previous calculation, we can finally get the expression energy efficiency of F-AP transmission mode as follows:

$$EE_{INS}^{F} = \frac{R_T N_C}{E_{CE}^{F} + E_{CT}^{F}}. \tag{4.14}$$

For brevity, the downlink inter-cell interference is expressed as I_f in the remaining sections. Similarly, the simplified expression of uplink inter-cell interference is I_d.

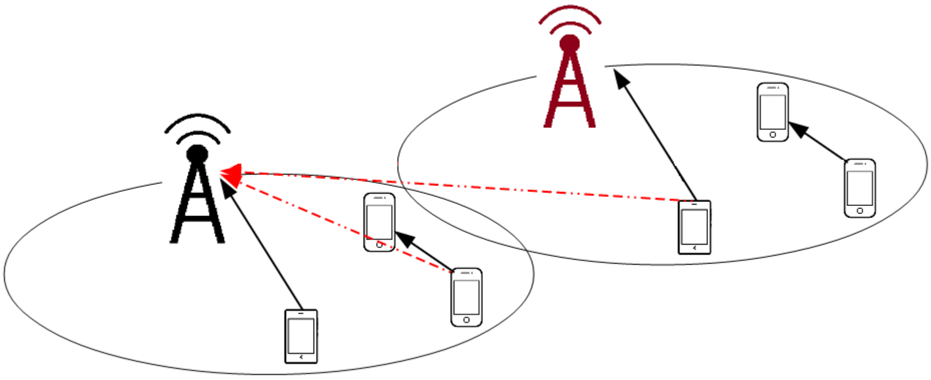

Figure 4.6: ChannelEstimation

4.2.2 D2D Mode

D2D Transmitter Selection

RUs that can be successfully served by an adjacent EU within the D2D transmission threshold distance L_d can select the D2D transmission mode, in which the requested content is provided by the assisted EU, which is called a D2D transmitter. After receiving the request from RU r, the fixed EU is selected by the F-AP, which acts as a D2D transmitter in the content transmission. Assuming that the corresponding F-AP already knows the channel state information of the D2D transmitters which are located in the coverage area. To implement direct D2D communication in this mode, the selected D2D transmitter first receives the F-AP's message, which includes the content request and the location of the RU. As the downlink control channel is used to transmit this message with N_{DC} symbols, the transmission power P_{DC}^{D} of F-AP k to send the message to selected EU e $(e \in \mathcal{T})$ can be expressed as

$$P_{DC}^{D}(d_{ke}) = \frac{(2^{R_T} - 1)(P_N + I_f)}{g(d_{ke})}, \tag{4.15}$$

where d_{ke} denotes the distance from F-AP k to the selected D2D transmitter which is EU e.

Channel Estimation

After the target D2D transmitter has received the message from F-AP k, it sends N_T training symbols to RU r for channel estimation. Consider that D2D communication is executed as an underlay of the uplink F-RAN system. Due to the threshold distance of L_d, which is defined, we assume that all of the EUs have the same maximum transmission power P_d ($P_d \leqslant P_f$). To implement a D2D communication link, an uplink sub-channel should be assisted. In other F-APs, the EU that occupies the same sub-channel can cause interference at RU r. For each F-AP, we assume that each sub-channel can only be allocated to at most one communication link. So, there is no interference within the coverage of F-AP k. Therefore, the selected D2D transmitter sends the training symbols with transmission power

$$P_T^D(d_{er}) = \frac{(1-2^{R_T})(P_N+I_d)'}{\bar{g}(d_{er})ln(1-p_{out})} \tag{4.16}$$

where d_{er} is the distance between RU r and EU e. Note that the mean channel power gain $\bar{g}(d_{er}) = 1/PLd_{er}^{\alpha}$. After the user requesting content has perfectly performed the channel estimation, it sends N_{FB} symbols back to the selected D2D transmitter with transmission power

$$P_{FB}^D(d_{er}) = \frac{(2^{R_T}-1)(P_N+I_d)}{g(d_{er})}. \tag{4.17}$$

Content Transmission

In the content transmission phase, the selected D2D transmitter can send the desired content with the minimum required transmission power based on the received channel state information. As mentioned, N_C symbols are transmitted to RU r. From the previous analysis, the transmission power is obtained as

$$P_{CT}^D(d_{er}) = \frac{(2^{R_T}-1)(P_N+I_d)}{g(d_{er})}. \tag{4.18}$$

Energy Efficiency

In order to calculate the energy efficiency, it is necessary to list the overhead for each phase at first. In D2D transmitter selection phase, N_{DC} symbols are sent to the target D2D transmitter EU e include the location and request of RU with transmission power P_{DC}^D. As for this, the energy consumption can be expressed as follows:

$$E_{DC}^{D}(d_{ke}) = P_{DC}^{D}(d_{ke})N_{DC}T_S \tag{4.19}$$

. Similarly, in the channel estimation phase, the overall energy power consumption can be divided into two parts: sending training symbols and the feedback of required CSI.

$$E_{CE}^{D}(d_{er}) = P_{T}^{D}(d_{er})N_T T_S + P_{FB}^{D}(d_{er})N_{FB}T_S. \tag{4.20}$$

Than the energy consumption for transmitting requested content from selected D2D transmitter EU e to the receiver RU r can be obtained:

$$E_{CT}^{D}(d_{er}) = P_{CT}^{D}(d_{er})N_C T_S. \tag{4.21}$$

Therefore, basing on the energy consumption we have listed, the instantaneous energy efficiency for direct D2D transmission mode can be expressed as follows:

$$EE_{INS}^{D} = \frac{R_T N_C}{E_{DC}^{D} + E_{CE}^{D} + E_{CT}^{D}}. \tag{4.22}$$

4.2.3 D2D Clustered Cooperative Beamforming Mode

As the D2D clustered cooperative beamforming transmission mode is chosen, the F-AP sends the request and location of RU to the selected CEU which can provide the content by making user of D2D cluster. In this transmission mode, the D2D cluster is composed by several idle D2D users. Each of idle D2D users acts as a half-duplex decode-and-forward relay. For implementing the cooperative beamforming and optimising the power consumption, channel estimation and relay selection are required. As for this, realising the D2D clustered cooperative beamforming requires following steps to forward the content (as shown in Fig. 4.7):

- **CEU selection**: F-AP finds the appropriate CEU which can transmit the content to the RU through the D2D cluster. The location and request are sent to this CEU by using downlink control channel.

- **Channel estimation**: Each relay node estimates the channel statement information to the CEU and RU, respectively.
- **Relay selection**: Based on the estimation of each channel power gain, K optimal relays will be selected for performing cooperative beamforming.
- **Content transmission**: The selected CEU broadcasts the requested content to the K selected relays, which then beamform the message to the requesting user. K relays will send the content with cooperative beamforming.

CEU Selection

Since the F-AP k receives the request from the RU r, the content provider should be selected by F-AP. For realising the D2D clustered cooperative beamforming transmission mode, the CEU b which can transfer the requested content with the D2D cluster is selected. As mentioned, N_{DC} symbols are transmitted from F-AP to the target D2D transmitter for sending RU r's request and location with transmission power P_{DC}^{CB}. By using the downlink control channels and cooperative radio resource management, the transmission power can be expressed as follows:

$$P_{DC}^{CB}(d_{kb}) = \frac{(2^{R_T}-1)(P_N+I_f)}{g(d_{kb})}, \tag{4.23}$$

where d_{kb} denotes the distance between the F-AP k and the selected CEU b in this transmission mode. Compared with the power consumption in signal transmission, the power consumed in the remaining circuitry can be neglected [87].

Channel Estimation

In order to implement cooperative beamforming and optimise the power consumption by relay selection, the power channel gain for each relays is required. As mentioned previously, since EUs not only have content caching functions but also have resource management and computing capabilities, we assume that the channel power gains between EU b and N relays are known. N relays are uniformly distributed around the selected CEU b within the distance of L_c. As for this, all of the channel power gains from D2D transmitter to multiple relays obey independent exponential distribution, which can be denoted as $g_{sj}(d_{sj})(j = 1, 2, 3...N)$ where d_{js} is denoted as the distance between relay RL_j and the source of cooperative beamforming CEU b. To simplify the description, the set of channel power gain between CEU b and N relays are expressed as: $\{g_{s1}, g_{s2}, g_{s3}, ..., g_{sN}\}$.

In this phase of process, the destination RU r broadcasts N_T training symbols to N relays in the D2D cluster with transmission power P_T^{DR}. Relays in the cluster estimates the channel power gains g_d. Making a assumption that decode and forwards relays in one same D2D cluster are located close to each other. As for this, the mean channel power gain between RU r and D2D cluster is given as: $\bar{g_d}(d_{dRL}) = 1/(PLd_{dRL}^{\alpha})$ with the distance d_{dRL} which means the approximate distance between D2D cluter and RU. Note that the RU occurs with a fixed data rate R_T (bits/symbols). Therefore, defined $2^{R_T} - 1$ as the value of threshold SINR. Besides, the outage probability p_{out} is assumed to be a constant in this transmission strategy[69]. To simplify the analysis, presumed that the destination RU r monopolises one uplink subchannel to broadcast the training symbol to N relays. Moreover, by considering the interference I_d and the additive white Gaussian noise P_N the transmission power is given as follows::

$$P_T^{DR}(d_{rRL}) = \frac{(1 - 2^{R_T}(P_N + I_d)}{\bar{g_d}(d_{rRL})ln(1 - p_{out})}. \tag{4.24}$$

To simplify the analysis reasonably, we assume that the channel estimation of each relay is perfect. Thus, each decode and forward relays can obtain the channel power gains between itself and the destination RU r. These channel power gains can be expressed as $g_{dj}(d_{sj})(j = 1, 2, 3...N)$.

Relay Selection

After the destination RU r has sent N_T training symbols to multiple relays in the cluster, assumed that there are M relays can decode the message successfully. By reference to [71], the time-based relays selection method is discussed in this work. Without the loss of generality, suppose that the number of successful decoding relays M is greater than 1. In order to minimize the overhead in this transmission mode, the number of selected relays K should be optimized. The optimal number K we desire can be changed by the distance between the D2D cluster and RU. In order to make the subsequent steps easier to be explained, a new sequence of M successful decoding relays is required. Basing on their channel power gains to the destination RU r, these M relays could be arranged in descending order where relay RL_1 has the highest channel power gain g_{d1}. The first K relays in this sequence are selected as the optimal relays. The new sequence of channel power gain can be generated as follows: $g_{d1} > g_{d2} > g_{d3}... > g_{dM}$. Followed by this, each relays starts a timer $t_i = \rho/g_{di}(i = 1, 2, 3...M)$ since it has received the notification symbol from D2D transmitter. Note that ρ is a constant value and each timer will not be changed after its expiration. For realising the cooperative beamforming in the data transmission phase, each selected relay should know the sum of the channel power gains $\sum_{i=1}^{K} g_{di}$. Defined that each correct decoding relay broadcasts a notification symbol with its channel sate information after the expiration of its timer.

As mentioned, the relay RL_1 with the shortest timer $t_1 = \rho/g_{d1}$ first broadcasts the notification symbol and be selected as the optimal relay. In the course of broadcasting, the remaining relays can overhear the message from RL_1, since we have assumed each of the relays belonging to this cluster locates close to each other. In order to guarantee all of the relays can receive this notification symbol, the transmission power P_{NT} of RL_1 has to be limited by data rate R_T and the corresponding radius L_r of D2D cluster.

$$P_{NT}(2L_r) = \frac{(1 - 2^{R_T})P_N}{\bar{g_d}(2L_r)ln(1 - p_{out})}. \tag{4.25}$$

Once $RL_i(i = 2, 3, 4...M)$ have received the notification symbol with unexpired timers, their timers are extended to $t_i + T_s$. As for this, the collisions caused by broadcasting notification symbols from remaining relays can be avoided. Based on the expression of timer, the relay with the second strongest value of g_{d2} reaches the expiration after RL_1 and broadcasts the notification symbol as the procedure discussed above. Followed by extending of the timers, the third best relays can be implemented by the same process.

Whenever the source of the cooperative beamforming CEU b receives a notification signal from one RL_i $(i = 1, 2, 3, ..., M)$, it calculates the power consumption of content transmission via i relays cluster. This procedure continues until the CEU receives the $(K + 1)$th notification symbol and the result indicates an increase in power consumption. In this case, the selected CEU broadcasts a feedback symbols to M relays. As for this, each relays knows that the selection of optimal relays has finished with K best relays. As soon as receiving the notification symbol from CEU, relay RL_{K+1} and the remaining $M - K$ relays with unexpired timers switch to the idle mode and keeps idling in the following steps. Therefore, only K selected relays could still be used after the relay selection phase. In addition, each of the optimal relays can get the channel power gains of other $K - 1$ selected relays in the processing of relay selection. Also, in this timer based relay selection, we assume that the influence of the time delay in the relay selection phase and the probability of collision are negligible.

Making an assumption that all of the relays broadcast in the same resource block, there is no other D2D pair in the same F-RAN which occupies this subchannel. For this reason, the relay RL_i broadcast the training symbol at the expiration of timer t_i under the same transmission power as RL_1. In considered the power consumption in the feedback of CEU b with transmission power:

$$P_{FB}^{CB}(d_{dRLK}) = \frac{(2^{R_T} - 1)P_N}{g_{dK}(d_{dRLK})}. \tag{4.26}$$

$g_{dK}(d_{dRLK})$ denotes the power channel gain between D2D transmitter CEU b and relay RL_K. The total transmission power consumption P_{RS} in the relay selection phase can be given as:

$$P_{RS} = (K + 1)P_{NT} + P_{FB}^{CB}. \tag{4.27}$$

Content Transmission

The content transmission phase for D2d clustered cooperative beamforming is composed of two main steps. In the first step, the selected transmitter CEU b broadcasts the content to K optimal relays with the power of P_T^{CB} which is given as:

$$P_T^{CB} = \frac{(2^{R_T} - 1)P_N}{g_{sKmin}}, \tag{4.28}$$

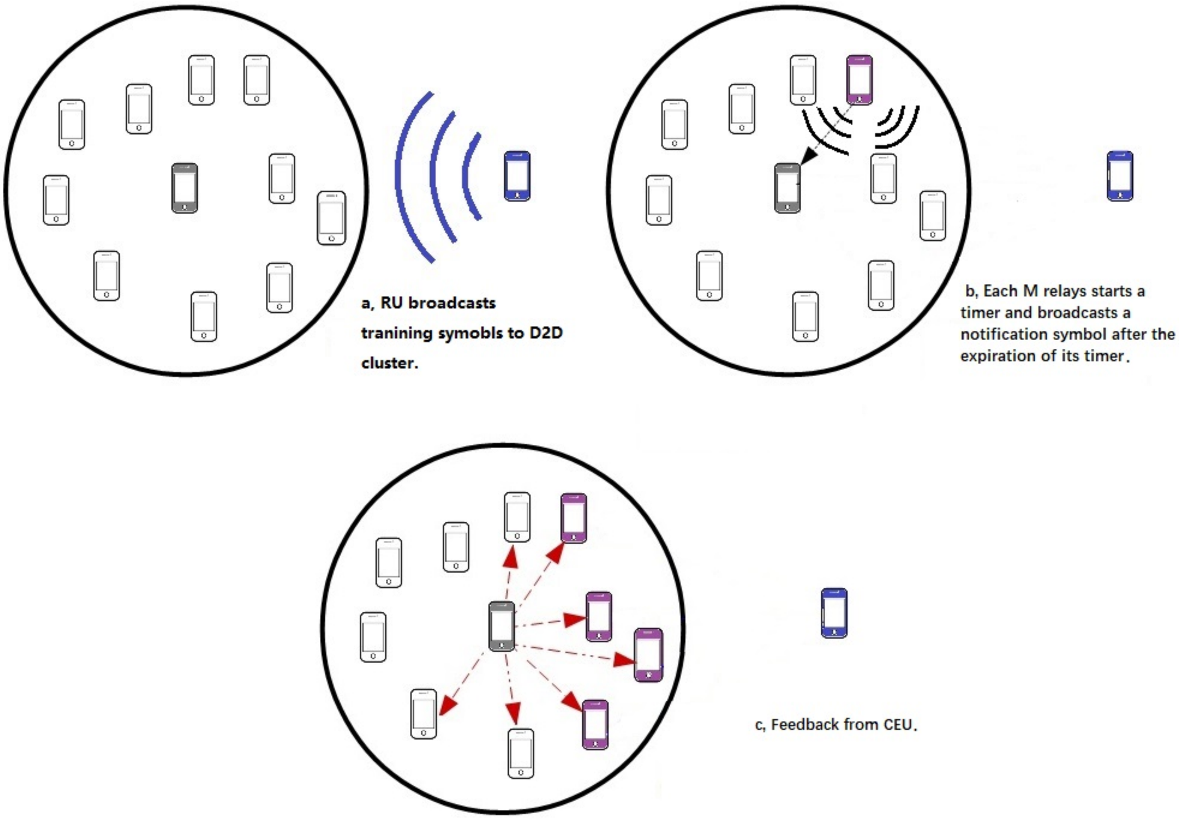
a, RU broadcasts tranining symobls to D2D cluster.
b, Each M relays starts a timer and broadcasts a notification symbol after the expiration of its timer.
c, Feedback from CEU.

Figure 4.7: Channel estimation and relay selection phase

where g_{sKmin} denotes the minimum channel power gain between CEU and K relays. Then, the received signal at the K best relays can be written as:

$$\mathbf{x} = \sqrt{P_T^{CB}\mathbf{g}_s}m + n_0, \tag{4.29}$$

where P_T^{CB} means the transmission power of the CEU which sends content m. Furthermore, n_0 denotes the noise and $\mathbf{g}_s$ denotes the channel power gain between the edge caching user and optimal relays. The vectors $\mathbf{g}_s$ can be expressed as follows:

$$\mathbf{g}_s = [g_{s1}, g_{s2}, g_{s3}...g_{sK}]^T. \tag{4.30}$$

After decoding the received information $\mathbf{x}$, the decode-and-forward relays re-encode the message to obtain $\mathbf{x}_e$ and re-transmit it to the content requesting user. Before multiple relays broadcast the re-coding information $\mathbf{x}_e$, each of them has to multiply with a beamforming coefficient w_i

for adjusting the amplification of the information at the receiver, where $i = 1, 2, 3...K$. We also assume that K selected relay cooperative information in sending the desired information to the receiver RU r can precisely delay their content transmission so as to achieve perfect synchronisation at the content requested user, which means that the coefficient w_i is a real value [56]. Let $K \times 1$ vector $\mathbf{t}$ act as the relay transmission signals, which can be expressed as:

$$\mathbf{t} = \mathbf{W}\mathbf{x}_e, \tag{4.31}$$

where $\mathbf{W}$ is denoted as a diagonal matrix which equals $diag([w_1^*, w_2^*, w_3^*...w_K^*])$ and where vector $\mathbf{w}^H = [w_1^*, w_2^*, w_3^*...w_K^*]$ means $\mathbf{W}$'s diagonal entries. The signal y which is received at the content requesting user is obtained as follows:

$$y = \sqrt{\mathbf{g}_s}^T \mathbf{t} + n_1 = \sqrt{\mathbf{g}_s}^T \mathbf{W}\mathbf{x}_e + n_1, \tag{4.32}$$

where $\mathbf{g} = [g_1, g_2, g_3...g_K]^T$ and n_1 denotes the sum of the noise. It is also assumed that $\mathbb{E}(\boldsymbol{x}_e \boldsymbol{x}_e^H) = 1w$, the total transmitted power can be simplified and expressed as follows:

$$P_K = \sum_{i=1}^{K} |w_i|^2. \tag{4.33}$$

With the condition of the SINR, denoting $(2^{R_T} - 1)$ as the threshold of the SINR, the inequation for cooperative beamforming can be given as follows:

$$\frac{\sum_{i=1}^{K} |w_i \sqrt{g_{di}}|^2}{(P_N + I_d)} \geqslant SINR_{min} = 2^{R_T} - 1. \tag{4.34}$$

where I_d is the inter-cell interference at RU r. According to previous assumptions, in cooperative beamforming, the selected relays reuse one uplink channel to reduce inter-cell interference. Based on Lagrangian multiplier techniques, the optimal allocation for relay i can be expressed as

$$|w_i| = \frac{\sqrt{g_{di}}}{\sum_{j=1}^{K} g_{dj}} \sqrt{T_c(P_N + I_d)}, \tag{4.35}$$

where $T_c = 2^{R_T} - 1$ denotes the threshold value to SINR. According to the weight of the optimal beamforming, the optimal transmission power from relay RL_i to RU is given as follows:

$$P_{CTi}^{CB} = (P_N + I_d)T_c \left(\frac{1}{\sqrt{g_{di}}} \sum_{j=1}^{K} g_{dj} \right)^{-2}. \tag{4.36}$$

4.2.4 Energy Efficiency

Based on the previous discussion, we can analyse the power consumption in each phases. The energy overhead in the CEU selection part can be expressed as:

$$E_{DS}^{CB}(d_{kb}) = P_{DC}^{CB}(d_{kb})N_{DC}T_s. \tag{4.37}$$

In the channel estimation phase, N_T training symbols are transmitted from the RU r to N relays. Therefore the energy consumption can be wrote as:

$$E_{CE}^{CB} = P_T^{DR}(d_{rRL})N_T T_s. \tag{4.38}$$

In the next step, in selecting optimal relays in the D2D cluster, timers are used in the selection process. The energy consumption is:

$$E_{RS}^{CB} = [(K+1)P_{NT}(2L_r)N_{NT} + P_{FB}^{CB}(d_{rRLK})N_{FB}]T_s, \tag{4.39}$$

where N_{NT} means the notification symbol. Finally, in the data transmission phase, energy consumption is:

$$E_{CT}^{CB} = (P_T^{CB} + \sum_{i=1}^{K} P_{CTi}^{CB})N_C T_s. \tag{4.40}$$

Based on the analyzing of the overhead as above, the instantaneous energy efficiency of cooperative beamforming transmission mode can be expressed as follows:

$$EE_{INS}^{CB} = \frac{R_T N_C}{E_{DS}^{CB} + E_{CE}^{CB} + E_{RS}^{CB} + E_{CT}^{CB}}. \tag{4.41}$$

4.3 Cache Hit Probability and Mode Selection Probability Analysis

In this section, we study and discuss two performance metric other than energy efficiency: cache hit probability and mode selection probability. In the wireless content caching analysis of F-RAN, caching hit probability $\mathbb{P}_H$ is one of the most important parameter which means the probability that one random requested content of one RU can be served. Since there are three different fog nodes that can be considered to provide the request, cache hit probability can be categorized[31]. Mode selection probability $\mathbb{P}_S$ means the probability that one transmission mode can be selected to serve the request according to the location base mode selection method and cache hit probability. In this section, we discuss the different cache hit probabilities and mode selection probabilities of each fog nodes and transmission mode.

4.3.1 Cache Hit Probability

F-AP

In this case, cache hit probability is denoted as the probability that one content requesting user can be served by the associated F-AP node. Note that each requesting user can select the nearest F-AP node as the content provider. According to the content caching policy mentioned, the probability expression is given as: $\mathbb{P}_H^F = p_{CC}^F = \sum_{i=1}^{S_F} p_i$.

Common Edge Caching Users

In the system model of pre-fetching phase, the content caching probability of common EUs is defined as p_{CC}^T. Besides, according to the property of the 2-D Poisson process, within the area of $A(L) = \pi L^2$, the probability that x random nodes which follow the PPP under density λ_X exist is given as $\mathbb{P}(x) = \frac{(\lambda_X \pi L^2)^x e^{-\lambda_X \pi r^2}}{x!}$. Setting $x = 0$, $L = L_d$ and $\lambda_X = \lambda_{td}$, we can obtain cache missing probability $\mathbb{P}_{Hmiss}^D = exp(-\pi\lambda_{td}L_d^2)$, which is defined as the probability that no DEUs can provide services within the distance of L_d [88]. We can further obtain the cache hit probability of DEUs:

$$\mathbb{P}_H^D = p_{CC}^T(1-\mathbb{P}_{Hmiss}^D) = p_{CC}^T(1-exp(-\pi\lambda_{td}L_d^2)). \tag{4.42}$$

Cooperative Beamforming Allowed Edge Caching Users

Presumed that the threshold service range of D2D clustered cooperative beamforming is L_{cr}. Carried out the similar derivative process step as above, we can get $\mathbb{P}_{Hmiss}^C = exp(-\pi\lambda_{tc}L_{cr}^2)$. Thus, the cache hit probability for CEUs is given as:

$$\mathbb{P}_H^C = p_{CC}^B(1-\mathbb{P}_{Hmiss}^C) = p_{CC}^B(1-exp(-\pi\lambda_{tc}L_{cr}^2)), \tag{4.43}$$

where p_{CC}^B is determined according to the content caching policy. If the probability-based caching policy is used for each CEU, then p_{CC}^B equals p_{CC}^{B1}, and $\mathbb{P}_H^C$ can be expressed as $\mathbb{P}_{HB1}^C = p_{CC}^{B1}(1-exp(-\pi\lambda_{tc}L_{cr}^2))$. When a sub-popularity caching policy is selected in the pre-fetching phase, we obtain $p_{CC}^B = p_{CC}^{B2}$ and $\mathbb{P}_{HB2}^C = p_{CC}^{B2}(1-exp(-\pi\lambda_{tc}L_{cr}^2))$.

4.3.2 Mode Selection Probability

F-AP Transmission Mode

Define $\mathbb{P}_S^F$ as the mode selection probability of F-AP transmission mode. To provide the PB content caching policy for CEUs, the probability that an RU can be served by DEUs and CEUs is given as

$$p_{CC}^T(1-\mathbb{P}_{Hmiss}^D) + p_{CC}^{B1}(1-\mathbb{P}_{Hmiss}^C)\mathbb{P}_{Hmiss}^D = \mathbb{P}_H^D + \mathbb{P}_{HB1}^C\mathbb{P}_{Hmiss}^D. \tag{4.44}$$

Thus, the mode selection probability for F-AP mode when PB content caching policy is allocated to CEUs is:

$$\mathbb{P}_{SPB}^F = 1-(\mathbb{P}_H^D + \mathbb{P}_{HB1}^C\mathbb{P}_{Hmiss}^D). \tag{4.45}$$

Similarly, when an SP content caching policy is considered in the fetching phase for each CEU, the probability for without selecting F-AP transmission mode can be expressed as

$$p_{CC}^{T}(1-\mathbb{P}_{Hmiss}^{D})+p_{CC}^{B2}(1-\mathbb{P}_{Hmiss}^{C})=\mathbb{P}_{H}^{D}+\mathbb{P}_{HB2}^{C}. \tag{4.46}$$

If we let $\mathbb{P}_S^{FSP}$ be the value of $\mathbb{P}_S^{F}$ when SP caching policy is applied to CEUs, then we can get

$$\mathbb{P}_S^{FSP}=1-(\mathbb{P}_{H}^{D}+\mathbb{P}_{HB2}^{C}). \tag{4.47}$$

.

D2D Transmission Mode

For the D2D transmission mode, we denote $\mathbb{P}_S^{DPB}$ as the probability when content set C_{B1} is cached in the caching space of CEUs. As we mentioned, CEUs can provide the content downloading service in two different transmission modes, in the case of using PB content caching policy, the probability of CEUs provide content directly is given:

$$p_{CC}^{B1}(1-exp(-\pi\lambda_{tc}L_d^2)). \tag{4.48}$$

Combined this equation with cache hit probability of DEUs, the expression of $\mathbb{P}_S^{DPB}$ can be calculated as follows:

$$\mathbb{P}_S^{DPB}=p_{CC}^{T}(1-exp(-\pi\lambda_{td}L_d^2))+p_{CC}^{B1}(1-exp(-\pi\lambda_{tc}L_d^2)). \tag{4.49}$$

Similarly, by applying the SP content caching policy to CEUs, the mode selection probability $\mathbb{P}_S^{DSP}$ can be given:

$$\mathbb{P}_S^{DSP}=p_{CC}^{T}(1-exp(-\pi\lambda_{td}L_d^2))+p_{CC}^{B2}(1-exp(-\pi\lambda_{tc}L_d^2)). \tag{4.50}$$

.

Table 4.2: Simulation parameters

Parameter	Value
Number of content items in cloud center S_C	1000
Number of content items in F-AP S_F	1000
Number of content items in EU S_U	100
Intensity of DEUλ_{td}	$0.5 \times 10^{-4} \sim 5 \times 10^{-4}$
Intensity of CEUλ_{tr}	$0.5 \times 10^{-4} \sim 5 \times 10^{-4}$
Maximum transmission power of F-AP P_f	26 dBm
Maximum transmission power of UE P_d	13 dBm
Content providing distance of DEU L_d	30 m [14]
Content providing distance of CEU L_c	30 m [14]
Radius of cluster L_r	20 m [84]
Noise power spectral density	-174 dBm/Hz [81]
Bandwidth B	15×10^3 Hz [81]
Data Rate R_T	4 bits/symbol
Path loss exponent α	4 [81]
Outage probability p_{out}	0.1 [71]

D2D Clustered Cooperative Beamforming Transmission Mode

In this case, the probability of selecting cooperative beamforming transmission mode with PB content caching policy is

$$\mathbb{P}_S^{CPB} = p_{CC}^{B1}(1 - exp(-\pi\lambda_{tc}L_{cr}^2)) - p_{CC}^{B1}(1 - exp(-\pi\lambda_{tc}L_d^2)). \tag{4.51}$$

When using an SP content caching policy, the mode selection probability is

$$\mathbb{P}_S^{CSP} = p_{CC}^{B2}(1 - exp(-\pi\lambda_{tc}L_{cr}^2)) - p_{CC}^{B2}(1 - exp(-\pi\lambda_{tc}L_d^2)). \tag{4.52}$$

4.4 Simulation and Results

We provide numerical results to analyse the impact of different content caching policies on the cache hit probability for F-AP and average energy efficiency. The effect of the density of DEUs, threshold distance of cooperative beamforming, and Zipf exponent are evaluated through simulation. Parameters used in the simulation are listed in Table 4.2. Set $N_T = 1$ symbol and $N_{FB} = 1$ symbols for channel estimation. For selecting EU in D2D transmission mode and

D2D clustered cooperative beamforming mode, $N_{DC} = 1$ symbol is used to provide the request and location of RU. In content transmission phase, we assume that the data package length for each content are where $N_C = 300$ symbols.

We plot the cache hit probability of F-APs by applying PB caching policy and SP caching policy for CEU with Variable value of λ_{td},L_{cr} and β. For energy efficiency evaluation, the results of changing β and L_{cr} are plotted.

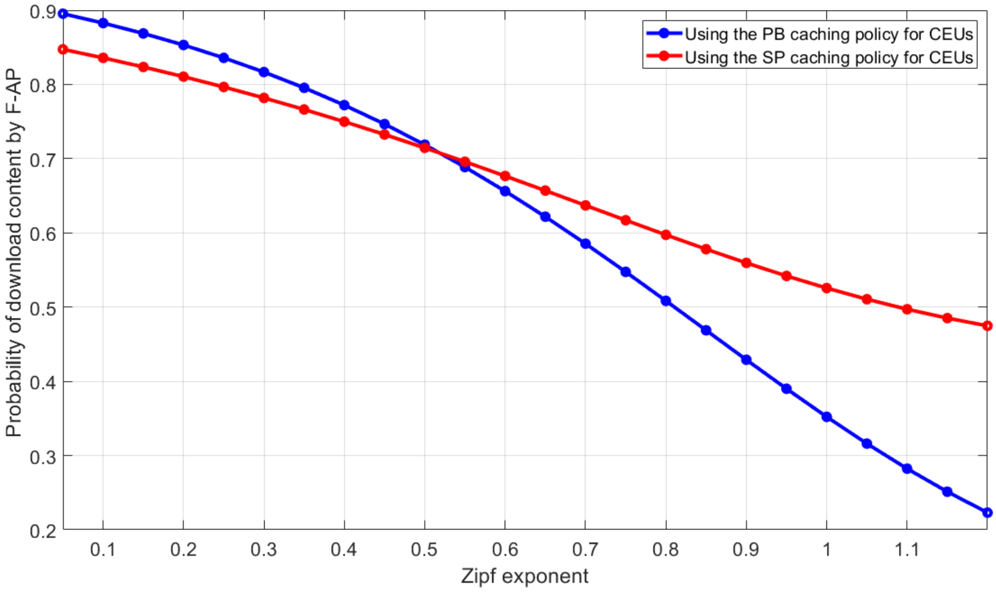

Figure 4.8: Mode selection probability of F-AP transmission over the value of Zipf exponent β1

Fig. 4.8 plots the mode selection probability of F-AP transmission versus Zipf exponent β for applying PB content caching policy and SP content caching policy. DEUs located with density $\lambda_{td} = 3 \times 10^{-4}m^{-2}$. The density of CEUs is $\lambda_{tc} = 0.6 \times 10^{-4}m^{-2}$. Threshold cooperative beamforming service distance $L_{cr} = 100m$. Since that only CEUs can select the content caching policy in pre-fetching phase, this figure indirectly shows the impact of using the CEUs on the content downloading service. Lower value of $\mathbb{P}_S^F$ means that EUs can provide more content downloading service. This is a decreasing tendency for $\mathbb{P}_S^F$ with increase value of β. This is due to the fact that the larger β provides a higher concentration of high-popularity contents which means the total probability of a content cached in the EUs being requested goes increase. Basing on the expression of $\mathbb{P}_S^F$, increasing p_{CC}^T and p_{CC}^B can makes more RUs be served by

EUs through D2D transmission and cooperative beamforming transmission. We can see that apply the SP content caching policy to CEUs can makes more EUs to serve the request when $\beta \leq 0.5$. Otherwise, when $\beta > 0.5$, applying PB content caching policy to CEUs can let CEUs provide more requests. This is because the lower value of β leads to smaller content caching probability p_{CC}^{T}. According to the expression of $\mathbb{P}_{H}^{D}$, the probability that a requested content needs to be served by the DEU decreases. As $p_{C}^{B1}C = p_{CC}^{T}$, the number of RUs who can serve by CEUs also be reduced. Only the RUs whose requested contents belong to $C_U = C_{B1}$ have the preconditions to be served by DEUS and CEUS when PB caching policy is selected. Since the number of CEUs in this case is much smaller than the number of DEUs, the help that D2D clustered cooperative beamforming can provide is reduced. On the other hand, due to content set C_{B2} is cached in the caching space of CEUs when SP caching policy is considered, CEUS can provide services to RUs who need to obtain the requested content by F-APs.

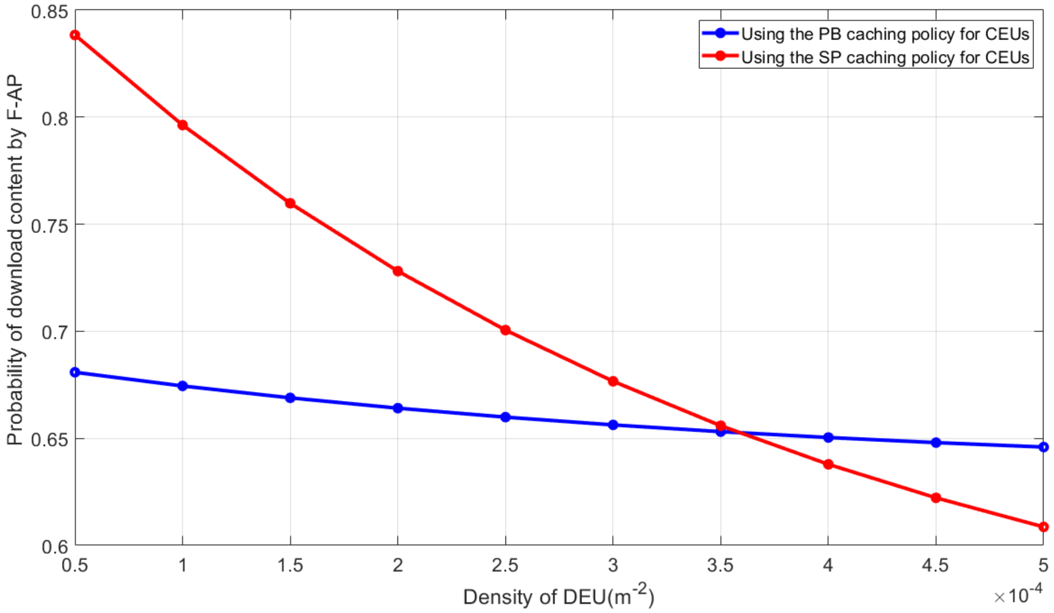

Figure 4.9: Mode selection probability of F-AP transmission over the density of DEUs λ_{td}

Fig. 4.9 compares the value of $\mathbb{P}_{S}^{F}$ between the PB caching policy and SP caching policy for different density of DEUs. In this case, $\beta = 0.6$, $\lambda_{tc} = 0.6 \times 10^{-4} m^{-2}$ and $L_{cr} = 100m$. From the figures, we can observe that with the increase of DEUs density λ_{td}, the probability of requests that need to be transmitted directly by the F-APs decreases faster when using the SP content caching policy for CEUs. This is due to the fact that as the density of DEUs increases, the number of services that CEUS with the content set C_{B1} can provide becomes smaller. On the other hand, while DEUs can service the majority of requests from $C_U = C_{B1}$, CEUS that can service RUs whose requested contents are in the content requested C_{B2} can more effectively reduce the transfer pressure on F-AP.

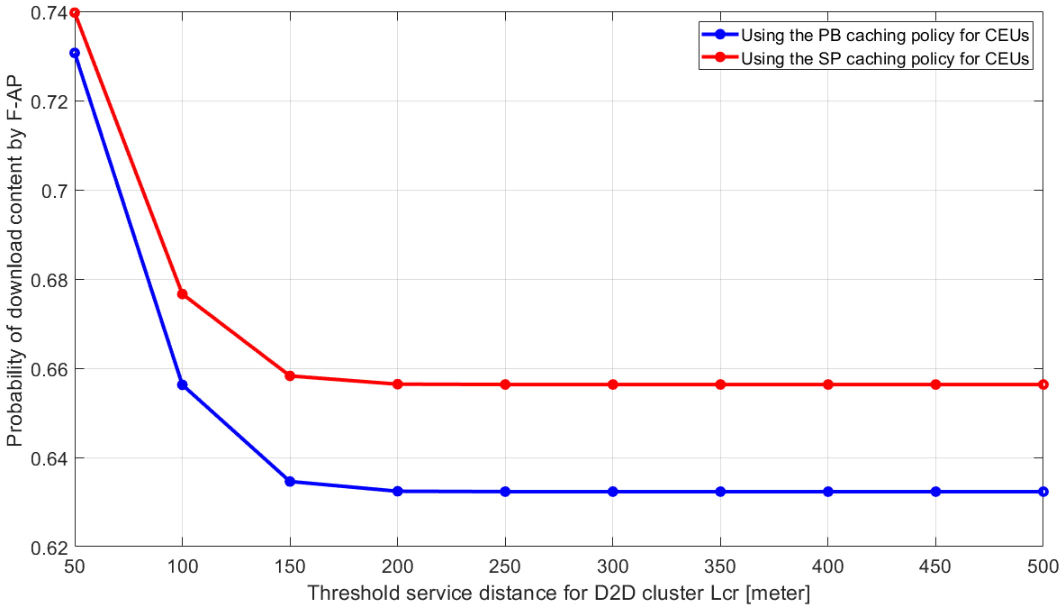

Figure 4.10: Mode selection probability of F-AP transmission over the threshold service distance of CEUs L_{cr}]

Fig. 4.10 plots the mode selection probability of F-AP transmission, with $\beta = 0.6$, $\lambda_{tc} = 0.6 \times 10^{-4} m^{-2}$, and $\lambda_{td} = 3 \times 10^{-4} m^{-2}$, for different values of L_{cr}. It is obvious from the figure that increasing the service scope of CEUs can effectively reduce the probability of using F-AP transmission mode when L_{cr} is less than 150 m. Since the density of CEUs does not change in this case, increasing the length of L_{cr} when the service radius is sufficient will no longer affect the usage probability of F-AP transmission.

Fig. 4.11 depicts the average energy efficiency per request over different Zipf exponents β and content caching policies for CEUs, where $\lambda_{td} = 3 \times 10^{-4} m^{-2}$, $\lambda_{tc} = 0.6 \times 10^{-4} m^{-2}$, and $L_{cr} = 100m$. Conforming to the observation regarding Fig. 4.8, the average energy efficiency increases with the Zipf exponent, which means that the probability of using F-AP transmission mode decreases. Also, it can be seen that with increasing β, the average energy efficiency using the PB caching policy increases and exceeds that of the SP caching policy. A larger Zipf

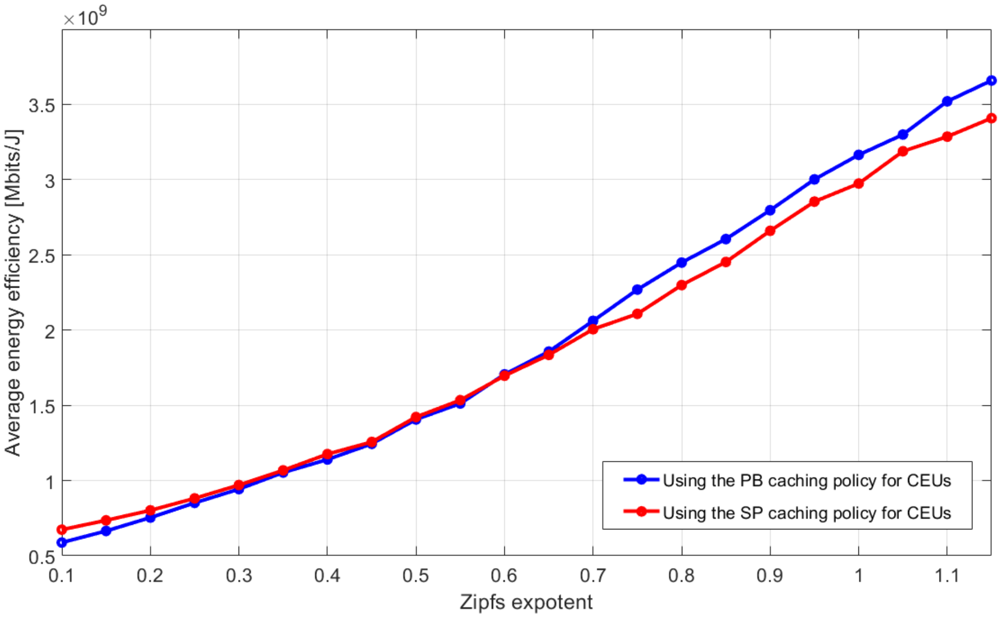

Figure 4.11: Average energy efficiency over Zipf exponent

exponent suggests that applying PB content caching policy to CEUs can provide more requests for RUs, and so it can reduce the energy consumption by using F-AP transmission, i.e., higher energy efficiency for each request.

A comparison of F-AP transmission and D2D clustered cooperative beamforming with PB and SP content caching policy is shown in Fig. 4.12. Obviously, we can observe that F-AP transmission mode can incur more overhead when providing content. According to the study of energy efficiency in Section 4.2, D2D direct transmission mode provides faster content downloading service and lower energy consumption than F-AP transmission mode. However, due to the limited content providing distance, the amount of RUs that can be directly serviced by EU in the F-AN system is limited. D2D cluster composed of CEU and relays can transmit the requested content to the corresponding RU by cooperative beamforming and has a longer service distance than D2D mode. At the same time, it has better energy efficiency than F-AP transmission mode. D2D cluster composed of CEU and relays can transmit the requested content to the corresponding RU by cooperative beamforming and has a longer service distance than D2D mode. Fig. 4.12 shows the impact of threshold service distance of D2D cluster on average energy efficiency. Meanwhile, the average energy efficiency without D2D cluster assisted is also given. According to the location based transmission mode selection we mentioned in

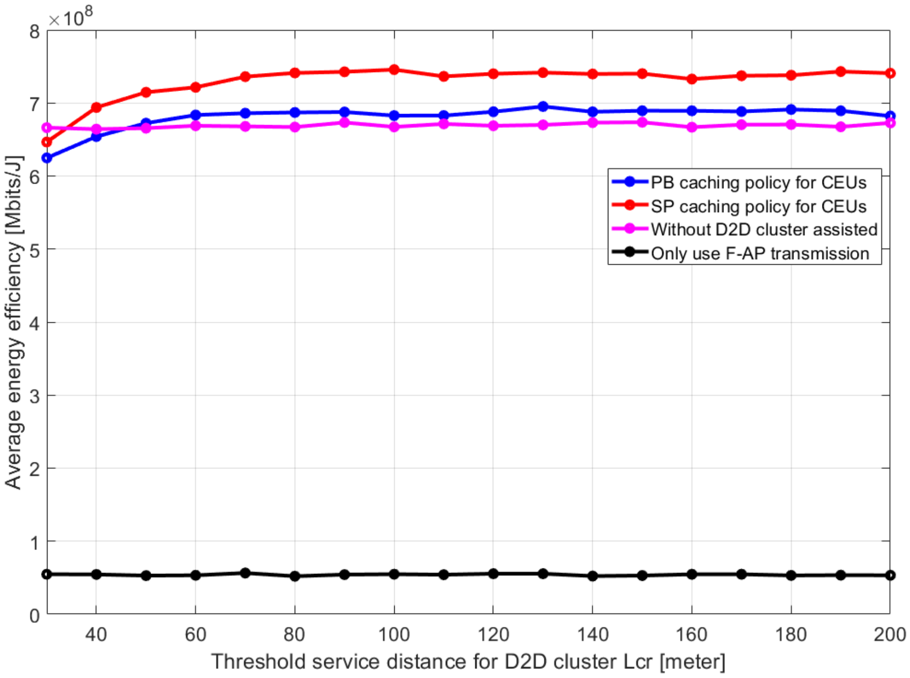

Figure 4.12: Average energy efficiency over the threshold service distance of CEUs L_{cr}

Section 4.1.3, the average energy efficiency is a fixed value in theory for the case that D2D cluster is not enable when $\beta = 0.2$, $\lambda_{td} = 3 \times 10^{-4} m^{-2}$, and $\lambda_{tc} = 0.6 \times 10^{-4} m^{-2}$. According to the energy efficiency analyation in Section 4.2, D2D direct transmission can provider higher energy efficiency than D2D cluster transmission at the same transmission distance. Therefore, when the threshold service distance L_{cr} is small, not enabling D2D cluster can make the content download costs less energy. As L_{cr}increases, the service distance of D2D cluster goes increase and more requested contents of RUs can be downloaded through D2D clustered cooperative beamforming mode. Since the usage of F-AP mode goes decrease in this case, the average energy efficiency can be optimized. Combined with Fig. 4.8, it is seen that if $\beta = 0.2$ and $\lambda_{td} = 3 \times 10^{-4} m^{-2}$, a small Zipf exponent and high density of DEUs allows the SP cache strategy to serve more RUs than the PB content caching policy, and reduces the number of RUs served by F-AP, improving the average energy efficiency. As for this, in Fig. 4.12, when $L_{cr} = 40$, using the SP content caching policy for CEUs can provide better average energy efficiency than not using D2D cluster for F-RAN system. At the same time, using the PB content caching policy does not provide better energy efficiency than not using D2D clusters. In this case, the value of L_{cr} needs to be further increased if using the PB caching policy for

CEUs consumes less energy in content download process than the case that the D2D clusters do not enabled in the F-RAN system. From Fig. 4.12, using PB content caching policy for CEUs can save the energy cost than the case that no D2D clusters.

4.5 Conclusion

We considered D2D clustered cooperative beamforming in an F-RAN system, while previous research focused on cooperative transmission for F-APs. We first defined a D2D cluster which has one CUE in the center with several relays uniformly distributed around. In Section 4.2 we proposed to considered the D2D clustered cooperative beamforming as the transmission mode for content providing. Than, we proposed a sub-popularity content caching policy for cooperative beamforming allowed edge caching users, which can be selected with the probability-based content caching policy to optimise performance. The providing content for RUs can be realized in threes different transmission modes: D2D mode, F-AP mode and D2D clustered cooperative beamforming mode.Next, A timer-based relay selection method and location-based transmission mode selection method were applied to analyse the average energy efficiency, cache hit probability, and mode selection probability. We proposed a sub-popularity content caching policy for cooperative beamforming allowed edge caching users, which can be selected with the probability-based content caching policy to optimise performance. The density of DEUs, popularity of content, and threshold service range of cooperative beamforming were considered as important factors influencing the optimization of energy efficiency, mode selection probability, and choice of caching policy of CEUs.According to the simulation results, an SP content caching policy is more energy-efficient than a PB content caching policy with a low Zipf exponent and high DEU density. Moreover, increasing the service scope of D2D clustered cooperative beamforming will not affect the choice of caching policy, but can optimise energy efficiency to some extent. Cooperative beamforming can generally provide a longer service range than direct D2D transmission, and consume less energy than F-AP transmission. It can adjust the cache policy according to the system environment to improve the energy-efficiency of the system. SP and PB content caching policies can be selected for CEUs based on the joint analysis of content popularity and DEU density.

Chapter 5

Content Caching Policy with Edge Caching User Classification in Fog RANs

In this Chapter, we present a new content caching policy for EUs, which is called edge caching user classification-based caching policy. In this caching policy, EUs are divided into two groups: EU group 1 and EU group 2. These two groups are assigned different sets of content. We study an optimized content caching policy selection algorithm for EUs to maximize the cache hit probability for EUs.

The F-RAN system model and the process of content providing service are described in Section 5.1. In Section 5.2, EU classification-based caching policy is presented. In addition, cache hit probability for EUs and average download delay for different caching policy are analyzed. The cache hit probability for EUs and average download delay are two metrics describing the performance of the content caching policy for EUs. In Section 5.3, an optimal cache hit probability maximization algorithm is developed using the adjustable feature of the edge caching user classification-based caching policy. The cache hit probability maximizaiton problem is formulated and the corresponding content caching policy selection method with EU classification is developed in Section 5.3. Numerous simulation results are presented and discussed in Section 5.4. Finally, we provide conclusions in Section 5.5.

5.1 System Model

5.1.1 Network Model

We consider one F-RAN model that consists of a cloud center, an F-AP, a number of EUs, and several content requesting users (RUs). As Fig. 5.1 shows, the F-AP is located at the center of its coverage area with a coverage radius of R_F, wherein all RUs and EUs can be served directly by the F-AP through a wireless link. The F-AP has a local storage size of N_f and can cache contents from the cloud center via a fronthaul link.

The EUs are distributed following a homogeneous Poisson point process (PPP) Φ_E with a density of γ_E. RUs are distributed following independent PPP Φ_R with a density of γ_R [14]. The numbers of EUs and RUs can be expressed as $K_{EF} = \gamma_E A_F$ and $K_{RF} = \gamma_R A_F$, respectively, where $A_F = \pi R_F^2$ is the F-AP coverage area. We assume that each EU or RU is equipped with one single antenna. Each EU has a maximum transmission distance L_E. Limited by L_E, the expected number of RUs within the communication range of an EU is given by $K_{RE} = \gamma_R A_E$, where $A_E = \pi L_E^2$ denotes the service area of an EU. Also, K_{EE} means the average number of EUs in the service range of one EU.

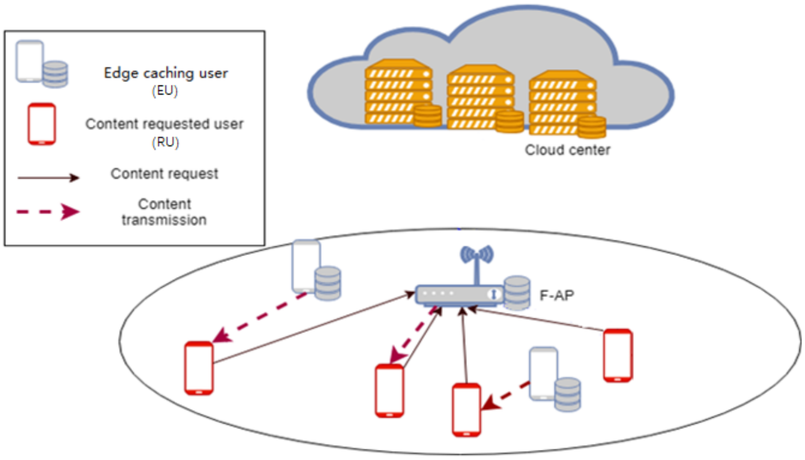

Figure 5.1: System model

5.1.2 Content Caching Phase

The cloud center periodically adjusts the popularity of contents by following user preferences. At the central cloud server, we define a finite content library $C_C = \{c_1, c_2, c_3..., c_M\}$ of size M, where $c_i(i \in \{1, ..., M\})$ denotes the ith most popular content file. We assume that the contents cached in F-AP and EUs are periodically updated. At the start of each period, the cloud center adjusts the popularity of contents according to the requests it collected during the previous period. Next, the F-AP updates its content set based on the new popularity of contents and its caching policy. To allow RUs to download content via edge caching, the EUs also need to complete content caching in the content caching phase. Each EU will be allocated the contents according to the content caching policy of EUs. For F-AP, we use the probability based caching policy, which means the F-AP only caches the most popularity contents from the cloud center. As a result, the content set in F-AP is given by $C_F = \{c_1, c_2, c_3..., c_{N_f}\}$. We assume that each EU has a local storage size of N_e, where $M > N_f > N_e$. The contents will be cached from the F-AP to EUs via wireless links. Without loss of generality, all the contents in the cloud center are assumed to have the same size of s bits. Following Zipf's law, the probability of an RU requesting content c_i is expressed as follows:

$$p_i = \frac{1/i^\beta}{\sum_{j=1}^{M} 1/j^\beta} \tag{5.1}$$

where β is the Zipf exponent that controls the popularity of content [86]. Most RUs would be interested in a few popular files with a high value of β [16]. According to each contents' popularity, two different caching policies for EUs will be discussed and compared in Section III.

5.1.3 Content Requesting Phase

All RUs in the F-AP coverage area A_F initially send their content requests to the F-AP. For simplicity, we assume the content stored in the F-AP can satisfy the content requests of all RUs in the system. Thus, there is no need for RUs to download contents directly from the cloud center. As a result, we consider two service modes for content delivery: the F-AP mode and the EU mode. If the requested content is to be delivered by an EU, the F-AP sends one message including information regarding the content request to the corresponding EU. Otherwise, the requested content is provided by the F-AP to the RU directly.

5.1.4 Content Downloading Phase

The content downloading phase begins with the F-AP receiving a content request from an RU. We assume that the F-AP has enough antennas to serve all RUs in the area A_F simultaneously. Consequently, there is no queue at the transmitter of the F-AP. Since an EU can serve only one RU at a time under the constraint of a single antenna, there is a queue of content requests at each EU.

Download delay is defined as the time period elapsed between the F-AP receiving a content request and the time the content is delivered. The time required for requested content to be sent from the content provider to the RU is called the service time. Thus, the function of the download delay for a content request is defined as *Download delay = Waiting time in queue + service time*. We assume the content request arrival for each RU follows an independent and identical Poisson process with arrival rate λ $(request/seconds)$. The arrival rate represents the number of requests per second generated by the RU. If an RU x makes a request, the serving F-AP will receive the request from RU x and decide how to provide the requested content to RU x. When there is an EU whose distance from RU is less than L and which can provide content required by RU x, the EU mode will be selected. Otherwise, the F-AP mode is selected. For the F-AP mode, the download delay is $t_f = \frac{s}{r_f}$, where r_f is denoted as the transmission rate from the F-AP to an EU. For the EU mode, the F-AP will identify and inform an EU to send the requested content to RU x via a D2D link. After that, this EU transmits the content to RU x directly. Accordingly, in the EU mode, i.e., one content request is served by an EU, the service time is composed of two parts: (i) the time required to transmit an instructive message containing information of the request from the F-AP to a selected EU, and (ii) the time required to transmit the requested content from the EU to the RU. We denote the length of the F-AP instruction message as s_f(bit) and the transmission rate between the EU and an RU as r_d(bps). Thus, the service time is $t_e = \frac{s_f}{r_f} + \frac{s}{r_d}$. Hence, the service rate for each EU is $\mu_e = 1/t_e$.

For the request queue model at each EU, there are two parameters: the expected number of request arrivals per unit time and the expected number of service completions per unit time. For each EU, the request queue can be modeled as an independent M/D/1 queue, where request arrivals follow a Poisson process and the service time is fixed [89]. Using the properties of an M/D/1 [90], we have:

- **Utilization factor** ρ: It is defined as $\rho = \frac{\lambda_{mean}}{\mu_{mean}}$, where λ_{mean} is the mean request arrival rate and μ_{mean} is the mean service rate of the EU.
- **Expected number of requests in the queue** L is given by $L = \rho + \frac{\rho^2}{2(1-\rho)}$.
- **Expected waiting time in the system for an individual content request** is given by: $W = \frac{L}{\lambda_{mean}}$.

5.2 Caching Policy and Performance Analysis

5.2.1 EU Classification-based Caching Policy

We consider two different EU caching strategies. The probability-based caching policy (PB caching policy) has been used many times in prior studies published in the literature[14][16]. In the PB caching policy, each EU is assigned the same set of the most popular contents from the F-AP. However, caching the same contents by each EU would limit the diversity of edge caching, and the storage capacity of an EU limits its service ability.

In order to enable more RUs to be served directly by EUs, we propose an EU classification (EUC)-based caching policy that is designed to increase the total amount of contents cached by EUs. More specifically, the EUs are divided into two groups: EU group 1 and EU group 2, which are assigned two different content sets and serve content requests simultaneously. We define α as the caching policy factor, which is the ratio of the number of EUs in group 1 to the total number of EUs. Hence, the density of EUs in the EU group 1 and EU group 2 can be expressed as: $\gamma_{E1} = \gamma_E \alpha$ and $\gamma_{E2} = \gamma_E(1-\alpha)$. In the PB caching policy, each EU caches the same content set[14]. The N_e contents with highest popularity compose this content set. In contrast, in the EUC-based caching policy, we set the N_e most popular contents as the first content set $C_{E1} = \{c_1, ..., c_{N_e}\}$, and set $C_{E2}\{c_{Ne+1}, ..., c_{2N_e}\}$ is defined as the second content set. Let EU group 1 denote the EUs that cache content set C_{E1}, and let EU group 2 denote the EUs that cache C_{E2}.

5.2.2 Performance Metrics

In order to analyze the performance of the two caching policies, we define the following performance metrics:

- **Cache hit probability** $\mathbb{P}_H$: $\mathbb{P}_H$ is defined as the probability that the content requested by an arbitrary RU is available at an EU or F-AP that is located within its maximum communication distance. Cache hit probability indicates the probability that a request for a content can be provided by the corresponding content provider. However, cache hit probability is not the download probability, because in the process of content download, the provider of content is affected by the choice of transmission model. Cache hit probability does not reflect this impact. Let $\mathbb{P}_{HE}$ denote the cache hit probability for the EU mode. Higher $\mathbb{P}_{HE}$ means that more EUs are available to serve the RU via a D2D communication link.

- **Average download delay** $\mathbb{D}$: In our system, the average download delay $\mathbb{D}$ is denoted as the download delay over a long time and over all of the RUs.

5.2.3 Cache Hit Probability $\mathbb{P}_H$

Probability-based caching policy

The content set in each EU can be expressed as: $C_E = \{c_1, ..., c_{N_e}\}$. The probability that the requested content is cached in an EU is:

$$Pr_E = \sum_{i=1}^{N_e} p_i = \frac{\sum_{i=1}^{N_e} 1/i^{\beta}}{\sum_{j=1}^{M} 1/j^{\beta}}. \tag{5.2}$$

Let an arbitrary RU be located at the center of a circle with radius L_E. This circular area is A_E. According to the property of a 2-D Poisson process, within the area A_E, the probability

Table 5.1: Key notations

Notation	Descriptions
EU	Edge caching user
RU	Content requested user
γ_E; γ_R	Density of EUs; of RUs
r_f; r_d	Data rate of using F-AP link: of using D2D link
ρ_{EU1}; ρ_{EU2}	Traffic load at EU in group 1; at EU in group 2
$\lambda(x)$	Traffic arrival rate of content requested user x
λ_{EU1}; λ_{EU2}	Rate of requests accepted by EU group 1; group 2
L_E; R_F	Content providing distance of EU; of F-AP
A_E; A_F	Coverage area of EU; of F-AP
Pr_{E1}; Pr_{E2}	Content caching probability for EU in group 1; group 2
α	Caching policy factor
K_{EE}	Average number of EUs in the service range of EU
K_{EF}; K_{RF}	Number of EUs; of RUs
s	Average content length
s_f	Length of instruction

that there exists m nodes that follow a PPP with density γ_E is given by the following:

$$\mathbb{P}(m) = \frac{(\gamma_E A_E)^m e^{-\gamma_E A_E}}{m!}. \tag{5.3}$$

If $m = 0$, eq.(5.3) gives the probability that no EU exists in the area A_E [88]. We can get the cache miss probability $\mathbb{P}_{miss} = \mathbb{P}(m = 0) = exp(-K_{EE})$. Let an RU be located at the center of a circle with radius L_E, then the probability of at least one EU being located in this circle is $1 - \mathbb{P}_{miss}$. Accordingly, the cache hit probability for the EU mode is:

$$\mathbb{P}_{HE} = Pr_E(1 - \mathbb{P}_{miss}). \tag{5.4}$$

EU Classification-based Caching Policy

For each EU, the content caching probability can be defined as the probability that one RU's desired content can be provided from its cache. The content caching probability for an EU in group 1 and group 2 can be obtained, respectively, as follows:

$$Pr_{E1} = \sum_{i=1}^{N_e} p_i = \frac{\sum_{i=1}^{N_e} 1/i^\beta}{\sum_{j=1}^{M} 1/j^\beta}, \tag{5.5}$$

$$Pr_{E2} = \sum_{i=1+N_e}^{2N_e} p_i = \frac{\sum_{i=1+N_e}^{2N_e} 1/i^{\beta}}{\sum_{j=1}^{M} 1/j^{\beta}}. \tag{5.6}$$

The probability of having at least one EU from group 1 or one EU from group 2 in the radius L_E can be expressed, respectively, as follows:

$$\begin{aligned} &1 - \mathbb{P}_{missE1} = 1 - \frac{(\alpha\gamma_E A_E)^m e^{-\alpha\gamma_E A_E}}{m!} \\ &= 1 - exp(-\alpha K_{EE}), \end{aligned} \tag{5.7}$$

$$\begin{aligned} &1 - \mathbb{P}_{missE2} = 1 - \frac{((1-\alpha)\gamma_E A_E)^m e^{-(1-\alpha)\gamma_E A_E}}{m!} \\ &= 1 - exp(-(1-\alpha)K_{EE}). \end{aligned} \tag{5.8}$$

Thus, by combining the definition of the caching hit probability, eq.(5.7), and eq.(5.8), we can draw the following conclusion:

$$\mathbb{P}_{HE1} = Pr_{E1}(1 - exp(-\alpha K_{EE})), \tag{5.9}$$

$$\mathbb{P}_{HE2} = Pr_{E2}(1 - exp(-(1-\alpha)K_{EE})). \tag{5.10}$$

From the above equations, the probability that any content download request can be provided by any type of EU can be expressed as $\mathbb{P}_{HE1} + \mathbb{P}_{HE2}$.

5.2.4 Average Download Delay $\mathbb{D}$

Probability-based Caching Policy

For EUs, the average request arrival rate in this policy should be considered first. As the total number of EUs is represented by K_{EF}, we denote $\mathbb{A}_{EU}$ as the set of EUs. Also, $\mathbb{A}_{RU}$ refers to the set of RUs. For RU $x \in \mathbb{A}_{RU}$, the rate of content requests arrivaing at EU $k \in \mathbb{A}_{EU}$ is a Poisson process with parameter $\lambda_k(x) = Pr_E \lambda(x) g_k(x)$, where $g_k(x)$ is an indicator function. $g_k(x) = 1$ means that EU k is the EU closest to RU x, where the distance between EU k and RU x is no more than the threshold value L_E. Otherwise, $g_k(x) = 0$. As mentioned, for RU x

we have $\lambda(x) = \lambda$. Therefore, the Poisson process parameter of the total content arrival rate at EU k can be expressed as: $\lambda_k = \int_{x \in \mathbb{A}_{RU}} \lambda_k(x)\, dx$.

Then, the traffic load at EU k can be derived with the two parameters $\rho_k = \frac{\lambda_k}{\mu_e}$. Following the property of an M/D/1 queue, as mentioned above, the average number of content in the EU k can be expressed as:

$$L_{EUS}(k) = \rho_k + \frac{\rho_k^2}{2(1-\rho_k)}. \tag{5.11}$$

Hence, for the entire EU service system, the average amount of content can be represented by $L_{EU} = \sum_{k \in \mathbb{A}_{EU}} L_{EUS}(k)$. The total request arrival rate in the EU service system followa the Poisson process, with $\lambda_{EU} = K_{RF}\mathbb{P}_{HE}\lambda$. Thus, according to Little's law [91], we have the following transmission latency equation: *Long-term average arrival request number = Long term request arrival rate * mean waiting time.* Therefore, based on the number of RUs, the average queue delay can be represented as $W_{EU} = \frac{L_{EU}}{\lambda_{EU}}$.

Finally, based on the cache hit probability in the probability-based caching policy, the average download delay for one RU is $\mathbb{D}_P = \mathbb{P}_{HE}W_{EU} + (1 - \mathbb{P}_{HE})t_f$.

EU Classification-based Caching Policy

We assume all *RU*s have the same traffic arrival rates in our system. For the cache hit probability for EU group 1 and EU group 2, the average rate of requests accepted by the two groups can be represented as $\lambda_{EU1} = K_{RF}\mathbb{P}_{HE1}\lambda$ and $\lambda_{EU2} = K_{RF}\mathbb{P}_{HE2}\lambda$.

Since the total number of EUs in group 1 and group 2 are expressed as αK_{EE} and $(1-\alpha)K_{EE}$, respectively. We denote $\mathbb{A}_{EU1}$ and $\mathbb{A}_{EU2}$ as the EU set for group 1 and group 2. Thus, for EU $m \in \mathbb{A}_{EU1}$ and EU $n \in \mathbb{A}_{EU2}$, we have:

$$\lambda_{EU1m} = \int_{x \in \mathbb{A}_{RU}} \lambda_{EU1m}(x)\, dx, \tag{5.12}$$

$$\lambda_{EU2n} = \int_{x \in \mathbb{A}_{RU}} \lambda_{EU2n}(x)\, dx, \tag{5.13}$$

where λ_{EU1m} and λ_{EU2n} refer to the average content arrival rate at EU m and EU n, respectively.

Therefore, the traffic load at each of them can be expressed as $\rho_{E1m} = \frac{\lambda_{EU1m}}{\mu_e}$ and $\rho_{E2n} = \frac{\lambda_{EU2n}}{\mu_e}$. According to the properties of the M/D/1 queue system, the average amount of contents in EU m and EU n are given as follows:

$$L_{EU1}(m) = \rho_{E1m} + \frac{\rho_{E1m}^2}{2(1-\rho_{E1m})}, \tag{5.14}$$

$$L_{EU2}(n) = \rho_{E2n} + \frac{\rho_{E2n}^2}{2(1-\rho_{E2n})}. \tag{5.15}$$

Thus, the average number of requests for each type of service group can be expressed as $L_{EU1} = \sum_{m\in\mathbb{A}_{EU1}} L_{EU1}(m)$ and $L_{EU2} = \sum_{n\in\mathbb{A}_{EU2}} L_{EU2}(n)$. Likewise, based on Little's law, the average waiting time for one content request in group 1 and group 2 can be expressed as $W_{EU1} = \frac{L_{EU1}}{\lambda_{EU1}}$ and $W_{EU2} = \frac{L_{EU2}}{\lambda_{EU2}}$, respectively. As a result, the average download delay for each request in the EUC based caching policy is $\mathbb{D}_E = \mathbb{P}_{HE1}W_{EU1} + \mathbb{P}_{HE2}W_{EU2} + (1-(\mathbb{P}_{HE1}+\mathbb{P}_{HE2}))t_f$.

5.3 Problem Formulation and Analysis

As the cache hit probability is an important parameter for quantifying the performance of content placement, it is the focus of our research. We want to maximize the cache hit probability of EUs by choosing the optimal caching policy. Meanwhile, changes in the average download delay during cache hit probability optimization will also be analyzed. If the EUC-based caching policy is selected, we will calculate the optimal α to maximum the cache hit probability for EUs. Otherwise, we select a PB caching policy.

According to the definition of α, the choice of caching policy can be determined by the value of α. When $\alpha = 1$, all EUs are classified as EU group 1. Each EU has the same content set C_E based on the popularity of contents and storage capacity of an EU. Since $C_E = C_{E1}$, the PB caching policy will be selected when $\alpha = 1$. Otherwise, $0 \leq \alpha < 1$, and EUs are divided into two groups, which means the EUC based caching policy can be considered in the system. We propose the value of α be optimized to maximize the cache hit probability $\mathbb{P}_{HE1} + \mathbb{P}_{HE2}$ when the EUC-based caching policy is selected. Thus, the cache hit probability maximization problem for the EUC-based caching policy can be formulated as follows:

$$\begin{aligned} \arg\max_{\alpha} \quad & \mathbb{P}_{HE1} + \mathbb{P}_{HE2} \\ & = Pr_{E1}(1 - exp(-\alpha K_{EE})) \\ & + Pr_{E2}(1 - exp(-(1-\alpha)K_{EE})) \end{aligned} \tag{5.16}$$

$$\text{s.t.} \quad C1: \quad \mathbb{P}_{HE1} + \mathbb{P}_{HE2} > \mathbb{P}_{HE} \tag{5.17}$$

$$C2: \quad 0 \leq \alpha < 1 \tag{5.18}$$

$$C3: \quad 0 < \rho_m < 1 - \zeta; \quad m \in \mathbb{A}_{EU1} \tag{5.19}$$

$$C4: \quad 0 < \rho_n < 1 - \zeta; \quad n \in \mathbb{A}_{EU2}. \tag{5.20}$$

$C1$ refers to the prerequisites for choosing the EUC-based caching policy, which means that compared to the PB caching policy, the EUC-based caching policy can increase the cache hit probability for EUs. Then, $C2$ shows the value range of α when the EUC-based caching policy is considered. Note that the proposed cache policy factor α can be explained while F-AP can collect the user density λ_E and popularity exponents of content β. In $C3$ and $C4$, ζ is denoted as a small positive constant for guaranteeing the stability of the M/D/1 queuing system. When the value of α satisfies both $C1$ and $C2$, the EUC based caching policy can be selected. According to $C1$, we have:

$$\begin{aligned} & Pr_{E1}(1 - exp(-\alpha K_{EE})) + Pr_{E2}(1 - exp(-(1-\alpha)K_{EE})) \\ & > Pr_{E1}(1 - exp(-K_{EE})). \end{aligned} \tag{5.21}$$

Since $0 < Pr_{E1}(1 - exp(-K_{EE})) < 1$, we set $e^{-K_{EE}} = U$ so that (21) can be simplified to $f(\alpha) = (Pr_{E1}U + Pr_{E2}) - Pr_{E1}U^{\alpha} - Pr_{E2}U^{1-\alpha} > 0$. We can get the zero points of equation $f(\alpha)$ are $U^{\alpha_{z1}} = e^{-K_{EE}}$ and $U^{\alpha_{z2}} = \frac{Pr_{E2}}{Pr_{E1}}$. Since $f(\alpha)$ is a concave function for α, the range of α for $C1$ is between $\alpha_{z1} = 1$ and $\alpha_{z2} = \frac{ln(Pr_{E2}/Pr_{E1})}{-K_{EE}}$. The policy confidence interval can help the F-AP to select the best caching policy for the system. When α_{z2} is no less than 1, the two conditions $C1$ and $C2$ cannot be met simultaneously. This means that, regardless of how α is adjusted, it is impossible to obtain a higher cache hit probability by choosing the EUC-based caching policy over the PB caching policy, so the PB caching policy is selected. In contrast, the EUC caching policy can be selected if α_{z2} is smaller than 1. Therefore, the caching policy can be selected by looking only at the value of α_{z2}.

Eq. (5.16) can be updated as follows:

$$\arg\min_{\alpha} \quad I(\alpha) = Pr_{E1}U^{\alpha} + Pr_{E2}U^{1-\alpha}. \tag{5.22}$$

Since $I(\alpha)$ is a convex function, the optimized value of α can be determined:

$$I^{'}(\alpha) = Pr_{E1}U^{\alpha}\ln(U) - Pr_{E2}U^{1-\alpha}\ln(U) = 0 \tag{5.23}$$

Thus, (16) can take the maximum value with $\alpha_{op} = \frac{ln(\sqrt{Pr_{E2}U/Pr_{E1}})}{-K_{EE}}$. Hence, we can derive the cache hit probability optimization as showed in Algorithm 4.

Algorithm 4 Cache Hit Probability Optimization Algorithm

```
Input: Number of EUs: $K_{EF}$; Number of RUs: $K_{RF}$; Traffic arrival rate: $\lambda$; length of content:
    $s$; length of instruction message: $s_f$; Coverage range of EU: $A_E$; Density of EU: $\lambda_E$; Set of
    available $\alpha$: $\mathbb{A}_\alpha = \emptyset$.
Output: Selected caching policy and $\alpha_{max}$ for maximizing cache hit probability of EUs.
Find the range of $\alpha$ values for EUC-based caching policy.
Calculate the upper limit of $\alpha$ using $U^{\alpha_{z1}} = e^{-K_{EE}}$ and $\alpha_{z1} = 1$ and $\alpha_{z2} = \frac{ln(Pr_{E2}/Pr_{E1})}{-K_{EE}}$.
if $\alpha_{z2} > 1$ is ture then
    $\alpha_{max} = 1$ and the PB caching policy is selected;
else
    for i = 1, 2, 3,..., $K_{EF}$ do
        $\alpha_i = \frac{i}{K_{EF}}$;
        if $\alpha_{z2} \leq \alpha_i \leq \alpha_{z1}$ then
            $\mathbb{A}_\alpha = \mathbb{A}_\alpha \cup \alpha_i$;
        else
            $\mathbb{A}_\alpha \cap \alpha_i = \emptyset$;
        end if
    end for
    Find the proposed $\alpha_{max}$ value from $\mathbb{A}_\alpha$;
    for $\alpha \in \mathbb{A}_\alpha$ do
        $D = |\alpha - \alpha_{op}|$;
        Find minimum $D$ with corresponding $\alpha$ as the
            $\alpha_{max}$ and the EUC cahing policy is selected;
    end for
end if
```

Table 5.2: Simulation parameters

Parameters	Value
Intensity of EUγ_E	$5 \times 10^{-4} \sim 10 \times 10^{-4}$
Intensity of RUγ_R	10×10^{-4}
Data rate of using F-AP link r_f	2Mbps[93][92]
Data rate of using D2D link r_d	20Mbps [93][92]
Average content length s	5Mbits
Length of instruction s_f	50Kbits
Content providing distance of EU L_E	30m[14]
Content providing distance of F-AP R_F	150m
Amount of content in F-AP N_f	1000
Amount of content in EU N_e	80

5.4 Simulation Results

In this section, we present numerical results that can be used to evaluate our proposed algorithm. The simulations were implemented in Matlab, and detailed parameters are provided in Table 5.2. Location based transmission mode selection is applied, as shown in Fig.5.2. We consider that the traffic arrival rate is $\lambda = 2$ requests for each user.

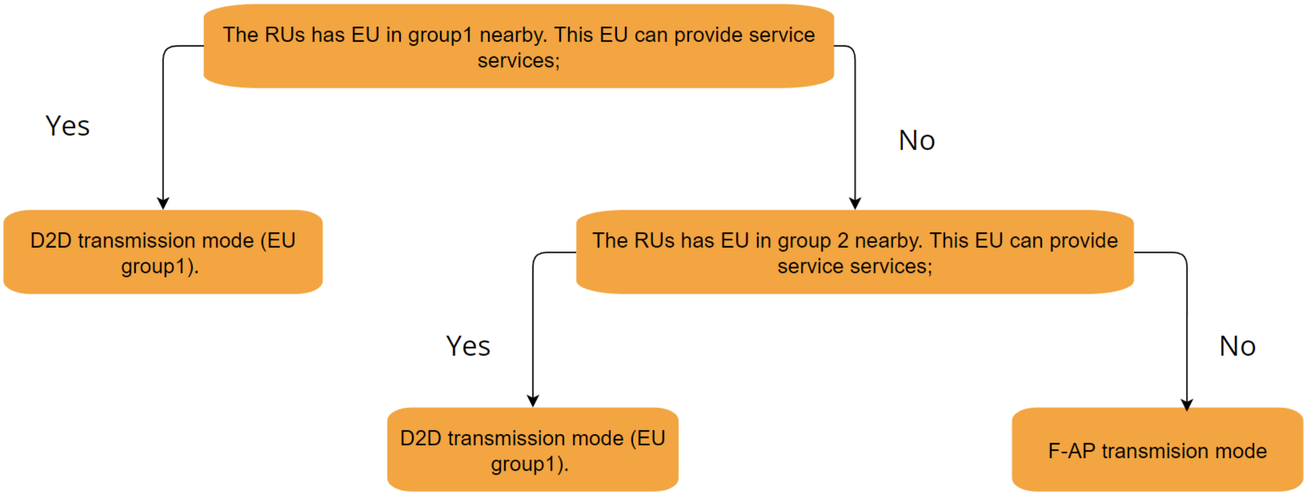
The RUs has EU in group1 nearby. This EU can provide service services;
Yes
No
D2D transmission mode (EU group1).
The RUs has EU in group 2 nearby. This EU can provide service services;
Yes
No
D2D transmission mode (EU group1).
F-AP transmision mode

Figure 5.2: Location based transmission mode selection

The relationship between cache hit probability and Zipf's law exponent β was simulated using $\gamma_E = 8 \times 10^{-4}$. As shown in Fig.5.3, the cache hit probability of EU increases as β is increased. According to the definition of β1, an increase in β concentrates the user's requests into content

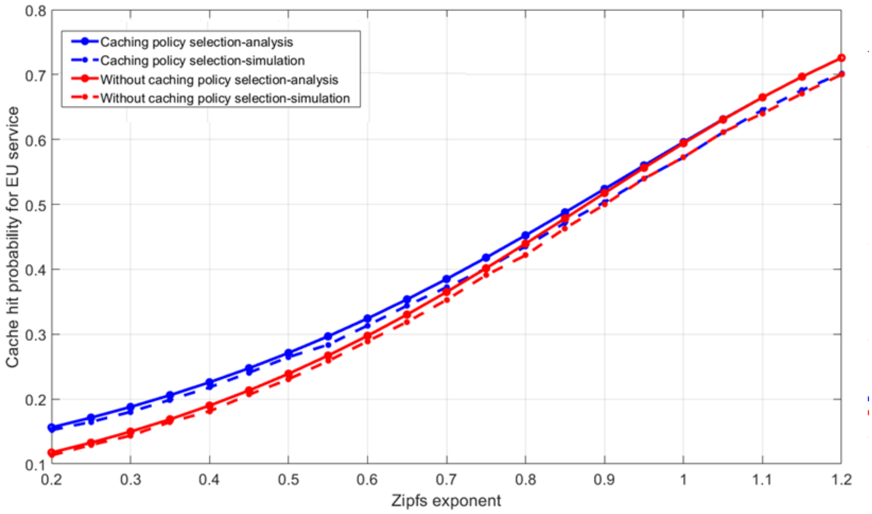

Figure 5.3: Cache hit probability for EUs vs. Zipf's exponent

set C_{E1}. The amount of popular contents being requested will be greatly increased. Therefore, the cache hit probability of the EUs increased with the higher value of Zipf's exponent β. As shown in Fig. 5.3, the proposed caching method can effectively improve cache hit probability of EUs when β is is relatively small. However, since the value of β gets larger, the lines for two caching methods will become a superposition. The EUC caching policy aims to increase the cache hit probability of the EU basing by grouping EUs reasonably. Since the value of Pr_{E1} is much higher than Pr_{E2} when β is quite large, setting up the EU group 2 will not provide optimisation for the cache hit probability. In this case, $\alpha_{op} = 1$, and the PB caching policy is selected. However, a small β leads to a small value of $Pr_{E1} - Pr_{E2}$, and the probability values of an RUs' requests from the C_{E1} and C_{E2} will be close. In this case, the two EU groups would allow for more RUs to be served directly by the EUs, thus increasing the cache hit probability.

Fig. 5.4 shows the cache hit probability and increasing intensity of EUs ($\gamma_E = 5 \times 10^{-4} \sim 10 \times 10^{-4}$) with $\beta = 0.5$. We find that the optimization effect of the proposed caching method will get better as the number of EUs increases. This is because the total service area becomes larger and more RUs can be covered by the EUs' service. Thus, more RUs can be directly served by EUs. Instead, low intensity EUs distribution will reduce the optimization of the cache hit probability of the proposed caching method.

Then, we further investigate the effect of Zipf's exponent β on the average download de-

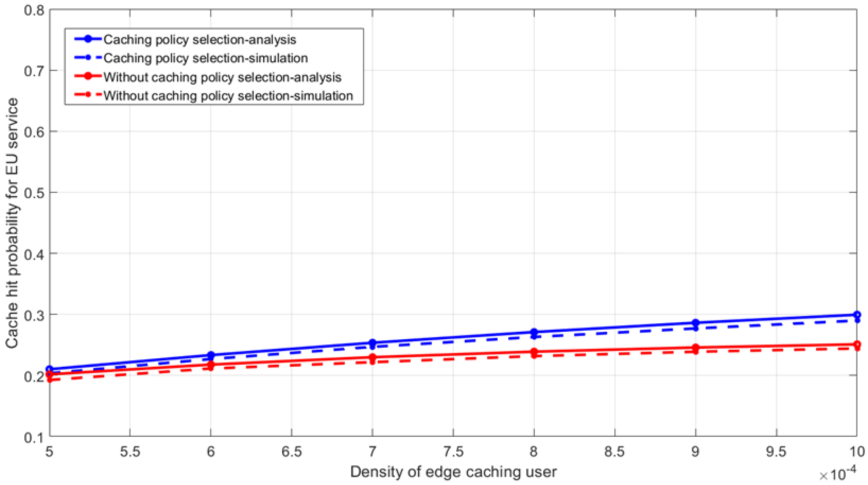

Figure 5.4: Cache hit probability for EUs vs. density of EUs

lay. Fig. 5.5 plots how the average download delay changes as Zipf's exponent β increases when α is given such that the cache hit probability is maximized. Basing on the definition in section 5.1.4, the download time delay is the sum of waiting time in queue and content service time where the time required to send the requested content from the content provider to the RU is called the service time. If the requested content is applied through F-AP transmission mode the download delay can calculated as: $Download \quad delay = 0 + \frac{s}{r_f} = 2.5 \quad (seconds)$. On the other hand, the download delay when EU is selected to provide the requested content can be expressed: $Download \quad delay = Waiting \quad time \quad in \quad queue + t_e = Waiting \quad time \quad in \quad queue + 0.2525 \quad (seconds)$, where $t_e = \frac{s_f}{r_f} + \frac{s}{r_d}$. Note that the values of r_f and r_d in Table 5.2 are easy to implement in practice [92]. Hence, compared with only using F-AP to transfer contents, it is obvious that EUs can reduce content download delay by participating in content transfer. By comparing the two different caching methods, one can see that the EUC caching policy allows more EUs to service users when the value of β is low. Thus, according to Fig. 5.5, when β less than 0.8, the average download delay can be optimized with the proposed cache hit probability optimization method. As the value of β goes increase, the distribution of user requests becomes more concentrated on popular content. This results in more EUs being allocated to content set C_{E1} than C_{E2} when using the EUC caching strategy as the value of β increases. Also, when $\alpha = 0.9$, the PB caching policy will be selected by the cache hit probability optimization method. This also affects the analysis of the average

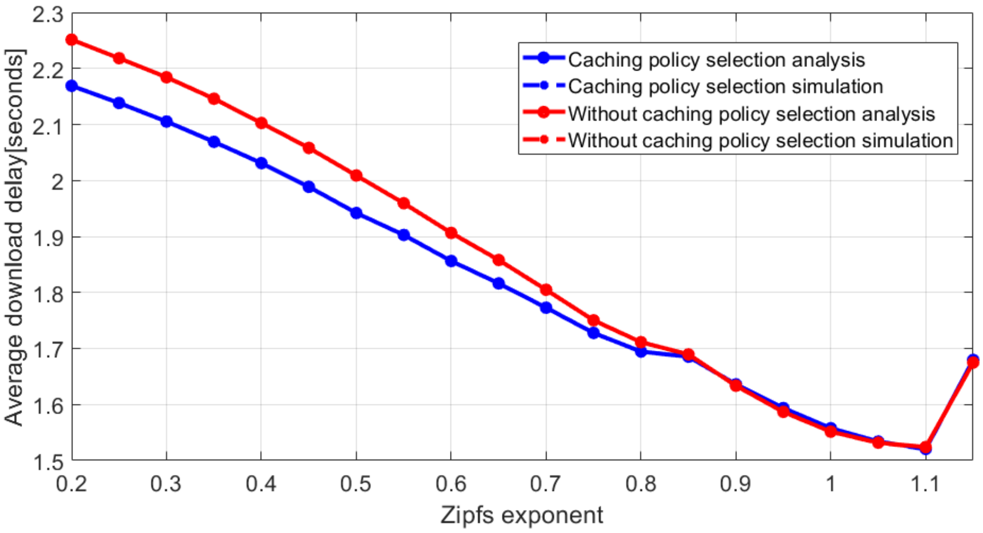

Figure 5.5: Average download delay vs. Zipf's exponent

download delay. After β is large enough, the two lines will overlap. In this case, the proposed method no longer provides download delay optimization.

As the β value increases, the probability that a random content request can be serviced by the EU increases. Fig.5.3 shows that as the Zipf exponent β increase from $\beta = 0.2$, the cache hit probability for EU service $\mathbb{P}_{HE}$ comes larger. Combining with the explanation of average request arrival rate at EU in chapter 5.2.4, the $\lambda_{EU} = K_{RF}\mathbb{P}_{HE}\lambda$ will increase with the growth of $\mathbb{P}_{HE}$. In this case, the average content arrival rate per EU goes increase and more content requirements can be served by EUs. Obviously, transferring contents directly via the EU is faster than accepting content via the F-AP. Due to this, the average download delay will decrease. However, larger value of $\mathbb{P}_{HE}$ means longer average request queue length and the longer mean waiting time W_{EU} which results in increasing average download delay when the request is served by EU. In conclusion, as the value of $beta$ increases, the number of requests that can be served by the EU service increases and the number of requests that need to be served by the F-AP service. This leads to the decreasing of the average download time. On the other hand, the increased request queue length and waiting time due to the growth of $beta$ can increase the average download delay when EU service is selected. According to Zipf's distribution, as $beta$ increases, the probability of user-generated requests being distributed in CE1 becomes higher and higher and finally approaches 1. As a result, the reduction in download latency due to reduced the usage of F-AP services will be smaller and smaller. At the same time, the effect of waiting time increase on average download delay will be larger. This results in the gradual growth of the request queue in each EU queuing system. One can see from Fig. 5.5 that, as β continues to increase after $\beta = 1.1$, the average download delay goes increase.

5.5 Conclusion

In this Chapter, we have provided a study of content caching policy for edge caching users in F-RAN system.

We first studied the content providing process for RUs which is consisted of content caching phase, content requesting phase and content downloading phase in Section 5.1. The process of providing content is described to explain the impact of a content caching policy on EUs. Then, aiming at maximizing the cache hit probability for EUs, we analyzed the cache hit probability of EUs and average download delay with two different content caching policies: the probability based caching policy content caching policy and the EU classification-based caching policy. The probability based caching policy has been used in the pervious content downloading studied. We proposed the EUC-based caching policy for EUs which is an adjustable caching policy to improve the performance. When the EUC content caching policy is selected, the number of content sets allocated for EUs during the pre-fetching phase changes to two sets.

Thus, EUs are divided into two groups. We defined α as a caching policy factor to represent the ratio of EUs numbers between the two groups. The cache hit probability of EUs can be changed by adjusting the value of α. By analyzing the EUC caching policy, we obtained an expression of α that can maximize the cache hit probability in terms of content popularity and number of EUs. The algorithm can provide the choice of caching policy for EU to optimize the cache hit probability according to the value of β, the number of EUs and RUs in the F-RAN.

Farther, in Section 5.3, we developed an algorithm for selecting the content caching policy that maximizes the cache hit probability of EUs. According to the numerical resutl, the cache hit probability is higher when β values are lower. Meanwhile, the caching policy can provide a lower average download delay. Future research will focus on developing a content caching policy select algorithm that attempts to also minimize the average download delay. In addition, the time-varying popularity of content will be considered in caching policy optimization.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

In this thesis, we have studied and analyzed performance, including average energy efficiency, cache hit probability, and average download delay, when D2D transmission and beamforming are used in an F-RAN. Energy efficiency for an F-RAN system with a D2D cluster and sub-popularity caching policy was analyzed. The mode selection for energy efficiency optimization in EU-assisted F-RAN and edge caching user classification to optimize download delay and cache hit probability was also studied.

In Chapter 3, we investigated the optimization problem of average energy efficiency for content downloading with greedy algorithm and mode selection method for users requesting content in EU-assisted F-RAN. The average energy efficiency per content request completed is the average energy efficiency. The energy consumption of the energy efficiency is the energy consumption of user transfer the content requirement to the F-AP and the energy consumption of content delivery. Compared to traditional F-RAN studies, not only fog radio access point is applied to provide contents for users, but we also considered using edge caching users to transmit requested contents directly through D2D communication. Therefore, according to the different transmission mode selection algorithm, the average energy efficiency will be different. To maximize the average energy efficiency of content downloading, we applied the greedy algorithm to optimize edge caching users allocation for D2D transmission and propose a transmission mode selection scheme for two different transmission modes. Simulation results indicated that, compared to a method that does not use the mode selection method, our method can improve energy efficiency. The optimization effect of mode selection is more obvious as the Zipf exponent increases. Furthermore, the optimization effect of increasing the number of D2D transmitters is limited by the number of RUs, but the average energy efficiency of downloaded content can still be optimized by using the mode selection method.

In Chapter 4, we studied and analyzed the cache hit probability and average energy efficiency per request for an F-RAN system that considers D2D clustered cooperative beamforming. Since D2D clustered cooperative beamforming can provide higher energy efficiency than F-AP mode and longer service range than D2D mode, it is introduced as an additional content download transmission mode for RU to select in this work to improve the average energy efficiency. In this thesis, the EU surrounded by idle D2D transmitter is called D2D cluster. Based on this, we defined a cooperative beamforming allowed EU (CEU) can send the content to a content requesting user (RU) through a D2D cluster. Meanwhile, the EUs which can only perform non-cooperative D2D communication is called D2D edge caching users (DEUs). In addition to the set of contents derived from the most popular contents, we also provide a relatively less popular set of contents called sub-popularity content set. An optional content caching policy called sub-popularity content caching policy is added for CEU during the pre-fetching phase for optimized the cache hit probability for EU. For D2D cluster, the timer-based relay selection was considered to realize cooperative beamforming in analyzing energy efficiency. By using energy efficiency expressions and timers that the relay sets internally the optimal relay selection can help D2D cluster to provide high energy efficiency with a wide service range. Since the transmission mode for content providing service can be allocated by F-AP, each user with content requirement must first send the request task to the F-AP. Also, duel to the channel estimation between content provider and corresponding content requested user is needed, in this study, energy efficiency calculation is divided into two parts: channel estimation with sending the content requirement and content transmission after transmission mode selection. Simulation results showed that, compared with PB content caching policy, our proposed policy can optimize energy efficiency when the Zipf exponent is small and the density of DEUs is high.

Since D2D clustered cooperative beamforming can consume less energy than F-AP transmission, a lower probability of mode selection for F-APs leads to better energy efficiency, as verified by simulation results. Moreover, we discovered that changing the service range of D2D clustered cooperative beamforming does not affect the choice of caching policy.

In Chapter 5, we investigated the impact of content caching policy on cache hit probability and average content download latency. The download delay is defined as the interval between receiving a content request at the F-AP and the content is delivered. Therefore, the download delay function for the content requirement is defined as *download delay = waiting time in the queue + service time.* To mitigate the impact of limited storage at each EU, we proposed an EU classification-based content caching policy for F-RAN, where EUs can be divided into two groups, each caching a different content set. EUC-based content caching policy can maximize the cache hit probability of EU by adjusting the ratio of two sets of EU using convex optimization. Then, combining with the PB caching scheme which is common use content caching policy in F-RAN study, to further optimize the cache hit probability of EU, we presented a caching policy selection algorithm that allows the F-AP to choose between the proposed EU classification-based content caching policy and the conventional probability-based caching policy. By modeling the content request queue at each EU as an independent M/D/1 queue model at EUs, we analyzed the average download delay under the proposed caching policy selection algorithm. The simulation results showed that the proposed algorithm can significantly increase the cache hit probability of EUs and reduce the average download delay for RUs.

Table 6.1: Brief presentation of the researchs in different chapters

Ch.	Transmission Modes	Analzed Parameters	Proposed Schemes
3	F-AP Mode, D2D Mode	Energy Efficiency	Greedy Algorithm Mode Selection
4	F-AP Mode, D2D Mode, Cooperative Beamforming Mode	Energy Efficiency	Cooperative Beamforming Mode, Sub-Popularity Caching Policy
5	F-AP Mode, D2D Mode	Cache Hit Probability, Transmission Delay	EU Classification Caching Policy, Cache Hit Probability- -Optimization Algorithm

6.2 Future Works

In this section, we briefly introduce several potential research directions for our future researches.

- **Dynamic content caching policy selection of EU with time-varying popularity of content in EU-assisted F-RAN**

  In this thesis, it was assumed that the time-invariant content popularity is known to F-APs when the Zipf exponent and number of content items are given. The mode selection method and content caching policy were discussed and designed under the known popularity of each item of content. In reality, the popularity of content changes over time, and the content caching policy should be adjusted accordingly. In future research, we intend to determine a reasonable edge caching policy based on the time-varying popularity of content. An online caching scheme with time-varying popularity of content has been studied, as well as online cache updating to a cloud center and cache replenishment for F-APs [20]. Most studies investigate the optimization of F-AP cache placement and generate only one optimized cache policy. Therefore, a dynamic caching policy and policy selection scheme for both F-APs and EUs with time-varying content popularity is an interesting research direction. A challenge in this study is to establish the expression of time-varying content probability. Building a scheme that adjusts the popularity of content is another challenge. The optimization of a content caching policy for both EUs and F-APs should be studied. For such challenges, we plan to take reinforcement learning as the optimization method

- **Optimization of content placement with user preference and cooperative beamforming via reinforcement learning in F-RAN**

  Different from chapters 3 and 4, we considered the case of an F-RAN system with multiple F-APs, where considering the user preferences of different scenarios when investigating content caching policy is an important learning area. An optimal edge caching policy has been proposed to maximize the cache hit probability [19], and online content popularity prediction and offline content popularity for user preference were studied. Deep

Q-learning has been used to learn the unknown content popularity, which changes with different F-APs [94]. It is a reasonable research direction that different content caching policies can be assigned to different F-APs. In such an environment, we consider not only the average download time for content across the system, but the fronthaul traffic load between the cloud and the F-AP. Considering that the content preference of the F-AP may change dynamically, the dynamic content caching policy of different F-APs also should be discussed. As the interactive relationship among F-APs needs to be defined, analysis of average download delay for the entire system also needs to capture the temporal dynamics of the F-AP and the preferences of users.

- **Average download delay optimization method**

  In the Chapter 5, we studied the optimisation method of cache hit probability at EUs based on EUC caching policy and the change of average download delay under this condition. We will focus later on the optimization of EUC content caching policy for average download delay. For calculating the average download delay for each request, according to Little's law, we have to give the average request queue length for each EU and sum it up. However, the choice of EU is an issue that needs to be discussed when the coverage of multiple EUs' services overlaps an existing RU within the overlapping scope. This is the challenge we face in averaging download delay optimization. It is essentially to review the literature from Pro. Andrews and to study the coverage probability and Voronoi tessellation [88]. Furthermore, the deep q-learning is proposed to solve this problem, where the choice of the EU can be seen as state information and the average download delay can be expressed as the reward for each state.

- **Average download delay optimization method**

  In chapter 5, we studied the optimization method of cache hit probability at EUs based on EUC caching policy and the change of average download delay under this condition. We intend to focus on the optimization of EUC content caching policy for average download delay. To calculate the average download delay for each request, according to Little's law, we must give the average request queue length for each EU and sum them. However, the choice of EU for a RU must be discussed when the coverage of multiple EUs' service is overlapped and existing this RU within the overlapping scope. This is the challenge in optimizing the average download delay. It is essential to review the literature from Andrews [88] and study the coverage probability and Voronoi tessellation. Deep Q-learning is proposed to solve this problem, where the choice of the EU can be seen as state information, and the average download delay can be expressed as the reward for each state.

Bibliography

[1] E. Ahmed, I. Yaqoob, A. Gani, M. Imran, M. Guizani, “Social-aware resource allocation and optimization for D2D communication,” *IEEE Wireless Communication.*, vol. 24, no. 3, pp. 122-129, June. 2017.

[2] A. Checko et al., “Cloud RAN for mobile networks—A technology overview,” *IEEE Commun. Surv. Tuts.*, 17, no. 1, pp. 405–426, 2015.

[3] M. Peng et al., “System architecture and key technologies for 5G heterogeneous cloud radio access networks,” *IEEE Network.*, vol.29, no. 2, pp. 6–14, Mar. 2015.

[4] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, “Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems,” *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[5] M. Peng et al., “Recent advances in underlay heterogeneous Nnetworks: Interference control, resource allocation, and self-srganization,” *IEEE Commun. Surveys and Tutorials.*, vol. 17, no. 2, pp. 700–729, Mar. 2015.

[6] M. Peng, S. Yan, K. Zhang, and C. Wang, “Fog-computing-based radio access networks: issues and challenges,” *IEEE Network* ., vol. 30, pp. 46 - 53, July. 2016

[7] M. Peng, C. Wang, V. Lau, and H. V. Poor, “Fronthaul-constrained cloud radio access networks: Insights and challenges,” *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152-160, Apr. 2015.

[8] M. Peng, Y. Li, T. Q. S. Quek, and C. Wang, “Device-to-device underlaid cellular networks under Rician fading channels,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4247-4259, Aug. 2014.

[9] F. Bonomi et al., “Fog Computing and its Role in the Internet of Things,”*Proc. Wksp. Mobile Cloud Computing.*, Helsinki, Finland, pp.13–16, Aug. 2012.

[10] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, “Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy effciencies,” *IEEEWireless Commun.*, vol. 21, no. 6, pp. 126-135, Dec. 2014.

[11] E. Bas¸tu˘g, M. Bennis, and M. Debbah, “Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks,”*IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[12] N. Golrezaei, A. Dimakis, and A. Molisch, “Scaling Behavior for Deviceto-Device Communications With Distributed Caching,”*IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, July 2014.

[13] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis,“Base-Station Assisted Device-to-Device Communications for High- Throughput Wireless Video Networks,”*IEEE Trans. Wireless Commun.*, vol. 13, no. 7, July 2014.

[14] S. Yan, M. Peng, M. Abana, and W. Wang, “An evolutionary game for user access mode selection in fog radio access networks,”*IEEE Trans. Wireless Commun.*, vol. 5, pp.2200-2210, Jan. 2017.

[15] X. Wang, S. Leng, K. Yang, “Social-aware edge caching in fog radio access networks,” *IEEE Access*, vol. 5, pp. 8492-8501, April. 2017.

[16] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, “Web caching and Zipf-like distributions: Evidence and implications,”*Proc. IEEE INFOCOM.*, New York, NY, USA, pp. 126-134, Mar. 1999.

[17] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, “I Tube, You Tube, Everybody Tubes: Analyzing The World’s Largest User Generated Content Video System,” *in Proc. ACM SIGCOMM Internet Measurement,* Oct. 2007.

[18] L.Lu, Y.Jiang, M.Bennis, Z.Ding, F.Zheng, and X.You, “Distributed Edge Caching via Reinforcement Learning in Fog Radio Access Networks,” *IEEE 89th Vehicular Technology Conference (VTC2019-Spring).*, May. 2019.

[19] Y. Jiang, M. Ma, M. Bennis, F. Zheng, and X. You, “A Novel Caching Policy with Content Popularity Prediction and User Preference Learning in Fog-RAN,” *IEEE Globecom Workshops (GC Wkshps).*, Dec. 2017.

[20] S.Azimi, O.Simeone, A.Sengupta, and R.Tandon, “Online Edge Caching and Wireless Delivery in Fog-Aided Networks With Dynamic Content Popularity,” *IEEE Journal on Selected Areas in Communications.*, vol. 36, pp.1189-1202, June. 2018.

[21] J.Liu, B.Bai, J.Zhang, K.Letaief, “Cache Placement in Fog-RANs: From Centralized to Distributed Algorithms,” *IEEE Transactions on Wireless Communications.*, vol. 16, pp.7039-7051, Aug. 2017.

[22] Y.Hu, Y. Jiang, M.Bennis, and F.Zheng, “Distributed Edge Caching in Ultra-Dense Fog Radio Access Networks: A Mean Field Approach,” *IEEE 88th Vehicular Technology Conference (VTC-Fall).*, ISSN. 2577-2465, Aug. 2018.

[23] I.Althamary, C.Huang, P.Lin, S.Yang, and C.Cheng, “Popularity-Based Cache Placement for Fog Networks,” *International Wireless Communications & Mobile Computing Conference (IWCMC).*, June. 2018.

[24] M. Tao, E. Chen, H. Zhou, W. Yu, “Content-centric sparse multicast beamforming for cache-enabled cloud RAN,” *IEEE Transactions on wireless communications*, vol. 15, no. 9, pp. 6118-6131, June. 2016.

[25] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, “Cluster content caching: An energy-efFcient approach to improve quality of service in cloud radio access networks,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207-1221, May. 2016.

[26] Y.Duan, J.Zhang, Member, and W.Xia “Energy Efficiency of Downlink C-RAN With Edge Caching and Fronthaul Compression,” *IEEE COMMUNICATIONS LETTERS.*, vol. 22, no. 12, pp.2527-2530, Dec.2018.

[27] D.Feng, L.Lu, Y.Yuan-Wu, G.Y.Li, S.Li, and G.Feng, “Device-to-device communications in cellular networks,” *IEEE Communications Magazine.*, vol. 52, no. 4, pp. 49–55, April 2014.

[28] J.Liu, N.Kato, J.Ma, K.Kadowaki, “Device-to-device communication in lteadvanced networks: A survey,” *IEEE Communications Surveys Tutorials.*, vol. 17, no. 4, pp. 1923–1940, Fourthquarter 2015.

[29] X.Hu, K Wong, and K.Yang “Wireless powered cooperation-assisted mobile edge computing,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2375–2388, Apr. 2018.

[30] G.Y.Li, Z.Xu, C.Xiong, C.Yang, S.Zhang, Y.Chen, and S.Xu, “Energy-efficient wireless communications: tutorial, survey, and open issues,” *IEEE Wireless Communications.*, vol. 18, no. 6, pp. 28–35, December 2011.

[31] X.Chen, M.Zhao, and Y.Cai, “Energy efficient content-centric beamforming in multicast fog radio access network,” *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP).*, Oct. 2017.

[32] Q.Li, Z.Xiang, P.Ren, and W.Li,“Content-Centric Transmission Design in Fog Radio Access Network With Partition-Based Caching,” *IEEE Access.*, vol. 7, pp.181994-182003, Dec. 2019.

[33] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, “On content-centric wireless delivery networks,” *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 118–125, Dec. 2014.

[34] Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2017-2022,”White Paper, Feb. 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/ visual-networking-index-vni/white-paper-c11-738429.html

[35] P. Rost et al., “Cloud technologies for flexible 5G radio access networks,” *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[36] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, “Multi-cell MIMO cooperative networks: A new look at interference,” *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[37] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, “Cache in the air: Exploiting content caching and delivery techniques for 5G systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[38] M.Lee, A. F. Molisch, “Caching Policy and Cooperation Distance Design for Base Station-Assisted Wireless D2D Caching Networks: Throughput and Energy Efficiency Optimization and Tradeoff” *IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS*, vol. 17, no. 11, pp. 7500–7514, Nov. 2018.

[39] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief. (Sep. 2015). “Backhaulaware caching placement for wireless networks.” [Online]. Available: https://arxiv.org/abs/1509.00558

[40] M. Tao, D. Gunduz, F. Xu and J. S. P. Roig, “Content caching and delivery in wireless radio access networks”, *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 4724-4749, Jul. 2019.

[41] M. Tao, E. Chen, H. Zhou, and W. Yu, “Content-centric sparse multicast beamforming for cache-enabled cloud RAN,” *IEEE Trans. Wireless Commun.*, to be published.

[42] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, “Hypergraph based analys is of clustered co-operative beamforming with application to edge caching,” *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 84–87, Feb. 2016.

[43] R. Tandon and O. Simeone, “Fog radio access networks: Fundamental latency tradeoffs,” in *Proc. IEEE Inf. Theory Appl. Workshop (ITA),* San Diego, CA, USA, Jan. 2016, pp. 1–5.

[44] S.-H Park, O. Simeone and S. Shamai, “Joint optimization of cloud and edge processing for fog radio access networks”, *2016 IEEE International Symposium on Information Theory (ISIT),* vol. 15, no. 11, pp. 7621-7632, Nov. 2016.

[45] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, “Recent Advances in Cloud Radio Access Networks: system architectures, key techniques, and open Issues,”*IEEE Commun. Surveys Tuts.*, pp. 1-1, Mar. 2016.

[46] M. Peng, C. Wang, V. Lau, and H. V. Poor, “Fronthaul- constrained cloud radio access networks: insights and challenges,”*IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152-160, Apr. 2015.

[47] F. Xu, K. Liu, and M. Tao, “Cooperative Tx/Rx caching in interference channels: A storage-latency tradeoff study,” *in Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, pp. 2034–2038, Jul. 2016.

[48] T. Han and N. Ansari, “User association in backhaul constrained small cell networks,”*Proc. IEEE WCNC 2015*, United States of America, pp. 1637-1642, Mar. 2015.

[49] L. Wei, R. Hu, Y. Qian, and G. Wu, “Enable Device-to-Device communications underlaying cellular networks: challenges and rsearch aspects,”*IEEE Communications Magazine.*, pp.90-96, June. 2014.

[50] Y. Zhang, Y. Xuy, M. Gao, Q. Zhang, H. Li, I. Ahmadz, and Z. Feng, “Resource Management in Device-to-Device Underlaying Cellular Network,”*IEEE Wireless Communications and Networking Conference.*, pp.1631-1636, March. 2015.

[51] Y. Mo, M. Peng, H. Xiang, Y. Sun, X. Ji, “Resource Allocation in Cloud Radio Access Networks with Device-to-Device communications,”*IEEE Access. SPECIAL SECTION ON DEPLOYMENT AND MANAGEMENT OF SMALL HETEROGENEOUS CELLS FOR 5G*, vol, 5, pp. 1250-1262, February. 2017.

[52] Z. Zhou, M. Dong, K. Ota, G. Wang, L. Yang, “Energy-Efficient resource allocation for D2D communications underlaying Cloud-RAN-Based LTE-A Networks,”*IEEE Internet of Things Journal.*, vol. 3, no. 3, pp. 428-438, November. 2015.

[53] A. Nosratinia, T. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol. 42, pp. 68–73, 2004.

[54] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity–part I: system description," *IEEE Trans. Wireless Commun.*, vol. 51, pp. 1927–1938, 2003.

[55] B. Rankov and A. Wittneben, "Distributed spatial multiplexing in a wireless network," *Proc. Asilomar Conf. on Signals, Systems, and Computers.*, pp. 1932–1937, 2004.

[56] A. E. Khandani, J. Abounadi, E. Modiano, and L. Zheng, "Cooperative routing in wireless networks," *Proc. Allerton Conf. on Commun., Control and Computing*, 2003.

[57] H. Ochiai, P. Mitran, H. V. Poor, and V. Tarokh "Collaborative beamforming for distributed wireless ad hoc sensor networks," *IEEE Trans. Signal Processing,*, vol. 53, pp. 4110–4124, Nov. 2005.

[58] A. Afana, V. Asghari, A. Ghrayeb, S. Affes, "Cooperative relaying in spectrum-sharing systems with beamforming and interference constraints," *2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications.*, pp. 429-433, Sept. 2012.

[59] A. Afana, A. Ghrayeb, V. Asghari, S. Affes, "Cooperative two-way selective relaying in spectrum-sharing systems with distributed beamforming," *Wireless Communications and Networking Conference (WCNC), 2013 IEEE.*, pp. 2976-2981, July. 2013.

[60] A. Wittneben, I. Hammerstroem, and M. Kuhn, "Joint cooperative diversity and scheduling in low mobility wireless networks," *Proc. Globecom.*, pp. 780–784, 2004.

[61] F. Onat et al., "Threshold selection for SNR-based selective digital relaying in cooperative wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4226–4237, Nov. 2008.

[62] Z. Zhou, S. Zhou, J.-H. Cui, and S. Cui, "Energy-efficient cooperative communication based on power control and selective single-relay in wireless sensor networks," *Wireless Communications, IEEE Transactions on.*, vol. 7, no. 8, pp. 3066–3078, August 2008.

[63] N. C. Beaulieu and J. Hu, "A closed-form expression for the outage probability of decode-and-forward relaying in dissimilar Rayleigh fading channels,"*IEEE Commun. Lett.*, vol. 10, no. 12, pp. 813–815, Dec. 2006.

[64] L. Li, M. Peng, C. Yang, and Y. Wu, "Optimization of base-station density for high energy-efficient cellular networks with sleeping strategies,"*IEEE Trans. Wireless Commun.*, vol. 65, no. 9, pp. 7501–7514, Sep. 2016.

[65] L. Wei, R. Q. Hu, Y. Qian, and G. Wu,L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Energy efficiency and spectrum efficiency of multihop device-to-device communications underlaying cellular networks,"*IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 367–380, Jan. 2016.

[66] A. Ibrahim, A. Sadek, W. Su, and K. Liu, "Cooperative communications with relay-selection: When to cooperate and whom to cooperate with?,"*IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2814–2827, Jul. 2008.

[67] A. Bletsas, A. Lippnian, and D. Reed, "A simple distributed method for relay selection in cooperative diversity wireless networks, based on reciprocity and channel measurements,"*Proc. IEEE 61st VTC—Spring*, vol. 3, pp. 1484–1488, May. 2005.

[68] J. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks,"*IEEE Trans. Inf. Theory.*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.

[69] R.Madan, N.Mehta, A.Molisch, and J.Zhang, "Energy-efficient cooperative relaying over fading channels with simple relay selection,"*IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3013–3025, Aug. 2008.

[70] Y. Zhang and R. Cheng, "Relay subset selection in cooperative systems with beamforming and limited feedback," *Wireless Communications, IEEE Transactions on.*, vol. 12, no. 10, pp. 5271–5281, Oct. 2013.

[71] B.Klaiqi, X.Chu, and J.Zhang, "Energy-efficient cooperative beamforming using timer based relay subset selection," *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, pp. 369–374, Mar. 2015,

[72] B.Klaiqi, X. Chu,and J. Zhang, “Energy efficiency of location-aware clustered cooperative beamforming without destination feedback,” *2015 IEEE International Conference on Communications (ICC).*, pp.2295-2300, June. 2017.

[73] A.K. Sadek, W. Su,and K.J.R. Liu, “Cluster cooperative communication in wireless networks,” *GLOBECOM ’05. IEEE.*,vol. 3, Dec. 2005.

[74] C. Choi, S. Park, D. Cho, “User-Cooperation Scheme based on Clustering for Energy Efficiency in Cellular Networks with D2D Communication,” *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC).*, pp.1365-1369, June. 2015.

[75] R. Ansari, S. Hassan, C. Chrysostomou “Energy Efficient Relay Selection in Multi-hop D2D Networks,” *2016 International Wireless Communications and Mobile Computing Conference.*, pp. 2376-6506, September. 2016.

[76] 3GPP TR 23.703 Version 0.4.1, “Study on architecture enhancements to support Proximity Services (ProSe),” *3GPP Technical Report,* June 2013.

[77] J.Y. Pan, M. Hsu, “Relay selection of relay-assisted Device-to-Device and uplink communication underlying cellular networks,” *Computing, Networking and Communications (ICNC), 2017 International Conference on* March. 2017.

[78] R. Ma, Y. Chang, H. Chen, C. Chiu, “On Relay Selection Schemes for Relay-Assisted D2D Communication in LTE-a systems,” *IEEE Vehicular Technology Society,*. March. 2017.

[79] W. Han, A. Liu, and V.t K. N. Lau, “PHY-caching in 5G wireless networks: design and analysis” *IEEE Communications Magazine*, vol. 54, no. 8, pp. 30-36, Aug. 2016.

[80] W.Teng, M.Sheng, K.Guo, and Z.Qiu, “ContentPlacement and User Association for Delay Minimization in Small Cell Networks,” *IEEE Transactions on Vehicular Technology* ., vol. 68, pp.10201-10215, Oct. 2019.

[81] 3GPP, *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation.* 3rd Generation Partnership Project (3GPP), TS 36.211 v10.0.0,, 2011.

[82] A. S. Avestimehr and D. N. C. Tse, "Outage Capacity of the Fading Relay Channel in the Low-SNR Regime,"*IEEE Transactions on Information Theory.*, vol. 53, pp.1401-1415, April. 2007.

[83] A. S. Avestimehr and D. N. C. Tse, "Outage-optimal relaying in the low SNR regime", *Proc. IEEE Int. Symp. Information theory*, pp. 941-945, 2005-Sep.

[84] L G.Lim, and L.J. Cimini, Jr., " Energy-Efficient Cooperative Beamforming in Clustered Wireless Networks," *IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.*, vol. 12, no. 3, pp.1376-1385, Mar. 2013.

[85] H.Min, J.Lee, S.Park, and D.Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks,"*IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp.3995-4000, Dec.2011.

[86] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks,"*IEEE Trans. Wireless Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[87] S.Cui, A.Goldsmith, and A.Bahai, "Energy-constrained modulation optimization,"*IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2349–2360, Sep. 2005.

[88] J.G.Andrews, F.Baccelli, and R.K.Ganti, "A tractable approach to coverage and rate in cellular networks,"*IEEE Trans. Commun.*, vol.59, no. 11, pp. 3122-3134, Nov. 2011.

[89] J. F. C. Kingman, "The first Erlang century—and the next,"Queueing Systems. 63: 3–4. doi:10.1007/s11134-009-9147-4. 2009.

[90] Robert S. R. Cahn, "Wide Area Network Design:Concepts and Tools for Optimization,"Morgan Kaufmann. p. 319. ISBN 1558604588. 1998.

[91] L. Kleinrock, *Queueing Systems*, vol. 1: Theory, vol. 2: Computer applications. Hoboken, NJ, USA: Wiley, 1975.

[92] S. Lederer, C. Müller, and C. Timmerer, "Dynamic adaptive streaming over HTTP dataset,"in *Proc. ACM 3rd Multimedia Syst. Conf.*, 2012, pp. 89–94.

[93] M. Lee, and A. F. Molisch, "Caching Policy and Cooperation Distance Design for Base Station-Assisted Wireless D2D Caching Networks: Throughput and Energy Efficiency Optimization and Tradeoff," *IEEE Transactions on Wireless Communications.*, vol.17, pp. 7500-7514, Nov. 2018.

[94] Y.Zhou, S.Yan, and M.Peng, "Content Placement with Unknown Popularity in Fog Radio Access Networks," *2019 IEEE International Conference on Industrial Internet (ICII).*, pp.361-367, Nov. 2019.