Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

UNIVERSITY OF DUBLIN TRINITY COLLEGE

PHD THESIS

Joint caching and communication for future wireless networks

Author: Ramy Amer

Supervisor: Nicola Marchetti

Co-Supervisor: M. Majid Butt

This this is submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

CONNECT Centre for Telecommunications Research,
Department of Electronic & Electrical Engineering

March 18, 2021

Declaration of Authorship

I, RAMY AMER, declare that this thesis titled, "Joint caching and communication for future wireless networks" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Abstract

Recently, caching the most popular content at network edges has emerged as a promising technique to avoid serving all requests from the core network through highly congested backhaul links. From the caching perspective, there are multiple types of network architectures, namely, caching on femtocells in small cell networks, caching on remote-radio-head (RRH) in cloud-radio-access-networks (CRANs), caching for mobile devices, and caching for unmanned aerial vehicle (UAV) networks. This thesis conducts a comprehensive study, analysis, and optimization of the problem of content caching on terrestrial mobile devices in various works, as well as aerial users (i.e., UAV-user equipments (UAV-UEs)).

In particular, in the first part of this thesis, we study the joint caching and communication problem for cache-enabled device-to-device (D2D) network. Key performance indicators such as offloading gain, throughput, average delay, energy consumption, and spectral efficiency are then characterized and optimized. For instance, We propose a novel D2D collaborative caching architecture, where popular files are cached in the users' surplus memory and then shared with others, either neighboring users in the same proximity or remote users in the same cell (denoted as cooperation). We show that allowing such content sharing helps reduce the network average delay per request. We also investigate the average per-request throughput for different caching schemes and conducted the scaling analysis for the average sum throughput. We then proposed a joint communication and caching optimization framework for clustered D2D networks. Joint channel access and probabilistic caching for optimizing the offloading gain and energy consumption is first conducted. We then proposed a novel joint bandwidth partitioning and caching scheme to minimize the average delay per content request. Furthermore,

we extended our work to study and optimize probabilistic caching for clustered D2D caching networks whose devices undergoing coordinated multipoint (CoMP) transmissions.

In the second part of the thesis, we turn our attention to a new wireless network architecture where ground base stations (BSs) deliver content and provide coverage to contemporary aerial users. We first propose a content delivery network for co-existing ground and UAV-UE. We particularly investigate the use of beamforming for simultaneous content delivery to an aerial user co-existing with multiple ground users. We then study the performance of aerial users under three-dimensional (3D) practical antenna configurations. We consider both static and mobile aerial users and characterize their performance. Finally, we investigate the use CoMP transmission along with caching to provide seamless connectivity to aerial users. We consider a network of clustered cache-enabled small BSs (SBSs) serving aerial users where requested content is cooperatively transmitted from collaborative ground SBSs. Scenarios with static and mobile aerial users are also considered.

Acknowledgements

I am indebted to my supervisors, Nicola Marchetti and Majid Butt, for their support and guidance over the previous four years. I have learnt so much during the PhD from them both, much of which extends far beyond academia. In particular, I wish to express my gratitude to Nicola for initially inducting me into the research world and introducing me to CONNECT.

Without question, the greatest aspect of doing this PhD has been the fantastic people with whom I have shared this journey. Although many people passed through the doors of CONNECT during my four years, and each made an impression in their own way, there are a few who deserve a special mention. Thank you to everyone for the memories: Indrakshi, Sandip, Harleen, Parna, Conor, Merim, Andrea, Andrei, Boris,, Andrew, Diarmuid, Stefan, Yi, Elma, Tom, both Lindas, Georgios, Marcelo, Marco, Irene, Francisco, Pedro, Harun, Jonathan, Joao, Nima, Erika, Maicon, Alan, and so many more. Finally thank you also to all the administrative staff, and in particular Catherine, for assisting with all my clueless enquiries.

I also owe a huge thank you to my friends and collaborators Dr Hesham Elsawy and Jacek Kibilda, who helped me find my feet in the second part of my PhD and left a lasting impression. Thank you to Ahmed Selim for his unquestionable knowledge (and questionable humour). A special thank you to Amr Elrasad, who sincerely helped me from day one in Dublin till the end of my PhD journey.

I would like to thank my parents for always being there. This PhD is the product of the never-ending encouragement they have always provided, and would not have been possible without them. Thank you also to my siblings and my cousins for spurring me on and giving me the belief to accomplish this. And last, but certainly not least, a huge thanks to my loving (and incredibly patient) wife and baby, Maha Amer and Ahmed Amer, who provided me with endless support and believed in me when I did not. I could not have done this without you both.

Contents

List of Figures

List of Tables

List of Acronyms

MNO mobile network operator

SDN software defined networking

QoS quality-of-service

LTE Long Term Evolution

HetNet heterogeneous networks

MAC medium access control

MIMO multiple-input multiple-output

UE user equipment

BS base station

mmWave millimeter wave

LoS line-of-sight

NLoS non-line-of-sight

NFV network function virtualization

C-RAN cloud radio access network

RAN radio access network

BBU baseband unit

RRH remote radio head

CSI channel state information

CoMP coordinated multipoint

D2D device-to-device

ICIC inter-cell interference coordination

PDF probability distribution function

KPI key performance indicator

SBS small base station

MBS macro base station

SCN small cell network

FIFO first in first out

VCC virtual cache center

UAV unmanned aerial vehicles

MPSQ multiclass processor sharing queue

EE energy efficiency

SIR signal-to-interference ratio

SINR signal-to-interference-plus-noise ratio

PPP Poisson point process

PCP Poisson cluster process

HCP hard-core placement

TCP Thomas cluster process

CPF caching popular files

GCA greedy caching algorithm

RC random caching

PC probabilistic caching

5G fifth generation

MEC mobile edge computing

AP access point

VoD video-on-demand

QoE quality-of-experience

CDN content delivery networks

F-RAN fog-radio access network

AR augmented reality

4C computing, caching, communication, and control

ABR Adaptive BitRate

ILP Integer Linear Program

UDN ultra-dense networks

ETSI European Telecommunications Standards Institute

PMF probability mass function

RV random variable

i.i.d. independently and identically distributed

CDF cumulative distribution function

PGFL probability generating functional

KKT Karush-Kuhn-Tucker

PGF point generating function

BPP binomial point process

IoT Internet-of-things

3D three-dimensional

2D two-dimensional

1D one-dimensional

CB conjugate beamforming

AU aerial user

GU ground user

CLT central limit theorem

SE spectral efficiency

MRT maximum ratio transmission

ZFBF zero-forcing beamforming

GBS ground base station

RWP random waypoint

AGL above ground level

w.r.t. with respect to

UB upper bound

LB lower bound

SCDP successful content delivery probability

HARP highest average received power

ULA uniform linear array

GPP Gaussian Poisson process

NCP nearest content provider

RSCP randomly-selected content provider

BCD block coordinate descent

3G third generation

4G fourth generation

For/Dedicated to/To my…

Chapter 1

Introduction

Unlike previous generations, which were primarily defined by their approach to the air interface and multiple access scheme (i.e. third generation (3G): Universal Mobile Telecommunications Service (UMTS)/Wideband Code Division Multiple Access (WCDMA) and fourth generation (4G): Long Term Evolution (LTE)/Orthogonal Frequency Division Multiple Access (OFDMA)), upcoming telecommunication networks, i.e., fifth generation (5G), are envisioned to be a very different type of network. This difference can be largely attributed to their capability of bringing the network infrastructure and application services at the edge, i.e., very close to the end users, e.g., small base stations (SBSs) in small cell networks (SCNs) and edge caching and task computation for content delivery and edge computing services, respectively.

1.1 Scope

The recent design improvement of mobile networks and devices has substantially enriched the mobile user experience, leading to a vast range of new wireless services, including edge caching, video-on-demand (VoD) streaming, web browsing, and social networks. This phenomenon has been further dominated by mobile video streaming, which currently accounts for almost 50 percent of mobile data traffic with an estimate of a 500-fold increase over the next few years [1]. This new challenge has encouraged mobile network operators (MNOs) to redesign their current networks into more advanced and sophisticated ones that can increase coverage, boost capacity, and bring contents near to the users in a cost effective way.

1.2 Outline

We divide the remainder of this thesis into eight chapters. Chapter 2 provides the foundation for the thesis by providing the setting and motivation for the vision of wireless caching for future mobile networks, and we survey opportunities and trends which pave the way towards this vision. Chapter 3 then examines the benefits of inter-cluster cooperation for D2D caching networks. Chapters 4 and 5 focus on the performance analysis and joint caching and communication for clustered D2D caching networks under a physical interference model.

In the second part of the thesis, i.e., particularly, Chapters 6-8, we turn our attention to a new network architecture wherein the content delivery and caching services are aimed towards the contemporary aerial users (also known as cellular-connected UAVs). Finally, Chapter 9 summarizes our main findings and concludes the thesis.

Next, we briefly summarize the work presented in each of the following chapters.

1.2.1 Chapter 2

In Chapter 2, we define the concept of wireless caching at the network edge, and review the diverse range of architectures and content caching schemes proposed in the literature. We then survey some of the architectures that combined edge caching and unmanned aerial vehicles (UAV) networks. Furthermore, we elaborate on the main challenges in the ongoing research pertaining wireless caching in future 5G networks. We then give the main motivation behind the work conducted in this thesis, and show how it fills a gap in the literature and solves several challenges pertaining the adoption of wireless caching for D2D networks and content delivery for networks with aerial users.

1.2.2 Chapter 3

In Chapter 3, we propose a novel inter-cluster cooperative architecture for D2D caching. We study a cellular network in which mobile devices can cache popular content and share it in the same proximity via D2D communication or using cellular

transmission if they are far apart. We characterize the network average delay per request from a queuing perspective and formulate the delay minimization problem. We then obtain a suboptimal optimal caching solution which is proven to be locally optimal within a factor ≈ 0.63 of the optimum.

1.2.3 Chapter 4

In Chapter 4, we present a comprehensive performance analysis and joint communication and caching optimization for a clustered D2D network. We assume mobile devices with a surplus memory which is exploited to proactively cache files from a known library, following a random probabilistic caching scheme. We optimize three key performance metrics, namely, offloading gain, energy consumption, and latency. We first maximize the offloading gain of the proposed network by jointly optimizing caching probability and channel access. We further formulate and solve the energy minimization problem and obtain the optimal caching scheme for the minimum energy consumption. Finally, we jointly optimize the caching scheme as well as bandwidth partitioning to minimize the weighted average delay per file request. We obtain closed-form solution for the bandwidth allocation and suboptimal solution for the caching sub-problem is also provided.

1.2.4 Chapter 5

In Chapter 5, we show the impact of cooperative communication on the performance of cache-enabled D2D networks. We develop a novel mathematical model based on stochastic geometry and an optimization framework for cache-assisted coordinated multi-point (CoMP) transmissions with clustered devices. We assume that desired contents that are not self-cached can be obtained via D2D CoMP transmissions from neighboring devices or, as a last resort, from the network. We characterize the offloading gain and rate coverage probability as functions of the system parameters. We show that cooperative transmission becomes more appealing in denser D2D caching networks and adverse interference conditions, which is the case of the imminent Internet-of-things (IoT) and massive machine type communications era.

1.2.5 Chapter 6

Motivated by the increasing importance of contemporary aerial users (also known as drone users or UAV-user equipments (UEs)), in Chapter 6, we extend our discussion to content delivery for co-existing ground and UAV-UEs. We investigate the use of conjugate beamforming (CB) for simultaneous content delivery to an UAV-UE co-existing with multiple ground users. We particularly considered a content delivery network of uniformly distributed massive multiple-input multiple-output (MIMO)-enabled ground base stations (BSs) serving both aerial and ground users through spatial multiplexing. We then investigate the effects of various system parameters such as antenna down-tilt angle, number of scheduled users, and number of antennas on the achievable performance.

1.2.6 Chapter 7

In Chapter 7, we study the performance of cellular-connected UAV-UEs under three-dimensional (3D) practical antenna configurations. We consider two scenarios, ones with static, hovering UAV-UEs and scenarios with mobile UAV-UEs. For both scenarios, we characterize the UAV-UE coverage probability as a function of the system parameters such as the number of antenna elements, density of BSs, and speed of UAV-UEs. and investigate the effects of the number of antenna elements on the UAV-UE achievable performance.

1.2.7 Chapter 8

In Chapter 8, we investigate the use of coordinated multipoint (CoMP) transmission along with caching to provide seamless connectivity to aerial users. We consider a network of clustered cache-enabled SBSs serving aerial users in which a requested content by an aerial user is cooperatively transmitted from collaborative ground SBSs. Scenarios with static and mobile UAV-UEs are considered. Under a maximum ratio transmission, we propose a novel framework that is then leveraged to derive upper and lower bounds on the UAV-UE coverage probability for both scenarios. For mobile UAV-UEs, we showed that not only the spatial displacements of UAV-UEs but also their vertical motions affect their handover rate and

coverage probability. Particularly, UAV-UEs that have frequent vertical movements and high direction switch rates, i.e., the rate at which an UAV-UE changes the direction of movement, are expected to have low handover probability and handover rate.

1.2.8 Chapter 9

Finally, in Chapter 9, we summarize the main insights of the thesis, and discuss the outlook and future directions for edge caching and computing-enabled future wireless networks.

Further detailed discussions will follow in the next chapters of this thesis. For the sake of an organized presentation, we postpone the detailed discussion of the related work to the respective technical chapters. That is, for each type of network addressed in this report, the related work is presented and the novelty is highlighted.

1.3 Contributions

In this section, we present the main research contributions made as part of this thesis. We state our general contribution in this thesis to be one of providing comprehension through the presentation of conceptual insights, as well as analytical and simulation-based results, of the challenges associated with enabling wireless caching networks. We group our contributions based on the chapter in which they appear.

1.3.1 Chapter 2

In this chapter, our main contribution is the provision of a comprehensive survey of the demands of future networks, the tools and techniques available to satisfy those demands, and the opportunities to deploy cache-enabled wireless networks. The contributions in this chapter relate to providing a vision for future wireless caching networks and highlighting the key research challenges for the adoption of such networks.

1.3.2 Chapter 3

Our chief contribution in this chapter lies in proposing and highlighting the benefits of inter-cluster cooperation for D2D cache-enabled networks in terms of the network latency and system throughput. The main technical contributions of this chapter are [2–4].

1.3.3 Chapter 4

The main contribution of this chapter is proposing a novel framework for joint content caching and communication for clustered D2D networks. The proposed framework is then leveraged to optimize several network and application key performance indicators (KPIs) such as latency, offloading gain, energy consumption, and rate coverage probability. The main technical contributions of this chapter are [5–7].

1.3.4 Chapter 5

Our contribution in this chapter relates to the adoption of cooperative transmission and content caching optimization for clustered cache-enabled D2D networks. While several papers have examined similar D2D caching architectures, our work is distinguished by a novel analysis and optimization for practical D2D caching model with the notion of device clustering and the underlying medium access control (MAC) and physical layers factored in. The main technical contributions of this chapter are [8, 9].

1.3.5 Chapter 6

In this chapter, we presented a novel framework for content delivery to one aerial user co-existing with multiple ground users. We have proposed MIMO beamforming for spatially multiplexing one aerial user and multiple ground users. The trade-off between the performance of aerial users and their ground counterparts and the effect of the antennas' down-tilt angle are investigated. The main technical contribution of this chapter is [10].

1.3.6 Chapter 7

Our main contribution in this chapter relates to studying the performance of vertically mobile aerial users under 3D antenna models. A new concept of *altitude handover* is quantified for the contemporary aerial users. The main technical contributions of this chapter are [11, 12].

1.3.7 Chapter 8

Our chief contribution in this chapter relates to proposing novel cache-assisted CoMP transmission model for aerial users while factoring in their inevitable 3D mobility model. Novel analysis is adopted to characterize key performance aspects of cellular-connected UAVs such as coverage probability, handover rate and handover probability. The main technical contributions of this chapter are [13, 14].

Other collaborative works in the areas of Terahertz and UAV communications are in [15–17].

1.4 Dissemination

This section lists accepted and under-review papers written during the PhD project. First, I state my contribution and the roles and contributions of the other authors. For the journal publications from (1) to (8), I led these works, defined the ideas and conducted all the analysis, carried out the literature survey and simulation, and wrote these papers. My supervisors Nicola Marchetti and Majid Butt helped me define the problems we solve and justify the importance of these problems and so on. For the other authors, in journals (1), (2), and (5), for example, our collaborator Walid Saad helped me define and formulate the problems to solve as he is an expert in the area of drone networks. The same applies for journals (4) and (6) where our collaborator Mehdi Bennis contributed and gave some advice in the area of wireless caching as he is an expert in this field. In Journal (6), our collaborator Eduard Jorswieck helped me in solving the different optimization problems. Finally our collaborators Hesham ElSawy and Jacek Kibilda helped me in the stochastic geometry analysis as they gave me very useful advice and guidance that facilitated

the analysis in my work. For the conference versions of these journals, i.e., conferences (2) and (4) to (7), the role of the authors are the same as the journal paper publications are extension of these conference papers.

1.4.1 Book Chapter:

- R. Amer, M. M. Butt, and N. Marchetti, "Caching at the Edge in Low Latency Wireless Networks," in book Wireless Automation as an Enabler for the Next Industrial Revolution, Wiley, 2019 (in press).

1.4.2 Journal Publications:

1. R. Amer, Walid Saad, B. Galkin, and N. Marchetti, "On the Performance of Mobile Cellular-Connected Drones Under Practical Antenna Configurations", in preparation for submission to *IEEE Transactions on Vehicular Technology*, 2020.

2. R. Amer, W. Saad, and N. Marchetti, "Efficient Mobility Support and Reliable Communications for Cellular-Connected Drones", to be submitted to *IEEE Communication Magazine*, 2020.

3. R. Amer, M. M. Butt, and N. Marchetti, "Optimizing Joint Probabilistic Caching and Channel Access for Clustered D2D Networks," in preparation for submission to *Journal of Communications and Networks*, 2020.

4. R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Performance analysis for wireless D2D caching with inter-cluster cooperation," in *IEEE Transactions on Wireless Communications*, 2018.

5. R. Amer, Walid Saad, and N. Marchetti, "Towards a Connected Sky: Performance of Beamforming with Down-tilted Antennas for Ground and UAV User Co-existence," in *IEEE Communications Letter*, July 2019.

6. R. Amer, H. ElSawy, M. M. Butt, Eduard A Jorswieck, Mehdi Bennis, Nicola Marchetti, "Optimized Caching and Spectrum Partitioning for D2D enabled Cellular Systems with Clustered Devices," in *IEEE Transactions on Communications*, 2020.

7. R. Amer, Walid Saad, and N. Marchetti, "Mobility in the Sky: Performance and Mobility Analysis for Cellular-Connected UAVs," in *IEEE Transactions on Communications*, 2020.

8. R. Amer, Hesham ElSawy, Jacek Kibilda, M. M. Butt, and N. Marchetti, "Performance Analysis and Optimization of Cache-Assisted CoMP for Clustered D2D Networks," in *IEEE Transactions on Mobile Computing*, Aug. 2020.

9. B. Galkin, R. Amer, E. Fonseca, Luiz A. DaSilva, and I. Dusparic, "REQIBA: Regression and Deep Q-Learning for Intelligent UAV Cellular User to Base Station Association" submitted to *IEEE JSAC SI-UAV-B5G*, 2020.

10. E. Fonseca, B. Galkin, R. Amer, Luiz A. DaSilva, and I. Dusparic, "Adaptive Height Optimization for Cellular-Connected UAVs using ReinforcementLearning" submitted to *IEEE Transactions on Vehicular Technology*, 2020.

1.4.3 Conference Publications:

1. B. Galkin, R. Amer, E. Fonseca, and Luiz A. DaSilva, "Intelligent UAV Base Station Selection in Urban Environments: A Supervised Learning approach" in Proc. of *IEEE 3rd 5G World Forum (5GWF)*, Sept. 2020.

2. R. Amer, Walid Saad, B. Galkin, and Nicola Marchetti, "Performance Analysis of Mobile Cellular-Connected Drones under Practical Antenna Configurations". in Proc. of *IEEE International Conference on Communications*, Dublin, Ireland, June 2020.

3. C. Chaccour, R. Amer, B. Zhou, and W. Saad, "On the Reliability of Wireless Virtual Reality at Terahertz (THz) Frequencies" in Proc. of *10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Spain, June 2019.

4. R. Amer, W. Saad, H. ElSawy, M. M. Butt, and N. Marchetti, "Caching to the sky: Performance analysis of cache-assisted CoMP for cellular-connected UAVs," in Proc. of *IEEE Wireless Communications and Networking Conference (WCNC)*, Morocco, April 2019.

5. R. Amer, H. ElSawy, J. Kibilda, M. M. Butt, and N. Marchetti, "Cooperative Transmission and Probabilistic Caching for Clustered D2D Networks," in Proc. of *IEEE Wireless Communications and Networking Conference (WCNC)*, Morocco, 2019.

6. R. Amer, M. M. Butt, M. Bennis, Hesham ElSawy, Jacek Kibilda, and N. Marchetti, "On Minimizing Energy Consumption for D2D Clustered Caching Networks," in Proc. of *IEEE Global Communications Conference GLOBECOM*, Abu Dhabi, UAE, Dec 2018.

7. R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Delay analysis for wireless D2D caching with inter-cluster cooperation," in Proc. of *IEEE Global Communications Conference (GLOBECOM)*, Singapore, Dec 2017.

Chapter 2

Background

2.1 Introduction

One of the promising approaches to meet the unprecedented traffic demands is the deployment of SCN [18]. Small cell networking is a novel networking paradigm based on the idea of deploying short-range SBS underlaying the already existing macrocellular network. Regarding the SCN, to date, the vast majority of research has dealt with issues related to self-organization, inter-cell interference coordination (ICIC), traffic offloading, energy efficiency, etc., see [18] and the references therein. These studies are carried out under the existing reactive networking paradigm wherein users' traffic requests are served upon their arrival or dropped in case of outages. Owing to these characteristics of the reactive network, the existing SCN paradigm does not help to accommodate the peak traffic demands from the network. This shortcoming is set to become more severe especially due to the surging number of connected devices and the advent of ultra-dense networks (UDN), which will continue to exhaust current cellular network infrastructures. Different from the evolutionary path of previous cellular generations that was based on spectral efficiency improvements, the most substantial amount of future system performance gains will be obtained by means of network infrastructure densification. By increasing the density of operator-deployed infrastructure elements, along with incorporation of user-deployed access nodes and mobile user devices acting as ”infrastructure prosumers”, i.e., when users' devices are employed in the communication process as a part of the network, it is expected that having one or more access nodes exclusively dedicated to each user will become

feasible. Although it is clear that UDN are able to take advantage of the significant benefits provided by proximal transmissions and increased spatial reuse of system resources, at the same time, large node density and irregular deployment introduce new challenges. These challenges are mainly due to the interference environment characteristics. Besides , the backhauling of dense small cell networks constitutes emerged against a bottleneck of their successful deployment. The increasing number of deployed small cells and the lack of high capacity backhaul links, represent limiting factors of the network densification gains. Therefore, it became of paramount importance to envisage a new networking paradigm that goes beyond current heterogeneous small cell deployments (i.e., networks of different types of SBSs with different transmission powers and capabilities) leveraging the latest developments in storage, context awareness, and social networking [19].

Cellular networks, increasingly the most essential aspect of the telecommunication infrastructure, are in a period of unprecedented change. Hence, incremental changes in designing and optimizing such reactive networks are becoming more and more outdated. Future cellular networks are expected to be smart in the sense that network nodes anticipate users' demands and utilize predictive abilities to reduce the traffic peak-to-average ratio, yielding significant network resource savings. Meanwhile, proactive caching in the wired Internet is well established and has been shown to reduce latency and energy consumption. Similar benefits can be expected by caching popular contents at the wireless edge, which can improve the network performance and accommodate the explosive demand for wireless data. Proactive caching either at users with D2D communications enabled, or at SBS to eliminate the backhaul bottleneck, is envisioned as a promising solution to satisfy the high demand for data and to alleviate the heavy burden on the core network. This is because content is available locally, instead of requiring redundant traverses across backhaul links [20–24].

2.2 Living on the Edge

The wireless caching network, as described in the seminal work [21], is proactive in essence and rooted in the fact that network nodes, including BSs and mobile

devices, would exploit users' context information, foresee users' demands, and leverage their predictive abilities to achieve significant resource savings and sustain acceptable quality-of-service (QoS) levels and keep low cost/energy expenditures [25]. This paradigm goes far beyond current cellular deployments, which have been designed mainly to serve dumb devices with very limited storage and processing power. In fact, current smartphones have become very sophisticated with considerably improved computing and storage capabilities. Therefore, under the proactive networking paradigm, network nodes could be assumed to track, learn, and build users' demand profiles to predict future requests, leveraging devices' capabilities and the vast amount of available data.

Recently, predictive analytics have received significant attention with respect to machine learning techniques to analyze billions of infrastructure logs to produce predictive and actionable information for outage prediction and content recommendation [26]. Leveraging these predictive capabilities, users can be scheduled in a more efficient manner, and resources are pre-allocated more intelligently by proactively serving peak-hour demands during off-peak times (e.g., at night). The proactive caching paradigm leverages the statistical traffic patterns and users' context information (i.e., file popularity distributions, location, velocity, and mobility patterns) aiming at a better prediction of when users' contents are requested, and matching that with the amount of resources needed, and at which network locations contents should be saved (cached). As a relevant practical example, online social networks (e.g., Facebook, Twitter) have become instrumental in users' content distribution [27]. As such, users tend to value highly recommended contents by friends or people with similar interests and are also likely to recommend them. Fig. 2.1 shows an example of a spatial network layer overlaid with the social network layer.

Now we are in a position to discuss different architectures for wireless caching networks. Caches can be installed in macro BS, SBS (say, pico or femto BS), relay nodes, and users' devices [28], see Fig. 2.2. A picocell is a small cellular base station typically covering a small area, such as in-building (offices, shopping malls, train stations, stock exchanges, etc.), while femtocell is defined as a small, low-power

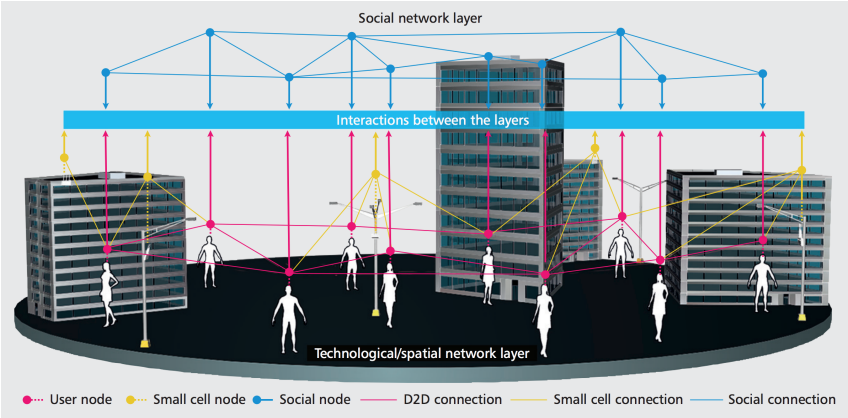


FIGURE 2.1: An illustration of an overlay of socially interconnected and technological/spatial network [21].

cellular base station, typically designed for use in a home or small business. Comparing to caching at the evolved packet core (EPC), caching at existing macro base station (MBS) and SBS essentially plays the role of replacing backhaul links, and hence alleviates backhaul congestion. Moreover, a new type of SBS with limited backhaul connections, called helpers (or relays) [20], can enable flexible and cost-effective deployments to deliver popular contents. Due to the fact that increasing cache size can increase the cache-hit probability, and hence lower the required backhaul capacity, there is a trade-off between cache size and backhaul capacity. Besides, caching contents at user terminals such as smartphones, tablets, and laptops has been applied as a technique to improve quality-of-experience (QoE) [29], and recently, it has also been proposed to offload wireless traffic. With a known content popularity, a BS can push the popular contents to all users via broadcast [30]. Also, with a known user preference, the BS can pre-download favorite contents to some users via unicast.

Future wireless networks will also witness a wide use of UAVs (also known as drones), either operating aerial BSs or aerial UE. In essence, UAVs have been the subject of concerted research over the past few years [31–36], owing to their autonomy, flexibility, and broad range of application domains. They have essentially been considered as key enablers of various applications including, but not limited to, military, surveillance and monitoring, telecommunications, delivery of medical supplies, and rescue operations. The unprecedented recent advances in drone technology make it possible to widely deploy UAVs, small aircrafts, balloons, and

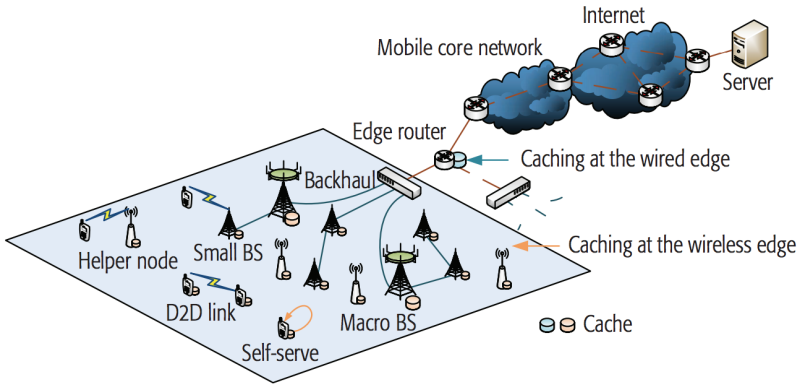


FIGURE 2.2: An illustration of local caching and content delivery at the wireless edge.

airships for wireless communication purposes [37]. In particular, if properly deployed and operated, UAVs can provide reliable and cost-effective wireless communication solutions for a variety of real-world scenarios.

On the one hand, drones can be used as aerial BSs that can deliver reliable, cost-effective, and on-demand wireless communications to desired areas. On the other hand, drones can function as aerial UEs, known as cellular-connected UAVs, in coexistence with ground users (e.g., delivery or surveillance drones). This exciting new avenue for the use of UAVs warrants a rethinking of the research challenges with wireless communications and networking being the primary focus, as opposed to control and navigation. In this thesis, we are mainly focused on content delivery and caching for such contemporary aerial users (i.e., cellular-connected UAVs). However, for completeness, later in this chapter, we review the prior work that incorporated caching for both aerial BSs and aerial UEs.

2.3 Different Approaches for Modeling and Analysis of Wireless Caching Networks

In this section, we detail different frameworks, perspectives, and possible classifications for wireless caching networks. The first subsection below is devoted to the discussion of three main architectures for wireless caching networks, namely, caching on SCN, caching over the cloud radio access network (C-RAN), caching among the mobile devices, and caching for UAVs and drone networks. Then, we

follow up by describing two main methodologies widely adopted in the analysis of wireless caching networks, namely, simple protocol model and physical interference model.

2.3.1 Wireless Caching Architecture

- Small Cell Networks

  The introduction of small cell base stations is viewed as a key paradigm to handle the increase of video traffic and improve the wireless capacity by bringing contents closer to the users. However, reaping the benefits of small cell deployments requires meeting several key challenges such as resource allocation and network modeling. For instance, owing to the cheap storage/memory prices and the fact that mobile video accounts for most of the internet traffic demands, one can leverage the use of storage at the small cell level to bring popular contents closer to the network edge (i.e., BSs) [38]. Indeed, one promising approach to improve the QoS of video transmission is through caching popular contents locally at the SBSs to alleviate peak traffic demands and minimize service delays. [38–43].

- Cloud Radio Access Networks

  While content delivery networks (CDN) have been recently enhanced to reduce Internet bandwidth consumption and associated delay/jitter of online video, such video content consumed by mobile devices must additionally travel through the wireless carrier core network and radio access network (RAN) before reaching the UE [44]. One promising approach to address the need for massive content distribution is to enable content caching within the wireless operators' networks, where popular contents are cached at the BS at the edge of the RAN [45]. C-RAN is an emerging architecture for wireless 5G system in which a centralized baseband unit (BBU) implements the baseband processing functionalities of a set of remote radio head (RRH), which are connected to the BBU by means of fronthaul links [46, 47]. Recently, an evolved network architecture, referred to as fog-radio access network (F-RAN), has been proposed, which enhances the C-RAN architecture by allowing the RRH

to be equipped with caching and signal processing functionalities [48–50]. This architecture is also referred to as a hybrid of cloud and fog processing in the literature [51]. As a cache-aided system, an F-RAN operates in two phases, namely the pre-fetching and the delivery phases [52–58]. Pre-fetching operates at a large time scale corresponding to the period in which content popularity remains constant. This time scale encompasses multiple transmission intervals. Based on the cached file messages, the delivery phase, instead, operates separately on each transmission interval. The fronthaul-aware design of the pre-fetching or delivery phases was studied in [52–56] under the assumption that the fronthaul links in an F-RAN are leveraged to convey to the RRH the requested content that is not present in the local caches.

- Device-to-Device

  Capitalizing on the fact that user demands are highly redundant, each user demand can be satisfied through local communication from the device's cache, without requiring a high throughput backhaul to the core network [59]. The concept of having helper nodes is pushed further by introducing the notion of wireless devices as helpers [20]. Recent years have seen an enormous proliferation of smartphones and tablets that have anywhere between 10 to 64 GB of storage (not to mention the 500 GB on typical laptop hard disks). By enabling D2D communications, the ensemble of wireless devices can become a distributed cache that allows a more efficient download of contents as compared to traditional networks without caching. The advantage of using wireless devices instead of fixed helper nodes lies in the small deployment costs and automatic upscaling of the capacity as the density of such devices increases. The drawback lies in the necessity to motivate users to participate in the caching process, and the randomness of the available throughput due to the decentralized and uncoordinated nature of D2D communication. There also exist some works focusing on the economic aspect of caching in wireless D2D networks. In such networks, the operators define a pricing scheme to motivate users to proactively download the most popular files and cache

them in their devices to serve other users' requests. In [60], the authors proposed a smart pricing scheme to maximize the benefit of the operator and minimize the charged price to the users. Via D2D communications, users can trade their cached files to minimize their expected payments. On the other hand, the operator defines a dynamic pricing model that differentiates off-peak and peak time periods to maximize its own benefit. If a user requests one of the files stored in neighbors' caches in the cluster, neighbors will handle the request locally through D2D communication; otherwise, the BS should serve the request. As a result, the probability of having more D2D communications among the users depends on what users store [20, 23, 28, 59, 61–64].

- Caching for UAVs and Drone Networks

  While caching at SBSs has emerged as a promising approach to improve users' throughput and to reduce the transmission delay, caching at static ground base stations may not be very effective in serving mobile users in case of frequent handovers (e.g., in ultra-dense networks with moving users). In this case, when a user moves to a new cell, its requested content may not be available at the new base station and, thus, the users cannot be served properly. To effectively service mobile users in such scenarios, each requested content needs to be cached at multiple base stations which is not efficient due to signaling overheads and additional storage usage. Hence, to enhance caching efficiency, there is a need to deploy flexible base stations that can track the users' mobility and effectively deliver the required contents.

  To this end, there are multiple envisioned scenarios in which UAVs, acting as flying base stations, can dynamically cache the popular contents, track the mobility pattern of the corresponding users and, then, effectively serve them [65–67]. In fact, the use of cache-enabled UAVs is a promising solution for traffic offloading in wireless networks. By leveraging user-centric information, such as content request distribution and mobility patterns, cache-enabled UAVs can be optimally moved and deployed to deliver desired services to users. Another advantage of deploying cache-enabled UAVs is that

the caching complexity can be reduced compared to a conventional static SBSs case. For instance, whenever a mobile user moves to a new cell, its requested content needs to be stored at the new base station. However, cache-enabled drones can track the mobility pattern of users and, consequently, the content stored at the drones will no longer require such additional caching at SBSs.

In practice, in a cache-enabled UAV system, a central cloud processor can utilize various user-centric information including users' mobility patterns and their content request distribution to manage the UAV deployment. In fact, such user-centric information can be learned by a cloud center using any previous available users' data. Then, the cloud center can effectively determine the locations and mobility paths of cache-enabled UAVs to serve ground users. This, in turn, can reduce the overall overhead of updating the cache content.

While performing caching with SBSs, content requests of a mobile user may need to be dynamically stored at different SBSs. However, cache-enabled UAVs can track the mobility pattern of users and avoid frequently updating the content requests of mobile users. Therefore, ground users can be effectively served by exploiting mobile cache-enabled UAVs that predict mobility patterns and content request information of users.

On the other hand, drones can also act as contemporary users of the wireless infrastructure. In particular, drone-users can be used for package delivery, surveillance, remote sensing, and virtual reality applications. Indeed, cellular-connected UAVs will be a key enabler of the IoT. For instance, for delivery purposes, drones are used for Amazon prime air drone delivery service, and autonomous delivery of emergency drugs [68]. The key advantage of drone-users is their ability to swiftly move and optimize their path to quickly complete their missions. To properly use drones as user equipments (i.e., cellular-connected drones (UAV-UEs)), there is a need for reliable and low-latency communication between drones and ground BSs.

To support a large-scale deployment of drones, a reliable wireless communication infrastructure is needed to effectively control the drones' operations while supporting the traffic stemming from their application services [69]. Beyond their need for ultra low latency and reliability, when used for surveillance purposes, drone-UEs will require high-speed uplink connectivity from the terrestrial network and from other UAV-BSs. In this regard, current cellular networks may not be able to fully support drone-UEs as they were designed for ground users whose operations, mobility, and traffic characteristics are substantially different from the drone-UEs. There are a number of key differences between drone-UEs and terrestrial users. First, drone-UEs typically experience different channel conditions due to nearly line-of-sight (LoS) communications between ground BSs and flying drones. In this case, one of the main challenges for supporting drone-UEs is significant LoS interference caused by ground BSs. Second, unlike terrestrial users, the on-board energy of drone-UEs is highly limited. Third, drone-UEs are in general more dynamic than ground users as they can continuously fly in any direction. Therefore, incorporating cellular-connected drone-UEs in wireless networks will introduce new technical challenges and design considerations.

Given such technical difficulties to provide seamless connectivity to the UAV-UEs, most of the ongoing work in the literature focuses mainly on addressing such challenges and reshaping the terrestrial wireless networks to effectively serve the UAV-UEs. To the best of our knowledge, only few works in the literature studied the role of wireless caching and content delivery for wireless networks serving UAV-UEs. For example, the authors in [70] proposed probabilistic caching is studied for ultra dense SCNs with terrestrial and aerial users. The work in [70] generally focuses on the successful download probability UEs from SBSs that cache the requested files under various caching strategies.

2.3.2 Modelling Wireless Caching Networks

The modeling the cache-enabled heterogeneous networks (HetNet), which have multiple types of low power radio access nodes in addition to the traditional macro-cell nodes, in the literature follows two main directions. The first line of work focuses on the fundamental throughput scaling results by assuming a simple protocol channel model [40, 64], known as the protocol model. As such, two devices can communicate with each other if they are within a certain distance. The second line of work, defined as the physical interference model, considers a more realistic model for the underlying physical layer [71, 72].

The physical interference model is based on the fundamental signal-to-interference ratio (SIR) metric, and therefore, is applicable to any wireless communication system. Modeling devices' locations as a Poisson point process (PPP) is widely employed in the literature, especially, in the wireless caching area [24, 71–74]. However, a realistic model for D2D caching networks requires to consider that a given device typically has multiple nearby devices, where any of them can potentially act as a serving device. This deployment is known as clustered devices deployment, which can be characterized by cluster processes [75], recently studied in [76–78].

2.4 Challenges in Wireless Caching Networks

Motivated by the above discussion, we now discuss there are some shortcomings in the models adopted in the literature. An example is the D2D caching network, where the two modeling approaches discussed above, have some limitations and are not practically viable. In particular, the simple protocol model, which is extensively adopted in the design and evaluation of communication protocols and scaling analysis, limits the D2D communication only to the intra-cluster transmission determined by a fixed distance. Besides, the physical interference model presumes that the devices are uniformly distributed, which is frequently modeled by a PPP. However, this assumption is not realistic in the sense that a given device typically has multiple proximate (neighboring) devices in the same region, wherein any of them can potentially act as a serving device, and other remote devices that are far apart from its proximity.

We already discussed the emergence of edge caching as a promising approach to alleviate the heavy burden on data transmission through caching and forwarding contents at the edge of wireless networks. However, deploying caches in such wireless networks poses many new challenges. As an example, many of the existing studies always treat storage and communication separately, despite of them being coupled. To explain this coupling, assume that there is a set of devices sharing a wireless channel with each other, and each device possesses a limited cache memory. Intuitively, we can expect that the devices that are accessing the channel less often tend to cache contents in a greedy way by caching the most popular content. On the other hand, if the devices can access the channel more frequently, they turn out to be more cooperative by caching different files to maximize the offloading gain. Therefore, to maximize the offloading gain of such a cache-enabled D2D communication system, both content placement and delivery need to be jointly designed. This implies that jointly considering content caching and scheduling policies is inevitable to improve the performance. This principle can be generalized to similar scenarios, e.g., joint caching and transmission, joint caching and resource allocation, joint caching and power control, etc.

As we already mentioned, this thesis also focuses on the content delivery and caching for UAV-UEs. One of the key challenges for maintaining reliable connectivity and content delivery to such UAV-UEs is the severe impact of handover. In fact, there is a crucial need for effective handover management mechanisms to deploy an aerial network of flying drone-UEs. Recall first that the handover is a key process in wireless networks in which user association changes in order to maintain the connectivity of mobile users. Handover management in UAV communications is significantly more challenging than traditional cellular networks due to the highly dynamic nature of drone-UEs. In particular, efficient handover mechanisms must be designed to accommodate 3D movements of the drone-UEs, while ensuring low-latency communications and control when serving drone-UEs. This handover design for flying devices must be done jointly with existing handover mechanisms for mobile ground users, such as vehicles. Moreover, for drone-UEs, all of the aforementioned challenges must also take into account the fact that ground base stations will have their antennas down-tilted to maximize coverage of ground users. As a

result, it is imperative to understand the impact of antenna tilt on the performance of UAV-UEs, while also studying how one can overcome this limitation via adaptive beamforming or new UAV-UE aware design of ground base stations. This will further lead to an underlying network infrastructure that is almost and suitable to deploy edge caching and content delivery services for the contemporary UAV-UEs.

2.5 Motivation and Objective

Motivated by the above discussion, we aim at developing a new caching architecture for HetNet which is shown to improve the network performance in terms of KPIs, such as offloading gain, energy efficiency, average delay, etc. Definitions of the KPIs are introduced in next chapters in the context of the presented studies. Further, we propose to study and optimize the performance of wireless caching networks, especially when we consider more realistic models, such as D2D caching with inter-cluster cooperation and spatially clustered point process model, as compared to the simple models widely adopted in the literature.

The main objective of this thesis is to develop a general framework that accounts for the joint optimization of caching and communication for HetNets aiming at improving the network performance. Particularly, in this work, we first envision a new D2D caching architecture by allowing D2D communication along with inter-cluster cooperation. Different from similar works where D2D communication is assumed to be only intra-cluster, we show that allowing file transfer between remote users through the BS acting as a relay, improves the network average delay and throughout, and boosts the coverage probability.

A comprehensive study and optimization of caching and communication under more practical assumptions for the user deployment is then conducted and the KPIs are optimized, where our model accurately captures the non-uniformity of the locations of the users (i.e., when users are grouped into clusters). As an example, the joint optimization of scheduling policy and content placement in a clustered D2D caching network is carried out. Content placement as well as bandwidth allocation are also jointly optimized to improve the KPIs in terms of the average delay and energy consumption. In this respect, we further studied the role of CoMP

transmissions and content caching for clustered D2D networks so as to boost the desired signal levels and improve the perceived QoE in dense D2D caching networks and adverse interference conditions.

In the second part of this thesis, we focused on enabling enhanced communication, content delivery, and caching for cellular-connected drones (a.k.a aerial users). We particularly investigated the use of MIMO beamforming and cooperative transmission to support reliable content delivery and communication for aerial users. In addition, we studied the impact of handover and 3D mobility of the UAV-UEs. Finally, we proposed a cache-assisted CoMP framework to enable efficient caching and seamless coverage to the UAV-UEs.

Further detailed discussions will follow in the next chapters of this thesis. For the sake of an organized presentation, we postpone the detailed discussion of the related work to the next chapters. That is, in the following chapters for each type of network addressed in this thesis, the related work is presented and the novelty is highlighted.

Chapter 3

Inter-cluster Cooperation for Wireless D2D Caching Networks

In this chapter, we propose a new architecture for D2D caching with inter-cluster cooperation. We study a cellular network in which devices cache popular files and share them with other devices either in their proximity via D2D communication or with remote devices using cellular transmission. We characterize the network average delay per request from a queuing perspective and formulate the delay minimization problem and show that it is NP-hard. Furthermore, We prove that the delay minimization problem is equivalent to the minimization of a non-increasing monotone supermodular function subject to a uniform partition matroid constraint. A computationally efficient greedy algorithm is then proposed which is proven to be locally optimal within a factor $(1 - e^{-1}) \approx 0.63$ of the optimum. We also analyze the average per request throughput for different caching schemes and conduct the scaling analysis for the average sum throughput. Finally, we show how throughput scaling depends on video content popularity when the number of files grows asymptotically large.

3.1 Introduction

From the discussion in the background chapter, it is now clear that caching the most popular content at various locations of the network edge helps alleviate the heavy burden on the highly congested backhaul links [21, 38, 79]. Four different architectures adopting the caching technology are discussed in the background chapter,

namely, caching on femtocells in small cell networks, caching on RRH in C-RAN, caching on mobile devices, and caching for drone networks [65, 67, 70]. In this chapter, we focus on device caching solely. The architecture of device caching exploits the large storage available in modern smartphones to cache multimedia files that might frequently be requested by the users. The users' devices exchange multimedia content stored on their local storage with nearby devices [80]. Since the distance between the requesting user and the caching user (a user who stores the file) will be small in most cases, D2D communication is commonly used for content transmission [80]. A representative set of such D2D-based works is [80–83]. For instance, Golrezaei *et al.* [81] proposed a novel architecture to improve the throughput of video transmission in cellular networks based on the caching of popular video files in base station-controlled D2D communication. The analysis of this network is based on the subdivision of a macrocell into small virtual clusters, such that one D2D link can be active within each cluster. Random caching is considered where each user caches files at random and independently, according to a certain caching distribution.

3.1.1 Motivation and Contribution

Motivated by the remarks from the above discussion, i.e., backhaul links being highly congested, the geometric distribution of the devices in clusters, we propose a novel D2D caching architecture with inter-cluster cooperation. We propose a system in which a user in a given cluster can search its requested files either in the local cluster or any of the remote clusters. We show that allowing inter-cluster collaboration via cellular communication achieves both user and system performance gains. From the user perspective, the average delay per request is reduced when downloading files from a remote cluster, instead of serving files from the core network. From the system perspective, the heavy burden on backhaul links is alleviated by decreasing the number of requests that are served directly from the core network. From a resource allocation perspective, similar to the work performed in [82, 83], we analyze the network average delay and throughput per user request for the proposed inter-cluster cooperative caching system under different caching schemes, and show how the network performance is significantly improved. To

the best of my knowledge, none of the works in the literature dealt with the performance analysis of D2D caching networks with inter-cluster cooperation. The main contributions of this chapter are summarized as follows:

- We study a D2D caching system with inter-cluster cooperation from a queueing theory perspective. We formulate the network average delay minimization problem in terms of cache placement. The delay minimization problem is then shown to be non-convex, and it can be reduced to a well-known 0 - 1 knapsack problem which is NP-hard.
- A closed-form expression of the network average delay is derived under the policy of caching popular files (CPF). Moreover, a locally optimal greedy caching algorithm (GCA) is proposed whose delay is within a factor $(1 - e^{-1})$ of the global optimum. Results show that the delay can be significantly reduced by allowing D2D caching with inter-cluster cooperation.
- We derive a closed form expression for the average throughput per request for the proposed inter-cluster cooperating scheme. Moreover, we conduct the asymptotic analysis for the average sum throughput when the content library size grows to infinity. The result of the scaling analysis shows that the upper bound for the network average sum throughput decreases when the library size increases asymptotically, and the rate of this decrease is controlled by the popularity of files.

3.2 Related Work

Different cooperation strategies in D2D networks are proposed in the literature. As an example, in [84], the authors proposed a cooperative D2D communications framework in order to combat the problem of congestion in crowded communication environments. The authors allowed a D2D transmitter to act as an in-band relay for a cellular link and at the same time transmit its data by employing superposition coding in the downlink. It is shown that cooperation between the cellular link and D2D transmitter helps increase the number of connections per unit area

with the same spectrum usage. In the area of D2D caching, the authors in [83] proposed an opportunistic cooperation strategy for D2D transmission by exploiting the caching capability at the devices to control the interference among D2D links. The authors considered an overlay inband D2D communication, divided the D2D devices into clusters, and assigned different frequency bands to cooperative and non-cooperative D2D links. The cluster size and bandwidth allocation are further optimized to maximize the network throughput.

The analysis of wireless caching networks from the resource allocation perspective is widely discussed in the literature. For instance, in [82], the authors showed how distributed caching and collaboration between devices and femtocells (helpers) can significantly improve throughput without suffering from the backhaul bottleneck problem common to femtocells. The authors also investigated the role of collaboration among devices - a process that can be interpreted as the mobile devices playing the role of helpers also. This approach allowed an improvement in the video throughput without the deployment of any additional infrastructure. Due to the dependence between content cache placement and resource allocation in wireless networks, the joint problem of caching and resource allocation is studied in many works. As an example, Zhang *et al.* in [85] proposed a single-hop D2D-assisted wireless caching network, where popular files are randomly and independently cached in the memory of end devices. The joint D2D link scheduling and power allocation problem is formulated to maximize the system throughput. Following a similar approach, Chen *et al.* in [72] studied the joint optimization of cache content placement and scheduling policies to maximize the so-called offloading probability. The successful offloading probability is defined as the probability that a user can obtain the desired file in the local cache or via a D2D link with data rate larger than a given threshold. The authors obtained the optimal scheduling factor for a random scheduling policy that controls interference in a distributed manner, and proposed a low complexity solution to compute caching distribution.

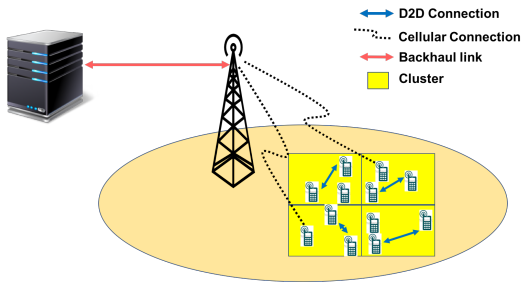


FIGURE 3.1: Schematic diagram of the proposed system model. A cellular cell is divided into square clusters, where devices in all clusters can download their requested files using D2D, cellular, or backhaul communication.

3.3 System Model and Assumptions

3.3.1 Network Model

In this subsection, we describe the proposed D2D caching network with inter-cluster cooperation. Fig. 3.1 illustrates the system layout. A cellular network consists of a SBS and a set of devices $\mathcal{U} = \{1, \ldots, n\}$ placed uniformly in the cell. The cell is divided into a set of equally sized clusters $\mathcal{K} = \{1, \ldots, K\}$. For mathematical convenience, we assume that the number of devices per cluster is $y = n/K$, as in [83] and the reference therein. Devices in the same cluster can communicate directly using low power high rate D2D communication in a dedicated frequency band for D2D transmission.

Each user $u \in \mathcal{U}$ requests a file f from a file library $\mathcal{F} = \{1, \ldots, m\}$ independently and identically, according to a given request probability mass function. It is assumed that each user can cache up to M files, and for the caching problem to be non-trivial, it is assumed that $M < m$. From the cluster perspective, there exists a cluster virtual cache center (VCC) formed by the union of devices' storage in the same cluster, which caches up to N files, i.e., $N = (n/K)M$.

We assume that the D2D communication does not interfere with communication between the BS and devices. We also assume that all D2D links share the same time-frequency transmission resource within one cell. Multiple transmissions on those resources are possible since the distance between requesting devices and devices with the stored file will typically be small. Furthermore, there should be no interference by other transmissions on an active D2D link. To achieve this, the cell

is divided into smaller areas, which we denoted as clusters. To avoid intra-cluster interference, only one such communication per cluster is allowed.[1] Devices in the same cluster are assumed to be served in a round-robin manner.

We define three modes of operation according to how a request for content $f \in \mathcal{F}$ is served:

1. **Local cluster mode (M_{lc} mode):** Requests are served from the local cluster. Files are downloaded from nearby devices via a single-hop D2D communication. In this mode, we neglect self-caching, i.e., the event when a user finds the requested file in its internal cache with zero delay. Within each cluster, the BS can help devices find their requested content by broadcasting signals containing the content replication ratio.

2. **Remote cluster mode (M_{rc} mode):** Requests are served from any of the remote clusters via inter-cluster cooperation. The BS fetches the requested content from a remote cluster, then delivers it to the requesting user by acting as a relay in a two-hop cellular transmission. The BS assists in content dissemination in the *"remote cluster mode"* by relaying the content between different clusters.

3. **Backhaul mode (M_{bh} mode):** Requests are served directly from the backhaul. The BS obtains the requested file from the core network via the backhaul link and then transmits it to the requesting user.

In each cluster, we assume that the stream of user requests are served sequentially based on first in first out (FIFO) criterion. The BS receives all requests and works as a coordinator to establish the file transfer between the requesting user (a user who requests the file) and the serving node (another user who caches the file or a caching server in the core network). The BS keeps track of which devices can communicate with each other and which files are cached on each device. Such BS-controlled D2D communication is more efficient and more acceptable to spectrum owners if the communication occurs in a licensed band as compared to traditional uncoordinated peer-to-peer communications [86]. To serve a request for file f in

[1] We adopt a simplified PHY-layer model in this work.

cluster $k \in \mathcal{K}$, first, the BS searches the VCC of cluster k. If the file is cached, it will be delivered from the local VCC (M_{lc} mode). We assume that the BS has all the information about cached content in all clusters, such that all file requests are sent to the BS, then the BS replies with the address of the caching user from whom the file will be retrieved.

If a file is not cached locally in cluster k but cached in any of the remote clusters, it will be fetched from a randomly chosen cooperative cluster (M_{rc} mode), instead of downloading it from the backhaul. Unlike multi-hop D2D cooperative caching discussed in [87], in this work cooperating clusters are assumed to exchange cached files using a two-hop cellular communication link through the BS, such that the D2D band is dedicated only to the intra-cluster communication. Hence, all the inter-cluster communication is performed in a centralized manner through the BS. Finally, if the requested file has not been cached in any cluster $j \in \mathcal{K}$ in the cell, it can be downloaded from the core network via the backhaul link (M_{bh} mode). The selection of the three modes of operation is conducted in a prioritized order from the local cluster, from the remote cluster, or finally from the core network through the backhaul link as a last resort.

Serving files sequentially according to the above three modes is based on the assumption that the BS has a capacity-limited wired backhauling, such that the average delay per request is decreased when allowing inter-cluster cooperation. Otherwise, if the backhaul is not a bottleneck, e.g., optical fiber or millimeter wave backhaul links are available, requests for files not cached in the local cluster are served directly from the core network through the high capacity backhaul link. The analysis in this chapter relies on a well-known grid-based clustering model [81], i.e., no specific underlying physical model or parameters are assumed. Therefore, the obtained design/results, e.g., design of caching scheme and the performance of the greedy algorithm, can be applied to similar scenarios with three prioritized paths (modes) for file downloading. For example, on-board devices, such as on a plane or a ship, can obtain requested files from neighboring devices via Bluetooth (local cluster mode), from a remote user through an access point [88] acting as a relay (remote cluster mode), or finally from the backhaul, which is the least preferred option. As another example, in the case of connecting devices through UAV [89],

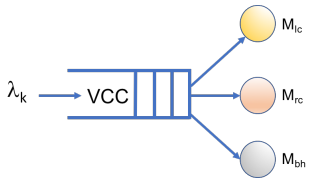


FIGURE 3.2: The devices' traffic model in a cluster k with cache center VCC is modeled as a multiclass processor sharing queue.

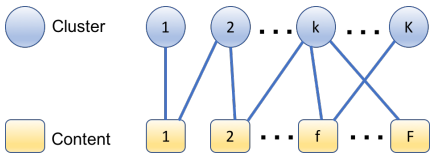


FIGURE 3.3: An example of the content cache placement modeled as a bipartite graph indicating how files are cached in clusters.

serving files can be prioritized as follows. A file is received from a neighboring user via D2D communication (local cluster mode), from a remote user through the UAV acting as a relay (remote cluster mode), or from the backhaul through the UAV as a last resort.

3.3.2 Content Placement and Traffic Characteristics

We use a binary matrix $\mathbf{C} = [c_{k,f}]_{K \times m}$ with $c_{k,f} \in \{0, 1\}$ to denote the cache placement in all clusters, where $c_{k,f} = 1$ indicates that content f is cached in cluster k. Fig. 3.2 shows the assumed devices' traffic model in a cluster k, modeled as an multiclass processor sharing queue (MPSQ) with arrival rate λ_k, and three serving processors representing the three transmission modes. According to the MPSQ definition [90], each transmission mode is represented by an M/M/1 queue with Poisson arrival rate and exponential service rate. A graphical interpretation of the content cache placement is shown in Fig. 3.3. The content caching policy is defined by a bipartite graph $\mathcal{Y} = (\mathcal{K}, \mathcal{F}, \mathcal{E})$, where edges $(k, f) \in \mathcal{E}$ denote that content f is cached in the VCC of cluster k.

If a user in cluster k requests a locally cached file f (i.e., $c_{k,f} = 1$), it will be served by the local cluster mode with an average rate R_D. However, if the requested file is not cached locally and cached in any of the remote clusters, i.e., when $c_{k,f} = 0$ and $\sum_{j \in \mathcal{K} \setminus \{k\}} c_{j,f} \geq 1$, it will be served by the remote cluster mode.

We denote the rate for the remote cluster mode by R_{WL}, accounting for the average sum transmission rate between the cooperating clusters through the BS. Accordingly, R_{WL} is shared between clusters simultaneously served by the remote cluster mode. Finally, requests for files that are not cached in the entire cell, i.e., when $\sum_{j=1}^{K} c_{k,f} = 0$, are served via the backhaul mode with an average sum rate R_{BH}. We assume that $R_{BH} << R_{WL}$, such that the part of the cellular rate allocated to the devices served by the backhaul mode is neglected for the delay analysis. R_{BH} is assumed to be the effective rate from the core network to the user using BS.

Due to traffic congestion in the core network and the transmission delay between cooperating clusters, we assume that the aggregate transmission rates for the above three modes are ordered such that $R_D > R_{WL} > R_{BH}$. We also assume that the content size S_f is exponentially distributed with mean $\overline{S}$ bits. Hence, the corresponding request service times of the three transmission modes also follow an exponential distribution with means $\tau_{lc} = \frac{\overline{S}}{R_D}$ sec, $\tau_{rc} = \frac{\overline{S}}{R_{WL}}$ sec, and $\tau_{bh} = \frac{\overline{S}}{R_{BH}}$ sec, respectively.

3.4 Problem Formulation

In this section, we characterize the network average delay on a per request basis from the global network perspective. Specifically, we study the request arrival rate and the traffic dynamics from a queuing theory perspective and get a closed form expression for the network average delay.

3.4.1 File Popularity Distribution

We assume that the popularity distribution of files in all clusters follows a Zipf's distribution with skewness order β [91]. However, it is assumed that the content may vary across clusters. This is inspired by the fact that, for instance, devices in a library may be interested in an entirely different set of files from the devices in a sports center. This assumption for the popularity distribution is extended from [23], where the authors explained that the scaling of popular files is sublinear with the number of devices.[2]

[2]The number of popular files increases with the number of devices with a rate slower than the linear polynomial rate, e.g., the logarithmic rate.

To illustrate, if user 1 and user 2 are interested in a set of files of size m_0, then, both of the users will be interested in a common subset of size $m_0/2$ files and the other subset (of size $m_0/2$) will be unique per each user. User 3, in turn, will share the interest in $2m_0/3$ files with user 1 or user 2, and will be interested in a unique subset of size $m_0/3$, etc. The union of all demanded (popular) files by n different users is hence $m = m_0(1 + \frac{1}{2} + \frac{1}{3} + \dots) = m_0 \sum_{i=1}^{n} \frac{1}{i} \approx m_0 \log n$. Therefore, the library size increases sublinearly with the number of devices. In this work, we assume that the scaling of the library size is sublinear with the number of clusters. The cell is divided into clusters with a small number of devices per cluster, such that devices in the same cluster are assumed to request files according to the same file popularity distribution function (i.e., devices in the same cluster are interested in the same set of popular files).

The probability that a file f is requested in cluster k, with m_0 highly demanded files in each cluster, follows a Zipf distribution written as [91],

$$P_{k,f} = \frac{(f - \frac{k-1}{k} m_0 a + (m - \frac{k-1}{k} m_0) b)^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}}, \tag{3.1}$$

where $a = \mathbb{1}(f > \frac{k-1}{k} m_0)$ and $b = \mathbb{1}(f \leq \frac{k-1}{k} m_0)$, $\frac{k-1}{k} m_0$ is the order of the most popular file in the $k-$th cluster, and $\mathbb{1}(.)$ is the indicator function. When $k = 1$, we get $P_{1,f} = \frac{(f)^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}}$ for the first cluster, which is the Zipf's distribution with the most popular file $f = 1$. For example, if $m_0 = 60$, then $P_{2,f} = \frac{(f-30a+(m-30)b)^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}}$ for the second cluster, which is the Zipf's distribution with the most popular file $f = \frac{m_0}{2} + 1 = 31$; also $f = \frac{2m_0}{3} + 1 = 41$ is the most popular file in the third cluster, and so on.

3.4.2 Arrival and Service Rates

The arrival rates for the three communication modes M_{lc}, M_{rc}, and M_{bh} in a cluster k are denoted respectively by $\lambda_{k,lc}$, $\lambda_{k,rc}$, and $\lambda_{k,bh}$ while the corresponding service rates are represented by μ_{lc}, μ_{rc}, and μ_{bh}. For the local cluster mode, we have

$$\lambda_{k,lc} = \lambda_k \sum_{f=1}^{m} P_{k,f}\, c_{k,f}\,, \tag{3.2}$$

where $\sum_{f=1}^{m} P_{k,f}\, c_{k,f}$ is the probability that the requested file is cached locally in cluster k. The corresponding service rate is $\mu_{lc} = \frac{1}{\tau_{lc}}$. For the remote cluster mode, the request arrival rate is defined as

$$\lambda_{k,rc} = \lambda_k \sum_{f=1}^{m} P_{k,f}\,(1 - c_{k,f}\,)\min\Big(\sum_{j \in \mathcal{K}\setminus\{k\}} c_{j,f}, 1\Big), \tag{3.3}$$

where $\min(\sum_{j\in\mathcal{K}\setminus\{k\}} c_{j,f}, 1)$ equals one only if the content f is cached in at least one of the remote clusters. Hence, $\sum_{f=1}^{m} P_{k,f}\,(1-c_{k,f}\,)\min(\sum_{j\in\mathcal{K}\setminus\{k\}} c_{j,f}, 1)$ is the probability that the requested file f is cached in any of the remote clusters given that it is not cached in the local cluster k. The corresponding service rate is $\mu_{rc} = \frac{1}{\tau_{rc} N_a}$, where N_a represents the number of cooperating clusters simultaneously served by the remote cluster mode, i.e, the number of clusters which share the cellular rate.

Finally, for the backhaul mode, the request arrival rate is written as

$$\lambda_{k,bh} = \lambda_k \sum_{f=1}^{m} P_{k,f} \prod_{k=1}^{K} (1 - c_{k,f}\,), \tag{3.4}$$

where $\sum_{f=1}^{m} P_{k,f} \prod_{k=1}^{K} (1 - c_{k,f}\,)$ is the probability that the requested file f is not cached entirely in the cell, so this content could be downloaded only from the core network. The corresponding service rate is $\mu_{bh} = \frac{1}{\tau_{bh} N_b}$, where N_b is defined as the number of clusters simultaneously served via the backhaul mode.

The traffic intensity of a queue is defined as the ratio of mean service time to mean inter-arrival time. We introduce ρ_k as a metric of the traffic intensity at cluster k as

$$\rho_k = \frac{\lambda_{k,lc}}{\mu_{lc}} + \frac{\lambda_{k,rc}}{\mu_{rc}} + \frac{\lambda_{k,bh}}{\mu_{bh}} \tag{3.5}$$

Similar to [92], we consider $\rho_k < 1$ as the stability condition, otherwise, the overall delay will be infinite. The traffic intensity at any cluster is closely related to the request arrival rate and the transmission rates of the three serving modes.

3.4.3 Network Average Delay

In [92], it is proven that the mean queue size for an MPSQ with arrival rate λ [sec^{-1}] and traffic intensity ρ, is

$$\rho + \frac{\lambda \sum_i \frac{\lambda_i}{\mu_i^2}}{1-\rho},$$

where λ_i and μ_i are respectively the arrival and service rates of a service group i. Given the fact that the average delay equals the mean queue size divided by the arrival rate, substituting the above expression to calculate the average delay per request in a cluster k yields

$$D_k = \frac{\rho_k}{\lambda_k} + \frac{\frac{\lambda_{k,lc}}{\mu_{lc}^2} + \frac{\lambda_{k,rc}}{\mu_{rc}^2} + \frac{\lambda_{k,bh}}{\mu_{bh}^2}}{1-\rho_k} \tag{3.6}$$

Based on the analysis of the delay in a single cluster, we derive the network weighted average delay per request as

$$D = \frac{1}{\lambda} \sum_{k=1}^{K} \lambda_k D_k, \tag{3.7}$$

where $\lambda = \sum_{i=1}^{K} \lambda_i$ denotes the overall user request arrival rate in the cell. We observe from (3.6) that the cluster per request delay D_k, and correspondingly the network average delay D, depend on the arrival rates of the three transmission modes, which are in turn functions of the content caching scheme. Because of the limited caching capacity on mobile devices, we would like to optimize the cache placement in each cluster to minimize the network weighted average delay per request. The delay optimization problem is then formulated as

$$\underset{c_{k,f}}{\text{minimize}} \quad D \tag{3.8}$$

$$\text{subject to} \quad \sum_{f=1}^{m} c_{k,f} \leq N, \tag{3.9}$$

$$c_{k,f} \in \{0,1\}, \tag{3.10}$$

where (3.9) and (3.10) are the constraints that the maximum cache size is N files per cluster, and the file is either cached entirely or is not cached, i.e., no partial

caching is allowed. The objective function in (3.8) is not a convex function of the cache placement elements $c_{k,f} \in \{0,1\}$. Moreover, this equation can be reduced to a well- known $0-1$ knapsack problem which is already proven to be NP-hard in [93].

Remark 3.4.3.1 ($N \geq m$)**.** In this case, the caching problem is trivial, i.e., there are no caching constraints. For any cluster k, $c_{k,f} = 1 \quad \forall f \in \mathcal{F}$ and $\sum_{f=1}^{m} c_{k,f} = m$. The optimal solution is obtained when all the files are cached in each cluster. All the requests are served internally from the local cluster via D2D communication.

In the next section, we analyze the network average delay under several caching policies. We further reformulate the optimization problem in (3.8) as a well-known structure that has a locally optimal solution within a factor $(1 - e^{-1})$ of the global optimum.

3.5 Proposed Caching Schemes

3.5.1 Caching Popular Files

In each cluster, the most popular files for the devices in the cluster are cached without repetition. Since popular files are different among clusters (but overlapped), applying CPF might end up replicating the same file in many clusters [86]. We assume that the request arrival rate λ_k is equal for all clusters.

Arrival Rate for D2D communication

The arrival rate of the D2D communication mode is given by

$$\lambda_{k,lc} = \lambda_k \sum_{f=\frac{k-1}{k}m_0+1}^{\frac{k-1}{k}m_0+N} P_{k,f}\,, \tag{3.11}$$

where $\sum_{f=\frac{k-1}{k}m_0+1}^{\frac{k-1}{k}m_0+N} P_{k,f}$ is the probability that the requested file is cached in the local cluster k, and $f = \frac{k-1}{k}m_0 + 1$ is the most popular file index for cluster k. As an example, for the first cluster, $\lambda_{1,lc} = \lambda_1 \sum_{f=1}^{N} P_{1,f}$.

Arrival Rate for Inter-cluster Communication

The arrival rate of the inter-cluster communication mode is given by

$$\lambda_{k,rc} = \lambda_k \sum_{j \in \mathcal{K} \setminus \{k\}} \sum_{f=c}^{\frac{j-1}{j} m_0 + N} P_{k,f} \,, \tag{3.12}$$

where c is defined as $\max\left(\frac{j-2}{j-1} m_0 + N + 1, \frac{j-1}{j} m_0 + 1\right)$. To explain, the inner summation $\sum_{f=c}^{\frac{j-1}{j} m_0 + N} P_{k,f}$ represents the probability that the requested file f is cached in a remote cluster $j \neq k$, where the cached files in the $j-$th cluster are indexed from $f = \frac{j-1}{j} m_0 + 1$ to $f = \frac{j-1}{j} m_0 + N$. c is defined such that a cached file in the remote clusters is counted only once when calculating $\lambda_{k,rc}$. The outer summation is the sum over all clusters except the local cluster k.

To compute the service rate of the remote cluster mode, μ_{rc}, we first need to obtain the number of cooperating clusters N_a since they share the cellular rate. As introduced in Section 3.4, N_a is a random variable representing the number of clusters served by the cellular communication whose mean is given by

$$\overline{N_a} = K \frac{\lambda_{k,rc}}{\lambda_k} = K \sum_{j \in \mathcal{K} \setminus \{k\}} \sum_{f=c}^{\frac{j-1}{j} m_0 + N} P_{k,f} \tag{3.13}$$

Arrival Rate for Backhaul Communication

The arrival rate of the backhaul communication mode is now calculated as

$$\begin{aligned} \lambda_{k,bh} &= \lambda_k \big(1 - (\lambda_{k,lc} + \lambda_{k,rc})\big) \\ &= \lambda_k \Big(1 - \big(\sum_{f=\frac{k-1}{k} m_0 + 1}^{\frac{k-1}{k} m_0 + N} P_{k,f} + \sum_{j \in \mathcal{K} \setminus \{k\}} \sum_{f=c}^{\frac{j-1}{j} m_0 + N} P_{k,f} \big)\Big) \end{aligned} \tag{3.14}$$

N_b is then obtained to calculate the backhaul service rate μ_{bh}. As alluded to in the definition of N_a, N_b is a random variable representing the number of clusters served via the backhaul link whose mean is given by

$$\overline{N_b} = K \frac{\lambda_{k,bh}}{\lambda_k} = K \Big(1 - \big(\sum_{f=\frac{k-1}{k} m_0 + 1}^{\frac{k-1}{k} m_0 + N} P_{k,f} + \sum_{j \in \mathcal{K} \setminus \{k\}} \sum_{f=c}^{\frac{j-1}{j} m_0 + N} P_{k,f} \big)\Big) \tag{3.15}$$

Obviously, we have $\lambda_k = \lambda_{k,lc} + \lambda_{k,rc} + \lambda_{k,bh}$. From (3.11), (3.12), and (3.14), the network average delay can be calculated directly from (3.7). The CPF scheme is computationally straightforward if the most popular content is known. Additionally, the CPF scheme is easy to implement in an independent manner since it is executed in a per cluster level regardless of the caching status of other clusters, which is different from the greedy algorithm proposed in the next subsection. However, it achieves high performance only if the popularity exponent β is large enough, i.e., when the content popularity distribution is skewed, since a small portion of content is highly demanded which can be cached entirely in each cluster.

3.5.2 Greedy Caching Algorithm

In this subsection, we introduce a computationally efficient caching algorithm. We prove that the minimization problem in (3.8) can be reformulated as a *minimization of a supermodular function* subject to *uniform partition matroid constraints*. This structure has a greedy solution which has been proven to be locally optimal within a factor $(1 - e^{-1})$ of the optimum [94–96].

We start with the definition of supermodular and matroid functions, then we introduce and prove some relevant lemmas.

Supermodular Functions

Let $\mathcal{S}$ be a finite ground set. The power set of the set $\mathcal{S}$ is the set of all subsets of $\mathcal{S}$, including the empty set and $\mathcal{S}$ itself. A set function g, defined on the powerset of $\mathcal{S}$ as g: $2^{\mathcal{S}} \rightarrow \mathbb{R}$, is supermodular if for any $A \subseteq B \subseteq \mathcal{S}$ and $x \in \mathcal{S} \setminus B$ we have [95]

$$g(A \cup \{x\}) - g(A) \leq g(B \cup \{x\}) - g(B) \tag{3.16}$$

To illustrate, let $g_A(x) = g(A \cup x) - g(A)$ denote the marginal value of an element $x \in \mathcal{S}$ with respect to a subset $A \subseteq \mathcal{S}$. Then, $\mathcal{S}$ is supermodular if for all $A \subseteq B \subseteq \mathcal{S}$ and for all $x \in \mathcal{S} \setminus B$, we have $g_A(x) \leq g_B(x)$, i.e., the marginal value of the included set is lower than the marginal value of the including set [95].

Matroid Functions

Matroids are combinatorial structures that generalize the concept of linear independence in matrices [95]. A matroid $\mathcal{M}$ is defined on a finite ground set $\mathcal{S}$ and a collection of subsets of $\mathcal{S}$ said to be independent. The family of these independent sets is denoted by $\mathcal{I}$ or $\mathcal{I}(\mathcal{M})$. It is common to refer to a matroid $\mathcal{M}$ by listing its ground set and its family of independent sets, i.e., $\mathcal{M} = (\mathcal{S}, \mathcal{I})$. For $\mathcal{M}$ to be a matroid, $\mathcal{I}$ must satisfy these three conditions:

- $\mathcal{I}$ is a nonempty set.
- $\mathcal{I}$ is downward closed; i.e., if $B \in \mathcal{I}$ and $A \subseteq B$, then $A \in \mathcal{I}$.
- If A and B are two independent sets of $\mathcal{I}$ and B has more elements than A, then $\exists e \in B \setminus A$ such that $A \cup \{e\} \in \mathcal{I}$.

One special case is a partition matroid in which the ground set $\mathcal{S}$ is partitioned into disjoint sets $\{S_1, S_2, \ldots, S_l\}$, where

$$\mathcal{I} = \{A \subseteq \mathcal{S} : |A \cap S_i| \leq k_i \text{ for all } i = 1, 2, \ldots, l\}, \tag{3.17}$$

for some given integers $k_1, k_2, \ldots, k_l$. One special case of the partition matroid is the uniform partition matroid in which $k_1 = k_2 = \cdots = k_l$.

Lemma 3.5.2.1. *The constraints in (3.9) and (3.10) can be rewritten as a uniform partition matroid on a ground set that characterizes the caching elements on all clusters.*

Proof. Please see Appendix A.1 for the proof. □

Lemma 3.5.2.2. *The objective function in equation (3.8) is a monotone non-increasing supermodular function.*

Proof. Please see Appendix A.2 for the proof. □

The greedy solution for this problem structure has been proven to be locally optimal within a factor $(1-e^{-1})$ of the optimum [94–96]. The GCA for the proposed D2D caching system with inter-cluster cooperation is illustrated in Algorithm 1, where S_k^f is an element denoting the placement of file f into the VCC of cluster

k. We first define the attributes of the system in the first line of the algorithm's pseudocode. We then initialize the cache memory of all clusters to zero. We set the number of iterations to be NK, which means that at each iteration, we cache one file in one cluster, resulting in caching N different files in K clusters after NK iterations. In each iteration, all combinations of caching a file $f \in \mathcal{F}$ in a cluster $k \in \mathcal{K}$ are tried, and the network service delay is calculated. A file f^* is chosen to be cached in the k^*-th cluster, which achieves the highest reduction in the network service delay.

The greedy algorithm is run at the BS level, and the BS then instructs the clusters' devices to cache the files according to the output of this algorithm. The deterministic caching approach (both CPF and GCA) can only be realized if the devices stay at the same locations for many hours. Otherwise, performance obtained with the deterministic caching strategy serves as a useful upper bound for more realistic schemes [86]. As examples of the greedy algorithm, the authors in [96] showed that the problem of optimal joint caching and routing can be formulated as maximization of a monotone submodular function subject to matroid constraints, and hence can be solved by the greedy algorithm. Also, the authors in [**mono**] showed that the delay minimization problem can be formulated as a minimization of a submodular function under matroid constraints, which can be solved by the greedy algorithm.

Algorithm 1: Greedy caching algorithm

Input : K, m, N, β, $\overline{S}$, R_D, $\overline{R_{WL}}$, $\overline{R_{BH}}$;
Initialization: $C \leftarrow (0)_{K \times F}$;
`/* Check if all clusters (devices' memories in each cluster) are fully cached. */`
while $\sum_{k=1}^{K} \sum_{f=1}^{m} c_{k,f} < NK$ **do**
  $(k^*, f^*) \leftarrow \text{argmax}_{(k,f)} D(C) - D(C \cup S_k^f)$;
  `/* File achieving highest marginal value is cached. */`
  $c_{k^*,f^*} = 1$;
end while
Output: Cache placement C;

3.6 Throughput Analysis

We have analyzed the per request average delay from the network perspective under different caching schemes. In this section, we conduct the per request throughput and throughput scaling analysis. We first characterize the per request throughput from the queuing theory perspective based on the analytical results of previous sections, then study the scaling of the average sum throughput when the number of files asymptotically goes to infinity.

3.6.1 Per request Throughput Analysis

In this subsection, we first formulate a condition on the traffic demand for the network to be stable, then we study the throughput per request from the cluster perspective. As introduced in Section 3.3.2, the content size S_f is assumed to have an exponential distribution with mean $\overline{S}$ [bits]. For a cluster $k \in \mathcal{K}$ whose devices' traffic is modeled as an MPSQ with three serving processors (transmission modes), the number of devices' requests in the queue that matches the j-th transmission mode is denoted by x_j, where $j \in \mathbb{D} := \{M_{lc}, M_{rc}, M_{bh}\}$. Denote $\mathbf{x} = (x_j)_{j\in\mathbb{D}}$ as the vector counting the numbers of devices' requests in the queue for each transmission mode $j \in \mathbb{D}$.

The process $\{X(t); t \geq 0\}$ describing the number of devices' requests served by the three serving processors (transmission modes) is then a continuous-time Markov process [90]. This process has a discrete state space $\mathbb{N}^{\mathbb{D}}$ and admits the following generator [90]:

$$\begin{cases} q(\mathbf{x}, \mathbf{x} + \epsilon_j) = \lambda_{k,j}, & \mathbf{x} \in \mathbb{N}^{\mathbb{D}}, j \in \mathbb{D}, \\ q(\mathbf{x}, \mathbf{x} - \epsilon_j) = \frac{R_j}{\overline{S}} \frac{x_j}{x_{\mathbb{D}}}, & \mathbf{x} \in \mathbb{N}^{\mathbb{D}}, j \in \mathbb{D}, x_j > 0, \end{cases}$$

where ϵ_j designates the vector of $\mathbb{N}^{\mathbb{D}}$ having coordinate 1 at position j and 0 elsewhere, and $x_{\mathbb{D}} := \sum_{j\in\mathbb{D}} x_j$. The first term of the above generator, $q(\mathbf{x}, \mathbf{x} + \epsilon_j)$, accounts for the arrival of a request that matches the $j-$th transmission mode while the second term, $q(\mathbf{x}, \mathbf{x}-\epsilon_j)$, accounts for serving a request by the $j-$th transmission mode.

Let $\mathbf{X} = (X_{lc}, X_{rc}, X_{bh})$ be the vector counting the number of devices' requests of each transmission mode at the steady state, and let $X_{\mathbb{D}} := \sum_{j\in\mathbb{D}} X_j$ be the total number of requests in the queue at the steady state. The average traffic demand ζ_j [bps] of each transmission mode $j \in \mathbb{D}$ in the $k-$th cluster is defined as [97]

$$\zeta_j = \lambda_{k,j}\overline{S}, \tag{3.18}$$

and the total traffic demand per cluster is then given by

$$\zeta = \sum_{j\in\mathbb{D}} \zeta_j \tag{3.19}$$

We now obtain the cluster critical traffic demand, beyond which the MPSQ is no longer stable. The constraint (3.5) that limits the traffic intensity ρ_k from the above to one can be rewritten as

$$\begin{aligned} \rho_k &= \frac{\lambda_{k,lc}}{R_D/\overline{S}} + \frac{\lambda_{k,rc}}{\overline{R_{WL}}/\overline{S}} + \frac{\lambda_{k,bh}}{\overline{R_{BH}}/\overline{S}} \leq 1, \\ &\frac{\zeta_{lc}}{R_D} + \frac{\zeta_{rc}}{\overline{R_{WL}}} + \frac{\zeta_{bh}}{\overline{R_{BH}}} \leq 1, \end{aligned} \tag{3.20}$$

by multiplying both sides by ζ and rearranging the terms, we get

$$\begin{aligned} \zeta &\leq \frac{\zeta}{(R_D^{-1}\zeta_{lc} + \overline{R_{WL}}^{-1}\zeta_{rc} + \overline{R_{BH}}^{-1}\zeta_{bh})}, \\ \zeta &\leq \zeta_c, \end{aligned} \tag{3.21}$$

where ζ_c [bps] is the critical traffic demand per cluster, beyond which the MPSQ loses its stability.

Lemma 3.6.1.1. *The steady state distribution of the total number of devices' requests in the MPSQ modeling the devices' traffic follows a geometric distribution with parameter* $p = 1 - \zeta/\zeta_c$.

Proof. This result can be deduced from [90] and the references therein, and the proof is omitted in this chapter to avoid repetition. □

As a direct result from Lemma 3.6.1.1, the mean number of total devices' requests in the MPSQ at the steady state is given by

$$\overline{N_q} = E[X_{\mathbb{D}}] = \frac{p}{1-p} = \frac{\zeta}{\zeta_c - \zeta} \tag{3.22}$$

At the steady state, the queue throughput is equal to the traffic demand ζ. Hence, the average throughput per request is defined as the ratio of the given queue throughput and the average number of devices' requests, i.e.,

$$\overline{r} = \frac{\zeta}{E[X_{\mathbb{D}}]} = \zeta_c - \zeta \tag{3.23}$$

3.6.2 Throughput Scaling Analysis

We conduct the scaling analysis of the average sum throughput when the number of files grows asymptotically to infinity, i.e., $m \to \infty$. We first define the outage probability for the proposed D2D cooperative caching system and then compare it with a clustered D2D caching system without inter-cluster cooperation [98]. The obtained formula of the outage probability is further approximated and then exploited in the throughput scaling analysis.

In the following, we shall implicitly ignore the non-integer effects when they are irrelevant for the scaling laws. For example, recalling that the network has node density n and it is divided into K clusters, the number of devices per cluster after integer rounding is denoted as y. Next, we conduct the analysis for the CPF scheme. Since the backhaul rate is considered much smaller than the rate of cellular and D2D communications, we assume that the throughput from the backhaul communication is negligible as compared to the cellular and D2D throughput.

Outage Probability

For a reference clustered D2D caching network without inter-cluster cooperation [98], the probability of no outage is defined as the probability that a randomly chosen user u can download a requested file from nearby devices in the same cluster [98]. Conversely, a user u is said to be in outage when its requested file is not cached within the allowed transmission range (i.e., not cached in a neighbor user

in the same cluster). In this cooperative clustered model, a user u is said to be in outage when the requested file is neither stored in the local cluster nor any of the remote clusters. We denote this outage probability as p_o, which also represents the percentage of devices who are in outage in relation to the total number of devices; the probability of no outage is then denoted as $1 - p_o$.

As stated before, the number of devices per cluster, denoted as y, equals (n/K). In addition, the probability of no outage, $1 - p_o$, can be calculated by determining the probability that a randomly chosen user u in cluster k is served via the local cluster or the remote cluster modes. The probability of no outage is therefore expressed as the sum of two terms, the first term is corresponding to the probability of serving requests from the local cluster, and the second term is the probability of being served from a remote cluster. From (3.11) and (3.12), and under the assumption of the CPF scheme, the probability of no outage is given by

$$1 - p_0 = \sum_{f=\frac{k-1}{k}m_0+1}^{\frac{k-1}{k}m_0+My} P_{k,f} + \sum_{j\in\mathcal{K}\setminus\{k\}} \sum_{f=c}^{\frac{j-1}{j}m_0+My} P_{k,f}\,, \tag{3.24}$$

where M is the maximum user cache size in files (the default is $M = 1$), and c is defined in (3.12). Substituting $P_{k,f}$ from (3.1), we obtain the result

$$1 - p_0 = \frac{\sum_{f=\frac{k-1}{k}m_0+1}^{\frac{k-1}{k}m_0+My} f^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}} + \sum_{j\in\mathcal{K}\setminus\{k\}} \frac{\sum_{f=c}^{\frac{j-1}{j}m_0+My} f^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}} \tag{3.25}$$

Due to the symmetry between clusters in terms of the cache content, cluster cache size, and the probability of being served from a remote cluster, we continue with the assumption that the user u is being served from the first cluster (i.e., $k = 1$) and the remote clusters (the potential cooperating clusters) are from $k = 2$ to $k = K = n/y$.

$$\begin{aligned} 1 - p_0 &= \frac{\sum_{f=1}^{My} f^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}} + \sum_{j=2}^{\frac{n}{y}} \frac{\sum_{f=c}^{\frac{j-1}{j}m_0+My} f^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}} \\ &= \frac{\sum_{f=1}^{My} f^{-\beta}}{\sum_{i=1}^{m} i^{-\beta}} + \frac{1}{\sum_{i=1}^{m} i^{-\beta}} \sum_{j=2}^{\frac{n}{y}} \sum_{f=c}^{\frac{j-1}{j}m_0+My} f^{-\beta} \end{aligned} \tag{3.26}$$

We now aim at deriving an approximated version of (3.26) by replacing the

summations with approximated integrals from [99], and then the obtained result is used later in the throughput scaling analysis. We have two approximations from [99],

$$\sum_{i=1}^{q} i^{-\alpha} \approx \int_{i}^{q+1} x^{-\alpha} dx = \frac{(q+1)^{1-\alpha}-1}{1-\alpha}, \tag{3.27}$$

and

$$\begin{aligned}\sum_{i=w+1}^{q-1} i^{-\alpha} &\approx \int_{w}^{q} x^{-\alpha} dx - \frac{w^{\alpha}+q^{\alpha}}{2}, \\ &= \frac{q^{1-\alpha}-w^{1-\alpha}}{1-\alpha} - \frac{w^{\alpha}+q^{\alpha}}{2}\end{aligned} \tag{3.28}$$

The above approximations are quite tight for small values of the popularity exponent, e.g., when $\beta < 1$. Substituting (3.27) and (3.28) into (3.26) yields

$$\begin{aligned}1-p_0 \approx & \frac{\frac{1}{1-\beta}(My+1)^{1-\beta} - \frac{1}{1-\beta}}{\frac{1}{1-\beta}(m+1)^{1-\beta} - \frac{1}{1-\beta}} + \\ & \frac{1}{\frac{1}{1-\beta}(m+1)^{1-\beta} - \frac{1}{1-\beta}} \sum_{j=2}^{\frac{n}{y}} \Big(\frac{\big(\frac{j-1}{j}m_0 + My + 1\big)^{1-\beta} - \big(c'\big)^{1-\beta}}{1-\beta} - \frac{\big(c'\big)^{\beta} + \big(\frac{j-1}{j}m_0 + My + 1\big)^{\beta}}{2} \Big), \\ & = \overline{p}_{0,nc} + \overline{p}_{0,wc}\end{aligned} \tag{3.29}$$

where c' =max($\frac{j-i}{j}m_0, \frac{j-2}{j-1}m_0 + My$), and $\overline{p}_{0,nc}$, $\overline{p}_{0,wc}$ represent respectively the probability of no outage for a non-cooperative system and the improvement (increase) in the probability of no outage due to the inter-cluster cooperation. In Fig. 3.4, we plot the outage probability of the proposed system with inter-cluster cooperation compared to a reference system without inter-cluster cooperation. We note that as the number of devices per cluster increases, the outage probability correspondingly decreases. That is attributed to the fact that the probability of obtaining the requested files from the local cluster increases with the number of devices per cluster.

Throughput Scaling Analysis

We now express the network average sum throughput, denoted as T_{sum}^{avg} (bps), as a function of the system parameters, namely, number of devices, library size, and popularity exponent. Based on the assumed interference model, only one D2D

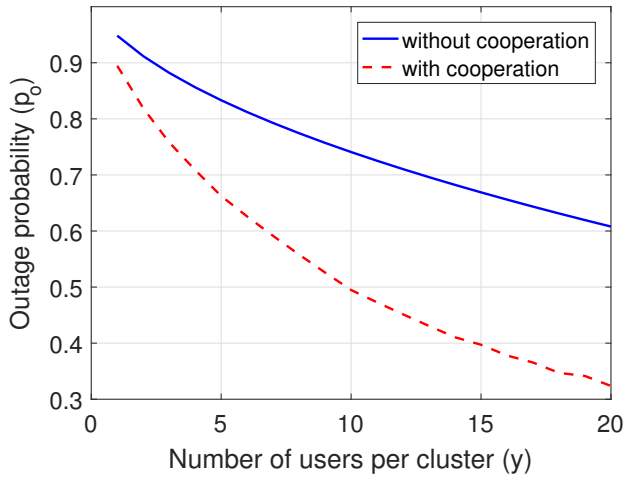


FIGURE 3.4: Outage probability of a D2D clustered caching system with cooperation compared to a reference system without cooperation [98] ($m = 108, n = 120, M = 1, m_0 = 60, \beta = 0.5$).

link can be active at any time in each cluster. Whenever there is an active D2D link within a cluster, we say the cluster is good.[3] We also assume that a D2D link is scheduled in any cluster whenever the opportunity arises, i.e., if the user to be served in a cluster requests a file not cached locally, the request is then served via the appropriate transmission mode (remote cluster or backhaul modes), meanwhile, another D2D link is scheduled inside the cluster. In addition to the D2D throughput, there is the cellular throughput from the remote cluster mode. So the instantaneous throughput T_{sum} (bps) can be written as

$$T_{sum} = \text{D2D throughput} + \text{Cellular throughput},$$

and the average sum throughput is obtained from

$$T_{sum}^{avg} = R_{D2D} E[L] + R_{WL} P_{rc}, \tag{3.30}$$

where L is the number of active D2D links, $E[L]$ is the expected number of active D2D links, which is approximately the expected number of good clusters, and P_{rc} is the probability of occurrence of cooperation between clusters. For notational

[3] In this article, and different from [98], we neglect the inter-cluster interference. We assume that any cluster can be active whenever there is a scheduled D2D link, regardless of the activity of all other clusters. This assumption makes the calculated throughput an upper bound for the actual throughput.

simplicity, we henceforth substitute R_{D2D} by C (bps) and R_{WL} by $k_1 C$ (bps), where $k_1 < 1$.

$$\begin{aligned} T_{sum}^{avg} &= C\big(E[L] + k_1 P_{rc}\big), \\ &\leq C\big(E[L] + k_1\big), \end{aligned} \tag{3.31}$$

where the above inequality holds because P_{rc} is a probability and cannot be greater than one. In particular, P_{rc} is tight with its upper bound for a large number of clusters and relatively uniform popularity distribution (i.e., not skewed). In the sequel, we calculate the expected number of good clusters $E[L]$.

Up to now, the cell is divided into $K = (n/y)$ virtual clusters, each of them with y uniformly distributed devices. As mentioned before, a cluster is good if at least one user requests a file that can be served from the locally cached content via D2D communication. Conversely, a cluster is not good if all y devices in the same cluster cannot serve their requests from the locally cached content, which occurs with probability $p_{0,nc}^{y}$ [63], where $p_{0,nc} = 1 - \overline{p}_{0,nc}$ is the probability that a randomly chosen user u in any cluster can not obtain a requested file from nearby devices in the same cluster. The probability of having a good cluster is then $1 - p_{0,nc}^{y}$. Therefore, we have the following

$$E[L] = \frac{n}{y}(1 - p_{0,nc}^{y}) \tag{3.32}$$

Substituting $p_{0,nc}$ from (3.29), and (3.32) into (3.31) yields

$$\begin{aligned} T_{sum}^{avg} &\leq C\Big(\frac{n}{y}(1 - p_{0,nc}^{y}) + k_1\Big), \\ &= C\frac{n}{y}\left(1 - (1 - \frac{\frac{1}{1-\beta}(My+1)^{1-\beta} - \frac{1}{1-\beta}}{\frac{1}{1-\beta}(m+1)^{1-\beta} - \frac{1}{1-\beta}})^{y}\right) + k_1 C \end{aligned} \tag{3.33}$$

Similar to [98] and [63], we define the quantity

$$\gamma = \frac{1-\beta}{2-\beta}, \tag{3.34}$$

where γ changes from 0 to $\frac{1}{2}$ when β changes from 1 to 0. These ranges of β and γ are interesting for the scaling analysis since they are reasonable in practice [98].

In the following, we conduct the scaling analysis for the regime when y changes sublinearly with m [98], i.e, $y = \rho m^{\gamma}$ for some constant ρ, and $\gamma \leq \frac{1}{2}$. We analyze the scaling of the upper bound for T_{sum}^{avg} when m asymptotically grows to infinity. Substituting $y = \rho m^{\gamma}$ into (3.33) yields

$$\begin{aligned} T_{sum}^{avg} &\leq C\frac{n}{\rho m^{\gamma}}\left(1-\left(1-\frac{\frac{1}{1-\beta}(My+1)^{1-\beta}-\frac{1}{1-\beta}}{\frac{1}{1-\beta}(m+1)^{1-\beta}-\frac{1}{1-\beta}}\right)^{\rho m^{\gamma}}\right)+k_1C, \\ &= C\frac{n}{\rho m^{\gamma}}\left(1-\left(1-M^{1-\beta}\rho^{1-\beta}F^{(1-\beta)(\gamma-1)}\right)^{\rho m^{\gamma}}\right)+k_1C, \\ &\overset{(a)}{=} C\frac{n}{\rho m^{\gamma}}\left(1-\left(1-M^{1-\beta}\rho^{1-\beta}m^{-\gamma}\right)^{\rho m^{\gamma}}\right)+k_1C, \end{aligned} \tag{3.35}$$

where (a) follows by using $(1-\beta)(\gamma-1) = -\gamma$, then we have

$$\begin{aligned} T_{sum}^{avg} &\leq k_1C + C\frac{n}{\rho m^{\gamma}}\Big(1-\big((1-\rho^{1-\beta}M^{1-\beta}m^{-\gamma})^{\rho^{-(1-\beta)}M^{-(1-\beta)}m^{\gamma}}\big)^{\rho^{2-\beta}M^{-(1-\beta)}}\Big), \\ &\overset{(b)}{=} k_1C + \frac{C}{\rho}\Big(1-\big(e^{-1}\big)^{\rho^{2-\beta}M^{-(1-\beta)}}\Big)\frac{n}{m^{\gamma}}, \end{aligned} \tag{3.36}$$

where (b) follows from $\lim_{x\to\infty}(1-x^{-1})^{x} = e^{-1}$, then we have

$$\begin{aligned} T_{sum}^{avg} &\leq \frac{C}{\rho}\Big(1-e^{-\rho^{2-\beta}M^{-(1-\beta)}}\Big)\frac{n}{m^{\gamma}}+k_1C, \\ &= \Theta\Big(\frac{n}{m^{\gamma}}\Big)+O(1) \end{aligned} \tag{3.37}$$

This result shows that:

- As the library size m increases, the upper bound for T_{sum}^{avg} decreases, since the probability of having active D2D links (good clusters) decreases.
- As γ increases, corresponding to the decrease of the popularity exponent β, the upper bound for T_{sum}^{avg} vanishes more rapidly with the library size m.
- The upper bound for T_{sum}^{avg} scales linearly with the number of devices n.[4]

The average sum throughput is plotted against the number of devices per cluster y in Fig. 3.5, for different values of β. We observe that there is an optimal value of y at which the throughput is maximized. First, the throughput increases with

[4]We use the standard Landau notation: $g(n) = O(g(n))$ denotes $g(n) \leq c_1 g(n)$ and $g(n) = \Theta(g(n))$ denotes $k_1 g(n) \leq g(n) \leq k_2 g(n)$, where c_1, k_1, and k_2 are real constants > 0.

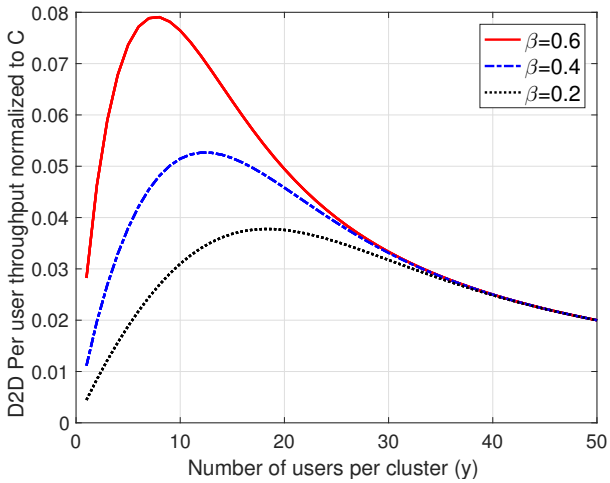


FIGURE 3.5: D2D per-user throughput of the cooperative system is plotted against the number of devices per cluster y at different values of the popularity exponent β (parameters as in [98], $n = 10,000$ devices, $m = 1000$ files , $m_0 = 200$ files).

the cluster size y. Then, as the cluster size increases, the outage probability decreases owing to the higher cache size per cluster. However, for larger cluster size, the throughput starts to decrease owing to the decrease in the number of clusters associated with the larger cluster size.

3.7 Results and Discussions

In this section, we evaluate the performance of the proposed inter-cluster cooperative architecture using simulation and analytical results. Results are obtained with the following parameters: $\lambda_k = 0.5$ requests/sec, $m_0 = 60$ files, $m = 108$ files, $\overline{S} = 4$ Mbits, $K = 5$ clusters, $n = 25$ devices, $M = 4$ files, and $N = 20$ files. $R_{WL} = 50$ Mbps and $R_{BH} = 5$ Mbps as in [94]. For a typical D2D communication system with transmission power of 20 dBm, transmission range of 10 m, and free space path loss model as in [23], we have $R_D = 120$ Mbps. As previously detailed, in our model, we assume a BS whose backhauling capacity is limited. In such a setup, the average service delay can be effectively reduced when leveraging both high data rate D2D communication (in-cluster) and inter-cluster cooperation through the base station as a relay. If this assumption does not hold, i.e., for high capacity backhaul links such as optical fiber or millimeter wave, requests for files not cached in the local

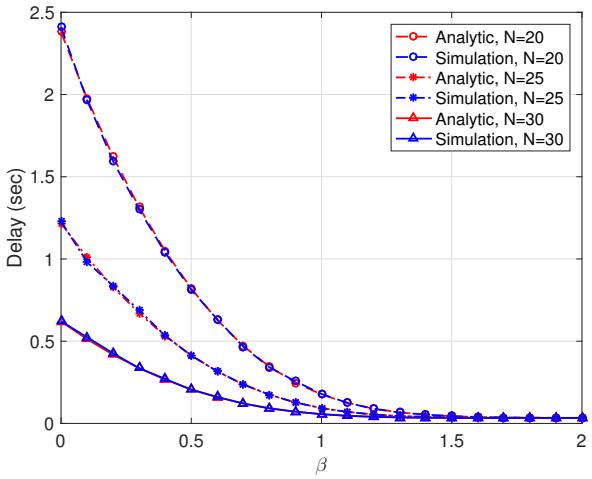


FIGURE 3.6: Network average delay versus popularity exponent β under the caching popular files scheme.

cluster will be served directly from the core network through the high capacity backhaul links.

In Fig. 3.6, we verify the accuracy of the analytical results of the network average delay under the CPF with inter-cluster cooperation. Monte-carlo simulation is adopted where one queue with Poisson arrival rate and three serving processors is modeled and the corresponding average service delay is numerically calculated. The theoretical and simulated results for the network average delay under the CPF scheme are plotted together, and they are consistent. We see that the network average delay is significantly improved by increasing the cluster cache size N. Moreover, as β increases, the average delay decreases. This is attributed to the fact that a small portion of content forms most of the requests that can be cached locally in each cluster and delivered via high data rate D2D communication.

In the following, we evaluate and compare the performance of various caching schemes. In Fig. 3.7, the proposed inter-cluster cooperative caching system is compared with a D2D caching system without cooperation under the CPF scheme. For a D2D caching system without cooperation, requests for files that are not cached in the local cluster are downloaded directly from the core network. For the sake of concise comparison, we define the delay reduction gain as

$$\text{Gain} = 1 - \frac{\text{Delay with inter-cluster cooperation}}{\text{Delay without inter-cluster cooperation}} \tag{3.38}$$

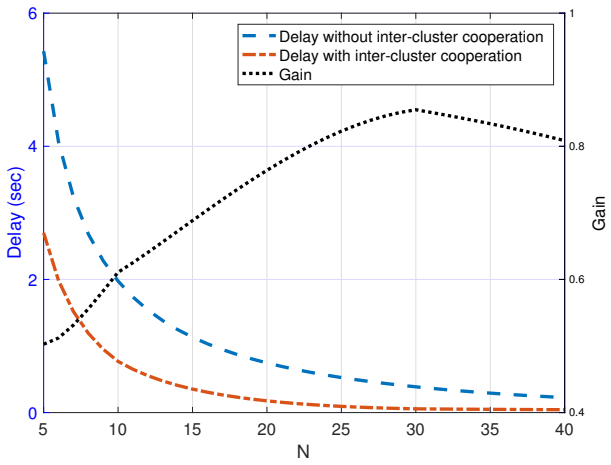


FIGURE 3.7: Network average delay (left hand side y-axis) and gain (right hand side y-axis) vs cluster cache size N.

Fig. 3.7 shows that, for a small cluster cache size, the delay reduction (gain) of the proposed inter-cluster cooperative caching is higher than 45% with respect to a D2D caching system without inter-cluster cooperation and greater than 80% if the cluster cache size is large.

Having discussed the benefits of adopting D2D communications with caching among the network devices, we next turn our attention to the cost associated with caching. This cost comes mainly from the energy consumed by the different devices to store and deliver the requested contents. There is an inherent tradeoff between the energy consumed for content delivery and the delay reduction attained from caching. To show the energy-delay reduction gain tradeoff among the devices, in Fig. 3.8, we plot the per-cluster energy consumption during the local and remote cluster modes and the gain attained from inter-cluster cooperation against the cluster cache size N. $P_{lc} = 20$ dBm, and $P_{rc} = 23$ dBm denote respectively the transmission power in the local cluster and remote cluster modes. In each transmission mode, the energy per request is the transmission power times the transmission duration. The transmission duration is given by the ratio of file size over the transmission rate. We see that the consumed energy during the local cluster transmission, i.e., D2D communication, monotonically increases with the cluster cache size N. With the increasing of N, more requests are served via the local cluster mode M_{lc}. For the consumed energy during the remote cluster transmission, we see that it

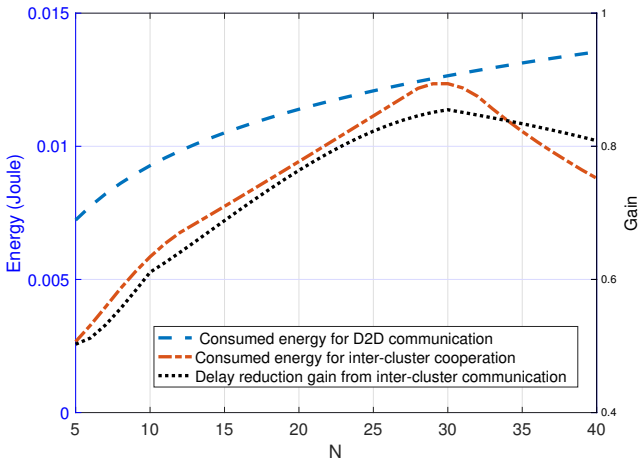
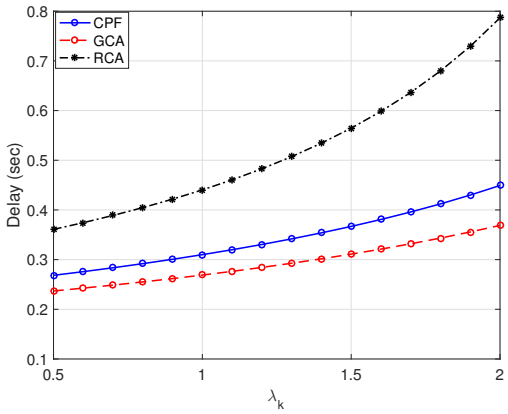


FIGURE 3.8: Energy consumption per cluster during the local and remote cluster transmissions (left hand side y-axis) and the gain attained from inter-cluster cooperation (right hand side y-axis) vs cluster cache size N.

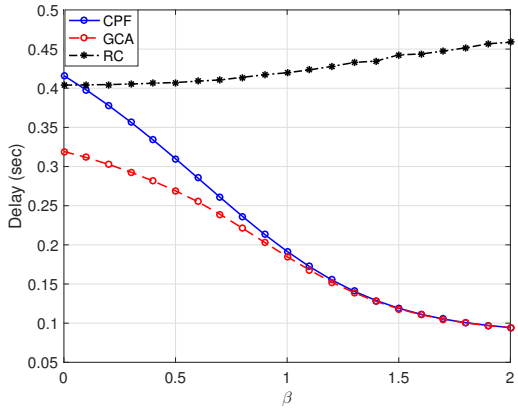
initially increases with N, then it decreases, and the same behavior is observed for the delay-reduction gain. This can be interpreted as follows. When N increases, the number of requests served from the remote clusters increases since the remote clusters' VCCs increase. When N becomes much larger, the local cluster cache becomes sufficiently large to serve most of the requests, as opposed to being served by the remote cluster mode.

For comparison purposes, Fig. 3.9 shows the average delay for the proposed caching schemes and random caching (RC) against various system parameters. Fig. 3.9(a) shows the network average delay plotted against the request arrival rate λ_k for three content placement techniques, namely, GCA, CPF, and RC.[5] In RC, content stored in clusters is randomly chosen from the file library. The most popular files are cached in the CPF scheme, and the GCA works as illustrated in Algorithm 1. We see that the average delay for all content caching strategies increases with λ_k since a larger request rate increases the probability of a longer waiting time for each request. It is also observed that the GCA, which is locally optimal, achieves significant performance gains over the CPF and RC solutions for the above setup. Fig. 3.9(b) shows that the GCA is superior to the CPF only for small values of the

[5]Here, we adopt different transmission rates from [23] and [94] to provide insights on the difference between the caching schemes, otherwise, the GCA is far superior to the other schemes, with negligible delay.



(a) Network average delay vs request arrival rate for three caching schemes, caching popular files, greedy caching algorithm, and random caching.

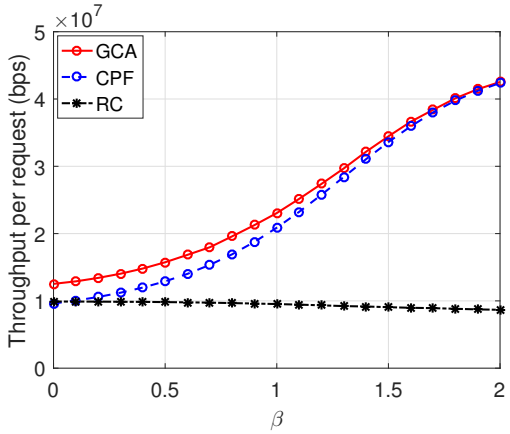


(b) Network average delay vs popularity exponent for three caching schemes, caching popular files, greedy caching algorithm, and random caching.

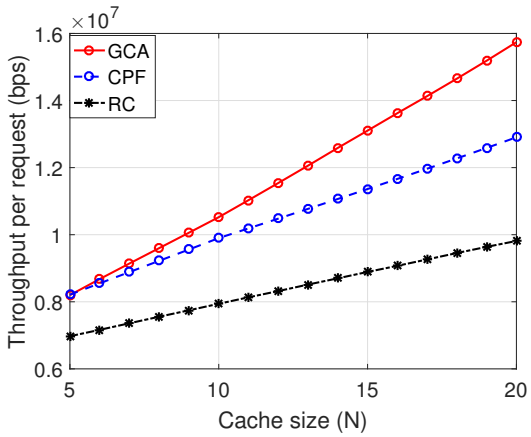
FIGURE 3.9: Evaluation and comparison of the average delay for the proposed caching schemes and random caching for various system parameters ($R_D = 50$ Mbps, $\overline{R_{WL}} = 15$ Mbps, $\overline{R_{BH}} = 10$ Mbps, $N = 20$, $\beta = 0.5$ for (a) and $\lambda_k = 0.5$ requests/sec for (b)).

popularity exponent β. If the popularity exponent β is high enough, CPF and GCA will achieve the same performance. When β increases, the CDF of the Zipf's distribution becomes more skewed. This implies that only a smaller portion of the files is highly demanded by the devices. The lower the number of files requested by the devices, the higher the probability of having such files cached in the clusters' VCCs. If all these files are cached locally in each cluster, the global minimum solution for the delay minimization problem is attained. This interpretation explains why when β increases, the CPF and GCA solutions converge to the global optimal solution. We also note that the CPF and RC schemes roughly achieve the same delay when $\beta = 0$. This stems from the fact that with $\beta = 0$, all files have equal popularity, and correspondingly, CPF is equivalent to RC. Moreover, RC fails to reduce the delay as β increases, since caching files at random results in a low probability of serving the requested files from local clusters.

Next, we turn my attention to the throughput results in Fig. 3.10. Fig. 3.10(a) plots the throughput per request as a function of the popularity exponent β for the three caching schemes. It is shown that the per request throughput monotonically increases with β for the CPF and GCA schemes, and shows a slight decrease for the RC scheme. When β increases for the GCA and CPF, the locally stored files form most of the devices' requests that can be delivered via high rate D2D communication. Conversely, for the RC scheme, which caches the files uniformly at random, the probability of having the requested files cached in the local clusters slightly decreases when the popularity of files becomes skewed (higher β). Due to the resulting lower probability of serving the requests from the local clusters, the throughput per request, in turn, slightly decreases owing to the lower probability of activating D2D links. In Fig. 3.10(b), the throughput per request is plotted against the cluster cache size N for the three caching schemes. It is noticed that for all the caching schemes, the per request throughput improves with the cluster cache size, and the GCA achieves the highest throughput. This can be explained by the fact that, with large cluster cache size, there is a high opportunity of exchanging cached content via the local cluster mode that exploits the high rate of the D2D communication.



(a) The per request throughput vs popularity exponent for three caching schemes, caching popular files, greedy caching algorithm, and random caching.



(b) The per request throughput vs cluster cache size for three caching schemes, caching popular files, greedy caching algorithm, and random caching.

FIGURE 3.10: Evaluation and comparison of the per request throughput for the proposed caching schemes and random caching for various system parameters ($R_D = 50$ Mbps, $\overline{R_{WL}} = 15$ Mbps, $\overline{R_{BH}} = 10$ Mbps, $\lambda_k = 0.5$ requests/sec).

3.8 Chapter Summary

In this chapter, we proposed a novel D2D caching architecture to reduce the network average delay. We studied a cellular network consisting of one SBS and a set of devices. The cell is divided into a set of equally-sized virtual clusters, where the devices in the same cluster exchange cache content via D2D communication, while the devices in different clusters cooperate by exchanging their cache content via cellular transmission. For the proposed system, we considered a special case wherein the backhaul link is overloaded or even not existing such that devices in different cluster can benefit from inter-cluster cooperation for content sharing and delivery. We formulated the delay minimization problem in terms of the content cache placement. However, the problem is NP-hard and obtaining the optimal solution is computationally hard. We then proposed two content caching policies, namely, CPF and greedy caching. By reformulating the delay minimization problem as a minimization of a non-increasing supermodular function subject to uniform partition matroid constraints, We showed that it could be solved using the proposed GCA scheme within a factor $(1 - e^{-1})$ of the optimum. Moreover, we conducted the throughput analysis to investigate the behavior of the average throughput per request under different caching schemes. We studied the scaling behavior of the average sum throughput when the library size asymptotically grows to infinity and show that the network average sum throughput decreases with the library size increase, and the rate of this decrease is controlled by the popularity exponent. We verified the analytical results by means of extensive simulations and the results showed that the network average delay could be reduced by 45%-80% by allowing inter-cluster cooperation.

Chapter 4

Optimizing Joint Probabilistic Caching and Communication for Clustered D2D Networks

The analysis of D2D caching networks based on a physical interference model is usually carried out by assuming that devices are uniformly distributed. However, this approach does not fully consider and characterize the fact that devices are usually grouped into clusters. Motivated by this fact, this chapter presents a comprehensive performance analysis and joint communication and caching optimization for a clustered D2D network. Devices are distributed according to a Thomas cluster process and are assumed to have a surplus memory which is exploited to proactively cache files from a known library, following a random probabilistic caching scheme. Devices can retrieve the requested files from their caches, from neighboring devices in their proximity (cluster), or from the BS as a last resort. Three key performance metrics are optimized in this chapter, namely, offloading gain, energy consumption, and latency. Firstly, we maximize the offloading gain, which is defined as the probability of downloading a requested content from the local cluster with certain target rate, by jointly optimizing channel access and caching probability. Secondly, we formulate and solve the energy minimization problem for the proposed model and obtain the optimal probabilistic caching for the minimum energy consumption. Finally, we jointly optimize the caching scheme as well as bandwidth allocation between D2D and BS-to-Device transmission to minimize the weighted average delay per file request. Employing the block coordinate descent

optimization technique, we propose an efficient iterative algorithm for solving the delay minimization problem. Closed-form solution for the bandwidth allocation sub-problem is also provided. Simulation results show significant improvement in the network performance reaching up to 10%, 17%, and 140% improvement in the offloading gain, energy consumption, and average delay, respectively, compared to the Zipf caching technique.

4.1 Introduction

As presented earlier, we showed that caching at the mobile devices significantly improves system performance by facilitating D2D communications, which enhance the spectrum efficiency and alleviate the heavy burden on backhaul links [40]. Modeling the cache-enabled heterogeneous networks, including SBS and mobile devices, follows two main directions in the literature. The first line of work focuses on the fundamental throughput scaling results by assuming a simple protocol channel model [40, 64], wherein two devices can communicate if they are within a certain distance from each other. The second line of work, which is relevant to the work in this chapter, defined as the physical interference model, considers a more realistic model for the underlying physical layer [71, 72].

4.1.1 Motivation and Contribution

In this chapter, we conduct comprehensive performance analysis and optimization of the joint communication and caching for a clustered D2D network, where the devices have unused memory to cache some files, following a random probabilistic caching scheme. My network model effectively characterizes the stochastic nature of channel fading and clustered geographic locations of devices. Furthermore, this chapter also puts emphasis on the need for considering the traffic dynamics and rate of requests when studying the delay incurred to deliver requests to devices. This work is the first in the literature that conducts a comprehensive spatial analysis of a doubly Poisson cluster process (PCP) (also called doubly PPP [75]) with the devices adopting a slotted-ALOHA random access technique to access a shared channel. The key advantage of adopting the slotted ALOHA access protocol is that

it is a simple yet fundamental MAC protocol, wherein no central controller exists to schedule the users' transmissions.[1] Also, we are the first to incorporate the spatio-temporal analysis in wireless caching networks by combining tools from stochastic geometry and queuing theory in order to analyze and minimize the average delay (see, for instance, [100–103]). The main contributions of this chapter are summarized below.

- We consider a Thomas cluster process (TCP) where the devices are spatially distributed as groups in clusters. The cluster centers are drawn from a parent PPP, and the cluster members are normally distributed around the centers, forming a Gaussian PPP. This organization of the parent and offspring PPPs forms the so-called doubly PPP.
- We conduct the coverage probability analysis where the devices adopt a slotted-ALOHA random access technique. We then jointly optimize the access probability and caching probability to maximize the cluster offloading gain. A closed-form solution of the optimal caching probability sub-problem is provided.
- The energy consumption problem is then formulated and shown to be convex and the optimal caching probability for the minimum energy consumption is also calculated.
- By combining tools from stochastic geometry as well as queuing theory, we minimize the per request weighted average delay by jointly optimizing bandwidth allocation between D2D and BS-to-Device communication and the caching probability. The delay minimization problem is shown to be non-convex. Applying the block coordinate descent optimization technique, the joint minimization problem is solved in an iterative manner.

[1] It is worth mentioning the adopted channel access model in this chapter, slotted ALOHA, can be extended to other channel models without lack of tractability. For example, in [7], we assumed that the cached content is downloaded from the nearest active device and the other devices can be active with a certain probability function.

- We validate the theoretical findings via simulations. Results show a significant improvement in the network performance metrics, namely, the offloading gain, energy consumption, and average delay as compared to other caching schemes proposed earlier in literature.

4.1.2 Related Work

Modeling devices' locations as a PPP is widely employed in the literature, especially in the wireless caching area [24, 71–74]. However, a realistic model for D2D caching networks requires that a given device typically has multiple proximate devices, where any of them can potentially act as a serving device. This is known as clustered devices deployment, which can be characterized by cluster processes [75]. Unlike the popular PPP approach, the authors in [76–78] developed a stochastic geometry based model to characterize the performance of content placement in the clustered D2D network. In [76], the authors discuss two strategies of content placement in a PCP deployment. First, when each device randomly chooses its serving device from its local cluster, and secondly, when each device connects to its k-th closest transmitting device from its local cluster. The authors characterize the optimal number of D2D transmitters that must be simultaneously activated in each cluster to maximize the area spectral efficiency. The performance of cluster-centric content placement is characterized in [77], where the content of interest in each cluster is cached closer to the cluster center, such that the collective performance of all the devices in each cluster is optimized. Inspired by the Matern hard-core point process, which captures pairwise interactions between nodes, the authors in [78] devised a novel spatially correlated caching strategy called hard-core placement (HCP) such that the D2D devices caching the same content are never closer to each other than the exclusion radius.

Energy efficiency in wireless caching networks is widely studied in the literature [24, 73, 74]. For example, an optimal caching problem is formulated in [73] to minimize the energy consumption of a wireless network. The authors consider a cooperative wireless caching network where relay nodes cooperate with the devices to cache the most popular files in order to minimize energy consumption. In [24], the authors investigate how caching at BSs can improve energy efficiency (EE)

of wireless access networks. The condition when EE can benefit from caching is characterized, and the optimal cache capacity that maximizes the network EE is found. It is shown that EE benefit from caching depends on content popularity, backhaul capacity, and interference level. The authors in [74] exploit the spatial repartitions of devices and the correlation in their content popularity profiles to improve the achievable EE. The EE optimization problem is decoupled into two related subproblems, where the first one addresses the issue of content popularity modeling, and the second subproblem investigates the impact of exploiting the spatial repartitions of devices. The authors derive a closed-form expression of the achievable EE and find the optimal density of active small cells to maximize the EE. It is shown that the SBS allocation algorithm improves the energy efficiency and hit probability. However, the problem of EE for D2D based caching is not yet addressed in the literature.

Recently, the joint optimization of delay and energy in wireless caching is conducted, see for instance [104–106]. The authors in [104] jointly optimize the delay and energy in a cache-enabled dense small cell network. The authors formulate the energy-delay optimization problem as a mixed integer programming problem, where file placement, device association with the small cells, and power control are jointly considered. To model the energy consumption and end-to-end file delivery delay tradeoff, a utility function linearly combining these two metrics is used as an objective function of the optimization problem. An efficient algorithm is proposed to approach the optimal association and power solution, which could achieve the optimal tradeoff between energy consumption and end-to-end file delivery delay. In [105], the authors showed that with caching, the energy consumption can be reduced by extending transmission time. However, this may lead to wasted energy if the device never needs the cached content. Based on the random content request delay, the authors study the maximization of EE subject to a hard delay constraint in an additive white Gaussian noise channel. It is shown that the EE of a system with caching can be significantly improved by increasing content request probability and target transmission rate compared with the traditional on-demand scheme, in which the BS transmits content file only after it is requested by the user. However, the problem of energy consumption and joint communication and caching for

clustered D2D networks is not yet addressed in the literature.

4.2 System Model

4.2.1 System Setup

We model the location of the mobile devices with a Thomas cluster process in which the parent points are drawn from a PPP Φ_p with density λ_p, and the daughter points are drawn from a Gaussian PPP around each parent point. In fact, the TCP is considered as a doubly PCP where the daughter points are normally scattered with variance $\sigma^2 \in \mathbb{R}^2$ around each parent point [75]. The parent points and offspring are referred to as cluster centers and cluster members, respectively. The number of cluster members in each cluster is a Poisson random variable with mean $\overline{n}$. The density function of the location of a cluster member relative to its cluster center is

$$f_Y(y) = \frac{1}{2\pi\sigma^2}\exp\Big(-\frac{\|y\|^2}{2\sigma^2}\Big), \qquad y \in \mathbb{R}^2 \tag{4.1}$$

where $\|.\|$ is the Euclidean norm. The intensity function of a cluster is given by $\lambda_c(y) = \frac{\overline{n}}{2\pi\sigma^2}\exp\big(-\frac{\|y\|^2}{2\sigma^2}\big)$. Therefore, the intensity of the process is given by $\lambda = \overline{n}\lambda_p$. We assume that the BSs' distribution follows another PPP Φ_{bs} with density λ_{bs}, which is independent of Φ_p.

4.2.2 Content Popularity and Probabilistic Caching Placement

We assume that each device has a surplus memory of size M designated for caching files. The total number of files is $N_f > M$ and the set (library) of content indices is denoted as $\mathcal{F} = \{1, 2, \ldots, N_f\}$. These files represent the content catalog that all the devices in a cluster may request, which are indexed in a descending order of popularity. The probability that the i-th file is requested follows a Zipf's distribution given by,

$$q_i = \frac{i^{-\beta}}{\sum_{k=1}^{N_f} k^{-\beta}}, \tag{4.2}$$

where β is a parameter that reflects how skewed the popularity distribution is. For example, if $\beta = 0$, the popularity of the files has a uniform distribution. Increasing

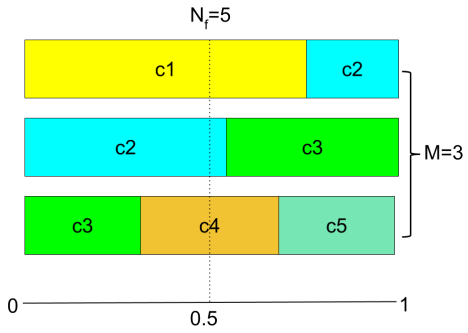


FIGURE 4.1: The cache memory of size $M = 3$ is equally divided into 3 blocks of unit size. Starting from content $i = 1$ to $i = N_f$, each content sequentially fills these 3 memory blocks by an amount b_i. The amounts (probabilities) b_i eventually fill all 3 blocks since $\sum_{i=1}^{N_f} b_i = M$ [107]. Then a random number $\in [0, 1]$ is generated, and content i is chosen from each block, whose b_i fills the part intersecting with the generated random number. In this way, in the given example, the contents $\{1, 2, 4\}$ are chosen to be cached.

β increases the disparity among the files popularity such that lower indexed files have higher popularity. By definition, $\sum_{i=1}^{N_f} q_i = 1$. We use Zipf's distribution to model the popularity of files per cluster.

D2D communication is enabled within each cluster to deliver popular content. It is assumed that the devices adopt a slotted-ALOHA medium access protocol, where each transmitter independently and randomly accesses the channel with the same probability p. This implies that multiple active D2D links might coexist within a cluster. Therefore, p is a design parameter that directly controls the intra-cluster interference, as described later in the next section.

A probabilistic caching model is assumed, where the content is randomly and independently placed in the cache memories of different devices in the same cluster, according to the same distribution. The probability that a generic device stores a particular file i is denoted as b_i, $0 \leq b_i \leq 1$ for all $i \in \mathcal{F}$. To avoid duplicate caching of the same content within the memory of the same device, we follow the caching approach proposed in [107] and illustrated in Fig. 4.1.

If a device caches the desired file, the device directly retrieves the content. However, if the device does not cache the file, it downloads it from any neighboring device that caches the file (henceforth called catering device) in the same cluster. According to the proposed access model, the probability that a chosen catering

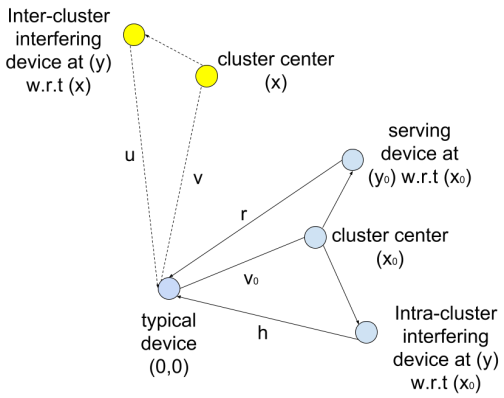


FIGURE 4.2: Illustration of the representative cluster and one interfering cluster.

device is admitted to access the channel is the access probability p. Finally, the device attaches to the nearest BS as a last resort to download the content which is not cached entirely within the device's cluster. We assume that the D2D communication is operating as out-of-band D2D. W_1 and W_2 denote respectively the bandwidth allocated to the D2D and BS-to-Device communication, and the total bandwidth of the system is denoted as $W = W_1 + W_2$. It is assumed that device requests are served in a random manner, i.e., among the cluster devices, a random device request is chosen to be scheduled and content is served.

In the following, we aim at studying and optimizing three important metrics, widely studied in the literature. The first metric is the offloading gain, which is defined as the probability of obtaining the requested file from the local cluster, either from the self-cache or from a neighboring device in the same cluster with a rate greater than a certain threshold. The second metric is the energy consumption which represents the dissipated energy when downloading files either from the BSs or via D2D communication. Finally, the latency which accounts for the weighted average delay over all the requests served from the D2D and BS-to-Device communication.

4.3 Maximum Offloading Gain

Let me define the successful offloading probability as the probability that a device can find a requested file in its own cache, or in the caches of neighboring devices

within the same cluster with D2D link rate higher than a required threshold R_0. Without loss of generality, we conduct the analysis for a cluster whose center is $x_0 \in \Phi_p$ (referred to as representative cluster), and the device which requests the content (henceforth called typical device) is located at the origin. We denote the location of the D2D-TX by y_0 w.r.t. x_0, where $x_0, y_0 \in \mathbb{R}^2$. The distance from the typical device (D2D-RX of interest) to this D2D-TX is denoted as $r = \|x_0 + y_0\|$, which is a realization of a random variable R whose distribution is described later. This setup is illustrated in Fig. 4.2. It is assumed that a requested file is served from a randomly selected catering device, which is, in turn, admitted to access the channel based on the slotted-ALOHA protocol. The successful offloading probability is then given by

$$\mathbb{P}_o(p, b_i) = \sum_{i=1}^{N_f} q_i b_i + q_i(1 - b_i)(1 - e^{-b_i \overline{n}}) \int_{r=0}^{\infty} f(r) \mathbb{P}(R_1(r) > R_0) \, \mathrm{d}r \,, \tag{4.3}$$

where $R_1(r)$ is the achievable rate when downloading content from a catering device at a distance r from the typical device with pdf $f(r)$. The first term on the right-hand side is the probability of requesting a locally cached file (self-cache) whereas the remaining term incorporates the probability that a requested file i is cached among at least one cluster member and being downloadable with a rate greater than R_0. To further clarity, since the number of devices per cluster has a Poisson distribution, the probability that there are k devices per cluster is equal to $\frac{\overline{n}^k e^{-\overline{n}}}{k!}$. Accordingly, the probability that there are k devices caching content i is equal to $\frac{(b_i \overline{n})^k e^{-b_i \overline{n}}}{k!}$. Hence, the probability that at least one device caches content i is 1-minus the void probability (i.e., $k = 0$), which equals $1 - e^{-b_i \overline{n}}$.

In the following, we first compute the probability $\mathbb{P}(R_1(r) > R_0)$ conditioning on the distance r between the typical and catering device, then we relax this condition. The received power at the typical device from a catering D2D-TX located at y_0 from the cluster center is given by

$$P = P_d g_0 \|x_0 + y_0\|^{-\alpha} = P_d g_0 r^{-\alpha} \tag{4.4}$$

where P_d denotes the D2D transmission power, $g_0 \sim \exp(1)$ is an exponential random variable which models Rayleigh fading and $\alpha > 2$ is the path loss exponent. Under the above assumption, the typical device sees two types of interference, namely, the intra-and inter-cluster interference. We first describe the inter-cluster interference, then the intra-cluster interference is characterized. The set of active devices in any remote cluster is denoted as $\mathcal{B}^p$, where p refers to the access probability. Similarly, the set of active devices in the local cluster is denoted as $\mathcal{A}^p$. Similar to (4.4), the interference from the simultaneously active D2D-TXs outside the representative cluster, at the typical device is given by

$$I_{\Phi_p^!} = \sum_{x \in \Phi_p^!} \sum_{y \in \mathcal{B}^p} P_d g_{yx} \|x + y\|^{-\alpha} \tag{4.5}$$

$$= \sum_{x \in \Phi_p^!} \sum_{y \in \mathcal{B}^p} P_d g_u u^{-\alpha} \tag{4.6}$$

where $\Phi_p^! = \Phi_p \setminus x_0$ for ease of notation, y is the marginal distance between a potential interfering device and its cluster center at $x \in \Phi_p$, $u = \|x + y\|$ is a realization of a random variable U modeling the inter-cluster interfering distance (shown in Fig. 4.2), $g_{yx} \sim \exp(1)$ are i.i.d. exponential random variables modeling Rayleigh fading, and $g_u = g_{yx}$ for ease of notation. The intra-cluster interference is then given by

$$I_{\Phi_c} = \sum_{\mathcal{A}^p} P_d g_{yx_0} \|x_0 + y\|^{-\alpha} \tag{4.7}$$

$$= \sum_{\mathcal{A}^p} P_d g_h h^{-\alpha} \tag{4.8}$$

where y is the marginal distance between the intra-cluster interfering devices and the cluster center at $x_0 \in \Phi_p$, $h = \|x_0 + y\|$ is a realization of a random variable H modeling the intra-cluster interfering distance (shown in Fig. 4.2), $g_{yx_0} \sim \exp(1)$ are i.i.d. exponential random variables modeling Rayleigh fading, and $g_h = g_{yx_0}$ for ease of notation. It is worth noting that the summation in the above is over the set of active devices in the same cluster, denoted by $\mathcal{A}^p$. From the thinning theorem [75], the set of active transmitters following the slotted-ALOHA medium access

forms a PPP Φ_c^p whose intensity is given by

$$\lambda_{cp} = p\lambda_c(y) = p\overline{n} f_Y(y) = \frac{p\overline{n}}{2\pi\sigma^2}\exp\Big(-\frac{\|y\|^2}{2\sigma^2}\Big), \quad y \in \mathbb{R}^2 \tag{4.9}$$

Assuming that the thermal noise is neglected as compared to the aggregate interference, the SIR at the typical device is written as

$$\gamma_r = \frac{P}{I_{\Phi_p^!} + I_{\Phi_c}} = \frac{P_d g_0 r^{-\alpha}}{I_{\Phi_p^!} + I_{\Phi_c}} \tag{4.10}$$

A fixed rate transmission model is adopted in this study, where each TX (D2D or BS) transmits at the fixed rate of $\log_2[1+\theta]$ bits/sec/Hz, where θ is a design parameter. Since the rate is fixed, the transmission is subject to outage due to fading and interference fluctuations. Consequently, the de facto average transmissions rate (i.e., average throughput) is given by

$$R = W\log_2[1+\theta]\mathrm{P_c}, \tag{4.11}$$

where W is the bandwidth, θ is the pre-determined threshold for successful reception, $\mathrm{P_c} = \mathbb{E}(\mathbf{1}\{\mathrm{SIR} > \theta\})$ is the coverage probability, and $\mathbf{1}\{.\}$ is the indicator function. The D2D communication rate under slotted-ALOHA access scheme is then given by

$$R_1(r) = pW_1\log_2\big(1+\theta\big)\mathbf{1}\{\gamma_r > \theta\}, \tag{4.12}$$

where W_1 is the D2D allocated bandwidth, and, accordingly, we refer to $W_2 = W - W_1$ as the BS-to-device bandwidth. The probability $\mathbb{P}(R_1(r) > R_0)$ is then

derived as follows.

$$\begin{aligned}\mathbb{P}(R_1(r) > R_0) &= \mathbb{P}\big(pW_1\log_2(1+\theta)\mathbf{1}\{\gamma_r > \theta\} > R_0\big)\\ &= \mathbb{P}\big(\mathbf{1}\{\gamma_r > \theta\} > \frac{R_0}{pW_1\log_2(1+\theta)}\big)\\ &\overset{(a)}{=} \mathbb{E}_{I_{\Phi_p^!}, I_{\Phi_c}}\Big[\mathbb{P}\big(\frac{P_d g_0 r^{-\alpha}}{I_{\Phi_p^!} + I_{\Phi_c}} > \theta\big)\Big]\\ &= \mathbb{E}_{I_{\Phi_p^!}, I_{\Phi_c}}\Big[\mathbb{P}\big(g_0 > \frac{\theta r^{\alpha}}{P_d}[I_{\Phi_p^!} + I_{\Phi_c}]\big)\Big]\\ &\overset{(b)}{=} \mathbb{E}_{I_{\Phi_p^!}, I_{\Phi_c}}\Big[\exp\big(\frac{-\theta r^{\alpha}}{P_d}[I_{\Phi_p^!} + I_{\Phi_c}]\big)\Big]\\ &\overset{(c)}{=} \mathscr{L}_{I_{\Phi_p^!}}\big(s = \frac{\theta r^{\alpha}}{P_d}\big)\mathscr{L}_{I_{\Phi_c}}\big(s = \frac{\theta r^{\alpha}}{P_d}\big) \end{aligned} \tag{4.13}$$

where (a) follows from the assumption that $R_0 < pW_1\log_2\big(1+\theta\big)$ always holds, otherwise, it is infeasible to get $\mathbb{P}(R_1(r) > R_0)$ greater than zero. (b) follows from the fact that g_o follows an exponential distribution, and (c) follows from the independence of the intra- and inter-cluster interference and the Laplace transform of them. In what follows, we first derive the Laplace transform of interference to get $\mathbb{P}(R_1(r) > R_0)$. Then, we formulate the offloading gain maximization problem.

Lemma 4.3.0.1. *The Laplace transform of the inter-cluster aggregate interference $I_{\Phi_p^!}$ evaluated at $s = \frac{\theta r^{\alpha}}{P_d}$ is given by*

$$\mathscr{L}_{I_{\Phi_p^!}}(s) = \exp\Big(-2\pi\lambda_p \int_{v=0}^{\infty}\Big(1 - \mathrm{e}^{-p\overline{n}\varphi(s,v)}\Big)v\,\mathrm{d}v\Big), \tag{4.14}$$

where $\varphi(s,v) = \int_{u=0}^{\infty}\frac{s}{s+u^{\alpha}}f_U(u|v)\,\mathrm{d}u$, and $f_U(u|v) = \mathrm{Rice}(u|v,\sigma)$ represents Rice's probability density function of parameter σ.

Proof. Please see Appendix B.1. □

Lemma 4.3.0.2. *The Laplace transform of the intra-cluster aggregate interference I_{Φ_c} evaluated at $s = \frac{\theta r^{\alpha}}{P_d}$ is given by*

$$\mathscr{L}_{I_{\Phi_c}}(s) = \exp\Big(-p\overline{n}\int_{h=0}^{\infty}\frac{s}{s+h^{\alpha}}f_H(h)\,\mathrm{d}h\Big) \tag{4.15}$$

where $f_H(h) = \mathrm{Rayleigh}(h, \sqrt{2}\sigma)$ represents Rayleigh's probability density function with a scale parameter $\sqrt{2}\sigma$.

Proof. Please see Appendix B.2. □

For the serving distance distribution $f(r)$, since both the typical device as well as a potential catering device have their locations drawn from a normal distribution with variance σ^2 around the cluster center, the serving distance has a Rayleigh distribution with parameter $\sqrt{2}\sigma$, and given by

$$f_R(r) = \frac{r}{2\sigma^2} \mathrm{e}^{\frac{-r^2}{4\sigma^2}}, \quad r > 0 \tag{4.16}$$

From (4.14), (4.15), and (4.16), the offloading gain in (4.3) is characterized as

$$\mathbb{P}_o(p, b_i) = \sum_{i=1}^{N_f} q_i b_i + q_i(1-b_i)(1-e^{-b_i \overline{n}}) \underbrace{\int_{r=0}^{\infty} \frac{r}{2\sigma^2} \mathrm{e}^{\frac{-r^2}{4\sigma^2}} \mathscr{L}_{I_{\Phi_p^!}}\big(s = \frac{\theta r^\alpha}{P_d}\big) \mathscr{L}_{I_{\Phi_c}}\big(s = \frac{\theta r^\alpha}{P_d}\big) \, \mathrm{d}r}_{\mathbb{P}(R_1 > R_0)}, \tag{4.17}$$

Hence, the offloading gain maximization problem can be formulated as

$$\textbf{P1:} \quad \max_{p, b_i} \quad \mathbb{P}_o(p, b_i) \tag{4.18}$$

$$\text{s.t.} \quad \sum_{i=1}^{N_f} b_i = M, \tag{4.19}$$

$$b_i \in [0, 1], \tag{4.20}$$

$$p \in [0, 1], \tag{4.21}$$

where (4.19) is the device cache size constraint, which is consistent with the illustration of the example in Fig. 4.1. Since the offloading gain depends on the access probability p, and p exists as a complex exponential term in $\mathbb{P}(R_1 > R_0)$, it is hard to analytically characterize (e.g., show concavity of) the objective function or find a tractable expression for the optimal access probability. In order to tackle this, we propose to solve **P1** by finding first the optimal p^* that maximizes the probability $\mathbb{P}(R_1 > R_0)$ over the interval $p \in [0, 1]$. Then, the obtained p^* is used to solve for the caching probability b_i in the optimization problem below. Since in the structure of **P1** p and b_i are separable, it is possible to solve numerically for p^* and then

substitute to get b_i^*.

$$\textbf{P2:} \quad \max_{b_i} \quad \mathbb{P}_o(p^*, b_i) \qquad (4.22)$$
$$\text{s.t.} \quad (4.19), (4.20)$$

The optimal caching probability is formulated in the next lemma.

Lemma 4.3.0.3. *$\mathbb{P}_o(p^*, b_i)$ is a concave function w.r.t. b_i and the optimal caching probability b_i^* that maximizes the offloading gain is given by*

$$b_i^* = \begin{cases} 1 & , v^* < q_i - q_i(1 - e^{-\overline{n}})\mathbb{P}(R_1 > R_0) \\ 0 & , v^* > q_i + \overline{n} q_i \mathbb{P}(R_1 > R_0) \\ \psi(v^*) & , \text{otherwise} \end{cases}$$

where $\psi(v^*)$ is the solution of v^* of (B.13) that satisfies $\sum_{i=1}^{N_f} b_i^* = M$.

Proof. Please see Appendix B.3. □

4.4 Energy Consumption

In this section, we formulate the energy consumption minimization problem for the clustered D2D caching network. In fact, significant energy consumption occurs only when content is served via D2D or BS-to-Device transmission. We consider the time cost c_{d_i} as the time it takes to download the i-th content from a neighboring device in the same cluster. Considering the size S_i of the i-th ranked content, $c_{d_i} = \overline{b}_i / R_1$, where R_1 denotes the average rate of the D2D communication. Similarly, we have $c_{b_i} = \overline{b}_i / R_2$ when the i-th content is served by the BS with average rate R_2. The average energy consumption when downloading files by the devices in the representative cluster is given by

$$E_{\text{av}} = \sum_{k=1}^{\infty} E(b_i|k)\mathbb{P}(n = k) \qquad (4.23)$$

where $\mathbb{P}(n = k)$ is the probability that there are k devices in the cluster x_0, and $E(b_i|k)$ is the energy consumption conditioning on having k devices within the

cluster x_0, written similar to [73] as

$$E(b_i|k) = \sum_{j=1}^{k} \sum_{i=1}^{N_f} \left[\mathbb{P}^d_{j,i} q_i P_d c_{d_i} + \mathbb{P}^b_{j,i} q_i P_b c_{b_i} \right] \tag{4.24}$$

where $\mathbb{P}^d_{j,i}$ and $\mathbb{P}^b_{j,i}$ represent the probability of obtaining the i-th content by the j-th device from the local cluster, i.e., via D2D communication, and the BS, respectively. P_d and P_b denote the device and BS transmission powers, respectively. Given that there are k devices per cluster, it is obvious that $\mathbb{P}^b_{j,i} = (1 - b_i)^k$, and $\mathbb{P}^d_{j,i} = (1 - b_i)\big(1 - (1 - b_i)^{k-1}\big)$.

The average rates R_1 and R_2 are now computed to get a closed-form expression for $E(b_i|k)$. From equation (4.11), we need to obtain the D2D coverage probability $\mathrm{P}_{\mathrm{c_d}}$ and BS-to-Device coverage probability $\mathrm{P}_{\mathrm{c_b}}$ to calculate R_1 and R_2, respectively. Given the number of devices k in the representative cluster, the Laplace transform of the inter-cluster interference is as obtained in (4.14). However, the intra-cluster interfering devices no longer represent a Gaussian PPP since the number of devices is conditionally fixed, i.e., not a Poisson random number as before. To facilitate the analysis, for every realization k, we assume that the intra-cluster interfering devices form a Gaussian PPP with intensity function given by $pkf_Y(y)$. Such an assumption is mandatory for tractability and is validated in the numerical section. From Lemma 4.3.0.2, the intra-cluster Laplace transform conditioning on k can be approximated as

$$\mathscr{L}_{I_{\Phi_c}}(s|k) \approx \exp\Big(- pk \int_{h=0}^{\infty} \frac{s}{s + h^\alpha} f_H(h)\, \mathrm{d}h \Big)$$

and directly, the D2D coverage probability is given by

$$\mathrm{P}_{\mathrm{c_d}} = \int_{r=0}^{\infty} \frac{r}{2\sigma^2} \mathrm{e}^{\frac{-r^2}{4\sigma^2}} \mathscr{L}_{I_{\Phi_p^!}}\big(s = \frac{\theta r^\alpha}{P_d}\big) \mathscr{L}_{I_{\Phi_c}}\big(s = \frac{\theta r^\alpha}{P_d} \big| k\big)\, \mathrm{d}r \tag{4.25}$$

With the adopted slotted-ALOHA scheme, the access probability p is computed over the interval [0,1] to maximize $\mathrm{P}_{\mathrm{c_d}}$ and, in turn, the D2D achievable rate R_1. Analogously, under the PPP Φ_{bs}, and based on the nearest BS association principle,

it is shown in [108] that the BS coverage probability can be expressed as

$$\mathrm{P}_{\mathrm{c_b}} = \frac{1}{{}_2F_1(1,-\delta;1-\delta;-\theta)}, \tag{4.26}$$

where ${}_2F_1(.)$ is the Gaussian hypergeometric function and $\delta = 2/\alpha$. Given the coverage probabilities $\mathrm{P}_{\mathrm{c_d}}$ and $\mathrm{P}_{\mathrm{c_b}}$ in (4.25) and (4.26), respectively, R_1 and R_2 can be calculated from (4.11), and hence $E(b_i|k)$ is expressed in a closed-form.

4.4.1 Energy Consumption Minimization

The energy minimization problem can be formulated as

$$\textbf{P3:} \quad \min_{b_i} \quad E(b_i|k) = \sum_{j=1}^{k}\sum_{i=1}^{N_f}\left[\mathbb{P}^d_{j,i} q_i P_d c_{d_i} + \mathbb{P}^b_{j,i} q_i P_b c_{b_i}\right] \tag{4.27}$$

$$\text{s.t.} \quad (4.19), (4.20)$$

In the next lemma, we prove the convexity condition for E.

Lemma 4.4.1.1. *The energy consumption $E(b_i|k)$ is convex if $\frac{P_b}{R_2} > \frac{P_d}{R_1}$.*

Proof. We proceed by deriving the Hessian matrix of E. The Hessian matrix of $E(b_i|k)(b_1,\ldots,b_{Nf})$ w.r.t. the caching variables is $\mathbf{H}_{i,j} = \frac{\partial^2 E(b_i|k)}{\partial b_i \partial b_j}$, $\forall i,j \in \mathcal{F}$. $\mathbf{H}_{i,j}$ a diagonal matrix whose i-th row and j-th column element is given by $k(k-1)S_i\Big(\frac{P_b}{R_2} - \frac{P_d}{R_1}\Big)q_i(1-b_i)^{k-2}$. Since the obtained Hessian matrix is full-rank and diagonal, $\mathbf{H}_{i,j}$ is positive semidefinite (and hence $E(b_i|k)$ is convex) if all the diagonal entries are nonnegative, i.e., when $\frac{P_b}{R_2} > \frac{P_d}{R_1}$. In practice, it is reasonable to assume that $P_b \gg P_d$, e.g., in [109], the BS transmit power is 100 fold the D2D power. □

As a result of Lemma 4.3.0.3, the optimal caching probability can be computed to minimize $E(b_i|k)$.

Lemma 4.4.1.2. *The optimal caching probability b_i^* for the energy minimization problem **P3** is given by,*

$$b_i^* = \left[1 - \Big(\frac{v^* + k^2 q_i S_i \frac{P_d}{R_1}}{k q_i S_i \big(\frac{P_d}{R_1} - \frac{P_b}{R_2}\big)}\Big)^{\frac{1}{k-1}}\right]^+ \tag{4.28}$$

where v^* satisfies the maximum cache constraint $\sum_{i=1}^{N_f}\left[1-\left(\frac{v^*+k^2q_iS_i\frac{P_d}{R_1}}{kq_iS_i\left(\frac{P_d}{R_1}-\frac{P_b}{R_2}\right)}\right)^{\frac{1}{k-1}}\right]^+ = M$, *and* $[x]^+ = max(x, 0)$.

Proof. The proof proceeds in a similar manner to Lemma 4.3.0.3 and is omitted. □

Proposition 4.4.1.1. By observing (4.28), we can demonstrate the effects of content size and popularity on the optimal caching probability. S_i exists in both the numerator and denominator of the second term in (4.28), however, the effect on numerator is more significant due to larger multiplier. The same property is observed for q_i. With the increase of S_i or q_i, the magnitude of the second term in (4.28) increases, and correspondingly b_i^* decreases. That is a content with larger size or lower popularity has smaller probability to be cached.

By substituting b_i^* into (4.23), the average energy consumption per cluster is obtained. In the remainder of the chapter, we study and minimize the weighted average delay per request for the proposed system.

4.5 Delay Analysis

In this section, the delay analysis and minimization are discussed. A joint stochastic geometry and queueing theory model is exploited to study this problem. The delay analysis incorporates the study of a system of spatially interacting queues. To simplify the mathematical analysis, we further consider that only one D2D link can be active within a cluster of k devices, where k is fixed. As shown later, such an assumption facilitates the analysis by deriving simple expressions. We begin by deriving the D2D coverage probability under the above assumption, which is used later in this section.

Lemma 4.5.0.1. *The D2D coverage probability of the proposed clustered model with one active D2D link within a cluster is given by*

$$\mathrm{P}_{\mathrm{c_d}} = \frac{1}{4\sigma^2 Z(\theta,\alpha,\sigma)}, \tag{4.29}$$

where $Z(\theta,\alpha,\sigma) = (\pi\lambda_p\theta^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)+\frac{1}{4\sigma^2})$.

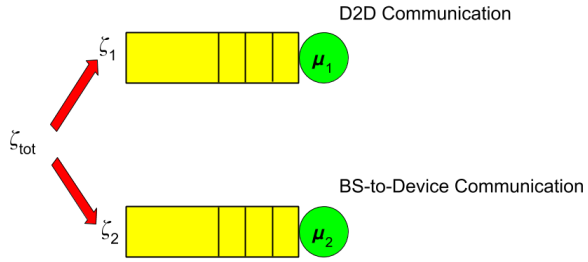


FIGURE 4.3: The traffic model of request arrivals and departures in a given cluster. Two M/G/1 queues are assumed, Q_1 and Q_2, that represent respectively requests served by the D2D and Base station-to-Device communication.

Proof. The result can be proved by using the displacement theory of the PPP [110], and then proceeding in a similar manner to Lemma 4.3.0.1 and 4.3.0.2. The proof is presented in Appendix B.4 for completeness. □

In the following, we firstly describe the traffic model of the network, and then we formulate the delay minimization problem.

4.5.1 Traffic Model

We assume that the aggregate request arrival process from the devices in each cluster follows a Poisson arrival process with parameter ζ_{tot} (requests per time slot). As shown in Fig. 4.3, the incoming requests are further divided according to where they are served from. ζ_1 represents the arrival rate of requests served via the D2D communication, whereas ζ_2 is the arrival rate for those served from the BSs. $\zeta_3 = 1 - \zeta_1 - \zeta_2$ denotes the arrival rate of requests served via the self-cache with zero delay. By definition, ζ_1 and ζ_2 are also Poisson arrival processes. Without loss of generality, we assume that the file size has a general distribution G whose mean is denoted as $\overline{S}$ MBytes. Hence, an M/G/1 queuing model is adopted whereby two non-interacting queues, Q_1 and Q_2, model the traffic in each cluster served via the

D2D and BS-to-Device communication, respectively. Although Q_1 and Q_2 are non-interacting queues because the D2D communication is assumed to be out-of-band, they are spatially interacted with similar queues in other clusters. To recap, Q_1 and Q_2 are two M/G/1 queues with arrival rates ζ_1 and ζ_1, and service rates μ_1 and μ_2, respectively.

4.5.2 Queue Dynamics

It is worth highlighting that the two queues Q_i, $i \in \{1, 2\}$, accumulate requests for files demanded by the clusters members, not the files themselves. First-in first-out (FIFO) scheduling is assumed where a request for content arriving first will be scheduled first either by the D2D or BS communication if the content is cached among the devices or not, respectively. The result of FIFO scheduling only relies on the time when the request arrives to the queue and is irrelevant to the particular device that issues the request. Given the parameter of the Poisson's arrival process ζ_{tot}, the arrival rates at the two queues are expressed respectively as

$$\zeta_1 = \zeta_{tot} \sum_{i=1}^{N_f} q_i \big((1 - b_i) - (1 - b_i)^k\big), \tag{4.30}$$

$$\zeta_2 = \zeta_{tot} \sum_{i=1}^{N_f} q_i (1 - b_i)^k \tag{4.31}$$

The network operation is depicted in Fig. 4.3, and described in detail below.

1. Given the memoryless property of the arrival process (Poisson arrival) along with the assumption that the service process is independent of the arrival process, the number of requests in any queue at a future time only depends upon the current number in the system (at time t) and the arrivals or departures that occur within the interval h.

$$Q_i(t + h) = Q_i(t) + \Lambda_i(h) - M_i(h) \tag{4.32}$$

where $\Lambda_i(h)$ is the number of arrivals in the time interval $(t, t + h)$, whose mean is ζ_i [sec^{-1}], and $M_i(h)$ is the number of departures in the time interval

$(t, t+h)$, whose mean is $\mu_i = \frac{\mathbb{E}(\mathbf{1}\{\mathrm{SIR}>\theta\})W_i \log_2(1+\theta)}{\overline{S}}$ [sec^{-1}]. It is worth highlighting that, unlike the spatial-only model studied in the previous sections, the term $\mathbb{E}(\mathbf{1}\{\mathrm{SIR} > \theta\})$ is dependent on the traffic dynamics since a request being served in a given cluster is interfered only from other clusters that also have requests to serve. What is more noteworthy is that the mean service time $\tau_i = \frac{1}{\mu_i}$ follows the same distribution as the file size. These aspects will be revisited later in this section.

2. $\Lambda_i(h)$ is dependent only on h because the arrival process is Poisson. $M_i(h)$ is 0 if the service time of the file being served $\epsilon_i > h$. $M_i(h)$ is 1 if $\epsilon_1 < h$ and $\epsilon_2 + \epsilon_1 > h$, and so on. As the service times $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent, neither $\Lambda_i(h)$ nor $M_i(h)$ depend on what happened prior to t. Thus, $Q_i(t+h)$ only depends upon $Q_i(t)$ and not the past history. Hence it is a is a continuous-time Markov chain (CTMC) which obeys the stability conditions in [111].

The following proposition provides the sufficient conditions for the stability of the buffers in the sense defined in [111], i.e., $\{Q_i\}$ has a limiting distribution for $t \to \infty$.

Proposition 4.5.2.1. The D2D and BS-to-Device traffic modeling queues are stable, respectively, if and only if

$$\lambda_1 < \mu_1 = \frac{\mathrm{P}_{\mathrm{c_d}} W_1 \log_2(1+\theta)}{\overline{S}} \tag{4.33}$$

$$\lambda_2 < \mu_2 = \frac{\mathrm{P}_{\mathrm{c_b}} W_2 \log_2(1+\theta)}{\overline{S}} \tag{4.34}$$

Proof. We show sufficiency by proving that (4.33) and (4.34) guarantee stability in a dominant network, where all queues that have empty buffers make dummy transmissions. The dominant network is a fictitious system that is identical to the original system, except that terminals may choose to transmit even when their respective buffers are empty, in which case they simply transmit a dummy packet. If both systems are started from the same initial state and fed with the same arrivals, then the queues in the fictitious dominant system can never be shorter than the queues in the original system. Similar to the spatial-only network, in the dominant system, the typical receiver is seeing an interference from all other clusters whether they have requests to serve or not. This dominant system approach yields

$\mathbb{E}(\mathbf{1}\{\text{SIR} > \theta\})$ equal to $\mathrm{P}_{\mathrm{c_d}}$ and $\mathrm{P}_{\mathrm{c_b}}$ for the D2D and BS-to-Device communication, respectively.[2] Also, the obtained delay is an upper bound for the actual delay of the system. The necessity of (4.33) and (4.34) is shown as follows: If $\zeta_i > \mu_i$, then, by Loynes' theorem [112], it follows that $\lim_{t\to\infty} Q_i(t) = \infty$ almost surely for all queues in the dominant network. □

Next, we conduct the analysis for the dominant system whose parameters are as follows. The content size has an exponential distribution with mean $\overline{S}$ [MBytes]. The service times also obey the same exponential distribution with means $\tau_1 = \frac{\overline{S}}{R_1}$ [second] and $\tau_2 = \frac{\overline{S}}{R_2}$ [second]. The rates R_1 and R_2 are calculated from (4.11) where $\mathrm{P}_{\mathrm{c_d}}$ and $\mathrm{P}_{\mathrm{c_b}}$ are from (4.29) and (4.26), respectively. Accordingly, Q_1 and Q_2 are two continuous time independent (non-interacting) M/M/1 queues with service rates $\mu_1 = \frac{\mathrm{P}_{\mathrm{c_d}} W_1 \log_2(1+\theta)}{\overline{S}}$ and $\mu_2 = \frac{\mathrm{P}_{\mathrm{c_b}} W_2 \log_2(1+\theta)}{\overline{S}}$ [sec^{-1}], respectively.

Proposition 4.5.2.2. The mean queue length L_i of the i-th queue is given by

$$L_i = \rho_i + \frac{2\rho_i^2}{2\zeta_i(1-\rho_i)}, \tag{4.35}$$

Proof. We can easily calculate L_i by observing that Q_i are continuous time M/M/1 queues with arrival rates ζ_i, service rates μ_i, and traffic intensities $\rho_i = \frac{\zeta_i}{\mu_i}$. Then, by applying the Pollaczek-Khinchine formula [113], L_i is directly obtained. □

The average delay per request for each queue is calculated from

$$D_1 = \frac{L_1}{\zeta_1} = \frac{1}{\mu_1 - \zeta_1} = \frac{1}{W_1\mathcal{O}_1 - \zeta_{tot}\sum_{i=1}^{N_f} q_i\big((1-b_i) - (1-b_i)^k\big)} \tag{4.36}$$

$$D_2 = \frac{L_2}{\zeta_2} = \frac{1}{\mu_2 - \zeta_2} = \frac{1}{W_2\mathcal{O}_2 - \zeta_{tot}\sum_{i=1}^{N_f} q_i(1-b_i)^k} \tag{4.37}$$

[2] It is worth noting that the adopted queuing model of this chapter is different from that of Chapter 3. Particularly, here we assumed that the incoming requests are accumulated in two different queues based on the serving mode, i.e., D2D of BS-to-device communications. In contrast, Chapter 3 assumed that all incoming requests are accumulated in the same queue, which has multiple processors with different serving rates. This approach of Chapter 3 assumes that all requests for contents are handled based on first come first served even if they might be served from different processors. Hence, the method adopted in Chapter 3 yields a higher average latency than the the method applied in Chapter.

where $\mathcal{O}_1 = \frac{\mathrm{P_{c_d}}\log_2(1+\theta)}{\overline{S}}$, $\mathcal{O}_2 = \frac{\mathrm{P_{c_b}}\log_2(1+\theta)}{\overline{S}}$ for notational simplicity. The weighted average delay D is then expressed as

$$\begin{aligned} D &= \frac{\zeta_1 D_1 + \zeta_2 D_2}{\zeta_{tot}} \\ &= \frac{\sum_{i=1}^{N_f} q_i\big((1-b_i)-(1-b_i)^k\big)}{\mathcal{O}_1 W_1 - \zeta_{tot}\sum_{i=1}^{N_f} q_i\big((1-b_i)-(1-b_i)^k\big)} + \frac{\sum_{i=1}^{N_f} q_i(1-b_i)^k}{\mathcal{O}_2 W_2 - \zeta_{tot}\sum_{i=1}^{N_f} q_i(1-b_i)^k} \end{aligned} \tag{4.38}$$

One important insight from the delay equation is that the caching probability b_i controls the arrival rates ζ_1 and ζ_2 while the bandwidth determines the service rates μ_1 and μ_2. Therefore, it turns out to be of paramount importance to jointly optimize b_i and W_1 to minimize the average delay. Subsequently, we formulate the delay minimization problem as

$$\textbf{P4:} \quad \min_{b_i, \mathrm{W_1}} \quad D(b_i, W_1) \tag{4.39}$$

$$\text{s.t.} \quad (4.19), (4.20)$$

$$0 \leq W_1 \leq W, \tag{4.40}$$

Although the objective function of **P4** is convex w.r.t. W_1, as derived below, the coupling of the optimization variables b_i and W_1 makes **P4** a non-convex optimization problem. Therefore, **P4** can not be solved directly using standard convex optimization techniques. By applying the block coordinate descent optimization technique, **P4** can be solved in an iterative manner as follows. First, for a given caching probability b_i, we calculate the bandwidth allocation subproblem. Afterwards, the obtained optimal bandwidth is used to update b_i. The optimal bandwidth for the bandwidth allocation subproblem is given in the next Lemma.

Lemma 4.5.2.1. *The objective function of **P4** in (4.39) is convex w.r.t. W_1, and the optimal bandwidth allocation to the D2D communication is given by*

$$W_1^* = \frac{\zeta_{tot}\sum_{i=1}^{N_f} q_i(\overline{b}_i - \overline{b}_i^k) + \varpi\big(\mathcal{O}_2 W - \zeta_{tot}\sum_{i=1}^{N_f} q_i \overline{b}_i^k\big)}{\mathcal{O}_1 + \varpi\mathcal{O}_2}, \tag{4.41}$$

where $\overline{b}_i = 1 - b_i$ and $\varpi = \sqrt{\frac{\mathcal{O}_1\sum_{i=1}^{N_f} q_i(\overline{b}_i - \overline{b}_i^k)}{\mathcal{O}_2\sum_{i=1}^{N_f} q_i \overline{b}_i^k}}$

Proof. The first derivative $D(b_i, W_1|k)$ can be written as

$$D(b_i, W_1|k) = \sum_{i=1}^{N_f} q_i(\bar{b}_i - \bar{b}_i^k)\big(\mathcal{O}_1 W_1 - \zeta_{tot}\sum_{i=1}^{N_f} q_i(\bar{b}_i - \bar{b}_i^k)\big)^{-1} + \sum_{i=1}^{N_f} q_i\bar{b}_i^k\big(\mathcal{O}_2 W_2 - \zeta_{tot}\sum_{i=1}^{N_f} q_i\bar{b}_i^k\big)^{-1},$$

The second derivative $\frac{\partial^2 D(b_i,W_1|k)}{\partial W_1^2}$ is hence given by

$$\frac{\partial^2 D(b_i, W_1|k)}{\partial W_1^2} = 2\mathcal{O}_1^2\sum_{i=1}^{N_f} q_i(\bar{b}_i - \bar{b}_i^k)\big(\mathcal{O}_1 W_1 - \zeta_{tot}\sum_{i=1}^{N_f} q_i(\bar{b}_i - \bar{b}_i^k)\big)^{-3} + 2\mathcal{O}_2^2\sum_{i=1}^{N_f} q_i\bar{b}_i^k\big(\mathcal{O}_2 W_2 - \zeta_{tot}\sum_{i=1}^{N_f} q_i\bar{b}_i^k\big)^{-}$$

The stability condition requires that $\mathcal{O}_1 W_1 > \zeta_{tot}\sum_{i=1}^{N_f} q_i(\bar{b}_i - \bar{b}_i^k)$ and $\mathcal{O}_2 W_2 > \zeta_{tot}\sum_{i=1}^{N_f} q_i\bar{b}_i^k$. Also, $\bar{b}_i \geq \bar{b}_i^k$ by definition. Hence, $\frac{\partial^2 D(b_i,W_1|k)}{\partial W_1^2} > 0$, and the objective function is a convex function of W_1. The optimal bandwidth allocation can be obtained from the Karush-Kuhn-Tucker (KKT) conditions similar to problems **P2** and **P3**, with the details omitted for brevity. □

Given W_1^* from the bandwidth allocation subproblem, the caching probability subproblem can be written as

$$\textbf{P5:} \quad \min_{b_i} \quad D(b_i, W_1^*) \tag{4.42}$$

$$\text{s.t.} \quad (4.19), (4.20) \tag{4.43}$$

The caching probability subproblem **P5** is a sum of two fractional functions, where the first fraction is in the form of a concave over convex functions while the second fraction is in the form of a convex over concave functions. The first fraction structure, i.e., concave over convex functions, renders solving this problem using fractional programming very challenging. Instead, we seek a heuristic yet efficient algorithm to solve for a suboptimal caching solution for **P5**. Particularly, we use the interior point method, implemented in Mathematica, to obtain a suboptimal caching probability $\boldsymbol{b}$, as done in [114]. The entire proposed algorithm to solve **P5** is presented in Algorithm 2 and works as follows.[3] Firstly, we start with an initial caching probability $\boldsymbol{b}_0$ and allocated bandwidth $W_1 = \frac{W}{2}$ to obtain a suboptimal caching solution based on the interior point method. Then, the obtained caching

[3]In the initialization of Algorithm 2, T_0 is first set to a large value, then, updated periodically based on the new $\boldsymbol{b}$ and W_1.

Algorithm 2: BCD algorithm for **P5**

Input : W, N_f, M, Υ_i, Υ_b, β, $\overline{S}$, θ, p, T_0;
Initialization: $\boldsymbol{b} \leftarrow \boldsymbol{b}_0$, $W_1 \leftarrow \frac{W}{2}$;
$(D) \leftarrow$ Eq. (4.38) $\leftarrow (\boldsymbol{b}_0, W_1)$;
while $D < D_0$ **do**
  /* Update D_0 with the calculated delay. */
  $D_0 = D$;
  $(\boldsymbol{b}) \leftarrow$ interior point method$(\boldsymbol{b}, W_1)$;
  $(W_1) \leftarrow$ Eq. (4.41) $\leftarrow (\boldsymbol{b})$;
  $(D) \leftarrow$ Eq. (4.38) $\leftarrow (\boldsymbol{b}, W_1)$;
end while
Output: $\boldsymbol{b}$, W_1;

probability $\boldsymbol{b}$ is used to update the bandwidth allocation in (4.41). The explained procedure, i.e., solving the two subproblems iteratively, is repeated until the value of **P5**'s objective function converges to a pre-specified accuracy. Importantly, the caching probability solution $\boldsymbol{b}$, given the optimal bandwidth W_1^*, depends on the initial value input to the interior point algorithm [115]. We use the Zipf caching as an initial point for this algorithm to obtain a suboptimal caching probability given the bandwidth calculated from (4.41).

4.6 Results and Discussions

The simulation setup is as follows. $W = 20$ MHz. $P_b = 43$ dBm, $P_d = 23$ dBm, $\sigma = 10$ m, $\beta = 1, \alpha = 4$, $N_f = 500$ files, $M = 10$ files, $\overline{n} = 5$ devices, $\lambda_p = 50$ clusters/km^2, $\overline{S} = 3$ Mbits, and $\theta = 0$ dB. This simulation setup will be used unless otherwise specified. We particularly assume a relatively small cache size per device and library size. These values, which are close to that used in [116] and [117], are reasonable in the study of communication and caching aspects of D2D content delivery networks. Other works in the literature, see, e.g., [118], considered a much larger size of file library, however, their objective was to conduct the scaling analysis of caching networks. In addition, any other values for the operating bandwidth W can be applied and the optimal bandwidth partitioning can be directly calculated from (4.41).[4]

[4]Typical values for the operating bandwidth is 10-20 MHz for Long Term Evolution (LTE) systems and higher than 50 MHz for 5G systems.

4.6.1 Offloading Gain Results

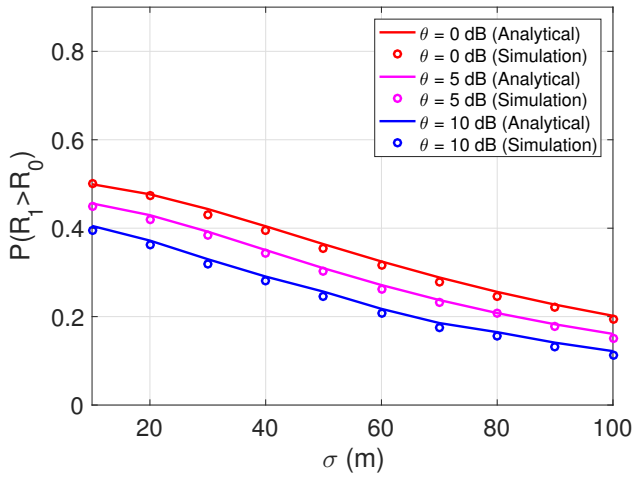
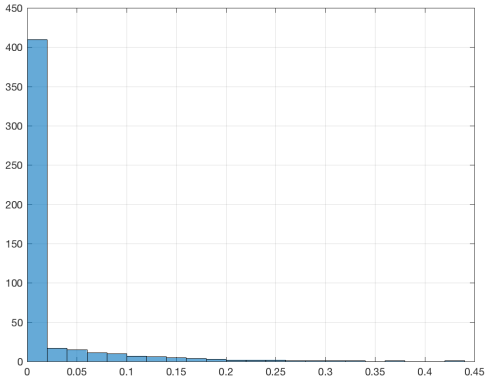
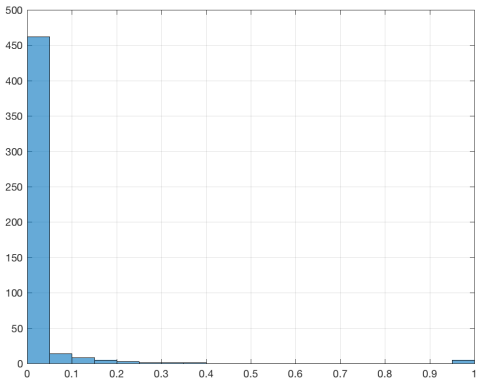


FIGURE 4.4: The probability that the achievable rate is greater than a threshold R_0 versus standard deviation σ.

In this subsection, we present the offloading gain performance for the proposed caching model. In Fig. 4.4, we verify the accuracy of the analytical results for the probability $\mathbb{P}(R_1 > R_0)$. Monte-carlo simulation is adopted where a clustered D2D network is realized and the simulation is run for thousands number of iterations, then, the average probability $\mathbb{P}(R_1 > R_0)$ is calculated based on the statistical average over all iterations. The theoretical and simulated results are plotted together, and they are consistent. We can observe that the probability $\mathbb{P}(R_1 > R_0)$ decreases monotonically with the increase of σ. This is because as σ increases, the serving distance increases and the inter-cluster interfering distance between the out-of-cluster interferers and the typical device decreases, and equivalently, the SIR decreases. It is also shown that $\mathbb{P}(R_1 > R_0)$ decreases with the SIR threshold θ as the channel becomes more prone to be in outage when increasing the SIR threshold θ. To show the effect of p on the caching probability, in Fig. 4.5, we plot the histogram of the optimal caching probability at different values of p, where $p = p^*$ in Fig. 4.5(a) and $p \neq p^*$ in Fig. 4.5(b). It is clear from the histograms that optimal caching probability b_i tends to be more skewed when $p \neq p^*$, i.e., when $\mathbb{P}(R_1 > R_0)$ decreases. This shows that file sharing is more difficult when p is not optimal. For example, if $p < p^*$, the system is too conservative owing to the small access probabilities. However, for $p > p^*$, the outage probability is high due to the aggressive interference.

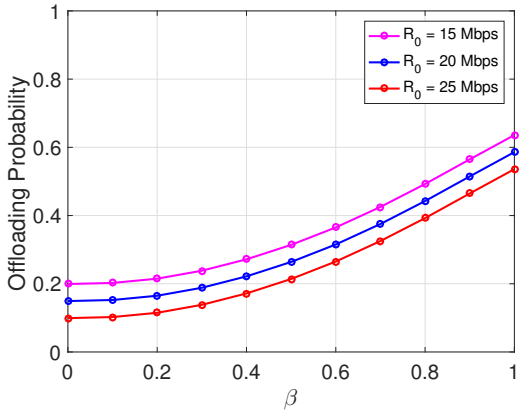


(a) $p = p^*$.

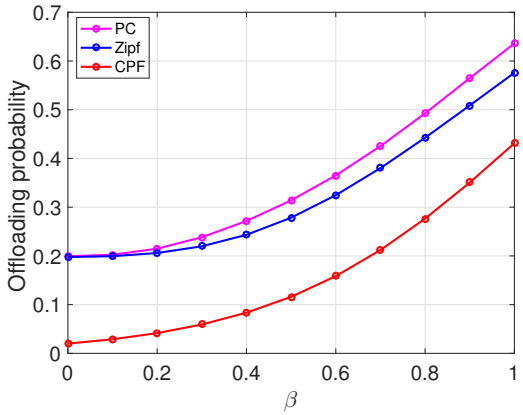


(b) $p \neq p^*$.

FIGURE 4.5: Histogram of the caching probability b_i when (a) $p = p^*$ and (b) $p \neq p^*$.



(a) The offloading probability versus the popularity of files β at different thresholds R_0.



(b) The offloading probability versus the popularity of files β under different caching schemes (probabilistic caching, Zipf, caching popular files).

FIGURE 4.6: The offloading probability versus the popularity of files β.

In such a low coverage probability regime, each device tends to cache the most popular files leading to fewer opportunities of content transfer between devices.

Last but not least, Fig. 4.6 manifests the prominent effect of the files' popularity on the offloading gain. In Fig. 4.6(a), we plot the offloading gain against β at different rate thresholds R_0. We note that the offloading gain monotonically increases with β since fewer files are frequently requested such that the files can be entirely cached among the cluster devices. Also, we see that offloading gain decreases with the increase of R_0 since the probability $\mathbb{P}(R_1 > R_0)$ decreases with R_0. In Fig. 4.6(b), we compare the offloading gain of three different caching schemes, namely, the proposed probabilistic caching (PC), Zipf caching (Zipf), and CPF. We can see that the offloading gain under the optimized PC scheme attains the best

performance as compared to other schemes. Also, we note that both PC and Zipf's schemes encompass the same energy consumption when $\beta = 0$ owing to the uniformity of content popularity.

4.6.2 Energy Consumption Results

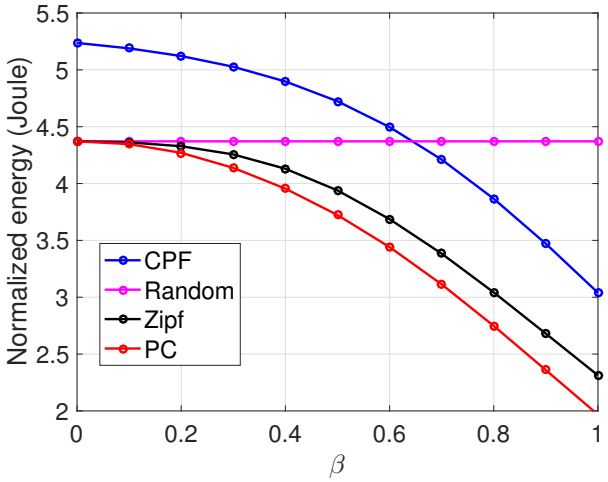


FIGURE 4.7: Normalized energy consumption versus popularity exponent β.

The results in this part are given for the energy consumption. Fig. 4.7 shows the energy consumption, normalized to the number of devices per cluster, versus β under different caching schemes, namely, PC, Zipf, CPF, and uniform random caching (random). We can see that the minimized energy under the proposed PC scheme attains the best performance as compared to other schemes. Also, it is clear that, except for the random uniform caching, the consumed energy decreases with β. This can be justified by the fact that as β increases, fewer files are frequently requested which are more likely to be cached among the devices under PC, CPF, and the Zipf caching schemes. These few files therefore are downloadable from the devices via low power D2D communication. In the random caching scheme, files are uniformly chosen for caching independently of their popularity.

We plot the normalized energy consumption per device versus the number of devices per cluster in Fig. 4.8. First, we see that the normalized energy consumption decreases with the number of devices. As the number of devices per cluster

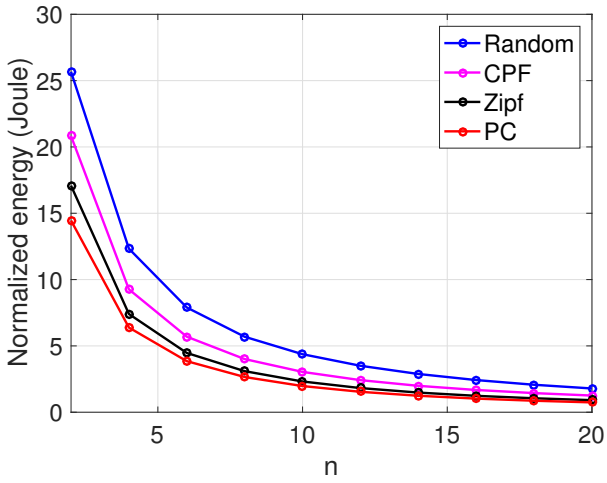


FIGURE 4.8: Normalized energy consumption versus number of devices per cluster.

increases, it is more probable to obtain requested files via low power D2D communication. When the number of devices per cluster is relatively large, the normalized energy consumption tends to flatten as most of the content becomes cached at the cluster devices. In addition, the optimized PC scheme is shown to achieve the lowest energy consumption compared to other caching methods.

4.6.3 Delay Results

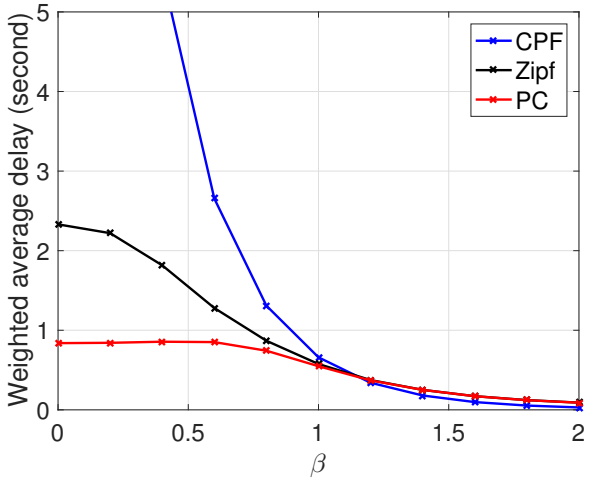


FIGURE 4.9: Weighted average delay versus the popularity exponent β.

In Fig. 4.9, we compare the average delay of three different caching schemes

PC, Zipf, and CPF. We can see that the jointly minimized average delay under PC scheme attains the best performance as compared to other caching schemes. This is driven by the joint optimization of content caching (queues' arrival rates) and bandwidth allocation (queues' service rate) so as to minimize the request service time. Also, we see that, in general, the average delay monotonically decreases with β when a fewer number of files undergoes the highest demand. Fig. 4.10 manifests the effect of the files' popularity β on the allocated bandwidth. Recall first that W_1^* and W_2^* refer to the D2D and BS-to-device allocated bandwidths, respectively. It is shown that optimal D2D allocated bandwidth W_1^* continues increasing with β. This can be interpreted as follows. When β increases, a fewer number of files become highly demanded. These files can be entirely cached among the devices. To cope with such a larger number of requests served via the D2D communication, the D2D allocated bandwidth needs to be increased.

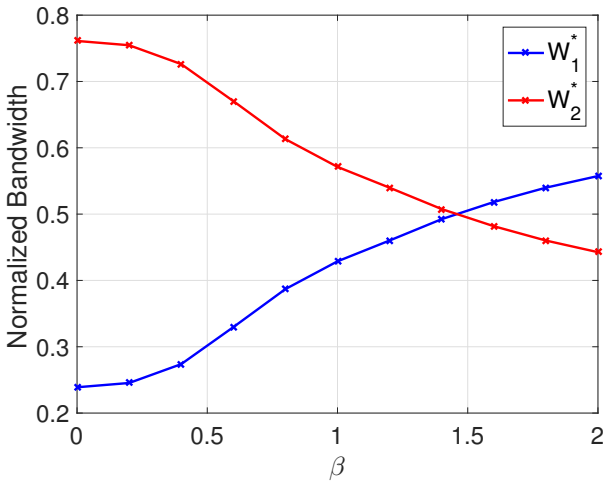


FIGURE 4.10: Normalized bandwidth allocation versus the popularity exponent β.

4.7 Chapter Summary

In this chapter, we conducted a comprehensive analysis of the joint communication and caching for a clustered D2D network with random probabilistic caching incorporated at the devices. We first maximized the offloading gain of the proposed system by jointly optimizing the channel access and caching probability. We solved

for the channel access probability numerically, and the optimal caching probability is then characterized. We showed that deviating from the optimal access probability p^* makes file sharing more difficult. More precisely, the system is too conservative for small access probabilities, while the interference is too aggressive for larger access probabilities. Then, we minimized the energy consumption of the proposed clustered D2D network. We formulated the energy minimization problem and show that it is convex and the optimal caching probability is obtained. We showed that a content with a large size or low popularity has a small probability to be cached. Finally, we adopted a queuing model for the devices' traffic within each cluster to investigate the network average delay. Two M/G/1 queues are employed to model the D2D and BS-to-Device communications. We then derived an expression for the weighted average delay per request. We observed that the average delay is dependent on the caching probability and bandwidth allocated, which control respectively the arrival rate and service rate for the two modeling queues. Therefore, we minimized the per request weighted average delay by jointly optimizing bandwidth allocation between D2D and BS-to-Device communication and the caching probability. The delay minimization problem is shown to be non-convex. Applying the block coordinate descent optimization technique, the joint minimization problem can be solved in an iterative manner. Results showed roughly up to 10%, 17%, and 140% improvement gain in the offloading gain, energy consumption, and average delay, respectively, compared to the Zipf caching technique.

Chapter 5

Performance Analysis for Cache-Assisted CoMP for Clustered D2D Networks

Caching at mobile devices and leveraging cooperative device-to-device (D2D) communications are two promising approaches to support massive content delivery over wireless networks while mitigating the effects of interference. To show the impact of cooperative communication on the performance of cache-enabled D2D networks, the notion of device clustering must be factored in to convey a realistic description of the network performance. In this regard, this chapter develops a novel mathematical model, based on stochastic geometry and an optimization framework for cache-assisted coordinated multi-point (CoMP) transmissions with clustered devices. Devices are spatially distributed into disjoint clusters and are assumed to have a surplus memory to cache files from a known library, following a random probabilistic caching scheme. Desired contents that are not self-cached can be obtained via D2D CoMP transmissions from neighboring devices or, as a last resort, from the network. For this model, we analytically characterize the offloading gain and rate coverage probability as functions of the system parameters, namely, density of clusters, number of devices, and the intra-cluster distance between devices. An optimal caching strategy is then defined as the content placement scheme that maximizes the offloading gain. For a tractable optimization framework, we pursue two separate approaches to obtain a lower bound and a provably accurate approximation of the offloading gain, which allows us to obtain optimized

caching strategies. Remarkably, if we replace the obtained expression for offloading gain with its lower bound, we can find a suboptimal caching strategy that is not only described via analytical formulas but can also show an improvement over the state-of-the-art caching schemes. Results reveal that cooperative transmission becomes more appealing in denser D2D caching networks and adverse interference conditions, which is the case of the imminent internet of things (IoT) and massive machine type communications era.

5.1 Introduction

Because of their respective advantages, both caching and cooperative transmission can be jointly adopted in many practical scenarios, e.g., social networks-related applications and proximity marketing. For example, ensuring reliable delivery of ultra-high-definition streaming, such as 360-degree virtual reality, over wireless networks is very challenging due to the stringent QoS requirements [119]. Leveraging D2D CoMP transmission of pre-downloaded frames from multiple devices to a common requesting device helps reduce communication delays and, in turn, improve the perceived QoS. Proximity marketing, which is a wireless content advertising system associated with a particular place, might be another use case that leverages both caching and CoMP transmission [120]. In particular, exploiting CoMP transmission to send pre-cached advertising content could lead to increasing the transmission range and mitigating interference among different operators of proximity marketing systems. Such cooperatively served contents are then delivered to the individuals who wish to receive them, provided that they have the necessary equipment to do so [120].

5.1.1 Motivation and Contribution

Motivated by the aforementioned discussion, it is important to study the role of cooperative transmission for cache-enabled D2D networks. In this chapter, we conduct performance analysis and statistical optimization of cache-assisted cooperative transmissions for a clustered D2D network. In particular, we characterize and optimize the offloading gain of a network of spatially clustered devices that adopt

CoMP transmission and probabilistic caching. By maximizing the statistically-averaged offloading gain, my approach efficiently provides optimal average performance on a long-time scale to reduce signaling and processing overheads [121]. Moreover, the proposed model effectively captures the stochastic nature of channel fading and the clustered, yet random, network topology aspects, which have not been studied in the literature, particularly in the context of caching and CoMP transmission.

5.1.2 Contributions

The main contributions of this chapter are summarized as follows:

- We propose a cooperative transmission scheme via D2D communications for clustered cache-enabled networks, whereby a device can be collaboratively served from multiple devices within the same cluster. We analytically characterize the offloading gain and rate coverage probability for the proposed network. The existence of multi-fold integration in the obtained expressions, however, yields the offloading gain maximization problem prohibitively complex.

- Using Taylor's series expansion, we can obtain a tractable lower bound on the rate coverage probability. Based on the derived bound, we prove that the interference seen by the typical device of a TCP is upper-bounded by that of a PPP whose density is the product of the TCP density of clusters times the average number of devices per cluster.

- To further improve tractability and computational efficiency, we approximate the cooperatively received signal by two components, which we refer to as nearest and mean received power terms. Using Chebyshev's inequality, we prove that this approximation is remarkably tight and helps turn the multi-fold integration to a single integral.

- We obtain two closed-form suboptimal caching solutions for the offloading gain maximization problem based on the adopted lower bound and approximation of the rate coverage probability. Simulation results quantify the performance gain from the CoMP transmission and show considerable improvements in the offloading gain under the optimized caching strategies compared to other benchmark caching techniques.

5.1.3 Related Works

The joint adoption of wireless caching and collaborative transmissions, where BSs (or devices) cooperatively serve a content, is widely adopted in literature [122–128]. For instance, the authors in [122] proposed a combined caching scheme whereby part of the cache space is reserved for caching the most popular content, that is then cooperatively served from multiple BSs. Moreover, in [123], the authors investigated the tradeoff between content diversity gain, i.e., serving diverse content, and cooperative gain, i.e., jointly transmitting the same content. In [124], the authors proposed cooperative transmissions for cache-enabled small cell networks to reduce the backhaul utilization cost and delay. Meanwhile, the authors in [125] employed content caching at wireless relays to improve the overall performance of collaborative relaying for a network consisting of one source, one destination, and multiple relay nodes.

Following a similar approach, employing cooperative content delivery in D2D caching networks is discussed in [126–128]. For instance, the authors in [126] proposed a multiple devices to a single device content delivery method via D2D communication. Moreover, an opportunistic cooperation strategy for D2D transmission is proposed in [127] to mitigate the interference among D2D links. Combining coded caching along with CoMP transmission is recently studied in [128], wherein redundantly stored data at caching helpers is utilized to combat wireless channel impairments due to channel fading and interference.

While interesting, the works in [126–128] did not consider the notion of device clustering, which is quite fundamental to the D2D network architecture [129] and [130]. In this regard, the authors in [76] developed a stochastic geometry-based

model to characterize the performance of content delivery in a clustered D2D network whose devices are distributed according to a PCP. Moreover, the authors in [77] proposed a cluster-centric content placement scheme where the content of interest is cached closer to the cluster center. Meanwhile, the authors in [116] proposed hybrid caching strategies to save the energy cost of D2D transmitters, where the location of these transmitters is modeled as a Gaussian Poisson process (GPP). However, while the clustering nature of D2D communication is considered in the prior works [76] and [77], these works assumed that contents are pre-cached, i.e., there was no study of the caching problem. Moreover, modeling clustered D2D networks by means of GPPs, as done in [116], is limited by two facts: (i) The distance between transmitting and receiving devices within the same cluster is not captured by this model. (ii) The number of devices per cluster is assumed to be constant, particularly, fixed to only one (or two) device(s) per cluster. Furthermore, the content popularity and caching schemes in [116] were assumed to be the same for all clusters. However, in practice, users in different clusters might have different file interest. For instance, users in a library might be interested in a different set of content from than users in a pub.

5.2 System Model

5.2.1 Network Model

We consider a D2D caching network in which devices are spatially distributed into disjoint clusters. The devices are assumed to have surplus memory that can be used to store content such as video files. Such a cached content is needed either for future use or to participate in content sharing with other devices within the same cluster. For this network, we model the location of the devices with a TCP composed of parent and daughter points. Let us denote the parent point process by $\Phi_p = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\}$, where $\boldsymbol{x}_i = \{x_1, x_2\} \in \mathbb{R}^2$, and $i \in \mathcal{N}$. Further, let (Φ_i) be a family of finite point sets representing the untranslated daughter Gaussian PPPs, i.e., untranslated clusters. The cluster process is then the union of the translated

clusters:

$$\Phi = \cup_{i\in\mathcal{N}} \boldsymbol{x}_i + \Phi_i. \tag{5.1}$$

We assume that the cluster centers are distributed according to a PPP Φ_p of density λ_p. We also assume that, for Gaussian PPPs, the cluster members are normally scattered with variance $\sigma^2 \in \mathbb{R}$ around their cluster centers [75]. As in Chapter 4, given the normal scattering of daughter points, the probability distribution function (PDF) of the cluster member location relative to its cluster center is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{2\pi\sigma^2}\exp\Big(-\frac{\|\boldsymbol{y}\|^2}{2\sigma^2}\Big), \qquad \boldsymbol{y}\in\mathbb{R}^2, \tag{5.2}$$

where $\boldsymbol{y}\in\mathbb{R}^2$ is the device location relative to its cluster center, $\|.\|$ is the Euclidean norm. If the average number of devices per cluster is $\bar{n}$, the cluster intensity will be:

$$\lambda_c(y) = \frac{\bar{n}}{2\pi\sigma^2}\exp\Big(-\frac{\|\boldsymbol{y}\|^2}{2\sigma^2}\Big), \qquad \boldsymbol{y}\in\mathbb{R}^2, \tag{5.3}$$

and the intensity of the entire process Φ will be $\lambda = \bar{n}\lambda_p$. We assume two kind of devices co-exist within the same cluster, namely, content clients and content providers, as done in [116]. In particular, the devices that can perform proactive caching and provide content delivery are called content providers while those requesting content, that also have caching capability, are called content clients.

5.2.2 Content Popularity and Probabilistic Caching

We assume that each device has a surplus memory of size M files designated for caching content from a known file library $\mathcal{F}$. The total number of files is $N_f > M$ and the set of content indices is denoted as $\mathcal{F} = \{1, 2, \ldots, N_f\}$. Similar to Chapter 4, we assume that the probability that the m-th content is requested follows the standard Zipf distribution, which is given by:

$$q_m = \left(m^\beta \sum_{k=1}^{N_f} k^{-\beta}\right)^{-1}, \tag{5.4}$$

where β is a parameter reflecting how skewed the popularity distribution is. Moreover, we assume that the content popularity may vary across clusters. For instance, users in a library may be interested in an entirely different set of files from the users in a sports center. Therefore, the Zipf distribution models the per-cluster popularity of files. Such a discrepancy of contents of interest can be captured by having different popularity indices β per different clusters. Ranks of popular contents can be also different among different clusters. This discrepancy of popular files implies that the content request and, consequently, caching design models vary across clusters. Such a cluster-specific popularity can be seen as a direct generalization of the individual user preferences that is modeled in [131].

The cluster-specific popularity model necessitates the design of content placement on a per-cluster basis. Hence, within each cluster, we assume a random content placement scheme in which file m is cached independently at each cluster device according to the probability b_m, with $0 \leq b_m \leq 1, \forall m \in \{1, \ldots, N_f\}$. To avoid duplicate caching of the same file within the memory of the same device, we follow the PC approach proposed in [132], which requires that $\sum_{m=1}^{N_f} b_m = M$. It is worth mentioning that the PC is a standard caching technique that is widely adopted in the literature [107, 133, 134].

5.2.3 Content Request and Delivery Model

Enabling seamless video delivery over cellular networks implies stringent QoS requirements. However, the performance of wireless networks, especially D2D communications, is limited by interference and the effects of small scale fading. Therefore, cooperative communication turns to be more appealing as a prominent interference mitigation tool. We hence allow multiple devices to jointly serve their cached content to a common device within the same cluster via non-coherent CoMP transmission. The underlying reason of assuming a non-coherent transmission is that it is hard to estimate the channel state information (CSI) for the D2D communications. We consider out-of-band D2D communication system, i.e., there is no cross-interference between the cellular network and D2D communication. All devices are equipped with a single transmit-receive isotropic antenna, and they have no CSI from the device they are serving. Furthermore, each D2D transmission uses

all the available bandwidth. Transmitted signals experience single-slope path loss with attenuation exponent $\alpha > 2$ and small scale fading, which we model as an independently and identically distributed (i.i.d.) complex Gaussian random variable (RV) with zero mean and unit variance.

Due to the cost of participating in content caching and delivery, e.g., battery consumption and memory utilization, not all *content providers* can be active in all time slots. Hence, within each cluster, we assume that content providers can be available for content delivery with probability $p \in [0, 1]$. In other words, among $\bar{n}$ average number of devices per cluster, on average only $p\bar{n}$ devices are willing to participate in content delivery and caching. Further, we assume a BS-assisted D2D link setup scheme, where the transmissions of different files in different clusters are orchestrated by the BS [135]. In detail, a *content client* first sends its request to its geographically closest BS, which knows the active content providers within the same cluster, their cached files, as well as their locations. If there are active content providers caching the requested file, the BS then establishes *direct CoMP D2D links* between the content client and the set of *active content providers* for this requested file. Requests for contents are assumed to be of negligible size, so that they are always successfully decoded at the BS.

Within each cluster, we assume a content client device whose distance to its cluster center is drawn from a Rayleigh distribution of scale parameter σ, according to the TCP definition [75]. Throughout time, content clients in different clusters may request files $i \in \mathcal{F}$ with a probability following the assumed *per-cluster Zipf distribution* in (5.4). Since each cluster has its own library, a given content client may either be served via D2D connection(s) from *(cooperative) active provider(s)* within the same cluster or, as a last resort, via the nearest geographical BS. To recap, under the proposed transmission and caching schemes, one content client per cluster is cooperatively served at a time from neighboring active content providers while being interfered only from active providers in other (remote) clusters (i.e., inter-cluster interference).

Notice that, according to the independent thinning theorem [75, Theorem 2.36], active devices within the same cluster form a Gaussian PPP Φ_{cp} whose intensity

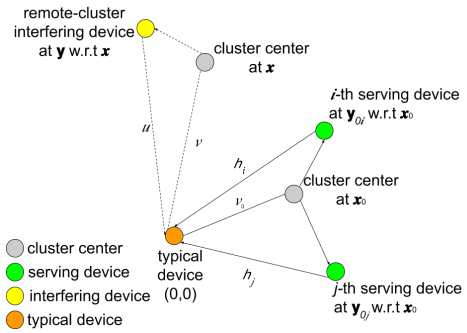


FIGURE 5.1: Illustration of the representative cluster and one interfering cluster, where $\{\boldsymbol{x}_0, \boldsymbol{y}_{0i}, \boldsymbol{y}_{0j}, \boldsymbol{x}, \boldsymbol{y}\} \in \mathbb{R}^2$ and $\{v_0, h_i, h_j, v, u\} \in \mathbb{R}$.

function is $\lambda_{cp}(\boldsymbol{y}) = p\lambda_c(\boldsymbol{y})$. Similarly, the set of active providers that cache a desired content m are modeled as a Gaussian PPP Φ_{cpm} with the intensity function given by $\lambda_{cpm}(\boldsymbol{y}) = b_m p\lambda_c(\boldsymbol{y})$. Hence, within each cluster, the number of active devices and the number of active devices caching content m are Poisson RVs of means $p\bar{n}$ and $b_m p\bar{n}$, respectively. By definition, $\Phi_{cpm} \subseteq \Phi_{cp} \subseteq \Phi_c$.

The most distinctive advantages of D2D caching networks are alleviating the burden of the backhaul links and improving the network spectral efficiency. To leverage these features, it is crucial to intelligently cache and deliver contents to maximize the percentage of offloaded traffic from the network core to the edge. The offloading gain is widely-adopted as a key performance metric to quantify this percentage [122, 123], and [116]. Specifically, the offloading gain is defined as the probability of obtaining a desired content either from the self-cache or via D2D communication with a received SIR greater than a target threshold. Hence, my target in the next sections is to characterize and maximize the offloading gain of the proposed CoMP-assisted D2D caching network.

5.3 Offloading Gain Characterization

Given stationarity of the parent process and independence of the offspring process, we can conduct the next analysis for the representative cluster, which is an arbitrary cluster whose center is located at $\boldsymbol{x}_0 \in \Phi_p$, and a *typical client*, which is a randomly selected member of the representative cluster that requests the content. Without loss of generality, we assume the typical client is located at the origin $(0, 0) \in \mathbb{R}^2$.

When some active content providers jointly transmit a desired content m, the signal received at the typical client consists of two main components: the desired signal that represents the joint non-coherent transmissions from active providers that cache content m in the representative (local) cluster, and the interference component that is created by other active providers in remote clusters. This can be formally stated as:

$$y_m = \underbrace{\sum_{\boldsymbol{y}_{0i} \in \Phi_{cpm}} \sqrt{\gamma_d} G_{\boldsymbol{y}_{0i}} \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|^{-\alpha/2} s_{\boldsymbol{y}_{0i}}}_{\text{desired signal}} + \underbrace{\sum_{\boldsymbol{x} \in \Phi_p^!} \sum_{\boldsymbol{y} \in \Phi_{cp}} \sqrt{\gamma_d} G_{\boldsymbol{y}} \|\boldsymbol{x} + \boldsymbol{y}\|^{-\alpha/2} s_{\boldsymbol{y}}}_{\text{inter-cluster interference}} + z,$$

where $i \in \{1, \ldots, |\Phi_{cpm}|\}$, $G_{\boldsymbol{y}_{0i}}$ denotes the power fading between an active provider at $\boldsymbol{y}_{0i} \in \Phi_{cpm}$ relative to its cluster center at $\boldsymbol{x}_0$ and the typical client, see Fig. 5.1; γ_d denotes the D2D transmission power, and $s_{\boldsymbol{y}_{0i}}$ is the symbol jointly transmitted by the active providers $\boldsymbol{y}_{0i} \in \Phi_{cpm}$. $\Phi_p^! = \Phi_p \setminus \{\boldsymbol{x}_0\}$ denotes the set of remote clusters, and $\Phi_{cp} \subseteq \Phi_c$ represents the set of active devices in a remote cluster centered at $\boldsymbol{x} \in \Phi_p^!$. Finally, z denotes the standard additive white Gaussian noise.

Note that representing the set of inter-cluster interferers as Φ_{cp} corresponds to the worst case interference scenario, when all active devices in a remote cluster are caching the required content of their own-cluster content client. This bound is in line with the analysis and the underlying network model, particularly, the assumption of different content popularity and placement per clusters. This is because, based on this bound, the inter-cluster interference power, and correspondingly, the per cluster cache design will be independent of the content demand in other clusters.

We assume that the system operates in the interference limited regime, i.e., the background noise is negligible compared to the interference and is hence ignored. Assuming unit power Gaussian symbols, the received SIR at the typical client when

downloading content m is given by

$$\mathrm{SIR}_m = \frac{\gamma_d \left| \sum_{\boldsymbol{y}_{0i} \in \Phi_{cpm}} G_{\boldsymbol{y}_{0i}} \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|^{-\alpha/2} \right|^2}{I_{\text{out}}}, \tag{5.5}$$

where I_{out} is the sum of interfering signal power associated with the downloading of content m, given by:

$$I_{\text{out}} = \gamma_d \Big| \sum_{\boldsymbol{x} \in \Phi_p^!} \sum_{\boldsymbol{y} \in \Phi_{cp}} G_{\boldsymbol{y}} \|\boldsymbol{x} + \boldsymbol{y}\|^{-\alpha/2} \Big|^2.$$

Finally, the offloading gain can be formally stated as:

$$\mathbb{P}_o(\boldsymbol{b}) = \sum_{m=1}^{N_f} q_m b_m + q_m (1 - b_m) \Upsilon_m, \tag{5.6}$$

where $\boldsymbol{b} = \{b_1, \ldots, b_m, \ldots, b_{N_f}\}$, and $\Upsilon_m = \mathbb{P}(\mathrm{SIR}_m > \vartheta)$ is the rate coverage probability for content m, i.e., the probability that the received SIR via CoMP transmission is larger than a target threshold ϑ, which we characterize in the sequel. In (5.6), the first term corresponds to the event of serving a desired content from local memory, i.e., self-cache [136]. The second term represents the joint event that the desired content is not locally cached while being cached and downloadable from active providers in the same cluster, with an SIR greater than the target threshold ϑ.

5.4 Rate Coverage Probability Analysis

Our objective in this section is to analytically characterize the offloading gain. In particular, we first derive the exact expression of $\mathbb{P}_o(\boldsymbol{b})$ as a function of the system parameters, namely, density of clusters, number of devices, and the intra-cluster distance between devices. Then, we seek lower bound and approximation of the rate coverage probability Υ_m that will result in easy-to-compute expressions of the offloading gain, and provide useful system design insights.

In the case of CoMP transmissions, active providers in the representative cluster jointly transmit the requested content to the typical client. The received power at

the typical client is the sum of the received signal powers from active providers, and hence, the rate coverage probability Υ_m is:

$$\Upsilon_m = \mathbb{P}\left[\frac{\gamma_d\Big|\sum_{\boldsymbol{y}_{0i}\in\Phi_{cpm}} G_{\boldsymbol{y}_{0i}}\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|^{-\alpha/2}\Big|^2}{I_{\text{out}}} \geq \vartheta\right]. \tag{5.7}$$

Since $G_{\boldsymbol{y}_{0i}}$ are i.i.d. complex Gaussian RVs, we get

$$\left|\sum_{\boldsymbol{y}_{0i}\in\Phi_{cpm}} \|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|^{-\alpha/2} G_{\boldsymbol{y}_{0i}}\right|^2 \sim \exp\left(\frac{1}{\sum_{\boldsymbol{y}_{0i}\in\Phi_{cpm}}\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|^{-\alpha}}\right). \tag{5.8}$$

Hence, from (5.7) and (5.8), we have

$$\begin{aligned}\Upsilon_m &= \mathbb{E}\Big[\exp\Big(-\frac{\vartheta(I_{\text{out}})}{\gamma_d S_{\Phi_{cpm}}}\Big)\Big] \\ &\overset{(a)}{=} \mathbb{E}\left[\mathcal{L}_{I_{\text{out}}}(t)\middle| S_{\Phi_{cpm}} = s_{\Phi_{cpm}}\right],\end{aligned} \tag{5.9}$$

where $S_{\Phi_{cpm}} = \sum_{\boldsymbol{y}_{0i}\in\Phi_{cpm}}\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|^{-\alpha}$ is a RV that can be physically interpreted as received signal power from the active providers devices $\boldsymbol{y}_{0i}\in\Phi_{cpm}$ subject to path loss only (as we already averaged over the fading based on the PDF in (5.8)), assuming normalized power. (a) follows from the Laplace transform of the interference I_{out} evaluated at $t=\frac{\vartheta}{\gamma_d s_{\Phi_{cpm}}}$. We derive the Laplace transform of interference in the following Lemma to compute the rate coverage probability, and correspondingly, the offloading gain.

Lemma 5.4.0.1. *Laplace transform of the inter-cluster interference, conditioned on a realization of the active providers for content m in the representative cluster, is given by*

$$\mathcal{L}_{I_{\text{out}}}(t) = \exp\Big(-2\pi\lambda_p\int_{v=0}^{\infty}\Big(1-e^{-p\bar{n}\zeta(v,t)}\Big)v\,\mathrm{d}v\Big), \tag{5.10}$$

where $t=\frac{\vartheta}{\gamma_d s_{\Phi_{cpm}}}$, $\zeta(v,t)=\int_{u=0}^{\infty}\frac{t\gamma_d}{u^\alpha+t\gamma_d}f_{U|V}(u|v)\,\mathrm{d}u$, $f_{U|V}(u|v)=\mathrm{Rice}(u;v,\sigma)$ is the Rician PDF modeling the distance $U=\|\boldsymbol{x}+\boldsymbol{y}\|$ between an interfering device at $\boldsymbol{y}$ relative to its cluster center at $\boldsymbol{x}\in\Phi_p$ and the origin $(0,0)$, conditioned on $V=\|\boldsymbol{x}\|=v$.

Proof. Please refer to Appendix C.1. □

We continue by characterizing the joint serving distance distribution. For a given realization $S_{\Phi_{cpm}} = s_{\Phi_{cpm}}$, let us assume that there are k active providers in the representative cluster. Let us also denote joint distances from the typical client (origin) to the k content providers in the representative cluster, centered at $\boldsymbol{x}_0$, as $\boldsymbol{H}_k = \{H_1, \ldots, H_k\}$. Then, conditioning on $\boldsymbol{H}_k = \boldsymbol{h}_k$, where $\boldsymbol{h}_k = \{h_1, \ldots, h_k\}$, the conditional (i.e., on k) PDF of the joint serving distance is denoted as $f_{\boldsymbol{H}_k}(\boldsymbol{h}_k)$. Hence, conditioning on k, we can express the rate coverage probability as

$$\Upsilon_{m|k} = \mathbb{E}\left[\mathcal{L}_{I_{\text{out}}}\Big(t = \frac{\vartheta}{\gamma_d \sum_{i=1}^{k} h_i^{-\alpha}}\Big)\middle| S_{\Phi_{cpm}} = s_{\Phi_{cpm}}\right], \tag{5.11}$$

where $s_{\Phi_{cpm}} = \sum_{i=1}^{k} h_i^{-\alpha}$, $h_i = \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|$, and $\boldsymbol{y}_{0i} \in \Phi_{cpm}$. Since a content provider i in the representative cluster centered at $\boldsymbol{x}_0$ has its coordinates in $\mathbb{R}^2$ chosen independently from a Gaussian distribution with standard deviation σ, then, by definition, the distance from such a content provider to the origin, denoted as $h_i = \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|$, has Rician distribution $f_{H_i|V_0}(h_i|v_0) = \text{Rice}(h_i; v_0, \sigma)$. Since also the content providers and the typical client have their locations sampled from a normal distribution with variance σ^2 relative to their cluster center $\boldsymbol{x}_0$, then, by definition, the statistical distance distribution between any two points, e.g., from the i-th content provider to the typical client, follows Rayleigh distribution with scale parameter $\sqrt{2}\sigma$, written as

$$f_{H_i}(h_i) = \text{Rayleigh}(h_i, \sqrt{2}\sigma) = \frac{h_i}{2\sigma^2} e^{-\frac{h_i^2}{4\sigma^2}}. \tag{5.12}$$

If the serving distances from the typical client to the different points of the cluster were independent from each other, $f_{\boldsymbol{H}_k}(\boldsymbol{h}_k)$ would simply be the product of k independent PDFs, each of which having $f_{H_i}(h_i) = \text{Rayleigh}(h_i, \sqrt{2}\sigma)$. However, there is a correlation between the serving distances due to the common factor $\boldsymbol{x}_0$ in the serving distance equation $h_i = \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|$ with $\boldsymbol{y}_{0i} \in \Phi_{cpm}$, see also Fig. 5.1. For analytical tractability, and similar to [76], we neglect this correlation. Hence, the conditional PDF of the joint serving distance $f_{\boldsymbol{H}_k}(\boldsymbol{h}_k)$ can be obtained from

$$f_{\boldsymbol{H}_k}(\boldsymbol{h}_k) = \prod_{i=1}^{k} \frac{h_i}{2\sigma^2} e^{-\frac{h_i^2}{4\sigma^2}}. \tag{5.13}$$

Conditioning on having k active content providers, i.e., $s_{\Phi_{cpm}} = \sum_{i=1}^{k} h_i^{-\alpha}$, the rate coverage probability will be:

$$\Upsilon_{m|k} = \int_{\boldsymbol{h}_k=\mathbf{0}}^{\infty} \mathscr{L}_{I_{\text{out}}}\left(\frac{\vartheta}{\gamma_d \sum_{i=1}^{k} h_i^{-\alpha}}\middle| k\right) f_{\boldsymbol{H}_k}(\boldsymbol{h}_k)\,\mathrm{d}\boldsymbol{h}_k\,. \tag{5.14}$$

Given that Φ_{cpm} is a Gaussian PPP, the number of active content providers for content m is a Poisson RV with mean $b_m p\bar{n}$. Therefore, the probability that there are k content providers is equal to $\frac{(p\bar{n}b_m)^k e^{-p\bar{n}b_m}}{k!}$. Invoking this along with (5.10), (5.13), and (5.14) into (5.6), $\mathbb{P}_o(\boldsymbol{b})$ is given as

$$\sum_{m=1}^{N_f} q_m \left(b_m + \big(1-b_m\big).\sum_{k=1}^{\infty} \frac{(p\bar{n}b_m)^k e^{-b_m p\bar{n}}}{k!} \int_{\boldsymbol{h}_k=\mathbf{0}}^{\infty} e^{-2\pi\lambda_p \int_{v=0}^{\infty}\left(1-e^{-p\bar{n}(1-\zeta(v,t))}\right)v\mathrm{d}v} \prod_{i=1}^{k} \frac{h_i}{2\sigma^2} e^{-\frac{h_i^2}{4\sigma^2}}\,\mathrm{d}\boldsymbol{h}_k \right). \tag{5.15}$$

Since the obtained expression in (5.15) involves multi-fold integrals and summations, this renders the calculation of the rate coverage probability computationally complex. Furthermore, the offloading gain maximization problem turns to be intractable. Therefore, in the sequel, we focus on tight bound and approximation of the rate coverage probability that will result in easy-to-compute expressions that also enable us to formulate a tractable optimization problem to maximize the offloading gain.

5.4.1 Lower Bound on Offloading Gain

Next, we obtain a tractable lower bound on the offloading gain based on an upper bound on the interference power.

Theorem 5.4.1.1. *Laplace transform of interference derived in (5.10) can be bounded by*

$$\mathcal{L}_{I_{\text{out}}}(t) \approx \exp\left(-\pi p\bar{n}\lambda_p t^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)\right), \tag{5.16}$$

and, correspondingly, a lower bound on the offloading gain $\mathbb{P}_o^{\sim}(\boldsymbol{b})$ is given by

$$\sum_{m=1}^{N_f} q_m \Big(b_m + \big(1-b_m\big).\sum_{k=1}^{\infty} \frac{(b_m p\bar{n})^k e^{-b_m p\bar{n}}}{k!} \int_{\mathbf{0}}^{\infty} e^{-\pi p\bar{n}\lambda_p \left(\frac{\vartheta}{\sum_{i=1}^{k} h_i^{-\alpha}}\right)^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)} \prod_{i=1}^{k} \frac{h_i}{2\sigma^2} e^{-\frac{h_i^2}{4\sigma^2}}\,\mathrm{d}\boldsymbol{h}_k \Big). \tag{5.17}$$

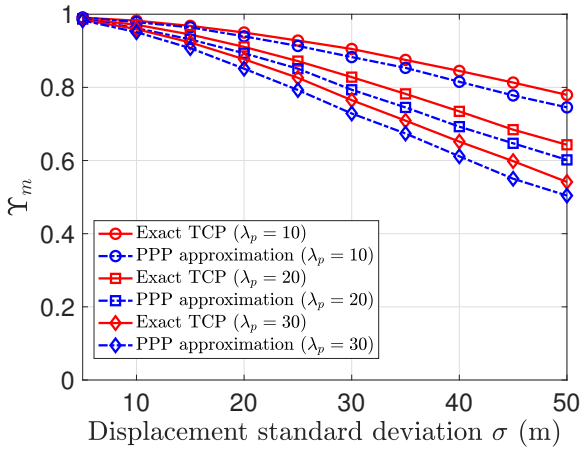


FIGURE 5.2: The lower bound on Υ_m based on (5.16) versus displacement standard deviation σ is plotted for various parent PPP densities λ_p ($\bar{n} = 20$, $p = 0.5$, $b_m = 0.5$). "Exact TCP" in the legend refers to the exact performance for the TCP while "PPP approximation" refers to the lower bound based on Theorem 5.4.1.1.

Remark 5.4.1.1. The obtained expression in (5.16) boils down to the Laplace transform of a PPP with intensity $\bar{n}\lambda_p$. This shows that the inter-cluster interference of a TCP with density of clusters λ_p and average number of devices per cluster $\bar{n}$, i.e., with intensity $\bar{n}\lambda_p$, is upper bounded by that of a PPP of the same intensity.

In Fig. 5.2, we plot the exact expression and its lower bound, based on (5.16), of the rate coverage probability versus displacement variance σ for various parent PPP Φ_p densities λ_p. The derived lower bound is considerably tight when both σ and λ_p are relatively small. Also, it is noticeable that Υ_m monotonically decreases with both σ and λ_p, which reflects the fact that the desired signal is weaker when the distance between content providers and the typical client is larger, and the effect of inter-cluster interference increases when the density of clusters increases, respectively. When λ_p and σ increase, the obtained lower bound becomes no longer tight, however, it still represents a reasonable bound on the exact Υ_m.

Having obtained a lower bound on the offloading gain, next, we seek a further approximation by replacing the desired signal component by a sum of two components, namely, nearest and mean components.

5.4.2 Serving Power Approximation

From (5.8), $S_{\Phi_{cpm}} = \sum_{\boldsymbol{y}_{0i}\in\Phi_{cpm}} \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|^{-\alpha}$ represents a RV that models the intended signal power from content providers $\boldsymbol{y}_{0i} \in \Phi_{cpm}$ subject to path loss only. Next, we adopt the so-called mean plus nearest approximation to approximate this intended signal power $S_{\Phi_{cpm}}$ as a sum of two terms. The first term is the received signal power from the nearest active provider while the second term is the average over the signal power received from all other active providers conditioning on the nearest serving distance $h_1 = \|\boldsymbol{x}_0 + \boldsymbol{y}_{01}\|$, where $\boldsymbol{y}_{01} = \operatorname{argmin}_{\boldsymbol{y}_{0i}\in\Phi_{cpm}}\{\|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|\}$. As we will see, this approximation yields an easy way to obtain the rate coverage probability while also being tight. This approach has been similarly adopted to circumvent intractable analysis in the stochastic geometry literature, see, e.g., [137]. Starting from the Laplace transform expression in (5.9), we approximate $S_{\Phi_{cpm}}$ as

$$S_{\Phi_{cpm}} \approx \left\|\boldsymbol{x}_0 + \boldsymbol{y}_{01}\right\|^{-\alpha} + \mathbb{E}\Big[S_{\Phi^{!}_{cpm}}|\boldsymbol{y}_{0i}\Big], \tag{5.18}$$

where $\Phi^{!}_{cpm} = \Phi_{cpm} \setminus \boldsymbol{y}_{01}$, and $S_{\Phi^{!}_{cpm}} = \sum_{\boldsymbol{y}_{0i}\in\Phi^{!}_{cpm}} \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|^{-\alpha}$. Next, we derive the distribution of nearest serving distance $h_1 = \|\boldsymbol{x}_0 + \boldsymbol{y}_{01}\|$. Then, we prove the concentration of the proposed approximation using Chebyshev's inequality. Finally, given the distance distribution to the nearest device $f_{H_1}(h_1)$, and a derived formula for $\mathbb{E}\left[S_{\Phi^{!}_{cpm}}\middle|H_1 = h_1\right]$, an approximation for $S_{\Phi_{cpm}}$ is obtained based on (5.18).

Lemma 5.4.2.1. *The PDF of the distance from the typical client to the nearest active provider in Φ_{cm} is given by*

$$f_{H_1}(h_1) = b_m p\bar{n} \int_{v_0=0}^{\infty} f_{V_0}(v_0) f_{H_1|V_0}(h_1|v_0) e^{-b_m p\bar{n}\int_0^{h_1} f_{H|V_0}(h|v_0)\mathrm{d}h}\,\mathrm{d}v_0\,, \tag{5.19}$$

which can be approximated by

$$f_{H_1}(h_1) \approx \frac{b_m p\bar{n} h_1 \exp\left(-b_m p\bar{n}\left(1 - \exp\left(\frac{-h_1^2}{4\sigma^2}\right)\right) - \frac{h_1^2}{4\sigma^2}\right)}{2\sigma^2}. \tag{5.20}$$

Proof. The proof is provided in Appendix C.3. □

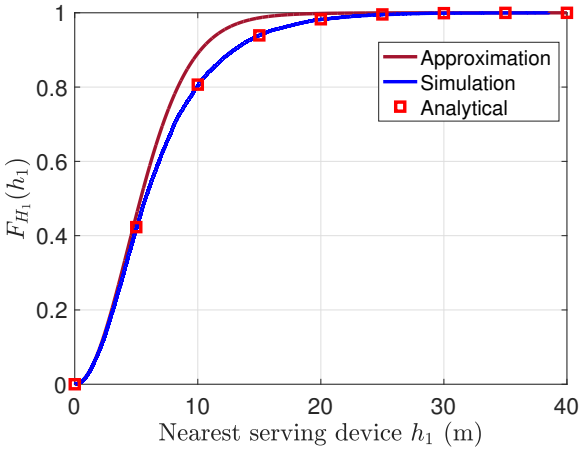


FIGURE 5.3: The derived nearest serving distance CDF in (C.4) is plotted and compared with simulation and Jensen's inequality-based approximation in (C.5) ($\bar{n} = 20$, $\sigma = 10\,\mathrm{m}$, $b_m = 0.5$, $p = 1$).

The accuracy of the derived cumulative distribution function (CDF) $F_{H_1}(h_1)$ in (C.4) and its approximation based on Jensen's inequality in (C.5) (see Appendix C.3) are verified in Fig. 5.3. It is clear from (5.20) that the distance to the nearest active content provider statistically decreases as b_m or p increase, i.e., when there is a high probability of having active and caching content providers within the local cluster. The distance is also more likely to decrease as $\bar{n}$ increases since a congested cluster has shorter distance between the content client and providers.

Next, we show that approximating the desired signal by its nearest and conditional mean components, see (5.18), yields an accurate yet tractable expression for the rate coverage probability and offloading gain.

Proposition 5.4.2.1. *For scenarios of practical interest, the proposed approximation for $S_{\Phi_{cpm}}$ in (5.18) is a tractable yet remarkably tight bound, and hence, it introduces a reasonable approximation for the rate coverage probability and offloading gain.*

Proof. The proof of the proposition relies on calculating the concentration bounds for $S_{\Phi^!_{cpm}}$. In other words, we will show that the RV $S_{\Phi^!_{cpm}}$ concentrates around its mean. For that purpose, we use Chebyshev's inequality that can be formulated as

$$\mathbb{P}\left(\left|S_{\Phi^!_{cpm}} - \mathbb{E}\left[S_{\Phi^!_{cpm}}\right]\right| > a\right) \leq \frac{\mathrm{Var}\left[S_{\Phi^!_{cpm}}\right]}{a^2}, \tag{5.21}$$

for $a > 0$, where $\mathrm{Var}\left[S_{\Phi^!_{cpm}}\right]$ is the variance of $S_{\Phi^!_{cpm}}$. We start by calculating the

conditional variance $\mathrm{Var}\left[S_{\Phi^!_{cpm}}|h_1\right]$ and mean $\mathbb{E}\left[S_{\Phi^!_{cpm}}|h_1\right]$ in the next two Lemmas.

Lemma 5.4.2.2. *The variance of the signal power received from all active providers except for the nearest device, subject to path loss only, and conditioned on the distance to the nearest active provider $H_1 = h_1$, is expressed as*

$$\mathrm{Var}\left[S_{\Phi^!_{cpm}}|H_1 = h_1\right] = b_m p\bar{n}\Gamma\left(-2\alpha+1, \frac{h_1^2}{4\sigma^2}\right), \tag{5.22}$$

Proof. The proof can be found in Appendix C.4. □

Lemma 5.4.2.3. *The average over the signal power received from all active content providers except for the nearest one, subject to path loss only, and conditioned on the distance to the nearest active provider $H_1 = h_1$, is expressed as*

$$\mathbb{E}\left[S_{\Phi^!_{cpm}}\middle|H_1 = h_1\right] = \frac{b_m p\bar{n}}{2\sigma^2}\left[\frac{\exp\left(-\frac{h_1^2}{4\sigma^2}\right)}{2h_1^2} - \frac{\Gamma(0, \frac{h_1^2}{4\sigma^2})}{8\sigma^2}\right]. \tag{5.23}$$

Proof. We can write the conditional mean as

$$\begin{aligned}\mathbb{E}\left[S_{\Phi^!_{cpm}}|H_1 = h_1\right] &= \mathbb{E}\Big[\sum_{\boldsymbol{y}_{0i}\in\Phi^!_{cpm}} \|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|^{-\alpha}\Big] \\ &\overset{(a)}{=} b_m p\bar{n}\int_{\mathbb{R}^2} \frac{1}{\|\boldsymbol{x}_0 + \boldsymbol{y}_{0i}\|^{\alpha}} f_{\boldsymbol{Y}_{0i}}(\boldsymbol{y}_{0i})\,\mathrm{d}\boldsymbol{y}_{0i}\,.\end{aligned} \tag{5.24}$$

where (a) follows from the mean and variance for PPPs [75, Corollary 4.8], along with the Gaussian PPP assumption $\Phi^!_{cpm}$. Following the same methodology as in Appendix C.4, the conditional mean can be directly obtained. Hence, Lemma 5.4.2.3 is proven. □

As an illustrative example, it is reasonable to assume a limited cache-size per device, which triggers $b_m < 1$, mean number of devices per cluster $\bar{n}$ from 5 to 10 devices, and small displacement standard deviation σ from $1\,\mathrm{m}$ to $10\,\mathrm{m}$. In such setup, we can observe the tightness of the approximation in Fig. 5.4. Particularly, we plot the term $\mathrm{Var}\left[S_{\Phi^!_{cpm}}\right]/a^2$, measuring how much $S_{\Phi^!_{cpm}}$ deviates from its mean, along with the CDF of nearest serving distance $F_{H_1}(h_1)$ versus the nearest

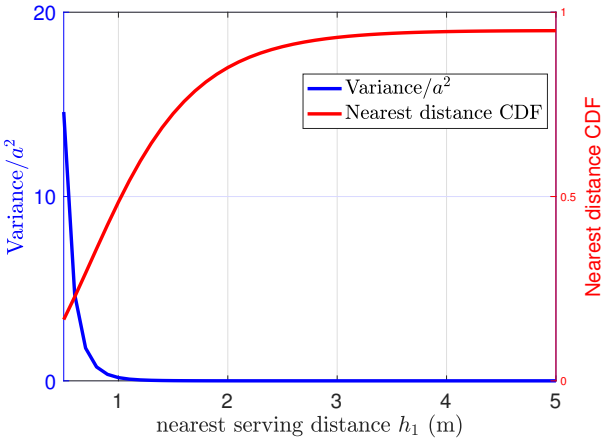


FIGURE 5.4: Nearest serving distance CDF $F_{H_1}(h_1)$ (right side y-axis) and Var $\left[S_{\Phi^!_{cpm}}\right]/a^2$ (left side y-axis) are plotted versus the nearest serving distance h_1 ($\sigma = 1\,\mathrm{m}$, $a = 1$, $b_m = 0.6$, $p = 0.5$, $\bar{n} = 10$).

distance h_1. From the figure, we first note that Var $\left[S_{\Phi^!_{cpm}}\right]/a^2$ is almost zero when the nearest active provider is farther than $\sigma = 1\,\mathrm{m}$, which happens with high probability (from the CDF $F_{H_1}(h_1)$). Moreover, Var $\left[S_{\Phi^!_{cpm}}\right]/a^2$ is larger than zero when the distance to the nearest active provider is shorter than σ, which happens with small probability (from the CDF $F_{H_1}(h_1)$). This shows that the "mean plus nearest" approximation can yield a remarkably tight bound on $S_{\Phi_{cpm}}$, and correspondingly on $\mathbb{P}_o(\boldsymbol{b})$ for scenarios of practical interest. Hence, the proof of Proposition 5.4.2.1 is complete. □

Piecing everything together, we get a tight approximation on $\mathbb{P}_o(\boldsymbol{b})$ as follows. We start with (5.9) with the substitution $s_{\Phi_{cpm}} \approx h_1^{-\alpha} + \mathbb{E}\left[S_{\Phi^!_{cpm}}|H_1 = h_1\right]$, where $\mathbb{E}\left[S_{\Phi^!_{cpm}}|H_1 = h_1\right]$ is derived in (5.23). Then, we proceed by calculating Laplace transform of the inter-cluster interference before averaging over the nearest distance h_1 using the nearest distance PDF $f_{H_1}(h_1)$ in (5.19). The approximated offloading gain is formally characterized in the next corollary.

Corollary 5.4.2.1. *A tight approximation of the offloading gain can be calculated from*

$$\mathbb{P}_o^{\approx}(\boldsymbol{b}) = \sum_{m=1}^{N_f} q_m \Big(b_m + (1-b_m) \int_{h_1=0}^{\infty} e^{-2\pi\lambda_p \int_{v=0}^{\infty} \left(1-e^{-p\bar{n}\zeta(v,t)}\right) v \mathrm{d}v} f_{H_1}(h_1) \mathrm{d}h_1 \Big) \tag{5.25}$$

$$\overset{(a)}{\approx} \sum_{m=1}^{N_f} q_m \Big(b_m + (1-b_m) p\bar{n} b_m \int_{h_1=0}^{\infty} \frac{h_1}{2\sigma^2} e^{-2\pi\lambda_p \int_{v=0}^{\infty} \left(1-e^{-p\bar{n}\zeta(v,t)}\right) v \mathrm{d}v} e^{-b_m p\bar{n}\left(1-e^{\frac{-h_1^2}{4\sigma^2}}\right) - \frac{h_1^2}{4\sigma^2}} \mathrm{d}h_1 \Big), \tag{5.26}$$

where

$$t = \frac{\vartheta/\gamma_d}{h_1^{-\alpha} + \frac{b_m p\bar{n}}{2\sigma^2}\left[\frac{e^{-\frac{h_1^2}{4\sigma^2}}}{2h_1^2} - \frac{\Gamma(0, \frac{h_1^2}{4\sigma^2})}{8\sigma^2}\right]}. \tag{5.27}$$

Proof. The above result follows from the nearest plus mean approximation in (5.18), along with the conditional mean expression obtained in (5.23); (a) follows from the approximated nearest serving distance PDF obtained in (5.20). □

Note that the exponential term inside the integral of (5.25) is a function of h_1 since $t = \frac{\vartheta}{\gamma_d s_{\Phi_{cpm}}}$. It is worth mentioning that with such an approximation, replacing $s_{\Phi_{cpm}}$ with its nearest plus conditional mean approximation converts the multi-fold integral over $\boldsymbol{h}_k$ in (5.15) to a single integral over h_1. Furthermore, the effect of having a random number of active providers is implicitly involved in the nearest distance PDF as well as in the conditional mean term. This explains why the condition of having k active content providers in the representative cluster no longer exists in the approximated rate coverage probability expression.

In Fig. 5.5, we plot the obtained formulas for the rate coverage probability Υ_m in (5.25) and (5.26) versus the displacement variance σ for various density of clusters λ_p. The derived bound is remarkably tight for the practical values of λ_p and σ. It starts to slightly diverge only for a system with considerably large λ_p and σ, which is an impractical scenario given that both the density of clusters and the intra-distance between devices are high. It is also clear that the approximated expression of the nearest distance PDF in (5.20) bounds that in (5.19) very tightly.

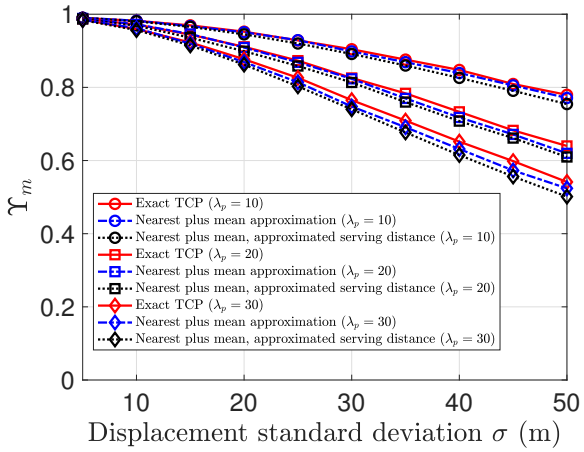


FIGURE 5.5: The derived approximations of Υ_m in (5.25) and (5.26) are plotted versus the displacement standard deviation σ for various parent PPP density λ_p ($\bar{n} = 20$, $p = 0.5$, $b_m = 0.5$). "Nearest plus mean approximation" in the legend refers to the performance based on the exact nearest serving distance PDF in (5.19).

It is worth mentioning that the importance of the obtained approximation $\mathbb{P}_o^{\approx}(\boldsymbol{b})$ in (5.25) is two-fold. Firstly, it provides an easy-to-compute approximation for the rate coverage probability, as shown in Fig. 5.5, and correspondingly, the offloading gain. Secondly, it allows us to solve for an optimized caching probability, by maximizing the approximated offloading gain, as will be clear shortly. In this regard, it is quite important to examine the achievable performance when a requested content is downloaded only from the nearest active provider. We refer to this as *nearest content provider (NCP) transmission scheme*, where its corresponding rate coverage probability is characterized in the next corollary.

Corollary 5.4.2.2. *The rate coverage probability for the NCP transmission scheme is expressed as*

$$\begin{aligned}\Upsilon_m &= \int_0^{\infty} e^{-2\pi\lambda_p \int_{v=0}^{\infty} \left(1-e^{-\bar{n}\zeta(v,t)}\right) v \mathrm{d}v} f_{H_1}(h_1) \mathrm{d}h_1 \\ &\approx \frac{b_m p \bar{n}}{2\sigma^2} \int_0^{\infty} h_1 e^{-2\pi\lambda_p \int_{v=0}^{\infty} \left(1-e^{-\bar{n}\zeta(v,t)}\right) v \mathrm{d}v} \times \\ &\qquad\qquad e^{-b_m p \bar{n}\left(1-e^{\left(\frac{-h_1^2}{4\sigma^2}\right)}\right) - \frac{h_1^2}{4\sigma^2}} \mathrm{d}h_1,\end{aligned}$$

where $t = \frac{\vartheta}{\gamma_d h_1^{-\alpha}}$, *and* $\zeta(v,t)$ *is defined in Lemma 5.4.0.1.*

Having obtained a lower bound and approximation of the rate coverage probability, next, we employ the obtained results to compute optimized caching probabilities $\boldsymbol{b}$ in order to maximize the offloading gain.

5.5 Optimized Caching Probabilities

We first formulate and solve the offloading gain maximization problem based on the mean plus nearest approximation in (5.26). Then we pursue another approach, based on the derived lower bound in (5.17), to obtain a low complexity solution.

5.5.1 Optimized Caching Based on the approximation

Here, we aim at maximizing the approximated offloading gain obtained in Corollary 5.4.2.1. We formulate the offloading gain maximization problem as

$$\textbf{P1:} \quad \max_{\boldsymbol{b}} \quad \mathbb{P}_o^{\approx}(\boldsymbol{b}) \tag{5.28}$$

$$\text{s.t.} \quad \sum_{n=1}^{N_f} b_m = M, \quad b_m \in [0,1] \tag{5.29}$$

Since the integral in the approximated offloading gain expression in (5.25) depends on the caching probability b_m, and b_m exists as a complex exponential term in the nearest serving distance PDF $f_{H_1}(h_1)$, it is hard to analytically characterize (e.g., show concavity of) the objective function or find a tractable expression for the caching probability b_m. In order to tackle this, similar to [114], we introduce a $\boldsymbol{b}$-independent integral by substituting the caching probability with an arbitrary caching probability $\boldsymbol{b}^0$. Denoting this $\boldsymbol{b}$-independent integral by $I^{\boldsymbol{b}^0}$, it can be easily verified that

$$\mathbb{P}_o^{\approx^{\boldsymbol{b}^0}}(\boldsymbol{b}) = \sum_{m=1}^{N_f} q_m \Big(b_m + (1-b_m) p \bar{n} b_m I^{\boldsymbol{b}^0} \Big), \tag{5.30}$$

is concave in $\boldsymbol{b}$. From the Karush-Kuhn-Tucker (KKT) conditions, the optimized caching probability maximizing $\mathbb{P}_o^{\approx^{\boldsymbol{b}^0}}(\boldsymbol{b})$ under the constraint (5.29) is given by

$$b_m^* = \Big[0.5 - \frac{v^* - q_m}{2 q_m p \bar{n} I^{\boldsymbol{b}^0}} \Big]^+, \tag{5.31}$$

where v^* satisfies the maximum cache constraint $\sum_{i=1}^{N_f}\left[0.5-\frac{v^*-q_m}{2q_m p\bar{n} I^{b^0}}\right]^+ = M$, and $[x]^+ = \max(x,0)$. For the arbitrary caching probability $\boldsymbol{b}^0$, a locally optimal caching probability can be adopted, which can be computed via, e.g., interior point method [115]. Nonetheless, we can increase the probability of finding the optimal solution of **P1** by using the interior point method with multiple random initial values and then picking the solution with the highest offloading gain. We refer to this caching probability in (5.31) as the solution from convex approximation. However, to obtain a caching policy of lower complexity, we maximize a special case of the lower bound $\mathbb{P}_o^{\sim}(\boldsymbol{b})$ in the sequel.

5.5.2 Optimized Caching Based on the Lower Bound

Although $\mathbb{P}_o^{\sim}(\boldsymbol{b})$ characterized in Theorem 5.4.1.1 is simpler to compute compared to $\mathbb{P}_o(\boldsymbol{b})$, it is still challenging to obtain the optimal caching probability due to the summation and multi-fold integration in (5.17). Similar to [123], we consider a special case when downloading content from one active content provider, for which the offloading gain maximization problem turns out to be convex.

One Content Provider

Next, we solve for the optimized caching probability when considering one serving content provider (instead of k in (5.17)). Starting from (5.17) with $t = \frac{\vartheta h^{\alpha}}{\gamma_d}$ for one serving provider, and the void probability of a Poisson RV, we get

$$\Upsilon_m = \big(1-\frac{(p\bar{n}b_m)^0}{0!}e^{-b_m p\bar{n}}\big)\times \int_{h=0}^{\infty} e^{-\pi p\bar{n}\lambda_p(\vartheta h^{\alpha})^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)}\frac{h}{2\sigma^2}e^{-\frac{h^2}{4\sigma^2}}\,\mathrm{d}h\,. \tag{5.32}$$

Solving the integral in (5.32), and substituting in (5.17), we get $\mathbb{P}_o^{\sim 1}(\boldsymbol{b})$ written as

$$\mathbb{P}_o^{\sim 1}(\boldsymbol{b}) = \sum_{m=1}^{N_f} q_m\Big(b_m + \big(1-b_m\big)\big(1-e^{-b_m p\bar{n}}\big)\frac{1}{\mathcal{Z}(\vartheta,\alpha,\sigma)}\Big), \tag{5.33}$$

where $\mathcal{Z}(\vartheta,\alpha,\sigma)=4\sigma^2\pi p\bar{n}\lambda_p\vartheta^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)+1$. Hence, the optimized caching probability can be computed by solving the following problem

$$\textbf{P2:}\quad \max_{\boldsymbol{b}}\quad \mathbb{P}_o^{\sim 1}(\boldsymbol{b}) \tag{5.34}$$

$$\text{s.t.}\quad \sum_{n=1}^{N_f} b_m = M,\quad b_m\in[0,1] \tag{5.35}$$

The optimal solution for **P2** is formulated in the following Lemma.

Lemma 5.5.2.1. *The lower bound on the offloading gain $\mathbb{P}_o^{\sim 1}(\boldsymbol{b})$ in (5.33) is concave w.r.t. the caching probability, and the optimal probabilistic caching $\underline{\boldsymbol{b}}^*$ for **P2** is given by*

$$\underline{b_m}^* = \begin{cases} 1 & ,v^* < q_m - \frac{q_m(1-e^{-p\bar{n}})}{\mathcal{Z}} \\ 0 & ,v^* > q_m + \frac{p\bar{n}q_m}{\mathcal{Z}} \\ \psi(v^*) & ,\text{otherwise,} \end{cases}$$

where $\psi(v^*)$ is the solution of

$$v^* = q_m + \frac{q_m}{\mathcal{Z}}\big(p\bar{n}(1-\underline{b_m}^*)e^{-p\bar{n}\underline{b_m}^*} - (1-e^{-p\bar{n}\underline{b_m}^*})\big),$$

that satisfies $\sum_{m=1}^{N_f}\underline{b_m}^* = M$, and $\mathcal{Z}=\mathcal{Z}(\vartheta,\alpha,\sigma)$ for ease of presentation.

Proof. It is easy to show the concavity of the objective function $\mathbb{P}_o^{\sim 1}(\boldsymbol{b})$ by confirming that Hessian matrix w.r.t. the caching variables is negative semi-definite. Also, the constraints are linear, which imply that the necessity and sufficiency conditions for optimality exist. The detailed proof of finding $\underline{\boldsymbol{b}}^*$ is omitted for brevity. □

It is worth mentioning that the optimal caching solution $\underline{\boldsymbol{b}}^*$ for **P2** is suboptimal relative to the caching solution of the original problem encompassing cooperative transmission. However, when substituted in (5.15), it provides useful insights into the system design and also attains considerable performance improvements over traditional caching schemes, as quantified in Section 5.6.

TABLE 5.1: Simulation Parameters

Description	Parameter	Value
Path loss exponent	α	4
SIR threshold	ϑ	$0\,\mathrm{dB}$
Density of cluster	λ_p	$30\,\mathrm{km}^{-2}$
Displacement standard deviation	σ	$30\,\mathrm{m}$
Average number of devices per cluster	$\bar{n}$	6
Library size	N_f	12
Device cache size	M	2
Access probability	p	0.5

5.6 Numerical Results

In this section, we evaluate the performance of the proposed cache-assisted CoMP transmission for clustered D2D networks. Unless otherwise specified, results are obtained for the parameters shown in Table 1, which represent typical values used in many previous works.[1] We refer to both solutions based on convex approximation in (5.31) and the suboptimal caching of Lemma 5 as optimized PC. We first verify the accuracy of the derived bound and approximation of the offloading gain. Then, we compare the achievable performance of the proposed PC and CoMP transmission, with conventional caching and transmission schemes.

5.6.1 Exact Offloading Gain Versus Approximation and Lower Bound

Fig. 5.6 verifies the accuracy of the obtained lower bound and approximation of the offloading gain in (5.17) and (5.25), respectively. It is clear that both the derived lower bound and approximation are tight to the exact offloading gain. Moreover, we observe that averaging over the request and caching probabilities, i.e., q_m and b_m, makes the adopted lower bound and approximation of the offloading gain tighter than those for the rate coverage probability Υ_m shown earlier in Fig. 5.2 and Fig. 5.5, respectively. As shown in Fig. 5.6, the offloading gain monotonically increases with the popularity of files β. This is because when β is large, only a small portion of content undergoes most of the demand, which can be cached among the cluster devices.

[1] In Table 1, we assume a relatively small size of the per device cache and file library. These values, which are close to those used in [123] and [116], are reasonable in the study of communication and caching aspects of D2D content delivery networks. Other works in the literature, e.g., [64], considered a much larger size of file library, however, their objective was to conduct the scaling analysis of caching networks.

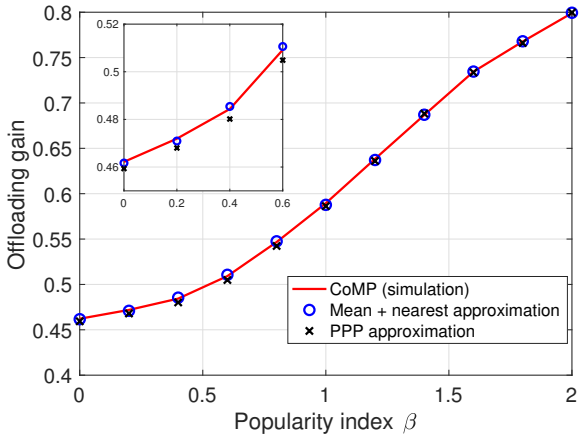


FIGURE 5.6: The exact offloading gain (simulation) based on CoMP transmission is compared to PPP-based lower bound ($\mathbb{P}_o^{\sim}(\boldsymbol{b})$), and mean plus nearest-based approximation ($\mathbb{P}_o^{\approx}(\boldsymbol{b})$), versus the popularity of files β under the Zipf caching scheme.

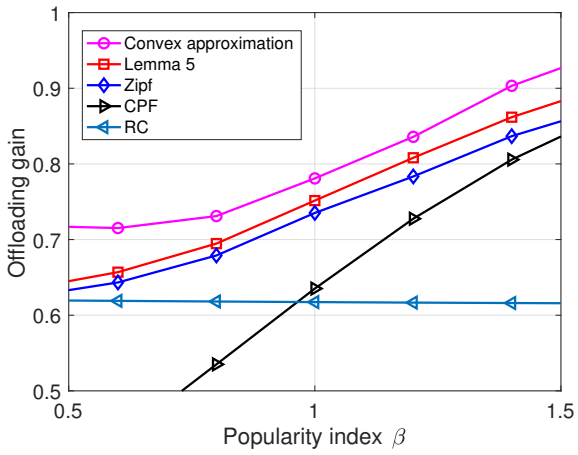
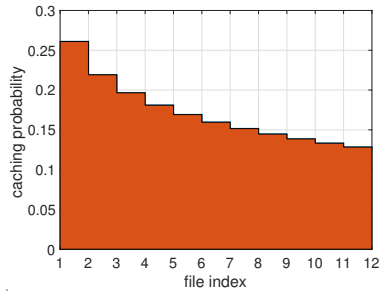


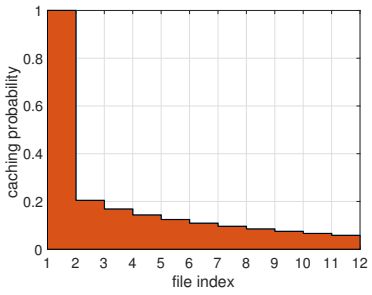
FIGURE 5.7: The offloading gain versus the popularity of files β under different caching schemes ($N_f = 40$, $M = 8$).

5.6.2 Comparison with Other Caching Schemes

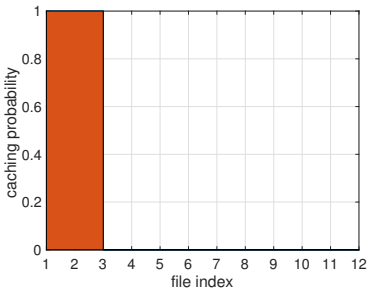
Fig. 5.7 compares the offloading gain of the proposed PC with other benchmark schemes, namely, Zipf caching (Zipf), CPF, and RC against the popularity of files β. For CPF, the M-most popular files are cached among each cluster device. Similarly, for RC, contents to be cached are uniformly chosen at random while for Zipf caching, the contents are chosen based on their popularity as in (5.4). Moreover, in Fig. 5.7, "convex approximation" refers to the caching solution characterized in (5.31), whereas "Lemma 5" refers to the caching probability characterized in Lemma 5. All the caching schemes are evaluated under CoMP transmissions. We



(a) Case one ($\sigma = 10\,\mathrm{m}$, $\lambda_p = 10\,\mathrm{km}^{-2}$)



(b) Case two ($\sigma = 30\,\mathrm{m}$, $\lambda_p = 30\,\mathrm{km}^{-2}$)



(c) Case three ($\sigma = 50\,\mathrm{m}$, $\lambda_p = 40\,\mathrm{km}^{-2}$)

FIGURE 5.8: Histogram of the caching solution from Lemma 5.5.2.1 is plotted for different network geometries ($\beta = 0.4$).

first see that the offloading gains under the optimized PC schemes outperform conventional caching schemes. Moreover, the PC based on convex approximation in (5.31) is superior to the suboptimal solution of Lemma 5.5.2.1. As β increases, except for RC, the offloading gain increases and gradually, the optimized PC, Zipf, and CPF schemes tend to achieve the same performance. This shows that when a small portion of files becomes highly demanded, i.e., for higher β, the optimal caching probability is attained via caching popular files among all cluster devices.

To show the prominent effect of the network geometry on the optimized caching probability, we plot the histograms of the solution of Lemma 5.5.2.1 for three different cases in Fig. 5.8. These three cases are ranging from a sparse network (small intra-cluster distance standard deviation σ and density of clusters λ_p), a relatively dense network (medium σ and λ_p), and a highly dense network (large σ and λ_p). Note that smaller σ results in higher desired signal power, while smaller λ_p yields lower inter-cluster interference power as clusters become sparser. The first case in Fig. 5.8(a) represents a sparse system with small values of λ_p and σ, i.e., sufficiently good transmission conditions. In this case, we see that the optimized

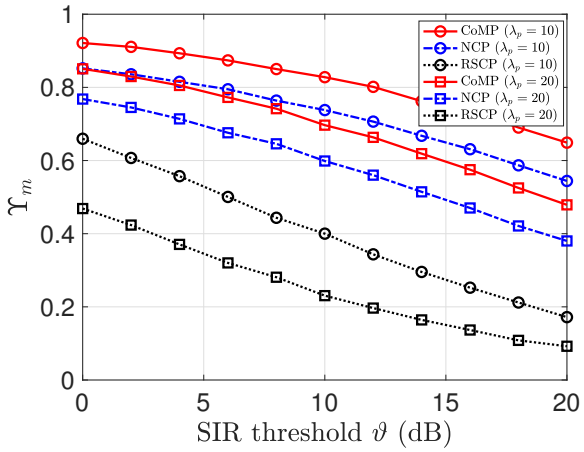


FIGURE 5.9: The rate coverage probability versus the SIR threshold ϑ for different transmission schemes ($\bar{n} = 20$, $b_m = 0.5$).

caching probability tends to be more uniform taking advantage of hitting a large number of files while being served in favorable transmission conditions. The second case in Fig. 5.8(b) represents a system with relatively good transmission conditions, i.e., medium values of σ and λ_p. It is clear from the histogram that the optimized caching solution tends to be more skewed than in the first case. The third case in Fig. 5.8(c) is then for a highly dense network with large values of both σ and λ_p. Clearly, the caching probability tends to be very skewed, which implies that caching popular files is an appropriate choice for such a highly dense network, i.e., a network with poor transmission conditions. Summing up, the results in Fig. 5.8 reveal interesting dependence of the optimized caching probability in Lemma 5.5.2.1 on the network geometry.

5.6.3 Comparison with Other Transmission Schemes

To quantify how much CoMP transmission can improve the achievable performance, we here compare the rate coverage probability Υ_m for three transmission schemes, namely, CoMP, NCP, and randomly-selected content provider (RSCP). Recall that for the NCP scheme, the requested content is served from the the nearest active provider to the content client within the same cluster while for the RSCP scheme, an active provider is chosen at random to serve the desired content. Firstly, Fig. 5.9 plots the exact rate coverage probability versus SIR threshold ϑ for different

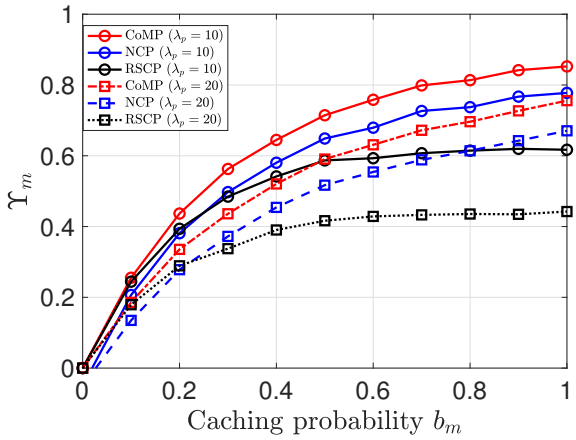


FIGURE 5.10: The rate coverage probability versus the caching probability b_m for different transmission schemes ($\bar{n} = 10$, $\vartheta = 5\,\mathrm{dB}$).

density of clusters λ_p. Intuitively, CoMP transmission achieves higher rate coverage probability than those of the other schemes. In particular, at high SIR threshold, allowing CoMP transmission can provide up to 300% improvement in the rate coverage probability compared to the RSCP scheme. Moreover, for all schemes, the rate coverage probability is seen to decrease as λ_p increases since higher interference power is encountered at the typical client when D2D clusters are denser.

In light of this comparison, Fig. 5.10 plots the rate coverage probability against the caching probability b_m for the three transmission schemes. As shown in Fig. 5.10, as the content availability increases, i.e., higher b_m, the rate coverage probability improves. Besides, Fig. 5.10 illustrates that while the rate coverage probability for the RSCP transmission scheme tends to flatten when b_m further increases, it keeps on increasing for CoMP and NCP transmission schemes. This is attributed to the fact that, for the NCP scheme, the serving distance is more likely to decrease with the increase of b_m, and hence the corresponding performance improves. Similarly, for the CoMP scheme, the transmission diversity improves with b_m since the average number of active and caching providers increases.

Finally, Fig. 5.11 illustrates the effect of the average number of devices per cluster $\bar{n}$ on the rate coverage probability. From Fig. 5.11, when $\bar{n}$ increases, the rate coverage probability for CoMP and NCP schemes increases. This is due to the fact that for the NCP scheme, the nearest serving distance is more likely to decrease

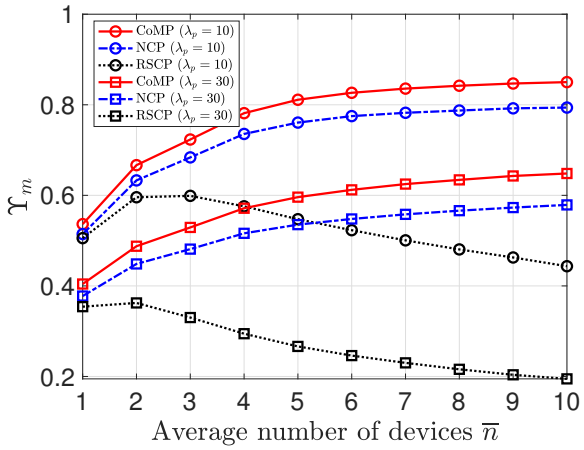


FIGURE 5.11: The rate coverage probability versus average number of devices per cluster for different transmission schemes ($\vartheta = 8\,\mathrm{dB}$, $b_m = 1$).

for congested clusters. Moreover, for the CoMP scheme, the transmission diversity also improves with $\bar{n}$. However, for the RSCP scheme, the rate coverage probability first increases driven by the increasing probability of finding a requested content within the local cluster. Then, the rate coverage probability turns to degrade as $\bar{n}$ further increases since, although the requested content can be found within the local cluster with high probability, the effects of interference also grows with $\bar{n}$. Furthermore, the rate coverage probability is shown to decrease for all schemes when λ_p increases driven by the growing effects of interference. Besides, while the performance of RSCP is very sensitive to the density of clusters, especially when the average number of devices is high, cooperative transmission attains substantially better performance. We hence conclude that for such highly dense D2D networks and adverse interference conditions, cooperative transmission becomes more appealing.

5.7 Conclusion

In this chapter, we have conducted performance analysis and content placement optimization for cache-assisted CoMP transmissions in clustered D2D networks. In particular, we have characterized the rate coverage probability and offloading gain as functions of the network parameters, namely, the density of clusters, average

number of devices per cluster, and the content popularity and placement schemes. Then, we have sought simple yet tight lower bound and approximation of the rate coverage probability and offloading gain. Based on the obtained results, we have showed that the inter-cluster interference of a TCP is upper bounded by that of a PPP of the same intensity. Moreover, we have formulated the corresponding offloading gain maximization problem and obtained optimized caching probabilities based on the proposed lower bound and approximation. Results showed that allowing CoMP transmission can attain up to 300% improvement in the rate coverage probability compared to the RSCP scheme. Finally, we conclude by showing that the proposed optimized PC results in a considerable improvement of the offloading gain over conventional caching schemes.

Chapter 6

Content Delivery to the Sky: Performance of Beamforming with Down-tilted Antennas for Ground and UAV User Co-existence

In the previous chapters, we have studied joint caching and communications for cache-enabled terrestrial networks, particularly, caching for D2D communications. Motivated by the increasing importance of contemporary aerial users (AUs) (i.e., drone users) and their wide range of applications, we extend our discussion in this chapter to caching and content delivery for co-existing ground and aerial users. In detail, providing reliable content delivery to aerial users (AUs) such as cellular-connected UAVs is a key challenge for tomorrow's cellular systems. In this chapter, the use of CB for simultaneous content delivery to an AU co-existing with multiple ground users (GUs) is investigated. In particular, a content delivery network of uniformly distributed massive MIMO-enabled ground BSs serving both aerial and ground users through spatial multiplexing is considered. For this model, the successful content delivery probability (SCDP) is derived as a function of the system parameters. The effects of various system parameters (such as antenna down-tilt angle, AU's altitude, number of scheduled users, and number of antennas) on the achievable performance are then investigated. Results reveal that whenever the AU's altitude is below the BS height, the antennas' down-tilt angles yield an inherent tradeoff between the performance of the AU and the GUs. However, if the

AU's altitude exceeds the BS height, down-tilting the BS antennas with a considerably large angle improves the performance of both the AU and the GUs.

6.1 Introduction

A tremendous increase in the use of UAVs in a wide range of applications, ranging from airborne BSs, delivery of goods, to traffic control, is expected in the foreseeable future [31] and [138]. To enable these applications, UAVs must communicate with one another and with ground devices. To enable such communications, it is imperative to connect UAVs, seen as AUs, to the ubiquitous wireless cellular network. Such cellular-connected UAVs have recently attracted attention in cellular network research in both academia and industry [13, 139–142] due to their ability to pervasively communicate. However, cellular networks have been designed to provide connectivity to GUs rather than AUs [139]. For instance, cellular-connected UAV communication possesses substantially different characteristics that pose new technical challenges which include: dominance of LoS interference and reduced ground base stations (GBSs) antenna gain [139].

6.1.1 Motivation and Contribution

The main contribution of this chapter is to propose a MIMO-CB approach that can improve the performance of cellular communication links for the AUs and effectively enhance the cellular system spectral efficiency (SE). We consider a network of one AU co-existing with multiple GUs that are being simultaneously served via massive MIMO-enabled GBSs. We introduce a novel analytical framework that can be leveraged to characterize the performance of the spatially multiplexed AU and GUs. Given the different channel characteristics and the corresponding precoding vectors among GUs and the AU, we first derive the gain of intended and interfering channels for both kind of users. We then analytically characterize the SCDP as a function of the system parameters. *To our best knowledge, this is the first work to perform a rigorous analysis of MIMO CB to simultaneously serve aerial and ground users.*

6.1.2 Related Works

The authors in [139] studied the feasibility of supporting drone operations using existing cellular infrastructure. Their results revealed that the favorable propagation conditions that AUs enjoy due to their altitude is also one of their strongest limiting factors since they are susceptible to LoS interference. Meanwhile, the authors in [141] minimized the UAV's mission completion time by optimizing its trajectory while maintaining reliable communication with the GBSs. In [142], through system simulations, the authors evaluated the performance of the downlink of AUs when supported by either a traditional cellular network, or a massive MIMO-enabled network with zero-forcing beamforming (ZFBF). In [13], the authors showed that cooperative transmission significantly improves the coverage probability for high-altitude AUs. However, while the works in [139, 141], and [13] have analyzed the performance of cellular-connected UAVs, their approaches can not be used to effectively improve the performance of AUs while enhancing SE by spatial multiplexing. Also, even though the work in [142] has proposed MIMO beamforming for an AU co-existing with multiple GUs, this work provides no analytical characterization of the performance of AUs or the impact of the antennas' down-tilt angles.

6.2 System Model

Consider a cellular network composed of massive MIMO-enabled BSs b_i distributed according to a homogeneous PPP Φ of intensity λ, where $\Phi = \{b_i \in \mathbb{R}^2, \forall i \in \mathbb{N}^+\}$. A three-sectored cell is associated with each BS, with each sector spanning an angular interval of $120\,^\circ$. Each sector has a large antenna array of M antennas each of which has a horizontal constant beamwidth of $120\,^\circ$, and vertical beamwidth θ_B. CB is employed to simultaneously serve a selected set $\mathcal{K}$ of K users. These K users are viewed as an AU that is scheduled with a set $\mathcal{K}_G$ of $K-1$ GUs, as done in [142]. This assumption is in line with the fact that the number of current GUs is much larger than the number of AUs. We assume that the GUs are located according to some independent stationary point process. BSs are deployed at the same height h_{BS} while AUs and GUs are at altitudes h_d and h_g, respectively, where

$h_d \gg h_g$. Given the symmetry of the problem, we consider the performance of the typical ground and aerial users located at $(0, 0, h_g)$, and $(0, 0, h_d)$, respectively. We also refer to the serving BS as *tagged BS*, which is the nearest BS to the origin $(0, 0) \in \mathbb{R}^2$, with d_{ig} and d_{id} being the distances from the GBS to the typical GU and AU, respectively.

For GUs, we consider i.i.d. quasi-static Rayleigh fading channels. The channel vector between the M antennas of tagged BS i and GU k is $\sqrt{\beta_{ik}}\boldsymbol{h}_{ik}$, where $\boldsymbol{h}_{ik} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \boldsymbol{I}_M)$ for $k \in \mathcal{K}_G$. σ^2 is the channel variance between each single antenna and user k, and $\boldsymbol{I}_M$ is the $M \times M$ identity matrix. $\beta_{ik} = d_{ig}^{-\alpha}$ defines the large-scale channel fading, where α is the large scale fading. We also assume that the GU channels are dominated by non-line-of-sight (NLoS) transmission. For the AU, we assume a wireless channel that is characterized by both large-scale and small-scale fading. For the large-scale fading, the channel between BS i and the AU includes LoS and NLoS components, which are considered separately along with their probabilities of occurrence [143]. For small-scale fading, we adopt a Nakagami-m_v model for the channel between each single antenna and the AU, as done in [13, 143, 144], with the following PDF:

$$f_{\Omega_v}(\omega, \eta) = \frac{2(\frac{m_v}{\eta})^{m_v}\omega^{2m_v-1}}{\Gamma(m_v)}\exp\big(-\frac{m_v}{\eta}\omega^2\big), \tag{6.1}$$

where $v \in \{l, n\}$, m_l and m_n are the fading parameters for the LoS and NLoS links, respectively, with $m_l > m_n$, and η is a controlling spread parameter. When $m_v = \eta = 1$, Rayleigh fading is recovered with an exponentially distributed instantaneous power, which is the case for GUs or AUs with no LoS communication. For Nakagami channels, we assume that the phase θ_{ng} is uniformly distributed in $[0, 2\pi]$, i.e., $\theta_{ng} \sim \mathcal{U}(0, 2\pi)$. Given that $\omega \sim$ Nakagami(m_v, η), it directly follows that the channel gain $\omega^2 \sim \Gamma(m_v, \frac{\eta}{m_v})$.

3D blockage is characterized by the fraction a of the total land area occupied by buildings, the mean number of buildings being ν per km^2, and the buildings' height modeled by a Rayleigh PDF with a scale parameter c. Hence, the probability of LoS when served from BS i, at a horizontal-distance r_i from the typical AU, is

given as [139]:

$$\mathbb{P}_l(r_i) = \prod_{n=0}^{p} \left[1 - \exp\Big(- \frac{\big(h_{\text{BS}} + \frac{h(n+0.5)}{p+1}\big)^2}{2c^2} \Big) \right], \tag{6.2}$$

where $h = h_d - h_{\text{BS}}$ and $p = \lfloor \frac{r_i\sqrt{a\nu}}{1000} - 1 \rfloor$. In our model, we assume that the AUs are deployed in an urban environment, where a and ν take relatively large values. Therefore, the large-scale channel fading for the AU is given by $d_{id}^{-\alpha_v}$, where $v \in \{l, n\}$, α_l and α_n are the path loss exponents for LoS and NLoS links, respectively, with $\alpha_l < \alpha_n$.

For a general user $k \in \mathcal{K}$ at altitude $h_k \in \{h_d, h_g\}$, the antenna directivity gain can be written similar to [139] as $G(r_i) = G_m$, for $r_i \in \mathcal{S}_{bs}$, and G_s, for $r_i \notin \mathcal{S}_{bs}$, where r_i is the horizontal distance to the BS, $\mathcal{S}_{bs}$ is formed by all the distances r_i satisfying $h_{\text{BS}} - r_i\tan(\theta_t + \frac{\theta_B}{2}) < h_k < h_{\text{BS}} - r_i\tan(\theta_t - \frac{\theta_B}{2})$, and θ_t and θ_B denote respectively the antenna down-tilt and beamwidth angles. Therefore, the antenna gain plus path loss will be

$$\zeta_v(r_i) = A_v G(r_i) d_i^{-\alpha_v} = A_v G(r_i) \big(r_i^2 + (h_k - h_{\text{BS}})^2\big)^{-\alpha_v/2},$$

where $d_i \in \{d_{ig}, d_{id}\}$, $v \in \{l, n\}$, and A_l and A_n are the path loss constants at a reference distance $d_i = 1\,\text{m}$ for LoS and NLoS, respectively. For the typical GU, $d_i = d_{ig}$, $h_k = h_g$ and, by NLoS assumption, $v = n$. Note that, since one AU is simultaneously scheduled with $K - 1$ GUs, the K scheduled users encounter independent small-scale fading. Also, for the $K - 1$ GUs, the small-scale fading is i.i.d. Moreover, for the AU, the impact of the channel spatial correlation can be significantly reduced using effective MIMO antenna design techniques, e.g., using antenna arrays whose elements have orthogonal polarizations or patterns [145]. Therefore, for analytical tractability, we ignore such spatial correlation.

Next, we introduce our proposed CB framework to spatially multiplex one AU and $K-1$ GUs. We develop a novel mathematical framework that can be leveraged to characterize the performance of aerial and ground users. This, in turn, allows us to quantify the impact of different system parameters, on the performance of AUs and GUs.

6.3 Content Delivery Analysis

We assume that perfect CSI is available at the tagged BS. Linear precoding in terms of CB creates a $K \times 1$ transmission vector $\boldsymbol{s}'$ for M antennas by multiplying the original data vector $\boldsymbol{s}$ by a precoding matrix $\boldsymbol{W}$: $\boldsymbol{s}' = \boldsymbol{W} \cdot \boldsymbol{s}$, where $[\boldsymbol{W}]_{M \times K}$ consists of the beamforming weights. Let $\boldsymbol{H}$ be the $M \times K$ channel matrix between M antennas of the tagged BS i and its K scheduled users, written as $\boldsymbol{H}_i = \Big[\boldsymbol{h}_{i1} \dots \boldsymbol{h}_{ik} \dots \boldsymbol{h}_{iK}\Big]$, where $\boldsymbol{H}_i \in \mathbb{C}^{M \times K}$, and $\boldsymbol{h}_{ik} \in \mathbb{C}^{M \times 1}$. For CB, tagged BS i creates a precoding matrix $\boldsymbol{W}_i = \Big[\boldsymbol{w}_{i1} \dots \boldsymbol{w}_{ik} \dots \boldsymbol{w}_{iK}\Big]$, with $\boldsymbol{w}_{ik} = \frac{\boldsymbol{h}_{ik}^H}{\|\boldsymbol{h}_{ik}\|}$, where each beam is normalized to ensure equal power assignment to each user [146]. Moreover, let $\boldsymbol{f}_{jk}$ be the channel between interfering BS j and typical user k. Neglecting thermal noise, the received signal at scheduled user k, denoted as y_{ik}, is given by

$$y_{ik} == P(r_i)\boldsymbol{h}_{ik}\boldsymbol{w}_{ik}s_{ik} + \sum_{\kappa \in \mathcal{K}_G} P(r_i)\boldsymbol{h}_{ik}\boldsymbol{w}_{i\kappa}s_{i\kappa} + \sum_{j \in \Phi^o}\sum_{l=1}^{K} P(u_j)\boldsymbol{f}_{jk}\boldsymbol{w}_{jl}s_{jl},$$

where $\Phi^o = \Phi \setminus \{i\}$. The first term in the above equation represents the desired signal from tagged BS i with $P(r_i) = \sqrt{\frac{P_t}{K}}\zeta_v(r_i)^{0.5}$ representing the received power and P_t denoting the BS transmission power. The second and third terms represent the intra- and inter-cell interference, denoted as I_{in} and I_{out}, respectively. The information signal intended for user k is denoted by a complex scalar s_{ik} with unit average power, i.e., $\mathbb{E}[|s_{ik}|^2] = 1$.

Since we assume both LoS and NLoS communications for the AU, with corresponding small-scale fading, we need to distinguish between the two communication paradigms. For the NLoS case, all the K users experience Rayleigh small-scale fading. For LoS communication, however, only the AU experiences Nakagami-m_l small-scale fading, where $m_l > 1$. We hence start by characterizing the gain of intended and interfering channels in Table 6.1.

The second and third columns in Table 6.1 list the marginal channel distributions, i.e., the channel from each single antenna to its intended receiver. We also use interfering BSs to refer to either intra- or inter-cell BS. The first row in Table 6.1 represents the intended channel gain for GUs. It is shown that the equivalent

TABLE 6.1: Channel gains for intended and interfering links.

No	Precoding for channel	Traverse through channel	Seen by	Intended	Channel gain
1	$\mathcal{CN}(0, \frac{\sigma^2}{2})$	$\mathcal{CN}(0, \frac{\sigma^2}{2})$	GU	Yes	$\Gamma(M, \sigma^2)$
2	Nakagami(m_v, η)	Nakagami(m_v, η)	AU	Yes	$\Gamma(m_v M, \frac{\eta}{m_v})$
3	$\mathcal{CN}(0, \frac{\sigma^2}{2})$	$\mathcal{CN}(0, \frac{\sigma^2}{2})$	GU	No	$\Gamma(1, \sigma^2)$
4	$\mathcal{CN}(0, \frac{\sigma^2}{2})$	Nakagami(m_v, η)	AU	No	$\Gamma(1, \eta)$
5	Nakagami(m_v, η)	$\mathcal{CN}(0, \frac{\sigma^2}{2})$	GU	No	$\Gamma(1, \sigma^2)$
6	Nakagami(m_v, η)	Nakagami(m_v, η)	AU	No	$\Gamma(1, \eta)$

channel gain from tagged BS to its associated GU follows $\Gamma(M, \sigma^2)$ [146]. Similarly, the second row represents the intended channel gain for the AU, which is the sum of M independent RVs, each of which follows $\Gamma(m_v, \frac{\eta}{m_v})$. Hence, its intended channel gain follows $\Gamma(m_v M, \frac{\eta}{m_v})$. The third row stands for the interference power caused by transmission of a single beam from an interfering BS to its associated GU when seen by the typical GU, which follows $\Gamma(1, \sigma^2)$ [146]. The fourth (fifth) row describes cases in which a single beam from an interfering BS to its associated GU (AU) is transmitted and seen by the typical AU (GU) (this case is not considered in this chapter as we assume a single AU). Similarly, the sixth row describes cases in which a single beam from an interfering BS to its associated AU is transmitted and seen by the typical AU. Next, we derive the channel gain for the fourth case, whereas the fifth and sixth cases follow in the same way.

Theorem 6.3.0.1. *Under the massive MIMO assumption, whenever a single beam from an interfering BS is received by the typical AU then, the interference channel gain will be given by* $\Gamma(1, \eta)$.

Proof. We write the interfering channel coefficient as

$$h_j = \boldsymbol{w}_{j\kappa} \boldsymbol{f}_{jk} = \frac{\boldsymbol{h}_{j\kappa}^H \boldsymbol{f}_{jk}}{\left\| \boldsymbol{h}_{j\kappa} \right\|} \triangleq \frac{\sum_{o=1}^{M} X_o \times Y_o}{\sqrt{\sum_{q=1}^{M} Z_q}} \tag{6.3}$$

$$\overset{(a)}{\triangleq} \frac{\sum_{o=1}^{M} X_o \times Y_o}{\sqrt{W}} \overset{(b)}{\triangleq} \frac{\sum_{o=1}^{M} X_o \times Y_o}{Q}, \tag{6.4}$$

where $X_o \sim \mathcal{CN}(0, \frac{\sigma^2}{2})$, $Y_o \sim \text{Nakagami}(m_v, \eta)$, $Z_q \sim \exp(\frac{1}{\sigma^2})$, $W \sim \Gamma(M, \frac{1}{\sigma^2})$, and $Q \sim \text{Nakagami}(M, \frac{M}{\sigma^2})$; (a) follows since W is a sum of M i.i.d. exponential RVs, hence it follows $\Gamma(M, \frac{1}{\sigma^2})$. (b) follows since Q equals the square root of the RV

$W \sim \Gamma(M, \frac{1}{\sigma^2})$, hence Q follows $\text{Nakagami}(M, \frac{M}{\sigma^2})$. Denoting the numerator of h_j as z, and writing z as sum of real and imaginary RVs:

$$\text{Re}(z) \triangleq \sum_{o=1}^{M} \Big(\underbrace{X_o \cos(\theta_{ng_o}) - X_o \sin(\theta_{ng_o})}_{\text{RV\#1}} \Big) \cdot \underbrace{Y_o}_{\text{RV\#2}}, \tag{6.5}$$

where, by assumption, $\theta_{ng_o} \sim \mathcal{U}(0, 2\pi)$. We hence have a sum of M i.i.d. RVs, each of which is the product of two independent RVs whose means and variances are $\{\mu_1, \mu_2\}$ and $\{\sigma_1^2, \sigma_2^2\}$, respectively. It can easily be shown that $\mu_1 = 0$ and $\sigma_1^2 = \frac{\sigma^2}{2}$. For large M, using the central limit theorem (CLT), we approximate the PDF of $\text{Re}(z)$ to $\mathcal{N}(\mu_{12}, \sigma_{12}^2)$, whose mean and variance are respectively $\mu_{12} = \mu_1 \mu_2 = 0$, and σ_{12}^2

$$\begin{aligned}
&= \sigma_1^2 \sigma_2^2 + \sigma_1^2 \mu_2^2 + \mu_1^2 \sigma_2^2 \\
&\overset{(a)}{=} \frac{\sigma^2}{2} \eta \Big(1 - \frac{1}{m_v} \Big(\frac{\Gamma(m_v + 0.5)}{\Gamma(m_v)} \Big)^2 \Big) + \frac{\sigma^2}{2} \Big(\frac{\Gamma(m_v + 0.5)}{\Gamma(m_v)} \Big(\frac{\eta}{m_v} \Big)^{0.5} \Big)^2 \\
&= \frac{\sigma^2 \eta}{2} - \frac{\sigma^2 \eta}{2 m_v} \Big(\frac{\Gamma(m_v + 0.5)}{\Gamma(m_v)} \Big)^2 + \frac{\sigma^2 \eta}{2 m_v} \Big(\frac{\Gamma(m_v + 0.5)}{\Gamma(m_v)} \Big)^2 = \frac{\sigma^2 \eta}{2},
\end{aligned} \tag{6.6}$$

where (a) follows from the mean and variance formulas for $\text{Nakagami}(m_v, \eta)$. For the denominator of (6.4), we use the Stirling approximation to approximate the PDF of Q by

$$f_\Omega(\omega, M, M/\sigma^2) = \frac{1}{\omega} \Big(\frac{\omega^2}{\frac{M}{\sigma^2} e^{\frac{\omega^2}{M/\sigma^2} - 1}} \Big)^M. \tag{6.7}$$

The fraction raised to the M-th power is smaller than one, and the integral of f_Ω is one (since it is a PDF). In fact, the factor raised to the M-th power is one only when $\omega = \frac{\sqrt{M}}{\sigma}$. Hence, for large M, from the CLT,

$$\text{Re}(h_j) \sim \frac{\mathcal{N}\left(0, \frac{\sigma^2 \eta}{2M}\right)}{\frac{\sigma}{\sqrt{M}}} \triangleq \mathcal{N}(0, \frac{\eta}{2}). \tag{6.8}$$

Similarly,

$$\text{Im}(h_j) \sim \mathcal{N}(0, \frac{\eta}{2}). \tag{6.9}$$

Hence, the channel gain $|h_j|^2 = \left(\sqrt{\mathrm{Re}\{h_j\}^2 + \mathrm{Im}\{h_j\}^2}\right)^2 \sim \Gamma(1,\eta)$. This completes the proof. □

Next, we derive the SCDP for the AU, which is defined as the probability of obtaining a requested content with SIR higher than a target SIR ϑ. This is an important performance metric that is widely studied in content delivery and caching networks [**chaccour2019reliability**, 3]. The same methodology can be applied to obtain the SCDP for GUs. We next index the AU as $k = 1$. Let $h_{iK} = \sum_{\kappa \in \mathcal{K}_G} |\boldsymbol{w}_{i\kappa}\boldsymbol{f}_{i1}|^2$ denote the intra-cell interference power. From Theorem 6.3.0.1, $|\boldsymbol{w}_{i\kappa}\boldsymbol{f}_{i1}|^2 \sim \Gamma(1,\eta)$. Neglecting the spatial correlation, we have h_{iK} representing sum of $K-1$ Gamma RVs, which yields $h_{iK} \sim \Gamma(K-1,\eta)$. Similarly, the inter-cell interference power $h_{jK} = \sum_{l=1}^{K} |\boldsymbol{w}_{il}\boldsymbol{f}_{j1}|^2 \sim \Gamma(K,\eta)$. Finally, according to the void probability of PPPs [75], the PDF of the horizontal-distance r to the *tagged BS* is $f_R(r) = 2\pi\lambda r e^{-\pi\lambda r^2}$.

Theorem 6.3.0.2. *The unconditional SCDP for the AU is given by*

$$\mathbb{P}_{\mathrm{c}} = \mathbb{P}\big(\mathrm{SIR} > \vartheta\big) = \int_{r=0}^{\infty} \Big[\mathbb{P}^{l}_{\mathrm{c}|r}\mathbb{P}_l(r) + \mathbb{P}^{n}_{\mathrm{c}|r}\mathbb{P}_n(r)\Big] f_R(r)\,\mathrm{d}r\,, \tag{6.10}$$

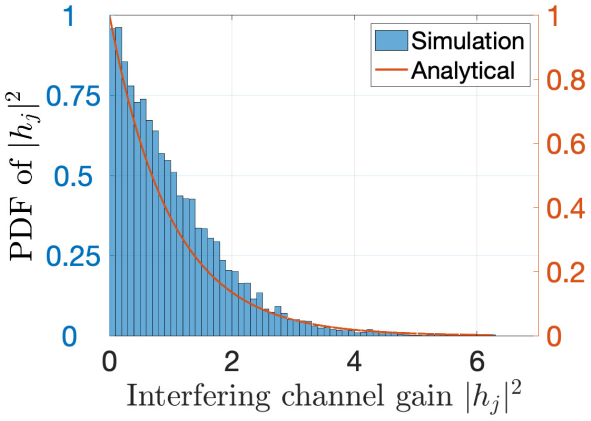
where $\mathbb{P}^{v}_{\mathrm{c}|r} = \|e^{\boldsymbol{T}_{M_v}}\|_1$, $\|.\|_1$ denotes the induced ℓ_1 norm, and $\boldsymbol{T}_{M_v}$ is the lower triangular Toeplitz matrix of size $M_v \times M_v$:

$$\boldsymbol{T}_{M_v} = \begin{bmatrix} t_0 & & & \\ t_1 & t_0 & & \\ \vdots & \vdots & \ddots & \\ t_{M_v-1} & \dots & t_1 & t_0 \end{bmatrix};$$

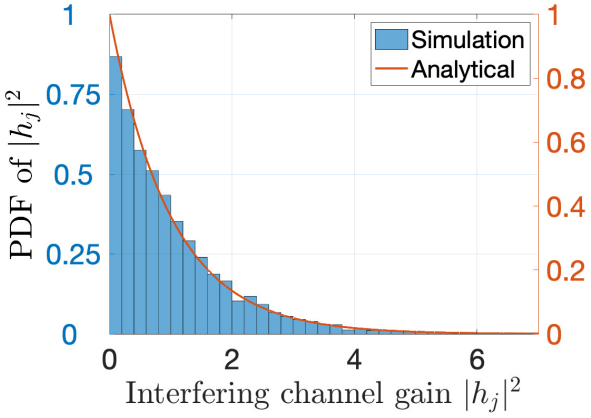
where $M_v = Mm_v$, and the non-zero entries for row i and column j are $t_{i-j} = \frac{(-s_v)^{i-j}}{(i-j)!}\varpi^{(i-j)}(s_v)$; $s_v = \frac{\vartheta K m_v}{\eta P_t \zeta_v(r)}$, $\varpi(s_v) = -(K-1)\log(1+s_v\eta P_v(r)^2) - 2\pi\lambda\int_{\nu=r}^{\infty}\big(1-\mathbb{P}_l(\nu)\delta_l(\nu,s_v) - \mathbb{P}_n(\nu)\delta_n(\nu,s_v)\big)\nu\,\mathrm{d}\nu$, and $\varpi^{(k)(s_v)} = \frac{d^k}{ds_v^k}\varpi(s_v)$; $\delta_l(\nu,s_v) = \big(1+s_v\eta P_l(\nu)^2\big)^{-K}$, and $\delta_n(\nu,s_v) = \big(1+s_v\eta P_n(\nu)^2\big)^{-K}$.

Proof. Please see Appendix D.1 □

Remark 6.3.0.1. The main merit of this representation, i.e., $\mathbb{P}^{v}_{\mathrm{c}|r} = \|e^{\boldsymbol{T}_{M_v}}\|_1$, is that it leads to valuable system insights. For example, the impact of the shape parameter



(a) Number of antennas $M = 4$



(b) Number of antennas $M = 32$

FIGURE 6.1: PDF of the interfering channel power.

$M_v = Mm_v$ on the intended channel gain $h_{iK} \sim \Gamma(Mm_v, \eta/m_v)$, which is typically related to the antenna size and the Nakagami fading parameter m_v, is clearly illustrated by the finite sum representation in the above. Although it is not tractable to obtain closed-form expressions for t_k (the entries populating $\boldsymbol{T}_{M_v}$), special cases of interest, e.g., LoS or NLoS communications, can lead to closed-form expressions, following [147].

Remark 6.3.0.2. When $K = 1$, only the AU is scheduled, i.e., maximal ratio transmission (MRT) beamforming. For MRT, the interfering channel gain is $\Gamma(1, \eta)$. Interestingly, this interfering channel gain is reduced as opposed to the typical Nakagami channel gain $\Gamma\left(m_l, \frac{\eta}{m_l}\right)$ when there is neither precoding nor MIMO transmission.

6.4 Numerical Results

For our simulations, we consider a network having the following parameters, unless otherwise specified. The number of antennas per sector is set to $M = 32$, which is a reasonable value for massive MIMO-enabled 5G BSs (minimum number of antennas is four for 5G). We also set $K = 4$, $\lambda = 1\,\mathrm{km}^{-2}$, $h_{\mathrm{BS}} = 55\,\mathrm{m}$, $h_g = 1\,\mathrm{m}$, $\alpha_l = 2.09$, $\alpha_n = 3.75, a = 0.6, \nu = 500\mathrm{km}^{-2}, c = 25, \vartheta = 10\,\mathrm{dB}, A_l = -41.1\,\mathrm{dB}, A_n = -32.9\,\mathrm{dB}, G_m = 10\,\mathrm{dB}, G_s = -3.01\,\mathrm{dB}, m_n = 1, m_l = 3, \eta = 1, \sigma^2 = 1$, $\theta_B = 45^\circ$, $\theta_t = 30^\circ$. These aforementioned values of the density of buildings (ν) and average buildings' heights (c) represent common values for urban environments. It is also worth highlighting that the results of this chapter are based on a simple antenna model where the antenna directivity gain has two distinct values, namely, main-lobe gain and side-lobe gain. This assumption is adopted for tractability. However, in the next chapter, we show that accounting for practical antenna models will yield a degraded performance as opposed to the simple antenna model. Hence, the results obtains in this chapter, e.g., the SCDP, represent an upper bound on the actual values.

In Fig. 6.1, we verify the accuracy of the obtained PDF of interfering channel gain $|h_j|^2$ (Table 6.1: row 4) in Theorem 6.3.0.1. Monte-carlo simulation is adopted where the overall channel gain is obtained from simulating links to ground and aerial users and accordingly computing and applying the precoding vectors. The overall channel gain histogram is then plotted and compared with the one from the formulated analytical expression. The figure shows that the derived PDF is quite accurate when M is sufficiently large as in Fig. 6.1(b), while for small M in Fig. 6.1(a), it still provides a reasonable approximation.

Next, we compare the SCDP of AUs with and without MIMO beamforming to GUs. Fig. 6.2 plots the SCDP as a function of the SIR threshold ϑ for the AU and the GUs.[1] For Fig. 6.2, the AU altitude is set to $90\,\mathrm{m}$, which is higher than the

[1]It is worth mentioning that, without lack of practicality, we choose the UAV-UE altitudes to reflect the fluctuations of the UAV-UE coverage probability with the antennas' down-tilt angle of the ground BSs. Generally speaking, the lowest possible altitude to consider for the UAV-UEs is zero altitude, i.e., at the ground level, which is the case for the UAV-UE during the take off and landing. In addition, the highest possible altitude might be limited by the surrounding environments and can be regulated so as not to interfere with other aircraft in the sky.

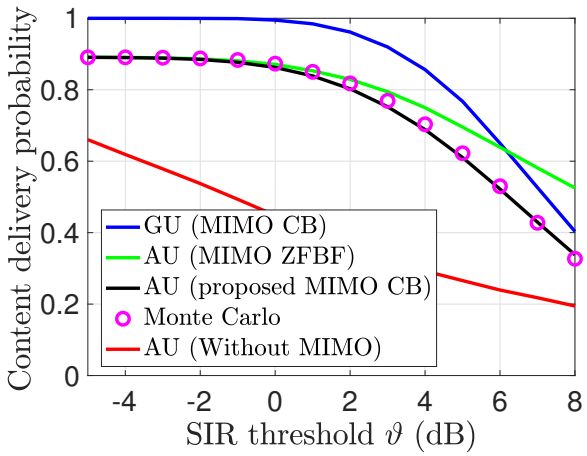


FIGURE 6.2: Effect of SIR threshold ($h_{\mathrm{BS}} = 30\,\mathrm{m}$, AU altitude $h_d = 90\,\mathrm{m}$)

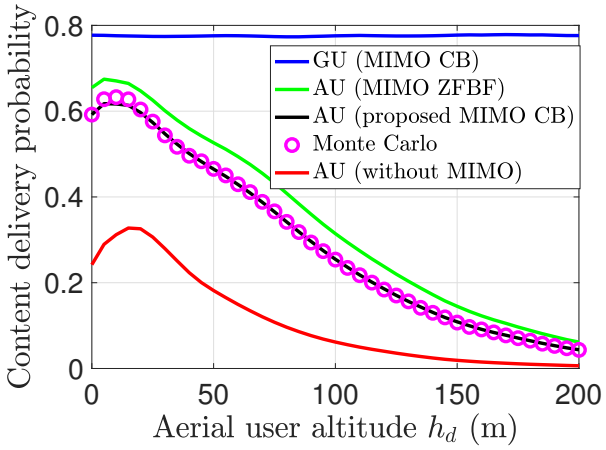


FIGURE 6.3: Effect of AU altitude ($h_{\mathrm{BS}} = 30\,\mathrm{m}$, $\vartheta = 5\,\mathrm{dB}$, $\lambda = 50\,\mathrm{km}^{-2}$)

heights of the ground BSs. Clearly, the achievable performance of GUs considerably outperforms that of an AU. This is because GUs have a superior propagation environment, driven by the down-tilted BS antennas in the desired signal side, and the NLoS interfering links. However, Fig. 6.2 also shows that the SCDP for the AU served by MIMO CB significantly outperforms that of the AU served by single-antenna GBSs. Moreover, although the ZFBF technique outperforms our proposed CB approach, the low complexity of CB and its associated performance gain over traditional single-antenna GBSs make it a good candidate to serve AUs. Fig. 6.3 shows the effect of AU altitude on the AU performance, with that of GUs plotted for comparison. Fig. 6.3 shows that the AU SCDP (for all transmission schemes) gradually increases with h_d up to a maximum value due to the larger LoS probability, before it decreases again due to the stronger LoS interference and higher large scale fading.

Next, we show the effect of the down-tilt angle θ_t on the performance of both the AU and the GUs, for different AU altitudes. The AU altitude is assumed between $0\,\mathrm{m}$ and $200\,\mathrm{m}$ so that the impact of the down-tilt angle on the performance of the AUs can be closely investigated. Recall that h_d and h_{BS} are the drone altitude and BS height, respectively. It is worth noting that the LoS probability from ground BSs to the AUs and the small fading parameters change with the altitude of the AUs. Such changes of the radio propagation parameters are effectively factored in from (6.2) where the LoS probability is calculated based on the current AU altitude and, accordingly, the small fading parameter is identified from (6.1). As illustrated in Fig. 6.4, for $h_d < h_{\mathrm{BS}}$, the performance of the AU is maximized at certain θ_t, and beyond that it starts to degrade. However, for GUs, their performance is maximized at a higher θ_t. Hence, adjusting the antennas' down-tilt angle yields a tradeoff between the performance of AUs and GUs owing to the difference in their altitudes. For $h_d > h_{\mathrm{BS}}$ in Fig. 6.5, the SCDP of the AU first decreases with θ_t to a minimum value, and then it increases again. This finding can be explained as follows: when θ_t is small, an AU at an altitude $h_d > h_{\mathrm{BS}}$ can be served from the main lobe of tagged BS while also experiencing high interference from the main lobe of other interfering BSs. Gradually, as θ_t increases, the worst performance is observed when the AU is no longer served from the main lobe of tagged BS antennas while

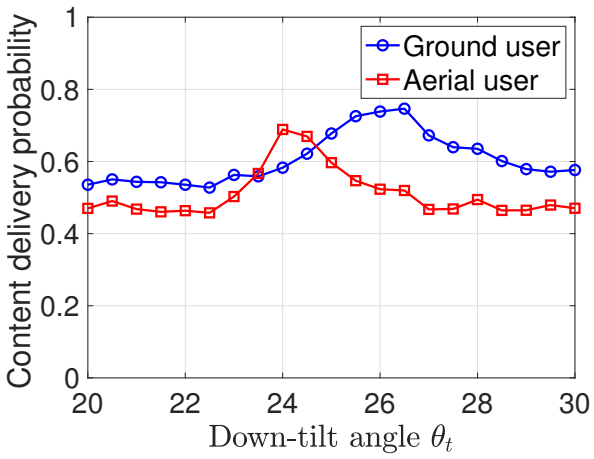


FIGURE 6.4: Effect of antenna down-tilt angle: AU altitude $h_d = 30\,\mathrm{m}$

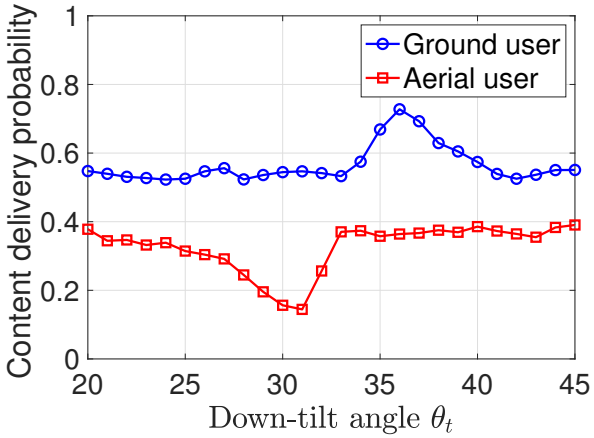


FIGURE 6.5: Effect of antenna down-tilt angle: AU altitude $h_d = 80\,\mathrm{m}$

still experiencing high interference from the main lobe of other BSs. Finally, for very large θ_t, both intended and interfering signals stem from the side-lobes, and hence the performance is improved again.

Next, we show the prominent effect of the number of scheduled users K and the number of antennas M on the network performance. Fig. 6.6 shows that the SCDP monotonically decreases for both AU and GU as K increases due to the effect of stronger interference. However, it is noticeable that increasing K highly degrades the AU performance compared to that of GUs. This stems from the fact that AUs are more sensitive to interference, which often exhibits LoS component.

In Fig. 6.7, we show the system spectral efficiency (SE) versus the number of scheduled users K. In this figure, $K = 1$ means that only the AU is scheduled.

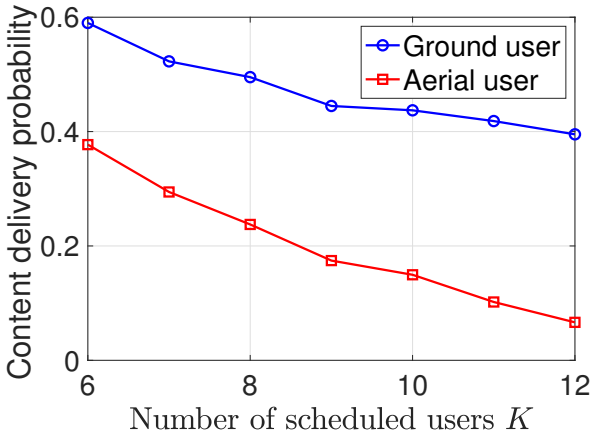


FIGURE 6.6: Effect of the number of scheduled users: number of antennas $M = 32$

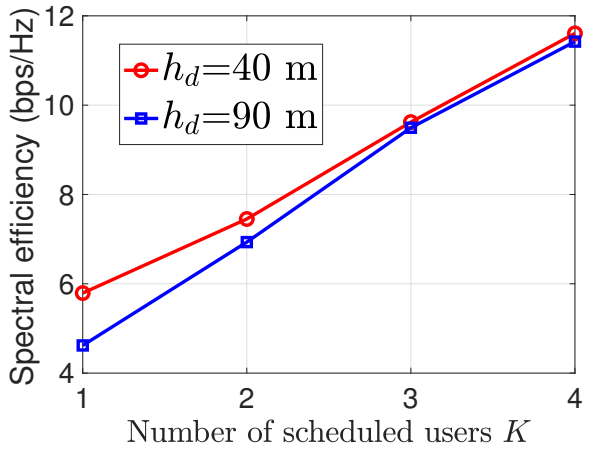


FIGURE 6.7: Effect of the number of scheduled users.: number of antennas $M = 32$

We show the system spectral efficiency for two different hovering altitudes of the AUs, namely, $40\,\mathrm{m}$ and $90\,\mathrm{m}$. Evidently, the overall system SE is shown to improve as K increases, which proves that spatially multiplexing one AU with the GUs significantly improves the system SE. This is driven by the concurrent transmission of the data of multiple users (i.e., one AU and $K-1$ GUs) over the same physical radio channel.

Fig. 6.8 shows that increasing the number of antennas M improves the SCDP for both users with nearly the same rate. This is essentially due to the enhanced transmission diversity when serving the aerial and ground users' from large number of antenna. This behavior shows that serving AUs from MIMO-enabled ground BSs will be an appealing approach to improve their cellular connectivity.

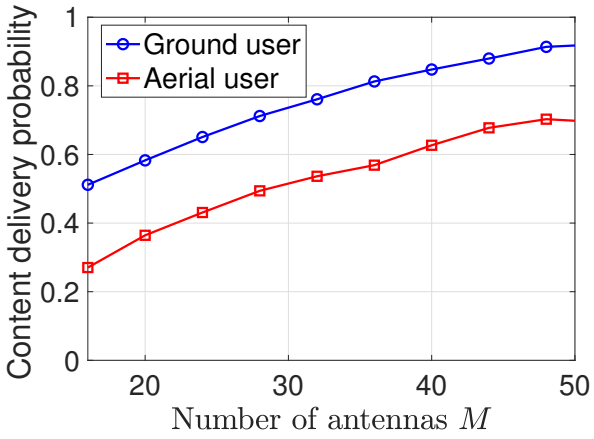


FIGURE 6.8: Effect of the number of antennas: number of users $K = 4$

6.5 Conclusion

In this chapter, we have proposed a novel CB framework for spatially multiplexing AUs and GUs. In order to analytically express the SCDP, we have derived the gain of intended and interfering channels. We have shown that exploiting CB from massive MIMO-enabled BSs to spatially multiplex an AU and GUs substantially improves the performance of the AU, in terms of SCDP. We have then shown that the down-tilt of the BS antennas leads to a tradeoff between the performance of AUs and GUs if the AU's altitude is below the BS height. Simulation results have shown the various properties of cellular communications when AUs and GUs co-exist, namely, the effect of the down-tilt angle on the achievable performance and the yielded serving and interfering channel gains when digital precoding is applied for different types of fading channels (i.e., Gaussian fading for GUs and Nakagami fading for AUs).

Chapter 7

Performance Analysis of Mobile Cellular-Connected Drones Under Practical Antenna Configurations

Providing seamless connectivity and content delivery to unmanned aerial vehicle user equipment (UAV-UE) is very challenging due to the encountered line-of-sight interference and reduced gains of down-tilted BS antennas. For instance, as the altitude of UAV-UEs increases, their cell association and handover procedure become driven by the side-lobes of the BS antennas. In this chapter, the performance of cellular-connected UAV-UEs is studied under 3D practical antenna configurations. Two scenarios are studied: scenarios with static, hovering UAV-UEs and scenarios with mobile UAV-UEs. For both scenarios, the UAV-UE coverage probability is characterized as a function of the system parameters, namely, the number of antenna elements, density of BSs, and speed of UAV-UEs. The effects of the number of antenna elements on the UAV-UE coverage probability and handover rate of mobile UAV-UEs are then investigated. Results reveal that the UAV-UE coverage probability under a practical antenna pattern is worse than that under a simple antenna model. Moreover, vertically-mobile UAV-UEs are susceptible to *altitude handover* due to consecutive crossings of the nulls and peaks of the antenna side-lobes.

7.1 Introduction

A tremendous increase in the use of UAVs (drones) in a wide range of applications, ranging from aerial surveillance and safety to product delivery, is anticipated in the foreseeable future [10, 13, 31, 139]. In such applications, UAVs need to communicate with each other as well as with ground UEs using wireless cellular connectivity. Cellular-connected UAVs have attracted attention in cellular network research in both industry and academia due to their ability to ubiquitously communicate [139–141]. However, the BS antennas of current cellular networks are tilted downwards to provide connectivity to ground UEs rather flying UAV-UEs [139]. Hence, UAV-UEs have to be served from the side-lobes of the BS antennas. Moreover, the UAV-UEs, especially at high altitudes, are dominated by LoS communication links. These key characteristics of the UAV-UE communications pose new technical challenges on their cell association, handover procedure, and the overall achievable performance [14].

7.2 Motivation and Contribution

The main contribution of this chapter is a rigorous analysis that provides an in-depth understanding of the performance of UAV-UEs under practical antenna configurations. In particular, we consider a network of ground BSs equipped with more than two antenna array elements to provide cellular connectivity to UAV-UEs. For this network, we characterize the coverage probability for two scenarios, specifically, static and mobile UAV-UEs. Moreover, we investigate the handover rate of mobile UAV-UEs in order to provide important design guidelines and understand novel handover aspects of UAV-UEs, such as the altitude handover. Our results show that the number of antenna elements controls the UAV-UE handover rate, while its impact on the coverage probability is marginal if the handover cost is low (i.e., when the handover does not yield excessive handover failure and dropped connections).

7.3 Related Works

In [148], the authors studied the feasibility of supporting drone operations using existing cellular infrastructure. It is shown that, under a simple antenna model, the cell association heavily depends on the availability of LoS links to the serving BS. In [143], similar results are verified for air-to-ground communication between UAV-BSs and ground UEs. Moreover, in [14], we showed that coordinated transmissions can effectively mitigate the effects of LoS interference of high-altitude UAV-UEs. While the works in [14, 143, 148] considered the possibility of LoS communication, they assumed a simple antenna configuration that is modeled as a step function of two gain values, namely, main- and side-lobe gains. However, practical antenna patterns resemble a sequence of main- and side-lobes with nulls between consecutive lobes. Such a 3D antenna pattern plays a crucial rule on the UAV-UE cell association and handover procedure, which is ignored in most of the prior works [14, 143, 148]. Performance analysis of UAV-UEs under 3D practical antenna models was done in the recent works [142, 149–151]. For instance, based on system-level simulations, the authors in [142] showed that the cell association of UAV-UEs is mainly dependent on the side-lobes. Moreover, in [149], the authors characterized the association probability and SIR under nearest-distance and maximum-power based associations.

The performance of mobile UAV-UEs under practical antenna patterns has been studied in recent works [150] and [151]. For instance, based on system-level simulation, the authors in [150] showed that, due to the LoS propagation conditions to many interfering cells, it is difficult for the UAV-UEs to establish and maintain connections to the network, which also leads to increased handover failure rates. Moreover, based on experimental trials in [151], the authors showed that drones are subject to frequent handovers once the typical flying altitude is reached. However, the results presented in these works are based on simulations and measurements. Also, although the work in [149] has characterized the signal-to-interference-plus-noise ratio (SINR) at UAV-UEs served from 3D antennas, there was no characterization of important performance metrics such as the coverage probability. Moreover, the work in [149] only considered a static UAV-UE scenario. As a first step

in the direction of understanding the behavior of mobile drones in practical network setups, this chapter aims to characterize the performance of static and mobile UAV-UEs under practical antenna configurations.[1]

7.4 System Model

7.4.1 Network Model

We consider a downlink transmission scenario from a terrestrial cellular network to cellular-connected UAV-UEs. We assume that ground BSs are distributed according to a two-dimensional (2D) homogeneous PPP $\Phi_b = \{b_i \in \mathbb{R}^2, \forall i \in \mathbb{N}\}$ with intensity λ_b. All BSs have the same transmit power P_t and are deployed at the same height h_{BS}. We consider a number of UAV-UEs that can be either static or mobile based on the application. Static UAV-UEs hover at a fixed altitude h_d, while mobile ones can make up and down movements within two altitude thresholds h_1, and h_2. We set $h_d = \frac{h_1+h_2}{2}$, i.e., h_d is the mean flying altitude of mobile UAV-UEs. As shown in Fig. 7.1, we consider high-altitude UAV-UEs where h_1, h_d, and h_2 are above the BS height h_{BS}. Each UAV-UE has a single antenna and receives downlink signals from a ground BS. Each ground BS is equipped with a directional antenna array composed of N_t vertically-placed elements. Two association schemes are considered for the UAV-UEs: Nearest and highest average received power (HARP) associations.

7.4.2 Channel Model

We consider a wireless channel that is characterized by both small-scale and large-scale fading. For the large-scale fading, the channel between a ground BS and an UAV-UE is described by the LoS and NLoS components, which are considered separately along with their probabilities of occurrence [139]. This assumption is apropos for such ground-to-air channels that often exhibit LoS communication [139] and [143]. For small-scale fading, we adopt a Nakagami-m_v model as done in [139] for the channel gain, whose PDF is given by: $f(\omega) = \frac{2m_v^{m_v}\omega^{2m_v-1}}{\Gamma(m_v)}e^{-m_v\omega^2}$. The fading

[1]While in this chapter we focus on the performance of vertically mobile UAV-UEs under practical antenna patterns, in Chapters 6 and 8, we assume simpler antenna models for tractability. In addition, in contrast to Chapters 6 and 8 where we aimed at improving the performance of UAV-UEs, the concept of altitude handover introduced in this chapter heavily relies on the underlying antenna patterns.

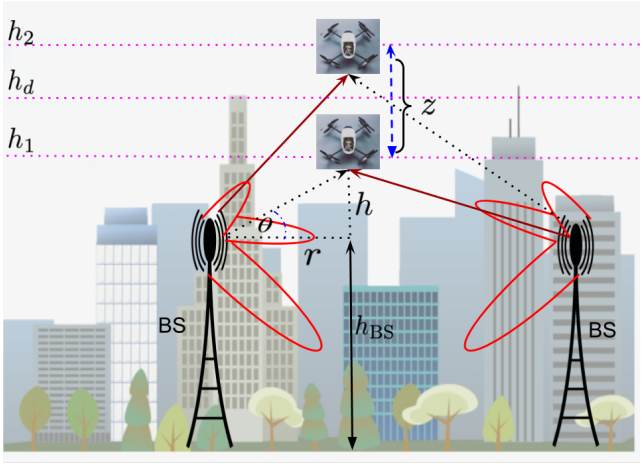


FIGURE 7.1: Illustration of the proposed system model in which 3D antenna-equipped ground BSs serve high-altitude static (or mobile) UAV-UEs. Here, h_1, h_2, and h_d refer to the minimum altitude, maximum altitude, and average altitude of a mobile UAV-UE, with h being the altitude difference $h_2 - h_1$. In addition, θ is the elevation angle and r is the horizontal distance between the UAV-UE and its serving ground BS.

parameter m_v is assumed to be an integer for tractability, where $v \in \{l, n\}$ accounts for LoS and NLoS communications, respectively. Given that $\omega \sim$ Nakagami(m_v), it directly follows that the channel gain power $\chi = \omega^2 \sim \Gamma(m_v, 1/m_v)$, which represents a Gamma RV whose shape and scale parameters are m_v and $\frac{1}{m_v}$, respectively. Hence, the PDF of channel power gain distribution will be $f(\chi) = \frac{m_v^{m_v}\chi^{m-1}}{\Gamma(m_v)}\exp\big(-m_v\chi\big)$.

Similar to Chapter 6, we assume a 3D blockage model that is characterized by the mean number of η buildings/km^2, the proportion a of the total land area occupied by buildings, and the height of buildings (modeled by a Rayleigh PDF with a scale parameter c). The probability of having a LoS communication from a BS at horizontal-distance r from an UAV-UE is hence given, similar to [148] and [143], as

$$\mathbb{P}_l(r) = \prod_{n=0}^{\max(0,o-1)} \left[1 - \exp\Big(-\frac{\big(h_{\text{BS}} + \frac{h(n+0.5)}{o}\big)^2}{2c^2}\Big)\right], \tag{7.1}$$

where h is the difference between the UAV-UE altitude and BS height and $o = \lfloor r\sqrt{a\eta}\rfloor$.

7.4.3 Antenna Model

The ground BSs are equipped with directional antennas of fixed radiation patterns and with a down-tilt angle φ. This is typically achieved by equipping the BS with a uniform linear array (ULA) of N_t vertically-placed elements, that are assumed to be omni-directional along the horizontal dimension [152]. Along the vertical dimension, the power radiation pattern is equal to the array factor times the radiation pattern of a single antenna. The N_t antenna elements are equally spaced with adjacent elements separated by half of the wavelength. With the down-tilt angle φ, the overall array gain in the direction θ is given by [149]:

$$G(\theta) = \underbrace{\frac{1}{N_t} \frac{\sin^2 \frac{N_t \pi}{2} \big(\sin(\theta) - \sin(\varphi)\big)}{\sin^2 \frac{\pi}{2} \big(\sin(\theta) - \sin(\varphi)\big)}}_{A_f(\theta)} \times \underbrace{10^{\min(-1.2(\frac{\theta}{\delta})^2, \frac{G_m}{10})}}_{G_e(\theta)},$$

where G_m gives the threshold for antenna nulls, $A_f(\theta)$ is the array factor of the ULA, $G_e(\theta)$ is the element power gain of each antenna along the vertical dimension, i.e., the elevation angle θ, and δ is the half power beamwidth. For simplicity, we assume that the gain $G_e(\theta)$ is a positive constant on the range of the elevation angle $\theta \in [0, \frac{\pi}{2}]$, and 0 otherwise (i.e., zero front-to-back power ratio as in [153]).[2] From Fig. 7.1, $\theta = \arctan(\frac{h}{r})$, where r is a realization of the RV R which represents the horizontal distance between a ground BS and the projection of an UAV-UE. If $G_e(\theta)$ is set to one, and for a zero down-tilt angle, the antenna array gain is simplified to:

$$G(r, h) = \frac{1}{N_t} \frac{\sin^2 \frac{N_t \pi}{2} \big(\sin(\arctan(\frac{h}{r}))\big)}{\sin^2 \frac{\pi}{2} \big(\sin(\arctan(\frac{h}{r}))\big)}. \tag{7.2}$$

Hence, the antenna gain plus path loss for the LoS and NLoS components will be: $\zeta_v(r) = A_v G(r, h) \big(r^2 + h^2\big)^{-\alpha_v/2}$, where $v \in \{l, n\}$, α_l and α_n are the path loss exponents, and A_l and A_n are the path loss constants at $\sqrt{r^2 + h^2} = 1\,\mathrm{m}$ for the LoS and NLoS, respectively.

Having defined our system model, next, we will study the performance of UAV-UEs for two scenarios: static and mobile UAV-UEs. For each scenario, we investigate the UAV-UE coverage probability under two association schemes, namely,

[2] Note that the elevation angle is bounded as $\theta \in [0, \frac{\pi}{2}]$ from the assumption of high-altitude UAV-UEs, i.e., $h_1, h_d, h_2 > h_{\mathrm{BS}}$.

nearest and HARP associations. The coverage probability is defined as the probability that the received SIR is higher than a target threshold ϑ.

7.5 Coverage Probability of Static UAV-UEs

We assume static UAV-UEs hovering at a fixed altitude h_d, hence, we have $h = h_d - h_{\mathrm{BS}}$ in (7.1). Given that a PPP is translation-invariant with respect to the origin, we conduct the coverage analysis for a UAV-UE located at the origin in $\mathbb{R}^2$, called the typical UAV-UE [75].

7.5.1 Nearest Association

To simplify the analysis, we only consider probabilistic LoS/NLoS links for interfering BSs, while the serving BS has a dominant LoS component. This is because, at high altitude, UAV-UEs will have an LoS-dominated channel toward nearby, serving BSs. However, UAV-UE interference at far-away BSs will not be LoS dominated. Moreover, the study of LoS-based association is considered in prior works, e.g., [143], while we are particularly focused on the impact of practical antennas.

We denote the horizontal distance from the typical UAV-UE to its ground BS by r_0. By the PPP assumption, $f_{R_0}(r_0) = 2\pi r_0 \lambda e^{-\pi\lambda r_0^2}$. Conditioning on $R_0 = r_0$, and neglecting the noise, the received SIR at the typical UAV-UE will be:

$$\Upsilon_{|r_0} = \frac{\chi_0 \zeta_l(r_0)}{I}, \tag{7.3}$$

where I is the aggregate interference, χ_0 is the Nakagami-m_l fading power, and $\zeta_l(r_0)$ represents the antenna gain plus path loss from the serving BS. The serving and interfering signals in (7.3) are normalized to the transmit power P_t. The UAV-UE coverage probability is characterized in the next Theorem.

Theorem 7.5.1.1. *The static UAV-UE coverage probability under nearest association is given by:*

$$\mathbb{P}_c = \int_{r_0=0}^{\infty} \mathbb{P}_{c|r_0} f_{R_0}(r_0)\, \mathrm{d}r_0\,, \tag{7.4}$$

where $\mathbb{P}_{c|r_0}$ is the UAV-UE conditional coverage probability:

$$\mathbb{P}_{c|r_0} = \sum_{i=0}^{m_l-1} \frac{(-\varpi_l)^i}{i!} \sum_{i_n+i_l=i} \frac{i!}{i_n!i_l!} \frac{\varpi_l^{i_n}}{i_n!} \frac{\partial^{i_n}}{\partial \varpi_l^{i_n}} \mathcal{L}_{I_n}(\varpi_l) \frac{\varpi_l^{i_l}}{i_l!} \frac{\partial^{i_l}}{\partial \varpi_l^{i_l}} \mathcal{L}_{I_l}(\varpi_l), \tag{7.5}$$

and $\varpi_l = \frac{\vartheta d_0^{\alpha_l} m_l}{A_l G(r_0,h)}$, $d_0 = \sqrt{h^2+r_0^2}$; $\mathcal{L}_{I_l}$ and $\mathcal{L}_{I_n}$ are the Laplace transforms of the LoS and NLoS interference, respectively. The Laplace transform of LoS interference is then given by

$$\mathcal{L}_{I_l}(\varpi_l) = e^{-\gamma(\varpi_l)}, \tag{7.6}$$

where

$$\gamma(\varpi_l) = 2\pi\lambda_b \sum_{j=\lfloor r_0\sqrt{a\eta}\rfloor}^{\infty} \mathbb{P}_l(\frac{j}{\sqrt{a\eta}}) \int_{\max(r_0,\frac{j}{\sqrt{a\eta}})}^{\frac{j+1}{\sqrt{a\eta}}} \Big(1 - (\frac{m_l}{m_l + \varpi_l \zeta_l(r)})^{m_l}\Big) r \, \mathrm{d}r \,,$$

and the Laplace transform of NLoS interference is calculated in a same manner.

Proof. The proof proceeds similar to [13, Theorem 1] and [143, Proposition 4]. □

Since it is hard to directly obtain insights from (7.4) on the effect of the practical antenna gain, several results based on (7.4) will be shown in Section 7.7 to provide key design guidelines. Moreover, we show that the coverage probability does not scale with N_t, which implies that increasing the number of antenna elements has a marginal effect on the coverage probability.

Corollary 7.5.1.1. *The coverage probability of static UAV-UEs does not scale asymptotically with the number of antenna elements.*

Proof. Please see Appendix E.1. □

7.5.2 Highest Average Received Power Association

Each UAV-UE is associated to the BS that delivers the highest average power, i.e., the effect of small scale fading is averaged. Hence, the received signal power depends on the path loss $d^{-\alpha_v}$ and antenna gain $G(r,h)$, which is, in turn, determined by the elevation angle θ. In Fig. 7.2, we plot the path loss, antenna gain, and the overall link gain, i.e., the path loss times antenna gain, versus the horizontal distance r. From (7.2), the antenna gain consists of $\frac{N_t}{2}$ lobes whose peaks increase with

the horizontal distance r. Fig. 7.2 also shows that the overall link gain (dashed line) consists of $\frac{N_t}{2}$ lobes whose maximum gains decrease as r increases.

It is worth highlighting that obtaining the serving distance PDF is not possible given the non-convexity of the locations of BSs that deliver the highest power along r (see Fig. 7.2). We hence propose a *novel geometry-based approximation* that leverages the relative symmetrical characteristics of the overall link gain. The key idea behind this approximation is to reproduce an equivalent PPP deployment of the BS locations in which the distance PDF to the BS delivering the highest average power (within each lobe) can be obtained. With this in mind, we can then obtain an expression approximating the UAV-UE coverage probability under HARP association. We particularly divide the space $r > 0$ into $\frac{N_t}{2}$ regions, each of which corresponds to the boundary of one lobe. Then, for each lobe, we consider the zone from its peak to the next null, assuming doubled density of BSs. For instance, in Fig. 7.2, BSs of density $\lambda = 2\lambda_b$ only exists within r_{pj} to r_{nj}, $j \in \{1, \ldots, \frac{N_t}{2}\}$. UAV-UEs then associate to the nearest BS from the reproduced deployment, which delivers the highest average power within its lobe. However, since this BS might not be the one that delivers the highest average power among all the BSs, the adopted method is an approximation.

From (7.2), there are $\frac{N_t}{2}$ nulls between the lobes that occur at elevation angles θ_{nj}, where $\frac{N\pi}{2}\sin(\theta_{nj}) = j\pi$. Hence, $\theta_{nj} = \arcsin(\frac{2j}{N})$ and the equivalent horizontal distances at these nulls will be $r_{nj} = h/\tan\big(\arcsin(\frac{2j}{N})\big)$. For instance, $\theta_{n0} = 0^\circ$ and $r_{n4} = \frac{h}{\tan(0)} = \infty$. To obtain the locations of peaks r_{pj}, we define the overall link gain as:

$$
\begin{aligned}
L(\theta) &= G(\theta) \times d^{-\alpha_v} \\
L(\theta) &= \frac{1}{N_t} \frac{\sin^2 \frac{N_t \pi}{2}\big(\sin(\theta)\big)}{\sin^2 \frac{\pi}{2}\big(\sin(\theta)\big)} \times \big(r^2 + h^2\big)^{-\alpha_v/2} \\
&\overset{(a)}{=} \frac{1}{N_t h^{\alpha_v}} \frac{\sin^2 \frac{N_t \pi}{2}\big(\sin(\theta)\big)}{\sin^2 \frac{\pi}{2}\big(\sin(\theta)\big)\big(\tan^{-2}(\theta) + 1\big)^{\alpha_v/2}},
\end{aligned}
$$

where (a) follows from $\tan(\theta) = \frac{h}{r}$. We find elevation angles θ_{pj} at which the peaks occur by taking the first derivative of $L(\theta)$ and setting it to zero. The first derivative

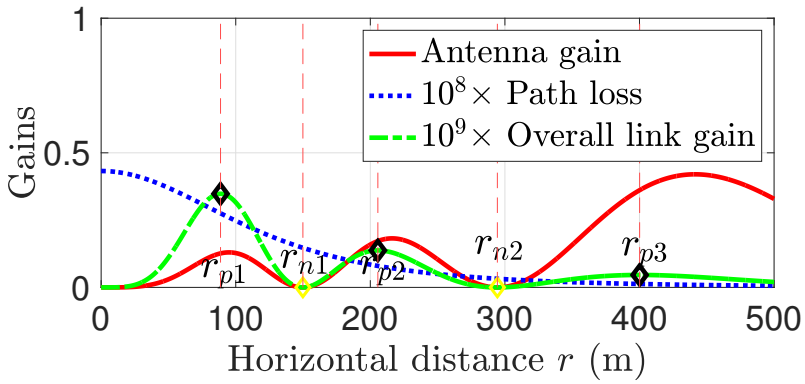


FIGURE 7.2: Illustration of the geometry-based approximation to calculate the UAV-UE coverage probability under HARP association.

is obtained as:

$$\frac{\partial L(\theta)}{\partial \theta} = (N_t-1)\sin(\frac{\pi(N_t+1)}{2}\sin(\theta)) + (N_t+1)\sin(\frac{\pi(1-N_t)}{2}\sin(\theta)) + \frac{\alpha_v\big(\cos(\frac{\pi(N_t-1)}{2}\sin(\theta)) - \cos(\frac{\pi(N_t+1)}{2}\sin(\theta)\big)}{\pi\cos(\theta)\tan(\theta)\big(1+\tan^2(\theta)\big)} = 0.$$

The roots of this equation are θ_{pj}, and equivalent distances are $r_{pj} = \frac{h}{\tan(\theta_{pj})}$. The density of BSs within $[r_{jp}, r_{jn}]$ is $\lambda = 2\lambda_b$, and zero otherwise. From the PPP definition, the probability that at least one BS exists within $[r_{pj}, r_{nj}]$ in the reproduced deployment is $1-e^{-\pi\lambda(r_{nj}^2-r_{pj}^2)}$. Hence, the piece-wise serving distance CDF will be

$$F_{R_m}(r_m) = \begin{cases} \Big(1-e^{-\pi\lambda(r_m^2-r_{p1}^2)}\Big)\Big(1-e^{-\pi\lambda(r_{n1}^2-r_{p1}^2)}\Big), & j=1 \\ \Big(1-e^{-\pi\lambda(r_m^2-r_{pj}^2)}\Big)\Big(1-e^{-\pi\lambda(r_{nj}^2-r_{pj}^2)}\Big)\times & \\ \prod_{i=1}^{j-1} e^{-\pi\lambda(r_{ni}^2-r_{pi}^2)}, & 1<j\leq\frac{N_t}{2} \\ 0, & \text{otherwise} \end{cases}$$

where the product term represents the probability that no BSs exist in the previous lobes. Hence, the serving distance PDF will be

$$f_{R_m}(r_m) = \frac{\partial F_{R_m}(r_m)}{\partial r_m} = \begin{cases} 2\pi\lambda r_m e^{-\pi\lambda(r_m^2-r_{p1}^2)}\Big(1-e^{-\pi\lambda(r_{n1}^2-r_{p1}^2)}\Big), & j=1 \\ 2\pi\lambda r_m e^{-\pi\lambda(r_m^2-r_{pj}^2)}\Big(1-e^{-\pi\lambda(r_{nj}^2-r_{pj}^2)}\Big)\times & \\ \prod_{i=1}^{j-1} e^{-\pi\lambda(r_{ni}^2-r_{pi}^2)}, & 1<j\leq\frac{N_t}{2} \\ 0, & \text{otherwise.} \end{cases}$$

It can be easily verified that $\int_{r_m=0}^{\infty} f_{R_m}(r_m) = 1$. Given $f_{R_m}(r_m)$, the approximated coverage probability under HARP association is characterized in the next corollary.

Corollary 7.5.2.1. *The coverage probability of static UAV-UEs under HARP association is approximated as:*

$$\mathbb{P}_c = \int_{r_m=0}^{\infty} \mathbb{P}_{c|r_m} f_{R_m}(r_m)\, \mathrm{d}r_m\,, \tag{7.7}$$

where $\mathbb{P}_{c|r_m}$ is calculated similar to $\mathbb{P}_{c|r_0}$ in Theorem 7.5.1.1, with $\varpi_l = \frac{\vartheta d_m^{\alpha_l} m_l}{A_l G(r_m,h)}$, $d_m = \sqrt{h^2 + r_m^2}$, and $\gamma(\varpi_l) = 2\pi\lambda_b \times \sum_{j=0}^{\infty} \mathbb{P}_l(\frac{j}{\sqrt{a\eta}}) \int_{\frac{j}{\sqrt{a\eta}}}^{\frac{j+1}{\sqrt{a\eta}}} \Big(1 - \big(\frac{m_l}{m_l + \varpi_l \zeta_l(r)}\big)^{m_l}\Big) r\, \mathrm{d}r$.

The proof of Corollary 7.5.2.1 follows from Theorem 7.5.1.1. As discussed previously, the cell association of UAV-UEs and their overall performance is essentially driven by the availability of LoS links and the encountered antenna gain. Thus far, we particularly showed that the coverage probability of static UAV-UEs heavily depends on the antenna pattern, but it does not scale with the number of antenna elements. Next, we study a scenario in which the UAV-UEs can be mobile.

7.6 Coverage Probability of mobile UAV-UEs

We consider vertically-mobile UAV-UEs that make frequent up and down movements with a fixed velocity $\bar{\nu}$ in the finite vertical region $[h_1, h_2]$. We refer to it as *vertical 1D random waypoint (RWP)* mobility model. Similar stochastic mobility models are adopted for UAV-BSs and UAV-UEs in the recent works [14, 154], and [155]. The proposed mobility model works as follows: Initially, at time instant t_0, the UAV-UE is at an arbitrary altitude h_0 selected uniformly from the interval $[h_1, h_2]$. Then, at next time epoch t_1, this UAV-UE at h_0 selects a new random waypoint h_1 uniformly in $[h_1, h_2]$, and moves towards it. Once the UAV-UE reaches h_1, it repeats the same procedure to find the next destination altitude and so on. Eventually, the steady-state altitude distribution converges to a non-uniform distribution $F_Z(z)$. Note that random waypoints refer to the altitude of a UAV-UE at each time epoch, which is uniformly-distributed in $[h_1, h_2]$, while vertical transitions are the differences in the UAV-UE altitude throughout its trajectory. While the random waypoints are independent and uniformly-distributed by definition,

the random lengths of vertical transitions are statistically dependent. This is because the endpoint of one movement epoch is the starting point of the next epoch. In [14], we showed that $f_Z(z) = \frac{h_1 z + h_2 z - h_1 h_2 - z^2}{\hbar^3/6}$, $\forall h_1 < z < h_2$, where $\hbar = h_2 - h_1$. It can be easily verified that the mean altitude is $\mathbb{E}[Z] = \frac{h_1+h_2}{2}$.

The coverage mobility of UAV-UEs is studied under nearest and HARP associations. Vertically-mobile UAV-UEs under nearest association maintain their connection to the nearest BSs. However, since UAV-UEs under HARP association connect to the BS delivering the highest average power, the serving BS might change with the UAV-UE altitude. This is because the elevation angle and, correspondingly, the BS antenna gain change with the UAV-UE altitude.

7.6.1 Nearest Association

For vertically mobile UAV-UEs, the distance to the nearest BS is denoted as $w_0 = \sqrt{r_0^2 + (z - h_{\rm BS})^2}$, with $w_0 \geq h_0$, and $h_0 = h_1 - h_{\rm BS}$. For simplicity, we use $f_Z(z)$ to obtain the steady state 3D distance PDF $f_{W_0}(w_0)$, rather than averaging over two RVs Z and R_0 in the coverage probability calculation. Since R_0 and Z are two independent RVs, $f_{R_0,Z}(r_0, z) = f_{R_0}(r_0) f_Z(z)$. Given that, we transform the two RVs R_0 and Z to W_0 and find $f_{W_0}(w_0)$, with the details omitted. The equivalent 3D distance PDF is obtained as

$$f_{W_0}(w_0) = 2\pi\lambda_b w_0 \Omega(h_1, h_2) e^{-\pi\lambda_b w_0^2}, \tag{7.8}$$

where $\Omega(h_1, h_2) = \frac{\psi(h_2)(1-\xi) - \psi(h_1)(\xi+1) + 2\kappa(h_1,h_2) - 2\kappa(h_2,h_1)}{2\pi\lambda_b^{3/2}\hbar^3/3}$, $\psi(x) = \text{erfi}\left(\sqrt{\pi}\sqrt{\lambda_b}(x - h_{\rm BS})\right)$, $\xi = 2\pi\lambda_b(h_1 - h_{\rm BS})(h_{\rm BS} - h_2)$, and $\kappa(x, y) = \sqrt{\lambda_b}(x - h_{\rm BS}) e^{\pi\lambda_b (y - h_{\rm BS})^2}$.

Since w_0 replaces both z and r_0, we set θ to $\arcsin(\frac{h}{w_0})$ in (7.2) to obtain the antenna gain $G(w_0, h)$, where $h = h_d - h_{\rm BS}$ and $h_d = \mathbb{E}[Z] = \frac{h_1+h_2}{2}$ is the mean flying altitude. The effect of the vertical mobility, i.e., altitude variation, and horizontal distance randomness are now captured by the RV W_0.

We characterize the coverage probability of mobile UAV-UEs under nearest association in the next corollary (whose proof follows Theorem 7.5.1.1).

Corollary 7.6.1.1. *The coverage probability of mobile UAV-UEs under nearest association is given by:*

$$\mathbb{P}_c = \int_{h_0}^{\infty} \mathbb{P}_{c|w_0} f_{W_0}(w_0)\, \mathrm{d}w_0\,, \tag{7.9}$$

where $\mathbb{P}_{c|w_0}$ *is calculated as in (7.5), with* $\varpi_l = \frac{\vartheta w_0^{\alpha_l} m_l}{A_l G(w_0,h)}$. *The Laplace transform of LoS interference is given by* $\mathcal{L}_{I_l}(\varpi_l) = e^{-\gamma(\varpi_l)}$, *where* $\gamma(\varpi_l) = 2\pi\lambda_b \sum_{j=j_0}^{\infty} \mathbb{P}_l(s) \int_{\max(w_0,s)}^{t} \Big(1 - \big(\frac{m_l}{m_l + \varpi_l A_l G(w,h) w^{-\alpha_l}}\big)^{m_l}\Big) w\, \mathrm{d}w$, $j_0 = \lfloor \sqrt{w_0^2 - h^2}\sqrt{a\eta} \rfloor$, $s = \frac{j}{\sqrt{a\eta + h^2}}$, $t = \frac{j+1}{\sqrt{a\eta + h^2}}$, *and* $\mathbb{P}_l(j)$ *is calculated from (7.1).*

The antenna gain effect on the coverage probability of mobile UAV-UEs can be interpreted in a similar way to the static scenario in (7.5.1.1). Particularly, conditioning on $W_0 = w_0$, the yielded expression holds the same insights as for static users, i.e., the coverage probability does not scale with N_t.

7.6.2 Highest Average Received Power Association

Mobile UAV-UEs under HARP association are prone to frequent *altitude handovers*. This happens when their trajectory crosses multiple peaks and nulls of the antennas' side-lobes. For instance, Fig. 7.1 shows that the UAV-UE at altitude h_1 associates to the right BS, while it turns to attach to the left BS at altitude h_2. This *altitude handover* negatively impacts the performance of UAV-UEs as it results in dropped connections. In fact, the elevation angle $\theta = \arctan\Big(\frac{z - h_{\mathrm{BS}}}{r_m}\Big)$ plays a crucial role on determining the serving BS. Hence, it is essential to average over the random distance R_m (as in Corollary 7.5.2.1) for every possible Z, to correctly account for the handover and cell selection. Motivated by this fact, we describe the coverage probability as stated in the next corollary.

Corollary 7.6.2.1. *The approximated coverage probability of mobile UAV-UEs under HARP association is described as:*

$$\mathbb{P}_c = \int_{z=h_1}^{h_2} \int_{r_m=0}^{\infty} \mathbb{P}_{c|r_m,z} f_{R_m}(r_m) f_Z(z)\, \mathrm{d}r_m\, \mathrm{d}z\,, \tag{7.10}$$

where $\mathbb{P}_{c|r_m,z}$ *is calculated as* $\mathbb{P}_{c|r_0}$ *in Theorem 7.5.1.1, with* $\varpi_l = \frac{\vartheta(r_m^2+(z-h_{\mathrm{BS}}))^{\alpha_l/2} m_l}{A_l G(r_m, z-h_{\mathrm{BS}})}$. *The Laplace transform of LoS interference is* $\mathcal{L}_{I_l}(\varpi_l) = e^{-\gamma(\varpi_l)}$, *where* $\gamma(\varpi_l) = 2\pi\lambda_b \sum_{j=0}^{\infty} \mathbb{P}_l(j) \int_{\frac{j}{\sqrt{a\eta}}}^{\frac{j+1}{\sqrt{a\eta}}} \Big(1 - \big(\frac{m_l}{m_l+\varpi_l \zeta_l(r)}\big)^{m_l}\Big) r \, \mathrm{d}r$.

To account for the UAV-UE mobility in the coverage probability expression, similar to [14] and [148], we consider a linear function that reflects the cost of handovers. Particularly, we define the mobile UAV-UE coverage probability as:

$$\mathbb{P}_c(\bar{\nu}, \lambda_b, \beta) = (1-\beta)\mathbb{P}_c + \beta \mathbb{P}(\bar{H}) \mathbb{P}_c, \tag{7.11}$$

where $\beta \in [0,1]$ represents the handover cost and $\mathbb{P}(\bar{H})$ is the probability of no handover. The handover probability is defined as the probability that an UAV-UE associates to a new BS rather than the serving BS after a unit time. It is clear from (7.11) that, if $\beta = 1$, the first term vanishes and the UAV-UE will be in coverage only if there is no handover associated with its mobility.

7.7 Numerical Results

For our simulations, we consider a network having the parameter values indicated in Table 7.1. Without lack of practicality, the selected values for the UAV-UE lowest and highest altitudes (h_1 and h_2) are set considerably above the heights of the BSs (h_{BS}) such that these UAV-UEs are serviced from the side-lobes of the ground BSs. This helps us investigate how the side-lobes of the ground BSs can affect the performance of both static and mobile UAV-UEs, which is one of the main objectives of this chapter. These lowest and highest altitudes would reflect the fluctuations of the received signal levels by UAV-UEs when crossing the peaks and nulls of the antennas' side-lobes.

7.7.1 Nearest Association

We first evaluate the performance of static and mobile UAV-UEs under the nearest association scheme. We compare the UAV-UE coverage probability under practical antenna patterns with that adopting simple antenna models, e.g., [14, 143, 148]. Particularly, high-altitude UAV-UEs are essentially served and interfered from the

TABLE 7.1: Simulation Parameters

Description	Parameter	Value
LoS and NLoS path-loss exponents	α_l, α_n	2.09, 3.75
LoS and NLoS path-loss constants	A_l, A_n	-41.1 dB, -32.9 dB
Number of antenna elements	N_t	8
LoS and NLoS fading parameters	m_l, m_n	3, 1
BS height and UAV-UE altitude	h_{BS}, h_d	30 m, 150 m
UAV-UE lowest and highest altitudes	h_1, h_2	140 m, 160 m
Mobile UAV-UE speed	$\bar{\nu}$	20 kmh
Environment blocking parameters	a, η, c	0.6, 500 km^{-2}, 30 m
Density of BSs and SIR threshold	λ_b, ϑ	50 km^{-2}, -15 dB

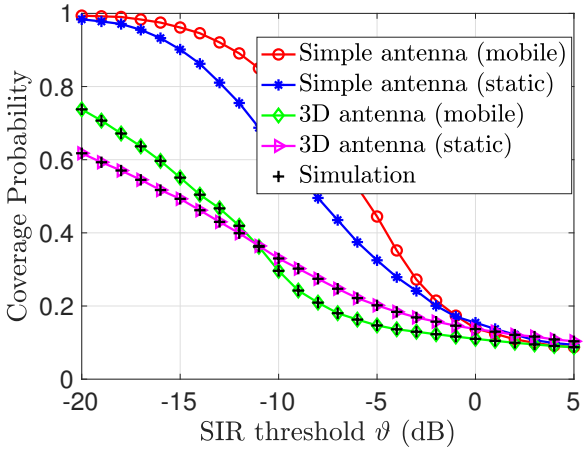


FIGURE 7.3: Coverage probability of static and mobile UAV-UEs under nearest association scheme versus the SIR threshold ϑ.

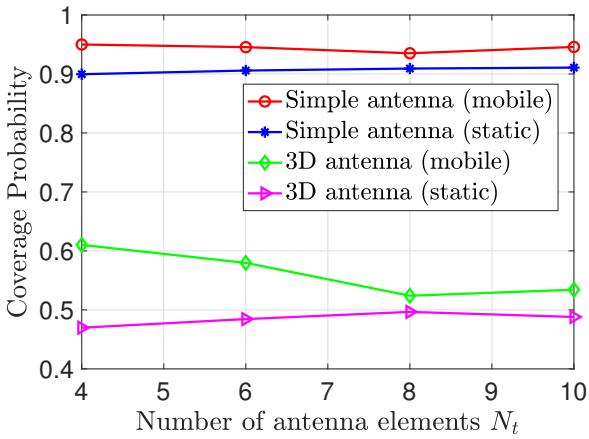


FIGURE 7.4: Coverage probability of static and mobile UAV-UEs under nearest association versus the number of antenna elements N_t.

antennas' side-lobes. Hence, for a simple antenna model, the antenna gain effect is normalized. Recall that the simple antenna model is defined as a step function of two gain values, namely, main- and sidelobe, while the practical antenna model

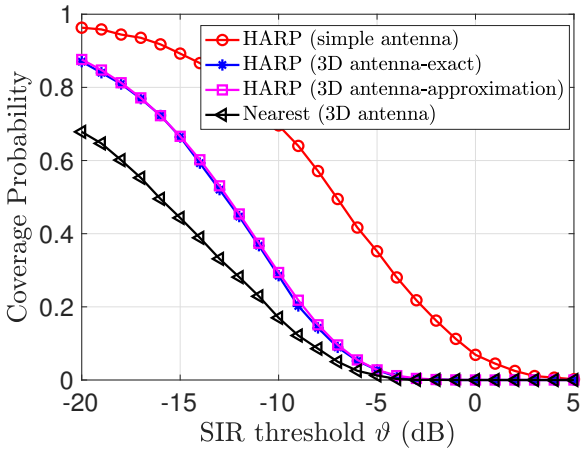


FIGURE 7.5: Coverage probability of static UAV-UEs under HARP association ($m_v = 1$, $N_t = 4$).

resembles a sequence of main- and side-lobes with nulls between consecutive lobes. Fig. 7.3 first shows that the performance attained under a simple antenna model is superior to that from a practical antenna pattern. This implies that adopting practical antenna models is vital to convey a realistic performance evaluation of the UAV-UEs. Fig. 7.3 also compares the performance of mobile UAV-UEs to static counterparts under nearest association. The effect of vertical mobility on the UAV-UE coverage probability is relatively marginal. This is attributed to the fact that there is no altitude handover associated with the vertical mobility as the UAV-UE maintains its connection with the nearest BS.

Fig. 7.4 investigates the effect of the number antenna elements N_t on the UAV-UE coverage probability. Fig. 7.4 shows that an increase in the number of antenna elements has an intangible effect on the coverage probability of static and mobile UAV-UEs under nearest association, hence verifying the claim in Corollary 7.5.1.1. This also can be interpreted by the fact that, while increasing N_t yields a higher number of lobes, the integrands in the coverage probability expression, which are functions of the antenna gains, constitute an overall area that does not significantly change with N_t.

7.7.2 HARP Association

Next, we discuss the performance of UAV-UEs under HARP association. Fig. 7.5 verifies the accuracy of the geometry-based approximation of Corollary 7.5.2.1.

Fig. 7.5 presents the exact expression and the approximation of the static UAV-UE coverage probability versus the SIR threshold ϑ. Clearly, the adopted approximation is relatively tight. Fig. 7.5 also shows that the coverage probability of static UAV-UEs under practical antenna patterns is much reduced compared to the achievable coverage probability from a simple antenna model. This result shows that the achievable performance under a simple antenna model is quite optimistic as compared to the actual achievable performance from under practical antenna configurations.

In Fig. 7.6, we plot the coverage probability of mobile UAV-UEs versus the number of antenna elements N_t, at different penalty costs β. Fig. 7.6 shows that the UAV-UE coverage probability monotonically decreases as both N_t and β increase. This can be interpreted by the fact that as long as N_t increases, the mobile UAV-UE becomes more susceptible to handovers, which are penalized by the cost β.

Finally, Fig. 7.7 investigates the relation between the handover rate H ($\sec^{-1}$) and the number of antenna elements N_t. The handover rate is numerically calculated from

$$H = \mathbb{E}\Big[\frac{\#\text{ handoffs per each movement}}{\text{movement length}}\Big] \times \bar{\nu},$$

where the movement is generated according to the adopted vertical mobility model, and $\bar{\nu}$ is the average speed of a mobile UAV-UE. Fig. 7.7 shows that the handover rate monotonically grows with the number of antenna elements. This is due to the higher number of nulls and peaks of the antenna vertical gain that the UAV-UE crosses through its trajectory. *From this result, we conclude that a larger number of antenna elements yields a higher rate of altitude handovers.*

7.8 Conclusion

In this chapter, we have studied the performance of UAV-UEs under practical antenna configurations. The coverage probability of static and mobile UAV-UEs is characterized as a function of the system parameters, namely, the number of antenna elements, density of BSs, and the UAV-UE altitude for both nearest and

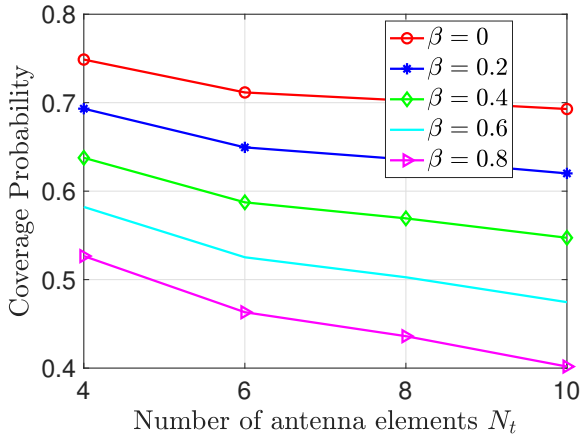


FIGURE 7.6: Coverage probability of mobile UAV-UEs under HARP association ($h_d = 100\,\text{m}$, $h_1 = 80\,\text{m}$, $h_2 = 120\,\text{m}$).

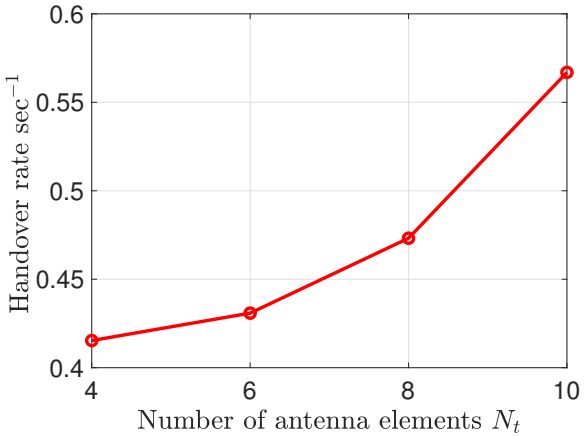


FIGURE 7.7: The handover rate versus the number of antenna elements ($h_d = 100\,\text{m}$, $h_1 = 80\,\text{m}$, $h_2 = 120\,\text{m}$).

HARP associations. The handover rate is also investigated to reveal the impact of practical antenna pattern on the cell association of mobile UAV-UEs. We have shown that the overall performance under practical antenna patterns is worse than that attained from a simple antenna model, which supports the fact that simple antenna models do not capture the real performance of the UAV-UEs. Moreover, for static UAV-UEs, or mobile UAV-UEs undergoing nearest association, the increase of the number of the antenna elements is shown to have a slight impact on their achievable performance. This is attributed to the fact that as these UAV-UEs do not tend to change their cell association, there is no accompanied altitude handover. Conversely, for HARP association, the coverage probability of mobile UAV-UE decreases as the number of antenna elements increases due to the excessive rate of

altitude handover that is penalized by the handover cost β.

Chapter 8

Caching to the Sky: Performance and Mobility Analysis for Cellular-Connected UAVs

Providing connectivity to aerial users, such as cellular-connected unmanned aerial vehicles (UAVs) or flying taxis, is a key challenge for tomorrow's cellular systems. In this chapter, the use of coordinated multi-point (CoMP) transmission along with caching for providing seamless connectivity to aerial users is investigated. In particular, a network of clustered cache-enabled small base stations (SBSs) serving aerial users is considered where requested content is cooperatively transmitted from collaborative ground SBSs. Two scenarios are studied: scenarios with static, hovering UAV-UEs and scenarios with mobile UAV-UEs. Under a maximum ratio transmission, a novel framework is developed and leveraged to derive upper and lower bounds on the UAV-UE coverage probability for both scenarios. Using the derived results, the effects of various system parameters such as collaboration distance, UAV-UE altitude, and UAV-UE speed on the achievable performance are studied. Results reveal that, for both static and mobile UAV-UEs, when the BS antennas are tilted downwards, the coverage probability of a high-altitude UAV-UE is upper bounded by that of ground UEs regardless of the transmission scheme. Moreover, for low signal-to-interference-ratio thresholds, it is shown that CoMP transmission can improve the coverage probability of UAV-UEs, e.g., from 28% under the nearest association scheme to 60% for an average of 2.5 collaborating BSs.

Meanwhile, key results on mobile UAV-UEs unveil that not only the spatial displacements of UAV-UEs but also their vertical motions affect their handover rate and coverage probability. In particular, UAV-UEs that have frequent vertical movements and high direction switch rates are expected to have low handover probability and handover rate. Finally, the effect of the UAV-UE's vertical movements on its coverage probability is marginal if the UAV-UE retains the same mean altitude.

8.1 Introduction

The past few years have witnessed a tremendous increase in the use of UAVs, popularly called drones, in many applications, such as aerial surveillance, package delivery, and even flying taxis [156] and [157]. Enabling such UAV-centric applications requires ubiquitous wireless connectivity that can be potentially provided by the pervasive wireless cellular network [31, 69, 158]. However, in order to operate cellular-connected UAVs using existing wireless systems, one must address a broad range of challenges that include interference mitigation, reliable communications, resource allocation, and mobility support [159]. Next, we review some of the works relevant to the cellular-connected UAV-enabled networks.

8.1.1 Motivation and Contributions

While there exist some approaches in the literature to improve the cellular connectivity for UAV-UEs [10, 160–164], none of these works studied the role of CoMP transmission as an effective interference mitigation tool to support the UAV-UEs. Moreover, these works only considered scenarios of static UAV-UEs. Furthermore, while the authors in [148, 151] studied the performance of mobile UAV-UEs, their results were based on system simulations and measurement trials. Particularly, a rigorous analysis for mobile UAV-UEs to quantify important performance metrics such as coverage probability and handover rate is still lacking in the current state-of-the-art. *To our best knowledge, this chapter provides the first rigorous analysis of CoMP transmission for both static and 3D mobile UAV-UEs, where a novel 3D mobility model is also provided.*

The main contribution of this chapter is a novel framework that leverages CoMP transmissions for serving cellular-connected UAVs, and develops a novel mobility model that effectively captures the 3D movements of UAV-UEs. We propose a maximum ratio transmission (MRT) scheme aiming to maximize the desired signal level at the intended UAV-UE, and, hence, the performance of cellular communication links for the UAV-UEs can be improved. In particular, we consider a network of disjoint clusters in which BSs in one cluster collaboratively serve one UAV-UE within the same cluster via coherent CoMP transmission. For this network, we consider two key scenarios, namely, static and mobile UAV-UEs. Using Cauchy's inequality and Gamma approximations, we develop a novel framework that is then leveraged to derive tight upper bound (UB) and lower bound (LB) on the UAV-UE coverage probability for both scenarios. Moreover, for mobile UAV-UEs, we analytically characterize the handover rate, and handover probability based on a novel 3D mobility model. We further quantify the negative impact of the UAV-UEs' mobility on their achievable performance.

Our results reveal that the achievable performance of UAV-UEs heavily depends on the UAV-UE altitude, UAV-UE speed, and the collaboration distance, i.e., the distance within which the UAV-UE is cooperatively served from ground BSs. Moreover, while allowing CoMP transmission substantially improves the UAV-UEs' performance, it is shown that their performance is still upper bounded by that of ground UEs due to the down-tilt of the BS antennas and the encountered LoS interference. Additionally, results on mobile UAV-UEs unveil that the spatial displacements of UAV-UEs jointly with their vertical motions affect their handover rate and handover probability. Moreover, while the UAV-UE spatial movements considerably impact its coverage probability (due to handover), the effect of the UAV-UE vertical displacements is marginal if the UAV-UE fluctuates around the same mean altitude. *Overall, cooperative transmission is shown to be particularly effective for high altitude UAV-UEs that are susceptible to adverse interference conditions, which is the case in a variety of drone applications.*

8.1.2 Related Works

Recently, cellular-connected UAVs have received significant attention, whereby UAVs as new UEs are integrated into the cellular network in order to ensure reliable and secure connectivity for the operations of UAV systems. However, it has been established that the dominance of LoS links makes inter-cell interference a critical issue for cellular systems with hybrid terrestrial and UAV-UEs. In this regard, extensive real-world simulations and fields trials in [140, 159, 165, 166] have shown that a UAV-UE, in general, has poorer performance than a ground UE. Due to the down-tilted BS antennas, it is found that UAVs at $40\,\mathrm{m}$ and higher, will be eventually served by the side-lobes of the BS antennas, which have reduced antenna gain compared to the corresponding main-lobes. However, UAV-UEs at $40\,\mathrm{m}$ and above experience favorable free-space propagation conditions. Interestingly, the work in [140] showed that the free-space propagation can make up for the BS antenna side-lobe gain reduction. However, this benefit of such a favorable LoS channel that UAV-UEs enjoy vanishes at high altitudes and turns to be one of their key limiting factors. This is because the free-space propagation also leads to stronger LoS interfering signals. Eventually, UAV-UEs at high altitudes potentially exhibit poorer communication and coverage compared to ground UEs [139, 140, 159, 165, 166].

While the works in [139, 140, 159, 165, 166] explored the feasibility of providing cellular connectivity for UAVs, they did not envision new techniques to improve their performance. In particular, UAVs, at high altitudes, have limited coverage and connectivity due to the encountered LoS interference and reduced antenna gains. Moreover, their cell association will be essentially driven by the side-lobes of BS antennas, which will lead to more handovers and handover failures for mobile UAV-UEs [140]. This necessitates the need to have sky-aware cellular networks that can seamlessly cover high altitudes UAV-UEs and support their inevitable mobility. Next, we review some recently-adopted techniques that aimed to provide reliable connectivity to the UAV-UEs.

Recently, various approaches have been proposed in [10, 160–164] in order to improve the cellular connectivity for UAVs using, e.g., massive MIMO, millimeter

wave (mmWave), beamforming, and power control. For instance, in [10], we proposed a MIMO conjugate beamforming scheme that can improve the cellular connectivity for UAV-UEs and enhance the system spectral efficiency. Moreover, the authors in [162] incorporated directional beamforming at aerial BSs to alleviate the strong LoS interference seen by their served UAV-UEs. However, while interesting, the works in [10, 160–164] only considered scenarios of static UAV-UEs. Moreover, they did not consider the use of cooperative communication through CoMP transmissions for UAV-UEs, which is a prominent interference mitigation tool that can diminish the effect of LoS interference. For instance, the work in [167] proposed CoMP transmissions from UAV-BSs in order to spatially multiplex signals from ground UEs and forward them to a central processor.

Unlike the static UAV assumptions in [10, 139, 160–164], the study of mobile UAVs has been conducted in [141, 148, 151, 154, 155, 168–171]. Prior works in the literature followed two main directions pertaining to trajectory design for mobile UAVs. The first line of work focuses on deterministic trajectories, whereby a UAV is assumed to travel between two deterministic, possibly known, locations [141, 168, 169]. This type of trajectories can be used for path planning and mission-related metrics' optimization, e.g., mission time and achievable rates. For instance, the authors in [141] studied the problem of trajectory optimization for a cellular-connected UAV flying from an initial location to a final destination. Moreover, the work in [169] proposed an interference-aware path planning scheme for a network of cellular-connected UAVs based on deep reinforcement learning.

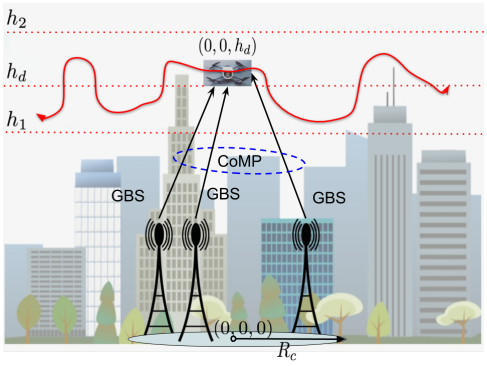
The second line of work in [154, 155, 170, 171] considers stochastic trajectories in which the movements of UAVs are characterized by means of stochastic processes. This type of trajectories is usually adopted in the study of communication and mobility-related metrics such as coverage probability and handover rate. For example, in [170], the authors proposed a mixed random mobility model that characterizes the movement process of UAVs in a finite 3D cylindrical region. The authors characterized the ground UE coverage probability in a network of one static serving aerial BS and multiple mobile interfering aerial BSs. The authors extended their work in [155] such that both serving and interfering aerial BSs are mobile. Meanwhile, the authors in [171] showed that an acceptable ground UE coverage

can be sustained if the aerial BSs move according to certain stochastic trajectory processes. Recently, the work in [154] characterized the performance of several canonical mobility models for UAV-BSs in an infinite drone network. However, while interesting, the proposed mobility models in [155, 170, 171] can only describe the motions of UAV-BSs deployed in a bounded cylindrical region in space. In contrast, cellular-connected UAV-UEs such as flying taxis and delivery drones can have very long trajectories that cross multiple areas served by different BSs. Moreover, the canonical mobility models in [154] are only suitable for 2D mobile UAV-BSs, while UAV-UEs are susceptible to inevitable 3D motions.

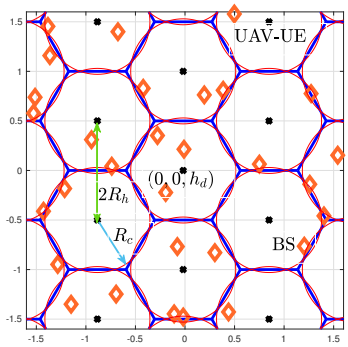
Ensuring reliable connectivity for such mobile UAV-UEs is of paramount importance for the control and operations of UAV systems. In this regard, the mobility performance of cellular-connected UAVs has been studied in recent works [148, 151]. In [148], the authors quantified the impact of handover on the UAV-UE throughput, assuming that no payload data is received during the handover procedures. Moreover, based on experimental trials, in [151], the authors showed that under the strongest received power association, drones are subject to frequent handovers once the typical flying altitude is reached. However, the results presented in these works are based on simulations and measurements while our work goes beyond this by analytically characterizing important performance metrics such as coverage probability and handover rate, which are still lacking in the current literature.

8.2 System Model

We consider a cache-enabled small cell network in which SBSs are distributed according to a homogeneous PPP $\Phi_b = \{b_i \in \mathbb{R}^2, \forall i \in \mathbb{N}^+\}$ with intensity λ_b. We consider a cellular-connected UAV-UE flying at an altitude h_d and located at $(0, 0, h_d) \in \mathbb{R}^3$, where h_d is the altitude of the UAV-UE. We assume a user-centric model in which the SBSs are grouped into disjoint clusters around UAV-UEs to be served [172]. A cluster is represented by a circle of radius R_c centered at the terrestrial projection of an UAV-UE, as shown in Fig. 8.1. The area of each cluster is then given by $A = \pi R_c^2$.



(a) Illustration of cooperative transmission



(b) Snapshot of a cluster-centric UAV-UE topology

FIGURE 8.1: Illustration of the proposed system model where BS cooperatively serve high-altitude UAV-UEs via CoMP transmission. UAV-UEs can be either hovering at a fixed altitude h_d or flying within minimum and maximum altitudes h_1 and h_2, respectively. In (b), the clusters are defined by a hexagonal grid, wherein BSs (orange diamonds) are distributed according to a homogeneous PPP and the UAV-UEs (black stars) are hovering above the centers of disjoint clusters.

SBSs belonging to the same cluster can cooperate to serve cached content to the UAV-UE whose projection on the ground is assumed at the cluster center. The UAV-UE can represent a conventional cellular-connected UAV or a passenger of a flying drone-taxi [156]. Due to the strong LoS-dominated interference at high altitudes, we allow multiple SBSs within a certain distance from the UAV-UE to cooperatively transmit a requested content that they previously cached.

8.2.1 Probabilistic Caching Placement

Each SBS has a surplus memory designated for caching content from a known file library. These files represent the content catalog that an UAV-UE may request, and

are indexed in a descending order of popularity. We adopt a random content placement policy in which each content f is cached independently at each SBS according to a probability c_f, $0 \leq c_f \leq 1$. Note that SBSs caching content f can be modeled as a PPP Φ_{bf} with the intensity function given by the independent thinning theorem as $\lambda_{bf} = c_f \lambda_b$ [75]. Similarly, SBSs not caching a content f can be modeled as another PPP $\Phi^{!}_{bf}$ with intensity function $\lambda^{\circ}_{bf} = (1 - c_f)\lambda_b$, where $\Phi_b = \Phi_{bf} \cup \Phi^{!}_{bf}$. The probability mass function (PMF) of the number of SBSs caching content f in a cluster is given by:

$$\mathbb{P}(n = \kappa) = \frac{(c_f \lambda_b A)^{\kappa} e^{-c_f \lambda_b A}}{\kappa!}, \tag{8.1}$$

which represents a Poisson distribution with mean $c_f \lambda_b A$.

8.2.2 Serving Distance Distributions

Under the condition of having κ caching SBSs in the cluster of interest, the distribution of such in-cluster caching SBSs will follow a binomial point process (BPP). This BPP consists of κ uniformly and independently distributed SBSs in the cluster.

The set of cooperative SBSs providing content f is defined as $\Phi_{cf} = \{b_i \in \Phi_{bf} \cap \mathcal{B}(0, R_c)\}$, where $\mathcal{B}(0, R_c)$ denotes the ball centered at the origin with radius R_c. Considering the UAV-UE located at the origin in $\mathbb{R}^2$, i.e., $(0, 0, h_d) \in \mathbb{R}^3$, the 2D distances from the cooperative SBSs to the UAV-UE are denoted by $\boldsymbol{R}_{\kappa} = [R_1, \ldots, R_{\kappa}]$. Then, conditioning on $\boldsymbol{R}_{\kappa} = \boldsymbol{r}_{\kappa}$, where $\boldsymbol{r}_{\kappa} = [r_1, \ldots, r_{\kappa}]$, the conditional PDF of the joint serving distances' distribution is denoted as $f_{\boldsymbol{R}_{\kappa}}(\boldsymbol{r}_{\kappa})$. The κ cooperative SBSs that cache a content f can be seen as the κ-closest SBSs to the cluster center from the PPP Φ_{bf}. Since the κ SBSs are independently and uniformly distributed in the cluster approximated by $\mathcal{B}(0, R_c)$, we have the PDF of the horizontal distance r_i from SBS i to the UAV-UE at $(0, 0, h_d)$ given as [75]

$$f_{R_i}(r_i) = \begin{cases} \frac{2r_i}{R_c^2}, & 0 \leq r_i \leq R_c, \\ 0, & \text{otherwise}, \end{cases}$$

for any $i \in \mathcal{K}_f = \{1, \ldots, \kappa\}$, where $\mathcal{K}_f$ is the set of SBSs that cache a content f within the ball $\mathcal{B}(0, R_c)$. From the i.i.d. property of BPP, the conditional joint PDF

of the serving distances $\boldsymbol{R}_\kappa = [R_1, \ldots, R_\kappa]$ is

$$f_{\boldsymbol{R}_\kappa}(\boldsymbol{r}_\kappa) = \prod_{i=0}^{\kappa} \frac{2r_i}{R_c^2}. \tag{8.2}$$

We consider a content delivery from ground SBSs having the same height h_{SBS} to an UAV-UE located at altitude h_d. The SBS vertical antenna pattern is directional and down-tilted for ground UEs. The vertical antenna beamwidth and down-tilt angle of the SBSs are denoted respectively by θ_B and θ_t. The side and main lobe gains of the antennas are denoted by G_s and G_m, respectively. Since the horizontal distance between the UAV-UE and SBS i is r_i, the communication link distance will be $d_i = \sqrt{r_i^2 + (h_d - h_{\text{SBS}})^2}$ for all $i \in \mathcal{K}_f$.

8.2.3 Channel Model

For the CoMP transmission between SBSs and the UAV-UE, we consider a wireless channel that is characterized by both large-scale and small-scale fading. For the large-scale fading, the channel between SBS i and the aerial user is described by the LoS and NLoS components, which are considered separately along with their probabilities of occurrence [173]. This assumption is apropos for such ground-to-air channels that are often dominated by LoS communication [139]. Therefore, the antenna gain plus path loss for each component, i.e., LoS and NLoS, will be

$$\zeta_v(r_i) = A_v G(r_i) d_i^{-\alpha_v} = A_v G(r_i) \big(r_i^2 + (h_d - h_{\text{SBS}})^2\big)^{-\alpha_v/2}, \tag{8.3}$$

where $v \in \{l, n\}$, α_l and α_n are the path loss exponents for the LoS and NLoS links, respectively, with $\alpha_l < \alpha_n$, and A_l and A_n are the path-loss constants at the reference distance $d_i = 1\,\text{m}$ for the LoS and NLoS, respectively. $G(r_i)$ is the total antenna directivity gain between SBS i and the aerial UE, which can be written similar to [174] as

$$G(r_i) = \begin{cases} G_m, & \text{for} \quad r_i \in \mathcal{S}_{bs}, \\ G_s, & \text{for} \quad r_i \notin \mathcal{S}_{bs}, \end{cases}$$

where $\mathcal{S}_{bs}$ is formed by all the distances r_i satisfying $h_{\text{SBS}} - r_i \tan(\theta_t + \frac{\theta_B}{2}) < h_d < h_{\text{SBS}} - r_i \tan(\theta_t - \frac{\theta_B}{2})$. In other words, the horizontal distance between a SBS and an

UAV-UE along with the antenna height, antenna beamwidth and down-tilt angles, and the altitude of this UAV-UE determine whether it is served by a mainlobe or sidelobe of a SBS antenna.

For the small-scale fading, we adopt a Nakagami-m model utilized in [139] and [174] for the channel gain, whose PDF is given by:

$$f(\omega) = \frac{2\frac{m}{\eta}^{m}\omega^{2m-1}}{\Gamma(m)}\exp\big(-\frac{m}{\eta}\omega^{2}\big), \tag{8.4}$$

where η is a controlling spread parameter, and the fading parameter m is assumed to be an integer for analytical tractability. Since the communication links between an UAV-UE and SBSs are LoS-dominated, e.g., suburban environments with $h_d > 40\,\mathrm{m}$ [140], it is assumed to have $m > 1$. Given that $\omega \sim$ Nakagami$(m, \eta/m)$, it directly follows that the channel gain power $\gamma = \omega^2 \sim \Gamma(m, \eta/m)$, where $\Gamma(k, \theta)$ is a Gamma RV with k and θ denoting the shape and scale parameters, respectively. Hence, the PDF of channel power gain distribution will be:

$$f(\gamma) = \frac{(\frac{m}{\eta})^{m}\gamma^{m-1}}{\Gamma(m)}\exp\big(-\frac{m}{\eta}\gamma\big). \tag{8.5}$$

3D blockage is characterized by the fraction a of the total land area occupied by buildings, the mean number of buildings e per km^2, and the buildings height modeled by a Rayleigh PDF with a scale parameter c. Hence, the probability of LoS of a caching SBS at a distance r_i from the UAV-UE is given by [175]:

$$\mathbb{P}_l(r_i) = \prod_{n=0}^{\max(p-1,0)}\left[1-\exp\Big(-\frac{\big(h_{\text{SBS}}+\frac{h(n+0.5)}{m+1}\big)^2}{2c^2}\Big)\right], \tag{8.6}$$

where $h = h_d - h_{\text{SBS}}$ and $p = \lfloor\frac{r_i\sqrt{ae}}{1000}\rfloor$. Different terrain structures and environments can be considered by varying the set of (a, e, c).

As discussed previously, the performance of a high-altitude UAV-UE is limited by the LoS interference they encounter. We hence propose a multi-SBSs cooperative transmission scheme aiming at mitigating inter-cell interference and improving the performance of such high-altitude aerial UEs. Under this setting, in the next section we develop a novel mathematical framework to characterize the performance of

TABLE 8.1: Mathematical Notations

Description	Notation
LoS and NLoS path-loss exponents	α_l, α_n
LoS/NLoS path-loss constants	A_l, A_n
Nakagami LoS/NLoS fading parameters	m_l, m_n
UAV-UE 3D cylindrical displacement	$\boldsymbol{X}_t$
UAV-UE velocity at epoch t	V_t
UAV-UE 3D transition length	U_t
UAV-UE 2D transition length	ρ_t
Mobility parameter	μ
UAV-UE average speed	$\bar{\nu}$
Vertical displacement	Z_t
Probability of handover	$\mathbb{P}(H)$
SIR threshold	ϑ
BSs' intensity	λ_b
Inter-cluster center and collaboration distances	$2R_h$, R_c
Antenna main and side-lobe gains	G_m, G_s
Gamma shape and scale parameters	K, θ
BS antenna height	h_{BS}
UAV-UE hovering altitude	h_d
UAV-UE upper and lower altitudes	h_1, h_2
Handover rate	H
Handover cost	β
Mean altitude of mobile UAV-UEs	$\mathbb{E}[Z_\infty]$

cache-assisted CoMP transmission for cellular-connected UAVs. The performance of UAVs is then contrasted to their terrestrial counterparts.

Having defined our system model, next, we will consider two scenarios: Static UAV-UEs and mobile UAV-UEs. For each scenario, we will characterize the coverage probability of high altitude UAV-UEs that are collaboratively served from BSs within their cluster. The performance of collaboratively-served UAV-UEs is then compared to their terrestrial counterparts and to UAV-UEs under the nearest association scheme. Moreover, we will characterize the handover rate for mobile UAV-UEs and quantify the negative impact of mobility on their achievable performance. For the reader's convenience, Table 8.1 summarizes our commonly-used notations hereinafter.

8.3 Coverage Probability Analysis

In this section, we characterize the network performance in terms of coverage probability. We assume that the SBSs have the same transmit power P_t. Without loss of

generality, we consider a typical UAV-UE located at $(0, 0, h_d) \in \mathbb{R}^3$. Conditioning on having κ SBSs serving a content f, the received signal at the UAV-UE will be:

$$P = \underbrace{\sum_{i=1}^{\kappa} P(r_i)\omega_i w_i X_f}_{\text{desired signal}} + \underbrace{\sum_{j \in \Phi_{bf}^{!} \cap \mathcal{B}(0,R_c)} P(u_j)\omega_j w_j Y_j}_{I_{\text{in}}} + \underbrace{\sum_{k \in \Phi_b \setminus \mathcal{B}(0,R_c)} P(u_k)\omega_k w_k Y_k}_{I_{\text{out}}} + \quad Z, \tag{8.7}$$

where the first term represents the desired signal from κ transmitting SBSs with $P(r_i) = \sqrt{P_t}\zeta_v(r_i)^{0.5}$, $v \in \{l, n\}$, ω_i being the Nakagami-m fading variable of the channel from SBS i to the aerial UE, w_i is the precoder used by SBS i, and X_f is the channel input symbol that is sent by the cooperating SBSs. The second and third terms represent the in-cluster interference I_{in}, and the out-of-cluster interference I_{out}, respectively. Y_j is the transmitted symbol from interfering SBS j and

$$P(u_j) = \begin{cases} P_l(u_j) = \sqrt{P_t}\zeta_l(u_j)^{0.5}, & \text{for} \quad \text{LoS}, \\ P_n(u_j) = \sqrt{P_t}\zeta_n(u_j)^{0.5}, & \text{for} \quad \text{NLoS}, \end{cases}$$

where u_j is the horizontal distance between interfering SBS j and the aerial UE. Z is a circular-symmetric zero-mean complex Gaussian RV modeling the background thermal noise. In-cluster interference occurs only for the case in which not all of the collaborative SBSs (within distance R_c) have the cached content (i.e., $c_f < 1$). In this case, the set of interfering SBSs will be characterized by $\Phi_b \setminus \Phi_{cf} = \Big\{ b_i \in \big\{\Phi_b \setminus \mathcal{B}(0, R_c)\big\} \cup \big\{\Phi_{bf}^{!} \cap \mathcal{B}(0, R_c)\big\} \Big\}$. For ease of notation, we denote $\big\{\Phi_{bf}^{!} \cap \mathcal{B}(0, R_c)\big\}$ as $\Phi_{cf}^{!}$. Otherwise, for $c_f = 1$, there is no in-cluster interference and the set of interfering SBSs will then be $\Phi_b \setminus \Phi_{cf} = \{b_i \in \Phi_b \setminus \mathcal{B}(0, R_c)\}$.

We assume that the CSI is available at the serving SBSs, i.e., the precoder w_i can be set as $\frac{\omega_i^*}{|\omega_i|}$, where ω_i^* is the complex conjugate of ω_i. Assuming that X_f, Y_j, and Y_k in (8.7) are independent zero-mean RVs of unit variance, and averaging over all LoS and NLoS configurations for the κ caching SBSs, the SIR at the UAV-UE will

then be:

$$\Upsilon_{|\boldsymbol{r}_\kappa} = \sum_{o=0}^{\kappa} \binom{\kappa}{o} \prod_{i=0}^{o} \mathbb{P}_l(r_i) \prod_{j=o+1}^{\kappa} \mathbb{P}_n(r_j) \cdot \frac{P_t \left| \sum_{i=1}^{o} \zeta_l^{1/2}(r_i)\omega_i + \sum_{j=o+1}^{\kappa} \zeta_n^{1/2}(r_j)\omega_j \right|^2}{I_{\text{in}} + I_{\text{out}}}. \tag{8.8}$$

In (8.8), we have $\left| \sum_{i=1}^{o} \zeta_l^{1/2}(r_i)\omega_i + \sum_{j=o+1}^{\kappa} \zeta_n^{1/2}(r_j)\omega_j \right|^2$ representing the square of a weighted sum of κ Nakagami-m RVs. Since there is no known closed-form expression for a weighted sum of Nakagami-m RVs, we use the Cauchy-Schwarz's inequality to get an upper bound on a square of weighted sum as follows:

$$\left| \sum_{i=1}^{o} \zeta_l^{1/2}(r_i)\omega_i + \sum_{j=o+1}^{\kappa} \zeta_n^{1/2}(r_j)\omega_j \right|^2 = \left(\sum_{i=1}^{\kappa} Q_i \right)^2 \leq \kappa \left(\sum_{i=1}^{\kappa} Q_i^2 \right), \tag{8.9}$$

where $Q_i = \zeta_v^{1/2}(r_i)\omega_i$, is a scaled Nakagami-$m$ RV, with $v \in \{l, n\}$ and $i \in \mathcal{K}_f$. Since $\omega_i \sim$ Nakagami$(m, \eta/m)$, from the scaling property of the Gamma PDF, $Q_i^2 \sim \Gamma\big(k_i = m, \theta_i = 2\eta\zeta_v(r_i)/m\big)$. To get a statistical equivalent PDF of a sum of κ Gamma RVs Q_i with different shape parameters θ_i, we adopt the method of sum of Gammas second-order moment match proposed in [176, Proposition 8]. It is shown that the equivalent Gamma distribution, denoted as $J \sim \Gamma(k, \theta)$, with the same first and second-order moments has the parameters $k = \Big(\sum_i k_i\theta_i \Big)^2 / \sum_i k_i\theta_i^2$ and $\theta = \sum_i k_i\theta_i^2 / \sum_i k_i\theta_i$. To showcase the accuracy of the second-order moment approximation in our case, for an arbitrary realization of the network, we plot in Fig. 8.2 the PDF of the equivalent channel gain. As evident from the plot, approximating a sum of κ Gamma RVs with an equivalent Gamma RV whose parameters are

$$k = \frac{m\big(\sum_i \zeta_v(r_i) \big)^2}{\sum_i \big(\zeta_v(r_i) \big)^2} \qquad \text{and} \qquad \theta = \frac{\eta \sum_i \zeta_v(r_i)}{m \sum_i \zeta_v(r_i)}, \tag{8.10}$$

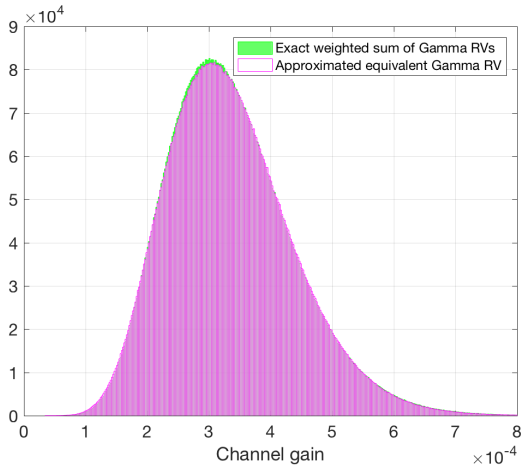


FIGURE 8.2: Monte Carlo simulation of the PDF of the equivalent gain of channels between cooperating SBSs and the aerial UE, including path loss and fading. A PPP realization of density $\lambda_b = 20$ SBS/km^2 is run for a simulated area of $20\,\mathrm{km}^2$ with $m = 3$ and $R_c = 200\,\mathrm{m}$.

is quite accurate. For tractability, we further upper bound the shape parameter k in (8.10):

$$k = m\frac{\Big(\sum_i \zeta_v(r_i)\Big)^2}{\sum_i \Big(\zeta_v(r_i)\Big)^2} \leq m\frac{\kappa\sum_i \Big(\zeta_v(r_i)\Big)^2}{\sum_i \Big(\zeta_v(r_i)\Big)^2} = m\kappa, \tag{8.11}$$

where $m\kappa$ is integer.

Next, we derive UB and LB expressions on the UAV-UE coverage probability. Our developed approach is novel in the sense that it adopts the Cauchy-Schwarz's inequality and moment match of Gamma RVs to derive an UB on the coverage probability, which is difficult to obtain exactly. A bound on the UAV-UE coverage probability conditioned on $\boldsymbol{R}_\kappa = \boldsymbol{r}_\kappa$ is expressed as:

$$\mathbb{P}_{\mathrm{c}|\boldsymbol{r}_\kappa} \overset{(a)}{\leq} \mathbb{P}\Big(\frac{\kappa P_t\big(\sum_{i=1}^{\kappa} Q_i\big)^2}{I_{\mathrm{out}}} > \vartheta\Big) \overset{(b)}{\approx} \mathbb{P}\Big(\frac{\kappa P_t J}{I_{\mathrm{out}}} > \vartheta\Big), \tag{8.12}$$

where (a) follows from the Cauchy-Schwarz's inequality, (b) follows from the Gamma approximation and rounding the shape parameter $K = m_l\kappa$, and ϑ is the SIR threshold. The coverage probability can be obtained as a function of the system parameters, particularly, the Nakagami fading parameter and collaboration distance, as stated formally in the following theorem.

Theorem 8.3.0.1. *An UB on the coverage probability of UAV-UEs cooperatively served from BSs within a collaboration distance R_c can be derived as follows:*

$$\mathbb{P}_{\mathrm{c}} = \sum_{\kappa=1}^{\infty} \mathbb{P}_{\kappa} \int_{\boldsymbol{r_\kappa}=\boldsymbol{R_c}}^{\infty} \mathbb{P}^{l}_{\mathrm{c}|\boldsymbol{r}_\kappa} \prod_{i=0}^{\kappa} \frac{2r_i}{R_c^2} \, \mathrm{d}\boldsymbol{r}_\kappa \,, \tag{8.13}$$

where $\mathbb{P}_{\kappa} = \frac{(2\sqrt{3}R_h^2\lambda_b c_f)^{\kappa} e^{-2\sqrt{3}R_h^2\lambda_b c_f}}{\kappa!}$ is the probability that there are κ in-cluster collaborating BSs. The conditional coverage probability is given by $\mathbb{P}^{l}_{\mathrm{c}|\boldsymbol{r}} = \|e^{\boldsymbol{T}_K}\|_1$, where $\|.\|_1$ represents the induced ℓ_1 norm and $\boldsymbol{T}_K$ is the lower triangular Toeplitz matrix:

$$\boldsymbol{T}_K = \begin{bmatrix} t_0 & & & \\ t_1 & t_0 & & \\ \vdots & \vdots & \ddots & \\ t_{K-1} & \dots & t_1 & t_0 \end{bmatrix};$$

$K = m_l\kappa$, $t_i = \frac{(-\varpi)^i}{(i)!}\Omega^{(i)}(\varpi)$, $\Omega^{(i)}(\varpi) = \frac{d^i}{d\varpi^i}\Omega(\varpi)_{|\boldsymbol{r}_\kappa}$, $\Omega(\varpi)_{|\boldsymbol{r}_\kappa} = -2\pi\lambda_b \int_{\upsilon=R_c}^{\infty} \Big(1 - \delta_l \mathbb{P}_l(\upsilon) - \delta_n \mathbb{P}_n(\upsilon)\Big)\upsilon \, \mathrm{d}\upsilon$, $\delta_l = \left(1 + \frac{\varpi P_l(\upsilon)^2}{m_l}\right)^{-m_l}$, $\delta_n = \left(1 + \frac{\varpi P_n(\upsilon)^2}{m_n}\right)^{-m_n}$, *and* $\varpi = \vartheta/\kappa P_t \theta$.

Proof. Please see Appendix F.1.[1] □

The main steps towards tractable coverage are summarized as follows [147]: We first derive the conditional log-Laplace transform $\Omega(\varpi)_{|\boldsymbol{r}_\kappa}$ of the aggregate interference. Then, we calculate the i-th derivative of $\Omega(\varpi)_{|\boldsymbol{r}_\kappa}$ to populate the entries t_i of the lower triangular Toeplitz matrix $\boldsymbol{T}_K$. The conditional coverage probability can be then computed from $\mathbb{P}^{l}_{\mathrm{c}|\boldsymbol{r}} = \|e^{\boldsymbol{T}_k}\|_1$.

Important insights on the coverage probability can be obtained from (8.13). First, if the collaboration distance R_c increases, both the probability $\mathbb{P}_{\kappa}$ and the integrand value in (8.13) increase, and, thus, the coverage probability grows accordingly. Furthermore, the effect of the BS density λ_b on the coverage probability is two-fold. On the one hand, the average number of BSs increases with λ_b as characterized by $\mathbb{P}_{\kappa}$, which results in a higher desired signal power. On the other hand, this advantage is counter-balanced by the increase in (LoS) interference power when λ_b increases, as captured in the decaying exponential functions

[1] Although there exists an infinite sum in (8.13), this sum vanishes for a small number of serving BSs that is determined by the collaboration distance R_c and the BSs' density λ_b.

in (F.3). Additionally, this compact representation, i.e., $\mathbb{P}^{l}_{\mathrm{c}|\boldsymbol{r}} = \|e^{\boldsymbol{T}_K}\|_1$, leads to valuable system insights. For instance, the impact of the shape parameter $K = \kappa m_l$ on the intended channel gain $\mathrm{Gamma}(K, \theta)$ is rigorously captured by the finite sum representation in (F.1) of Appendix F.1, which is typically related to the collaboration distance R_c and the Nakagami fading parameter m_l.

Next, we derive an LB on the coverage probability, which will lead to closed-form expressions for t_k, i.e., the entries populating $\boldsymbol{T}_K$ in (8.13). Given the high-altitude assumption of UAV-UEs, we will consider a special case when interfering BSs have dominant LoS communications to the typical UAV-UE, i.e, $\mathbb{P}_l(\upsilon) = 1$ and $\mathbb{P}_n(\upsilon) = 0$ in (8.13). Since this case results in higher interference power, this yields the derived coverage probability LB.

Corollary 8.3.0.1. *An LB on the coverage probability of the UAV-UEs can be computed from (8.13), where $\mathbb{P}^{l}_{\mathrm{c}|\boldsymbol{r}} = \|e^{\boldsymbol{T}_K}\|_1$, and the entries of $\boldsymbol{T}_K$ are given in closed-form expressions as*

$$t_k = \pi\lambda_b R_{ch}^2 \Big(\boldsymbol{1}\{k=0\} - c_{k2} F_1(k+m_l, k-\delta_l; k+1-\delta_l; -\varpi\varsigma R_{ch}^{-\alpha_l/2} m_l) \Big), \quad (8.14)$$

where $c_k = \frac{\delta_l a_k \Gamma(k+m_l) m_l^{-k}}{(\delta_l - k)\Gamma(k+1)\Gamma(m_l)}$, $a_k = (\varpi\varsigma R_{ch}^{-\alpha_l/2})^k$, $\varsigma = P_t A_l G_s$, $\delta_l = \frac{2}{\alpha_l}$, $R_{ch}^2 = R_c^2 + h^2$, $\boldsymbol{1}\{.\}$ is the indicator function, and ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the ordinary hypergeometric function.

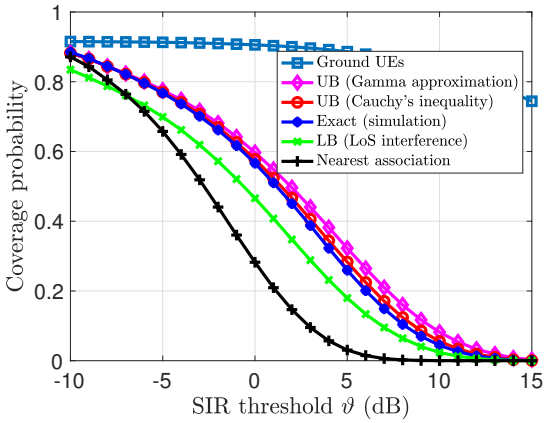
Proof. Please see Appendix F.2. □

For comparison purposes, next, we derive the UAV-UE coverage probability under the nearest association scheme.

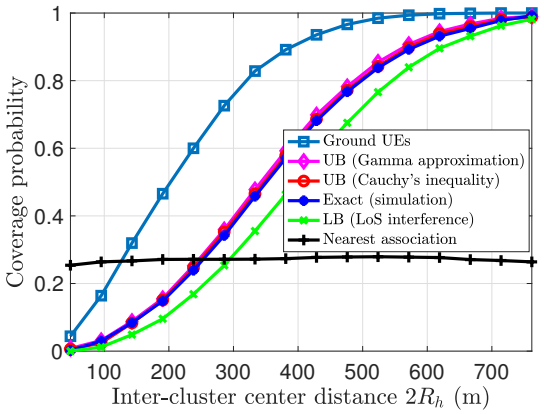
Corollary 8.3.0.2. *The coverage probability of the UAV-UEs under the nearest association scheme is:*

$$\mathbb{P}_{\mathrm{c}} = \int_0^\infty \mathbb{P}^{l}_{\mathrm{c}|r_0} f_{R_0}(r_0)\, \mathrm{d}r_0\,, \quad (8.15)$$

where $\mathbb{P}^{l}_{\mathrm{c}|r_0} = \|e^{\boldsymbol{T}_{m_l}}\|_1$, $\boldsymbol{T}_{m_l}$ is defined as $\boldsymbol{T}_K$ in (8.13), with $\Omega(\varpi) = -2\pi\lambda_b \int_{\upsilon=r_0}^\infty \Big(1 - \delta_l \mathbb{P}_l(\upsilon) - \delta_n \mathbb{P}_n(\upsilon)\Big)\upsilon\, \mathrm{d}\upsilon$, $\varpi = \frac{\vartheta m_l}{P_t \zeta_l(r_0)}$, and $f_{R_0}(r_0) = 2\pi\lambda_b r_0 e^{-\pi\lambda_b r_0^2}$ is the 2D serving distance PDF.



(a) Inter-cluster half distance $R_h = 190\,\mathrm{m}$



(b) SIR threshold $\vartheta = 0\,\mathrm{dB}$

FIGURE 8.3: The derived upper and lower bounds on the coverage probability of UAV-UEs are plotted versus the SIR threshold ϑ and collaboration distance R_c: $\lambda_b = 20\,\mathrm{km}^{-2}$, $R_{\mathrm{sim}} = 20\,\mathrm{km}^2$, $\alpha_l = 2.09$, $\alpha_n = 3.75$, $h_{\mathrm{BS}} = 30\,\mathrm{m}$, $m_l = 3$, $A_l = 0.0088$, $A_n = 0.0226$, $h_d = 120\,\mathrm{m}$, $a = 0.3$, $b = 300\,\mathrm{km}^{-2}$, and $c = 20\,\mathrm{m}$.

Proof. The proof follows directly from [139] and Theorem 8.3.0.1, and hence is omitted. □

To verify the accuracy of our proposed approach, in Fig. 8.3, we show the theoretical UB and LB on the coverage probability of the UAV-UEs, and simulation of the UB based on (8.9). Monte-carlo simulation is adopted where the overall channel gain is obtained from simulating the aerial users and accordingly calculating the coverage probability of the UAV-UEs. Fig. 8.3(a) shows that the Cauchy's inequality-based UB is remarkably tight. Moreover, although the obtained UB expression in (8.13) is less tight, it still represents a reasonably tractable bound on the

exact coverage probability. Hence, (8.13) can be treated as a proxy of the exact result. Recall that (8.9) is based on an UB on a square of a sum of $\mathrm{Nakagami}$-m_l RVs while the expression in (8.13) goes further by two more steps. First, we approximate the sum of Gamma RVs to an equivalent Gamma RV. Then, we round the shape parameter of the yielded Gamma RV to an integer $m_l\kappa$. Finally, the LB based on (8.14) can be also seen as a relatively looser bound than the UBs. As evident from Fig. 8.3, allowing CoMP transmission significantly enhances the coverage probability, e.g., from 28% for the baseline scenario with nearest serving BSs to 60% at $\vartheta = -5\,\mathrm{dB}$ (for an average of 2.5 cooperating BSs). In Fig. 8.3, the performance of UAV-UEs is also compared to that of their ground counterparts experiencing Rayleigh fading and NLoS communications. We assume that the BSs' antennas are ideally down-tilted accounting for the ground UEs, i.e., the antenna gains for desired and interfering signals are G_m and G_s, respectively. Under such a setup, we observe that the coverage probability of ground UEs substantially outperforms that of UAV-UEs, especially at high SIR thresholds. Fig. 8.3(b) shows the prominent effect of the collaboration distance R_c on the coverage probability of ground and UAV-UEs. We can see that for both kind of UEs, the coverage probability monotonically increases with R_c since more BSs cooperate to serve the UEs when R_c increases. Moreover, due to the down-tilt of the BSs' antennas and LoS-dominated interference for UAV-UEs, the coverage probability of ground UEs outperforms that of the UAV-UEs. However, we can see that the rate of coverage probability improvement with R_c, i.e., the slope, is slightly higher for the UAV-UE. This can be interpreted by the fact that increasing R_c yields more LoS signals on the desired signal side and subtracts them from the interference. Conversely, for ground UEs, the transmission is dominated by NLoS signals and Rayleigh fading.

To show the effect of content availability, i.e., content caching, in Fig. 8.4, we plot the coverage probability versus the SIR threshold ϑ for different c_f. We observe that the coverage probability decreases as the caching probability c_f decreases. This stems from the fact that the average number of caching SBSs decreases as c_f decreases. This in turn reduces the cooperative transmission gain. Note that the value c_f is, in fact, a parameter that can be designed based on various factors such as the memory size of SBSs, the popularity of files, and file library size.

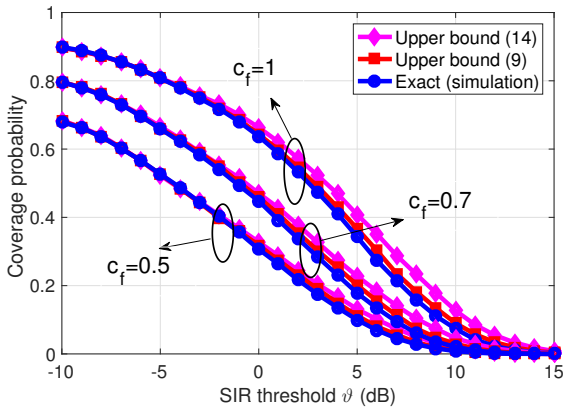


FIGURE 8.4: Coverage probability versus SIR threshold ϑ for different content caching probability c_f.

Having characterized the performance of static UAV-UEs, next, we turn our attention to applications in which the UAV-UEs can be mobile. It is anticipated that mobile UAV-UEs will span a wide variety of applications, e.g., flying taxis and delivery drones. Hence, it is quite important to ensure reliable connections in the presence of UAV-UE mobility by potentially mitigating the LoS interference through CoMP transmissions. Moreover, unlike the ground UEs that can only move horizontally, UAV-UEs can fly in 3D space. Hence, a 3D mobility model is essential to convey a realistic description of the performance of mobile UAV-UEs. As a first step in this direction, we develop a novel 3D RWP mobility model that effectively captures the vertical displacement of UAV-UEs, along with their typical 2D spatial mobility. The use of RWP mobility is motivated by its simplicity and tractability that is widely adopted in the mobility analysis in cellular networks [177–180]. Moreover, as we will discuss shortly, it has tunable parameters that can be set to appropriately describe the mobility of different mobile nodes, ranging from walking or driving users to 3D UAVs, [155] and [180]. For simplicity of analysis, we next assume that the required content is always available at the ground BS, i.e., we assume that $c_f = 1$.

8.4 3D Mobility and Handover Analysis

Recently, the support of mobile drones such as UAV-UEs and UAV-BSs has been explored in 3GPP standardization efforts [152]. Particularly, there has been an

increasing interest in the community to characterize the effect of the mobility of drones on their performance, whether they are cellular-connected UAVs or aerial BSs [148, 151, 154, 155, 170, 171]. Motivated by this, we develop a novel stochastic 3D mobility model for multiple mobile UAV-UEs in random networks. Our proposed model provides a rich set of combinations to study various real-world UAV-UE deployment scenarios. Particularly, the proposed mobility framework could effectively model scenarios in which the information about the locations of UAVs exhibits randomness or uncertainty. For instance, in applications such as target scanning for reconnaissance, search and rescue, or random arrival of package delivery drones, the precise information about different strategic locations of multiple drones might be unavailable [155]. Moreover, in applications that rely on mobile, autonomous UAV-UEs in which drones intelligently update their locations based on the dynamics of their environments, the trajectories of the UAVs can neither be pre-planned nor fully deterministic [169]. This indeed resembles another scenario that is effectively characterized by stochastic mobility models.

Next, we illustrate the various elements of our proposed model. Then, we characterize the handover rate and handover probability for mobile UAV-UEs. Since we introduce the first study on 3D mobile UAV-UEs, for completeness, we consider two cases: UAV-UEs under CoMP transmissions, and UAV-UEs served by the nearest ground BS.

In the classical 2D mobility model, the spatial motion is considered only through a displacement and an angle. However, for the UAV-UE, due to the mission requirements, and environmental and atmospherical conditions, the UAV-UEs must change their altitude and make vertical motions. For instance, due to variations in the altitudes of buildings, UAV-UEs might have frequent up and down displacements along their trajectories. Indeed, the vertical motion is always associated with the take-off and landing of UAV-UEs. This inherently triggers the concept of 3D mobility in 3D space.[2]

First, recall that in a classical RWP mobility model [177–180], the movement trace of a node (e.g., the UAV-UE) can be formally described by an infinite sequence

[2]We assume that the UAV-UEs are sparsely deployed such that there are no imposed constraints on the trajectories of different UAV-UEs. The analysis of multiple trajectories with such constraints is interesting but beyond the scope of this chapter.

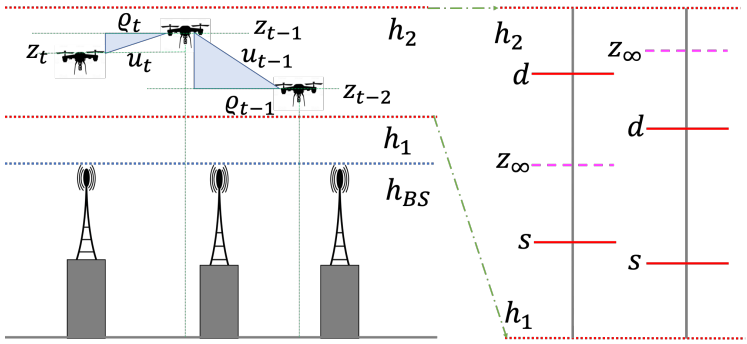


FIGURE 8.5: A sample trace of the proposed 3D RWP mobility model, and an illustration of the 1D RWP vertical mobility of [155].

of tuples: $\{(\boldsymbol{X}_{t-1}, \boldsymbol{X}_t, V_t)\}$, $\forall \boldsymbol{X}_t \in \mathbb{R}^3$, and $t \in \mathbb{N}$, where t is the movement epoch and $\boldsymbol{X}_t = (\varrho_t, \phi_t, z_t)$ is the 3D cylindrical displacement of the UAV-UE at epoch t, see Fig. 8.5. During the t-th movement epoch, $\boldsymbol{X}_{t-1}$ denotes the starting waypoint, $\boldsymbol{X}_t$ denotes the target waypoint, and V_t is the UAV-UE speed that follows a generalized PDF $f_{V_t}(\nu_t)$. Given the current waypoint $\boldsymbol{X}_{t-1}$, the next waypoint $\boldsymbol{X}_t$ is chosen such that the included angle ϕ_t between the projection of the vector $\boldsymbol{X}_{t-1} - \boldsymbol{X}_t$ on the x-y plane and the abscissa is uniformly distributed on $[0, 2\pi]$. We define the transition length as the Euclidean distance between two 3D successive waypoints, i.e., $u_t = \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\| = \sqrt{\varrho_t^2 + (z_t - z_{t-1})^2}$. Furthermore, the vertical displacement between two consecutive points, i.e., the change in z-axis, is also distributed according to a RV. We also let φ_t be the acute angle of $u_t = \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|$ relative to the horizontal line ρ_t.

The selection of waypoints is assumed to be independent and identical for each movement epoch, and there is no pause time at these waypoints [178]. Particularly, similar to [180], the horizontal transition lengths $\{\rho_1, \rho_2, \dots\}$ are chosen to be i.i.d. with the CDF $F_{\rho_t}(\varrho_t) = 1 - \exp(-\pi\mu\varrho_t^2)$, i.e., the spatial transition lengths are Rayleigh distributed in $\mathbb{R}^2$ with mobility parameter μ. The corresponding displacement PDF is hence $f_{\rho_t}(\varrho_t) = \frac{\partial F_{\rho_t}(\varrho_t)}{\partial \varrho_t} = 2\pi\mu\varrho_t e^{-\pi\mu\varrho_t^2}$. As also done in [155] and [170], we adopt a uniform distribution for the vertical displacement, however, the analysis for generalized PDFs can readily follow. In particular, we assume that Z_t is uniformly distributed on $[h_1, h_2]$, i.e., $Z_t \sim \mathcal{U}(h_1, h_2)$ and $f_{Z_t}(z_t) = \frac{1}{h_2 - h_1}, \forall h_1 \leq z_t \leq h_2$. We henceforth refer to $\hbar = h_2 - h_1$ as altitude difference. Since the major restrictions of all drones' operations are their flying altitudes,

it is reasonable to assume that Z_t is bounded by h_1 and h_2. For instance, UAVs cannot fly higher than certain altitudes (above ground level (AGL)) that are typically chosen below the cruising altitude of manned aircrafts. The UAVs also have an inherent minimum altitude of zero AGL. However, due to mission requirements as well as environmental and atmospherical conditions, it is reasonable to assume $h_1 > 0$. We further assume that $h_1 > h_{\mathrm{BS}}$ for a high altitude UAV-UEs scenario. For simplicity, we assume that the UAV-UE moves with a constant speed $\bar{\nu}$. However, the analysis for generalized PDFs $f_{V_t}(\nu_t)$ of the speed can be done by using the same methodology. Finally, for the 3D displacement, we have $u_t = \sqrt{\varrho_t^2 + (z_t - z_{t-1})^2}$.

Under the proposed mobility model, several mobility patterns can be captured by choosing different mobility parameters and altitude thresholds. For example, larger values of μ statistically imply that the 2D and, consequently, 3D transition lengths are shorter. This means that the movement direction switch rates are higher. These values of the mobility parameter appropriately describe the motion of UAV-UEs frequently traveling between nearby hovering locations such as for the use case of aerial surveillance cameras. In contrast, a smaller μ statistically implies that the 2D and 3D transition lengths are longer and the corresponding movement direction switch rates are lower. These values of μ would be suitable to describe the motion of UAV-UEs traveling large distances such as for the use case of flying taxis and delivery drones. As a special case, if there is no change in the UAV-UE altitude along its trajectory, i.e., $h_1 = h_2 = h_d$, then the UAV-UE would move along straight lines in random directions with a constant speed, which is perhaps the simplest mobility model. This model is known to provide performance bounds and useful insights in wireless networks [154] and [181]. This model has also been recently adopted to model drone mobility in 3GPP studies related to drone networks [152]. In contrast, if the vertical mobility is solely considered, our model can efficiently describe the UAV-UE movements during take off and landing [11] and [182].

Given the independence assumption between Z_t and ρ_t, we obtain their joint PDF from $f_{\rho_t,Z_t}(\varrho_t, z_t) = f_{\rho_t}(\varrho_t) f_{Z_t}(z_t) = \frac{2\pi\mu\varrho_t}{h} e^{-\pi\mu\varrho_t^2}$, $\forall h_1 \leq z \leq h_2, 0 \leq \varrho \leq \infty$. Consequently, the PDF of the 3D displacement $f_{U_t}(u_t)$ is readily obtained in the next lemma.

Lemma 8.4.0.1. *The PDF of the 3D transition lengths U_t is given by $f_{U_t}(u_t) = 2\pi\mu u_t e^{-\pi\mu u_t^2}\Omega(\mu,\hbar)$, where $\Omega(\mu,\hbar) = \frac{\pi\hbar\sqrt{\mu}\text{erfi}\left(\sqrt{\pi\mu}\hbar\right)-e^{\pi\mu\hbar^2}+1}{\pi\mu\hbar^2}$, and $\text{erfi}(.) = \frac{-2i}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$.*

Proof. We can reach this result by transforming the RVs Z_t, Z_{t-1}, and ρ_t to U_t, where $u_t = \sqrt{\varrho_t^2 + (z_t - z_{t-1})^2}$, with the details omitted. □

Remark 8.4.0.1. If $h_1 = h_2$, it can be easily verified that $\lim_{\hbar\to 0}\Omega(\mu,\hbar) \to 1$, and $f_{U_t}(u_t) = 2\pi\mu u_t e^{-\pi\mu u_t^2}$. This shows that if the UAV-UE moves only along a horizontal plane, the PDF of the 3D displacement distance is reduced to its 2D counterpart, which verifies the correctness of the obtained 3D displacement distribution $f_{U_t}(u_t)$.

Having described the various elements of our proposed 3D RWP model, our immediate objective is to characterize the handover rate and handover probability for mobile UAV-UEs under CoMP transmissions and nearest association.

8.4.1 Handover Rate and Handover Probability for Nearest Association

Assume that a mobile UAV-UE is located at $\boldsymbol{X}_{t-1}$ and let $\boldsymbol{X}_{t-1}$ and $\boldsymbol{X}_t$ be two arbitrary successive waypoints. The handover rate is defined as the expected number of handovers per unit time. Hence, inspired from [180], we can compute the handover rate as follows. We first condition on an arbitrary position of the mobile UAV-UE $\boldsymbol{X}_t = \boldsymbol{x}_t$, and a given realization of the Poisson-Voronoi tessellation Φ_b. Subsequently, the number of handovers will be equal to the number of intersections of the UAV-UE trajectory and the boundary of the Poisson-Voronoi tessellation. Then, by averaging over the spatial distribution of $\boldsymbol{X}_t$ and the distribution of Poisson-Voronoi tessellation, we derive the expected number of handovers. Alternatively, we notice that the number of handovers is equivalent to the number of intersections of the Poisson-Voronoi tessellation and the *horizontal projection of the segment $\left[\boldsymbol{X}_{t-1}, \boldsymbol{X}_t\right]$ on the x-y plane*. Therefore, following [180, 183, 184], the expected number of handovers during one movement epoch will be: $\mathbb{E}\big[N\big] = \frac{2}{\pi}\sqrt{\frac{\lambda_b}{\mu}}$. The handover rate is then the ratio of the expected number of handovers during one movement $\mathbb{E}\big[N\big]$ to the mean time of one transition movement $\mathbb{E}\big[T\big]$. Since we have $\mathbb{E}[T] = \mathbb{E}[\frac{U_t}{V}] = \frac{\mathbb{E}[U_t]}{\bar{\nu}} = \frac{\Omega(\mu,\hbar)}{2\sqrt{\mu\bar{\nu}}}$, where $\mathbb{E}[U_t] = \frac{\Omega(\mu,\hbar)}{2\sqrt{\mu}}$, then, the handover rate

will be:

$$H = \frac{\mathbb{E}[N]}{\mathbb{E}[T]} = \frac{2}{\pi}\sqrt{\frac{\lambda_b}{\mu}}\Bigg/\frac{\Omega(\mu,\hbar)}{2\sqrt{\mu\bar{\nu}}} = \frac{4\bar{\nu}\sqrt{\lambda_b}}{\pi\Omega(\mu,\hbar)}. \tag{8.16}$$

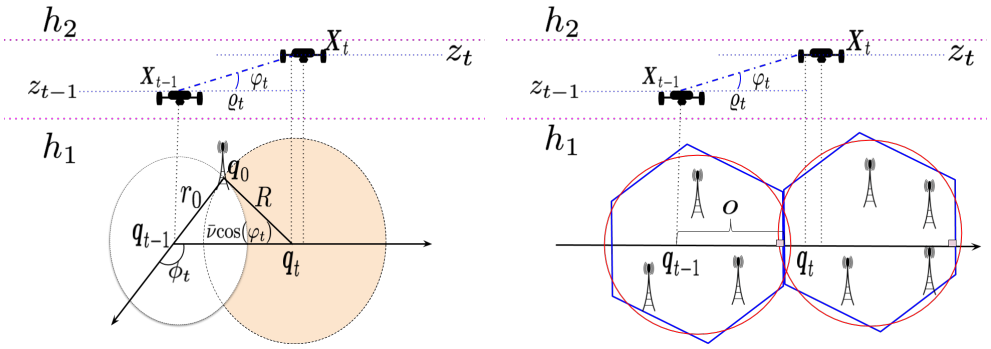
Remark 8.4.1.1. Unlike the handover rate for 2D RWP [180, 183, 184], H in (8.16) is a function of the mobility parameter μ through $\Omega(\mu,\hbar)$. This captures the fact that, in the case of an UAV-UE, since each stochastically generated horizontal displacement is accompanied with a vertical one, the handover rate depends on μ that affects the vertical displacement switch rates.

Next, to characterize the coverage probability of mobile UAV-UEs, we use the concept of *handover probability*. Similar to [184] and [185], given the current location of a mobile UAV-UE, the handover probability is defined as the probability that there exists a BS closer than the serving BS after a unit time. From Fig. 8.6(a), for two arbitrary consecutive waypoints $\boldsymbol{X}_{t-1} = (\varrho_{t-1}, \phi_{t-1}, z_{t-1})$ and $\boldsymbol{X}_t = (\varrho_t, \phi_t, z_t)$, the horizontal speed of the UAV-UE from waypoint $\boldsymbol{X}_{t-1}$ to waypoint $\boldsymbol{X}_t$ is a RV whose realization is $V_h = \bar{\nu}\cos(\varphi_t)$, where $\varphi_t = \arccos\Big(\frac{\varrho_t}{\sqrt{\varrho_t^2+(z_t-z_{t-1})^2}}\Big)$. It is also assumed that the angle ϕ_t is taken with respect to the direction of connection as shown in Fig. 8.6(a). Define $\boldsymbol{q}_{t-1}$ and $\boldsymbol{q}_t$ in $\mathbb{R}^2$ as the horizontal projections of $\boldsymbol{X}_{t-1}$ and the location reached by the UAV-UE after a unit time, respectively. Fig. 8.6(a) illustrates that the UAV-UE is first associated with its nearest BS located at $\boldsymbol{q}_0$, i.e., there are no BSs in the ball of radius $r_0 = \|\boldsymbol{q}_{t-1} - \boldsymbol{q}_0\|$ centered at $\boldsymbol{q}_{t-1}$. Using the law of cosines, $\boldsymbol{q}_t$ is at distance $R = \sqrt{r_0^2 + (\bar{\nu}\cos(\varphi_t))^2 + 2r_0(\bar{\nu}\cos(\varphi_t))cos(\phi_t)}$ from the BS located at $\boldsymbol{q}_0$.[3]

The handover occurs only if another BS becomes closer to $\boldsymbol{q}_t$ than the serving BS located at $\boldsymbol{q}_0$, i.e., when there is at least one BS in the shaded area in Fig. 8.6(a). Therefore, given $\{r_0, z_{t-1}, z_t, \varrho_t, \phi_t\}$, the conditional probability of handover is $\mathbb{P}(H|r_0, z_{t-1}, z_t, \varrho_t, \phi_t)$

$$\begin{aligned}
&= \mathbb{P}\Big(\mathcal{B}(\boldsymbol{q}_t, R) \setminus \mathcal{B}(\boldsymbol{q}_{t-1}, r_0) > 0 | r_0, z_{t-1}, z_t, \varrho_t, \phi_t\Big) \overset{(a)}{=} 1 - e^{-\lambda_b\left|\mathcal{B}(\boldsymbol{q}_t,R)\setminus\mathcal{B}(\boldsymbol{q}_{t-1},r_0)\right|} \\
&= 1 - e^{-\pi\lambda_b(R^2-r_0^2)} = 1 - e^{-\pi\lambda_b\left(r_0^2+(\bar{\nu}\cos(\varphi_t))^2+2r_0\bar{\nu}\cos(\varphi_t)cos(\phi_t)-r_0^2\right)},
\end{aligned} \tag{8.17}$$

[3]Since the UAV-UE starts from waypoint $\boldsymbol{X}_{t-1}$, we assume that it does not change its direction in a time shorter than the unit time. Hence, q_t is assumed to be within the segment $[\boldsymbol{X}_{t-1}, \boldsymbol{X}_t]$ in Fig. 8.6.



(a) Nearest association handover scenario (b) Inter-CoMP handover scenario

FIGURE 8.6: The probability of handover is computed based on the network geometry.

where (a) follows from the void probability of PPP. $\mathcal{B}(\boldsymbol{q}_t, R)$ represents the ball with radius R centered at $\boldsymbol{q}_t$ and $\mathcal{B}(\boldsymbol{q}_{t-1}, r_0)$ is excluded from $\mathcal{B}(\boldsymbol{q}_t, R)$ since the BS located at $\boldsymbol{q}_0$ is the nearest BS to $\boldsymbol{q}_{t-1}$. Finally, averaging over Z_{t-1}, Z_t, ρ_t, and ϕ_t, where $\phi_t \sim \mathcal{U}(0, 2\pi)$, we get

$$\mathbb{P}(H|r_0) = 1 - \mathbb{E}_{\rho_t, Z_t, Z_{t-1}, \phi_t}\left[e^{-\pi\lambda_b\left((\bar{\nu}\cos(\varphi_t))^2 + 2r_0(\bar{\nu}\cos(\varphi_t))cos(\phi_t)\right)}\right]. \tag{8.18}$$

For the special case in which the UAV-UE moves radially away from the serving BS, i.e., $\phi_t = 0$, next, we obtain a tractable yet accurate UB on the conditional handover probability. This assumption is reasonable, particularly, if the UAV-UE follows a horizontally-direct path subject only to vertical fluctuations due to mission, environmental, and atmospherical conditions.

Lemma 8.4.1.1. *An UB on the conditional probability of handover is given by*

$$\mathbb{P}(H|r_0) = 1 - e^{-\frac{2\lambda_b r_0 \bar{\nu}}{\sqrt{\pi}\mu\hbar^2}\psi(\mu,\hbar)} e^{-\pi\lambda_b\zeta(\mu,\hbar)}, \tag{8.19}$$

where $\psi(\mu,\hbar) = \pi\hbar^2\mu G^{2,2}_{2,3}\left(\hbar^2\pi\mu \middle| \begin{matrix} \frac{1}{2}, \frac{1}{2} \\ 0, 1, -\frac{1}{2} \end{matrix}\right) - G^{2,2}_{2,3}\left(\hbar^2\pi\mu \middle| \begin{matrix} 1, \frac{3}{2} \\ 1, 2, 0 \end{matrix}\right)$, $G^{m,n}_{p,q}$ *denotes the Meijer G function, defined as*

$$G^{m,n}_{p,q} = \left(x \middle| \begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix}\right) = \frac{1}{2\pi i}\int \frac{\prod_{j=1}^{m}\Gamma(b_j + s)\prod_{j=1}^{n}\Gamma(1 - a_j + s)}{\prod_{j=n+1}^{p}\Gamma(a_j + s)\prod_{j=m+1}^{q}\Gamma(1 - b_j + s)} x^s \, \mathrm{d}s\,, \tag{8.20}$$

and $\zeta(\mu,\hbar)=\bar{\nu}^2\Big(1-\frac{2\pi\mu}{\hbar^2}\int_0^{\hbar}(\hbar-p)p^2e^{\pi\mu p^2}\Gamma\left(0,\pi p^2\mu\right)\mathrm{d}p\Big)$.

Proof. Please see Appendix F.3. □

From (8.19), it is intuitive to see that $\mathbb{P}(H|r_0)$ increases with $\bar{\nu}$ and λ_b because there will be a higher probability of handover when the UAV-UE speed is higher, and the network is denser. Moreover, $\mathbb{P}(H|r_0)$ decreases as the term $\mu\hbar^2$ increases. This reveals important insights on the effect of the altitude difference $\hbar$ and the density μ. Particularly, the handover probability decreases when the UAV-UE jointly has higher direction switch rates (higher μ) and larger altitude difference $\hbar$. Next, we obtain the handover rate for UAV-UEs under CoMP transmissions.

8.4.2 Inter-CoMP Handover Rate and Handover Probability

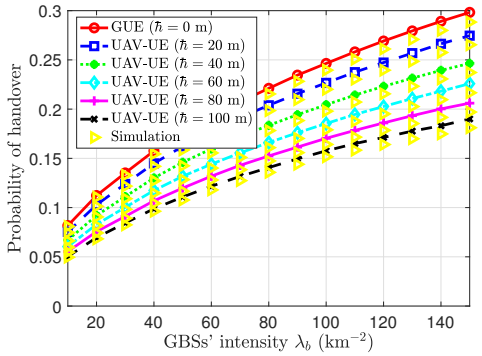
We define the number of handovers $\mathbb{E}[N]$ as the number of intersections of the horizontal projection of the segment $\left[\boldsymbol{X}_{t-1},\boldsymbol{X}_t\right]$ and the boundaries of disjoint clusters whose inter-cluster center distance is $2R_h$, as discussed in Section 8.2. The hexagonal cell has six sides of length $\ell=\frac{2R_h}{\sqrt{3}}$. Following the Buffon's needle approach for hexagonal cells [184], we next obtain the inter-CoMP handover rate.

Proposition 8.4.2.1. *The inter-CoMP handover rate for a network of disjoint clusters whose inter-cluster center distance is* $2R_h$ *is given by*

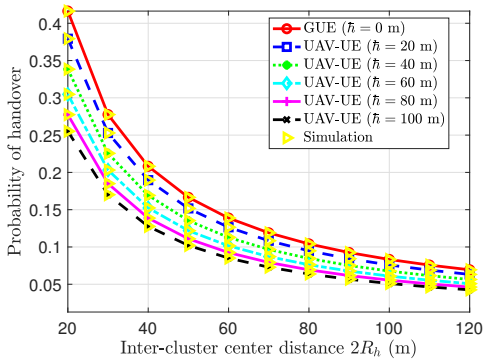
$$H=\frac{\mathbb{E}[N]}{\mathbb{E}[T]}=2\frac{\pi\hbar^2\mu G_{2,3}^{2,2}\Big(\hbar^2\pi\mu\Big|\begin{matrix}\frac{1}{2},\frac{1}{2}\\0,1,-\frac{1}{2}\end{matrix}\Big)-G_{2,3}^{2,2}\Big(\hbar^2\pi\mu\Big|\begin{matrix}1,\frac{3}{2}\\1,2,0\end{matrix}\Big)}{\pi\sqrt{\pi}R_h\hbar^2\mu}\bar{\nu}. \tag{8.21}$$

Proof. Please see Appendix F.4. □

We now characterize the UAV-UE inter-CoMP handover probability. To keep the analysis simple, we consider a special case in which the UAV-UE moves perpendicularly to the inter-cluster boundaries, as shown in Fig. 8.6(b). As discussed in Section 8.4.1, this is a practical assumption for a UAV-UE that follows a horizontally straight path subject only to vertical fluctuations. Moreover, since these boundaries represent virtual borders between disjoint clusters, the assumption that such boundaries are in a direction perpendicular to the UAV-UE trajectory is quite



(a) Nearest association scenario



(b) CoMP transmission scenario

FIGURE 8.7: The probability of handover is plotted versus network parameters for nearest association and CoMP transmission schemes ($\bar{\nu} = 30\,\text{kmh}$, $\mu = 300\,\text{km}^{-2}$, $h_1 = 100\,\text{m}$).

reasonable. Hence, the UAV-UE moves by a horizontal distance $\bar{\nu}\cos(\varphi_t)$ in a unit of time in a direction perpendicular to the inter-cluster boundaries. A handover occurs if the traveled distance in the horizontal direction is larger than the distance o to the cluster side, which is a realization of RV O whose PDF is $f_O(o)$. Conditioning

on $O = o$, the handover probability is formally stated as:

$$\begin{aligned}
\mathbb{P}(H|o) &= \mathbb{P}\big(\frac{\bar{\nu}\varrho_t}{\sqrt{\varrho_t^2 + (z_t - z_{t-1})^2}} > o\big) = \mathbb{P}\big(\varrho_t > \frac{o(z_t - z_{t-1})}{\sqrt{\bar{\nu}^2 - o^2}}\big) \\
&= \mathbb{E}_{Z_t, Z_{t-1}} \int_{\varrho_t = \frac{o(z_t - z_{t-1})}{\sqrt{\bar{\nu}^2 - o^2}}}^{\infty} 2\pi\mu\varrho_t e^{-\pi\mu\varrho_t^2} \, \mathrm{d}\varrho_t \qquad (8.22) \\
&\overset{(a)}{=} \mathbb{E}_{Z_t, Z_{t-1}} e^{-\pi\mu\left(\frac{o^2(z_t - z_{t-1})^2}{\bar{\nu}^2 - o^2}\right)} \\
&\overset{(b)}{=} \mathbb{E}_p e^{-\pi\mu p^2 \frac{o^2}{\bar{\nu}^2 - o^2}} \\
&\overset{(c)}{=} \frac{1}{\hbar^2} \frac{\hbar\sqrt{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}\mathrm{erf}\left(\sqrt{\pi}\hbar\sqrt{\frac{o^2}{\bar{\nu}^2 - o^2}}\sqrt{\mu}\right) + \frac{e^{-\pi\mu\hbar^2\frac{o^2}{\bar{\nu}^2 - o^2}} - 1}{\pi}}{\frac{\mu o^2}{\bar{\nu}^2 - o^2}} \qquad (8.23) \\
&= \frac{1}{\hbar} \frac{\mathrm{erf}\left(\sqrt{\pi}\hbar\sqrt{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}\right)}{\sqrt{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}} + \frac{1}{\pi\hbar^2} \frac{e^{-\pi\mu\hbar^2\frac{o^2}{\bar{\nu}^2 - o^2}} - 1}{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}, \qquad (8.24)
\end{aligned}$$

where (a) follows from solving the integral of (8.22), (b) follows from change of variables $p = z_t - z_{t-1}$, with $f_P(p) = \frac{\hbar - |p|}{\hbar^2}, -\hbar \leq p \leq \hbar$, and (c) follows from taking the expectation with respect to (w.r.t.) p. Note that O is a uniform RV in $[0, 2R_h]$ that models the distance between the UAV-UE and the inter-cluster boundaries, see Fig. 8.6(b). Averaging over O given that $f_O(o) = \frac{1}{2R_h}, 0 < o < 2R_h$, we get $\mathbb{P}(H)$. However, we observe that if $o > \bar{\nu}$, the handover probability will be zero since the UAV-UE can not travel the distance o in a unit of time, hence, we have

$$\mathbb{P}(H) = \frac{1}{2R_h\hbar} \int_{o=0}^{\bar{\nu}} \frac{\mathrm{erf}\left(\sqrt{\pi}\hbar\sqrt{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}\right)}{\sqrt{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}} + \frac{1}{2\pi R_h\hbar^2} \int_{o=0}^{\bar{\nu}} \frac{e^{-\pi\mu\hbar^2\frac{o^2}{\bar{\nu}^2 - o^2}} - 1}{\frac{\mu o^2}{\bar{\nu}^2 - o^2}}. \qquad (8.25)$$

Fig. 8.7 verifies the accuracy of the obtained handover probabilities. Fig. 8.7(a) presents the handover probability versus BSs' intensity λ_b under the nearest association scheme. The figure shows that the obtained UB in (8.19) is quite tight. It is also noted that as long as the UAV-UE has frequent vertical movements, i.e., larger $\hbar$, the handover probability is lower since the effective horizontal travelled distance becomes shorter. The handover probability also monotonically increases with λ_b since a higher rate of handover occurs for denser networks. Fig. 8.7(b) shows the inter-CoMP handover probability versus the inter-cluster center distance $2R_h$. The

handover probability monotonically decreases with R_h since a lower rate of handover is anticipated when the cluster size increases. Next, we will use our proposed RWP model to obtain the coverage probability of 3D mobile UAV-UEs.

8.5 Coverage Probability of Mobile UAV-UEs

Next, we will employ the handover probabilities in (8.19) and (8.25) to quantify the coverage probability of mobile UAV-UEs under the nearest association and CoMP transmissions, respectively. It is worth highlighting that (8.13) represents the probability that a static UAV-UE is in coverage with neither mobility nor handover considered. However, mobile UAV-UEs are susceptible to frequent handovers that would negatively impact their performance. For instance, handover typically results in dropped connections and causes longer service delays. In fact, higher handover rates lead to a higher risk of QoS degradation.

To account for the user mobility, similar to [184–186], we consider a linear function that reflects the cost of handovers. Under this model, the UAV-UE coverage probability can be defined as:

$$P_c(\bar{\nu}, \mu, \beta) = \mathbb{P}(\Upsilon \geq \vartheta, \bar{H}) + (1-\beta)\mathbb{P}(\Upsilon \geq \vartheta, H), \tag{8.26}$$

where the first term represents the probability that the UAV-UE is in coverage and no handover occurring. The second term is the probability that the UAV-UE is in coverage and handover occurs penalized by a handover cost, where $\beta \in [0,1]$ represents the probability of connection failure due to handover. The coefficient β, in effect, measures the system sensitivity to handovers, which highly depends on the hysteresis margin and ping-pong rate [183–186]. Our goal is to obtain the coverage probability of a mobile UAV-UE for a given handover penalty β [184]. After some manipulations, we can rewrite (8.26) as

$$P_c(\bar{\nu}, \mu, \beta) = (1-\beta)\mathbb{P}(\Upsilon \geq \vartheta | r_0) + \beta\mathbb{P}(\Upsilon \geq \vartheta, \bar{H} | r_0). \tag{8.27}$$

To obtain $P_c(\bar{\nu}, \mu, \beta)$, we first need to calculate the statistical distribution of the UAV-UE altitude for our proposed 3D mobility model. As shown in Fig. 8.5, the

3D mobility model defines the vertical movement of the UAV-UE in a finite region $[h_1, h_2]$, *referred to as vertical one-dimensional (1D) RWP* as in [155]. Initially, at time instant t_0, the UAV-UE is at an arbitrary altitude h_0 selected uniformly from the interval $[h_1, h_2]$. Then, at next time epoch t_1, this UAV-UE at h_0 selects a new random waypoint h_1 uniformly in $[h_1, h_2]$, and moves towards it (along with the spatial movement characterized by $f_{\rho_t}(\varrho_t)$). Once the UAV-UE reaches h_1, it repeats the same procedure to find the next destination altitude and so on. After a long running time, the steady-state altitude distribution converges to a *nonuniform distribution* $F_{Z_\infty}(z_\infty)$ [177], where Z_∞ is a RV representing the steady state vertical location of the UAV-UE. Note that random waypoints refer to the altitude of a UAV-UE at each time epoch, which is uniformly-distributed in $[h_1, h_2]$, while vertical transitions are the differences in the UAV-UE altitude throughout its trajectory. While the random waypoints are independent and uniformly distributed by definition, the random lengths of vertical transitions are not statistically independent. This is because the endpoint of one movement epoch is the starting point of the next epoch. In [177], it is shown that $F_{Z_\infty}(z_\infty) = \frac{\mathbb{E}[L_{z_\infty}]}{\mathbb{E}[L]}$, where L_{z_∞} and L denote the length $\|z_\infty - h_1\|$, and the entire movement length at any given epoch, respectively. From [177], we have $\mathbb{E}[L] = \frac{\hbar}{3}$ and $\mathbb{E}[L_{z_\infty}]$ can be similarly derived from:

$$\mathbb{E}[L_{z_\infty}] = \int_{s=h_1}^{h_2} \int_{d=h_1}^{h_2} l_{z_\infty}(s,d) f_S(s) f_D(d) \,\mathrm{d}d \,\mathrm{d}s\,, \tag{8.28}$$

where s and d refer to the source and destination of a movement, respectively; $f_S(s) = f_D(d) = \frac{1}{\hbar}$, $h_1 \le s, d \le h_2$, see the right side of Fig. 8.5. $l_{z_\infty}(s,d)$ denotes the value of the random variable L_{z_∞} if $S = s$ and $D = d$. Because of the symmetry of s and d, it is sufficient to restrict the calculation to epochs with $s < d$, and then multiply the result by a factor of two. A necessary condition for $l_{z_\infty}(s,d) \neq 0$ is that $s \le z_\infty$. From Fig. 8.5, if $d \le z_\infty$, we have $l_{z_\infty}(s,d) = d - s$, however, if $d > z_\infty$, we get $l_{z_\infty}(s,d) = z_\infty - s$, which yields

$$\begin{aligned}\mathbb{E}[L_{z_\infty}] &= \frac{2}{\hbar^2} \int_{s=h_1}^{z_\infty} \int_{d=s}^{z_\infty} (d-s) \,\mathrm{d}d \,\mathrm{d}s + \frac{2}{\hbar^2} \int_{s=h_1}^{z_\infty} \int_{d=z_\infty}^{h_2} (z_\infty - s) \,\mathrm{d}d \,\mathrm{d}s \\ &= \frac{2}{\hbar^2} \big(- \frac{h_1^3}{6} + \frac{h_1^2 h_2}{2} - h_1 h_2 z_\infty + \frac{h_1 z_\infty^2}{2} + \frac{h_2 z_\infty^2}{2} - \frac{z_\infty^3}{3} \big). \end{aligned} \tag{8.29}$$

Therefore, the PDF of Z_∞ is given by

$$\begin{aligned} f_{Z_\infty}(z_\infty) &= \frac{\partial F_{Z_\infty}(z_\infty)}{\partial z_\infty} \\ &= \frac{\partial}{\partial z_\infty} \frac{\mathbb{E}[L_{z_\infty}]}{\mathbb{E}[L]} \\ &= \frac{h_1 z_\infty + h_2 z_\infty - h_1 h_2 - z_\infty^2}{\hbar^3/6} \quad \forall h_1 < z_\infty < h_2, \end{aligned} \tag{8.30}$$

and the corresponding mean is given by $\mathbb{E}[Z_\infty] = \frac{1}{2\hbar^3}\left(h_2^4 - h_1^4 + 2h_1^3 h_2 - 2h_1 h_2^3\right)$. Next, we will use the derived PDF $f_{Z_\infty}(z_\infty)$, along with the probability of handover from the previous section, to fully characterize $P_c(\bar{\nu}, \mu, \beta)$ under the nearest association and CoMP transmissions.

8.5.1 Coverage Probability for Nearest Association

Next, we derive the coverage probability of a mobile UAV-UE under the nearest association scheme. Observing (8.27), for a given β, we must compute $\mathbb{P}(\Upsilon \geq \vartheta, \bar{H}|r_0)$ to obtain $P_c(\bar{\nu}, \mu, \beta)$. The former probability is basically the joint event of being in coverage and no handover occurs. We adopt the tight UB on the handover probability obtained in (8.19), where $\mathbb{P}(\bar{H}, r_0) = 1 - \mathbb{P}(H, r_0)$ is the conditional probability of no handover. Unlike static UAV-UEs, under the 3D RWP model, both the altitude of the UAV-UE and the horizontal distance R_0 to the nearest BS are RVs. Since R_0 and Z_∞ are two independent RVs, we have $f_{R_0, Z_\infty}(r_0, z_\infty) = f_{R_0}(r_0) f_{Z_\infty}(z_\infty)$.

We assume that the UAV-UE has an arbitrary long trajectory that passes through nearly all SIR states. Therefore, the average SIR through a randomly selected UAV-UE trajectory is inferred from a stationary PPP analysis. This assumption, which is adopted in [184–186] for ground UEs, is practically reasonable for mobile UAV-UEs such as flying taxis and delivery drones that typically have sufficiently long trajectories. Given the handover probability in (8.19) and the linear function in (8.27), the coverage probability under the nearest association scheme is given below.

Theorem 8.5.1.1. *The coverage probability of a 3D mobile UAV-UE associated with its nearest BS is*

$$P_c(\bar{\nu},\mu,\beta) = 2(1-\beta)\pi\lambda_b \times \int_{h_1}^{h_2}\int_0^{\infty} r_0 e^{-\pi\lambda_b r_0^2}\mathbb{P}^l_{\mathrm{c}|r_0,z_\infty} f_{Z_\infty}(z_\infty)\,\mathrm{d}r_0\,\mathrm{d}z_\infty + 2\beta\pi\lambda_b e^{-\pi\lambda_b\zeta(\mu,\hbar)} \times \int_{h_1}^{h_2}\int_0^{\infty} r_0 e^{-\pi\lambda_b r_0^2} e^{-\frac{2\lambda_b r_0 \bar{\nu}}{\sqrt{\pi}\mu\hbar^2}\psi(\mu,\hbar)}\mathbb{P}^l_{\mathrm{c}|r_0,z_\infty} f_{Z_\infty}(z_\infty)\,\mathrm{d}r_0\,\mathrm{d}z_\infty\,, \tag{8.31}$$

where $\mathbb{P}^l_{\mathrm{c}|r_0,z_\infty} = \|e^{\boldsymbol{T}_{m_l}}\|_1$, $\boldsymbol{T}_{m_l}$ *is defined as* $\boldsymbol{T}_K$ *in (8.13), with* $\Omega(\varpi)_{|r_0,z_\infty} = -2\pi\lambda_b\int_{\upsilon=r_0}^{\infty}\big(1-\delta_l\mathbb{P}_l(\upsilon)-\delta_n\mathbb{P}_n(\upsilon)\big)\upsilon\,\mathrm{d}\upsilon$, $\delta_l = \left(1+\frac{\varpi P_t A_l G_s(\upsilon^2+z_\infty^2)^{-\alpha_l/2}}{m_l}\right)^{-m_l}$, $\delta_n = \left(1+\frac{\varpi P_t A_n G_s(\upsilon^2+z_\infty^2)^{-\alpha_n/2}}{m_n}\right)^{-m_n}$, *and* $\varpi = \frac{\vartheta m_l}{P_t A_l G_s(r_0^2+z_\infty^2)^{-\alpha_l/2}}$; $\psi(\mu,\hbar)$ *and* $\zeta(\mu,\hbar)$ *are given in Lemma 8.4.1.1.*

Proof. The first term in (8.31) is obtained directly from (8.27) and Corollary 8.3.0.2, where the UAV-UE altitude h_d is replaced with the RV z_∞ whose PDF is given in (8.30). Additionally, the second term in (8.31) represents the joint event of no handover and being in coverage, which is computed based on $\mathbb{P}(H|r_0,\phi)$ in (8.19). □

It is clear from (8.31) that, if $\beta = 1$, the first term vanishes and the UAV-UE will be in coverage only if there is no handover associated with its mobility. This is because the handover will always cause connection failure. Moreover, since it is hard to directly obtain insights from (8.31) on the effect of the altitude z_∞ and the altitude difference $\hbar$, several numerical results based on (8.31) will be shown in Section 8.6 to provide key practical insights. Next, we similarly derive the coverage probability of a mobile UAV-UE under CoMP transmission.

8.5.2 Coverage Probability for CoMP Transmission

Similar to Section 8.5.1, we employ the handover probability in (8.25) and the linear function in (8.27) to obtain the coverage probability under CoMP transmission. The probability of inter-cluster handover $\mathbb{P}(H)$ is derived in (8.25) assuming that the UAV-UE moves perpendicular to the cluster boundaries. To compute $\mathbb{P}(\Upsilon \geq \vartheta, \bar{H})$ in (8.27), the joint PDF of the serving distances needs to be characterized given the random location of the UAV-UE along its trajectory. However, for tractability, we consider the joint serving distances when the UAV-UE horizontal projection is at

the cluster center. Therefore, the obtained performance can be seen as an UB on the performance of a randomly located UAV-UE. This assumption is in line with prior work [41] and the analysis for static UAV-UEs, where we sought an UB on the coverage probability.

Since $\boldsymbol{R}_\kappa = [R_1, \ldots, R_\kappa]$ and Z_∞ are independent RVs, their joint PDF is $f_{\boldsymbol{R}_\kappa, Z_\infty}(\boldsymbol{r}_\kappa, z_\infty) = f_{\boldsymbol{R}_\kappa}(\boldsymbol{r}_\kappa) f_{Z_\infty}(z_\infty)$. Given (8.25) and (8.27), an UB on the coverage probability of a mobile UAV-UE under CoMP transmissions is obtained in the next theorem.

Theorem 8.5.2.1. *An UB on the coverage probability of a 3D mobile UAV-UE cooperatively served via CoMP transmission from BSs within a collaboration distance R_c is given by:*

$$\mathbb{P}_{\mathrm{c}} = \big(1 - \beta + \beta \times \mathbb{P}(\bar{H})\big) \sum_{\kappa=1}^{\infty} \mathbb{P}_\kappa \int_{h_1}^{h_2} \int_{\boldsymbol{r}_\kappa = \boldsymbol{R}_c}^{\infty} \mathbb{P}^l_{\mathrm{c}|\boldsymbol{r}_\kappa, z_\infty} f_{Z_\infty}(z_\infty) \prod_{i=0}^{\kappa} \frac{2 r_i}{R_c^2} \, \mathrm{d}\boldsymbol{r}_\kappa \, \mathrm{d}z_\infty \,, \tag{8.32}$$

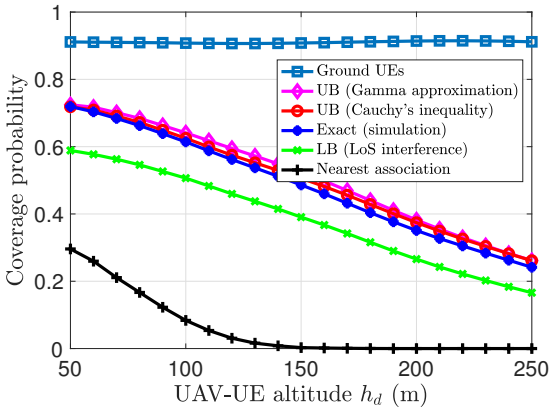
where $\mathbb{P}(\bar{H}) = 1 - \mathbb{P}(H)$ from (8.25), $\mathbb{P}^l_{\mathrm{c}|\boldsymbol{r}_\kappa, z_\infty} = \|e^{\boldsymbol{T}_K}\|_1$, and $\boldsymbol{T}_K$ is as defined in (8.13), with $\Omega(\varpi)_{|\boldsymbol{r}_\kappa, z_\infty} = -2\pi\lambda_b \int_{\upsilon=R_c}^{\infty} \Big(1 - \delta_l \mathbb{P}_l(\upsilon) - \delta_n \mathbb{P}_n(\upsilon)\Big) \upsilon \, \mathrm{d}\upsilon$, $\delta_l = \Big(1 + \frac{\varpi P_t A_l G_s (\upsilon^2 + z_\infty^2)^{-\alpha_l/2}}{m_l}\Big)^{-m_l}$, $\delta_n = \Big(1 + \frac{\varpi P_t A_n G_s (\upsilon^2 + z_\infty^2)^{-\alpha_n/2}}{m_n}\Big)^{-m_n}$, $\varpi = \frac{\vartheta}{\kappa P_t \theta}$, $\theta = \frac{\sum_i^\kappa \zeta_l(r_i)^2}{m_l \sum_i^\kappa \zeta_l(r_i)}$, and $\zeta_l(r_i) = A_l G_s \big(r_i^2 + z_\infty^2\big)^{-\alpha_l/2}$.

Proof. The proof follows from (8.27) and Theorem 8.3.0.1, and is analogous to Theorem 8.5.1.1. □

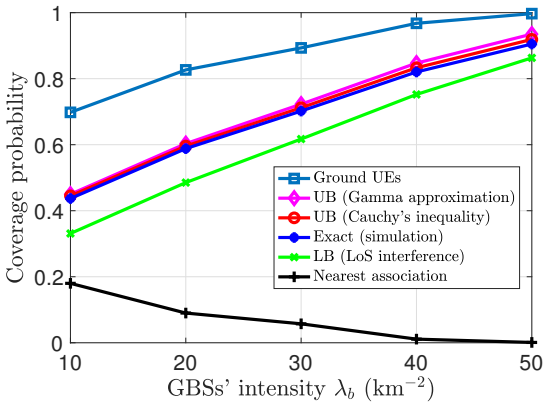
The effect of β on the coverage probability in (8.32) can be interpreted in a similar way to the nearest association scheme in (8.31). Moreover, conditioning on $Z_\infty = z_\infty$, and for a given β in (8.32), the yielded expression holds the same insights as for static UAV-UEs. In particular, what the Nakagami fading parameter m_l, antenna down-tilting angle, and the collaboration distance R_c entail for the performance of mobile UAV-UEs is similar to the that of static UAV-UEs. Finally, a simple lower bound on the mobile UAV-UE coverage probability can be obtained similar to Corollary 8.3.0.1, with the details omitted.

TABLE 8.2: Simulation Parameters

Description	Parameter	Value	Description	Parameter	Value
LoS path-loss exponent	α_l	2.09	SIR threshold	ϑ	0 dB
NLoS path-loss exponent	α_n	3.75	BSs' intensity	λ_b	20 BSs/km^2
LoS path-loss constant	A_l	-41.1 dB	Inter-cluster center distance	$2R_h$	380 m
NLoS path-loss constant	A_n	-32.9 dB	Antenna main-lobe gain	G_m	10 dB
Nakagami fading parameter (LoS)	m_l	3	Antenna side-lobe gain	G_s	-3.01 dB
Nakagami fading factor (NLoS)	m_n	1	BS antenna height	h_{BS}	30 m
Area fraction occupied by buildings	a	0.3	UAV-UE altitude	h_d	120 m
Density of buildings	η	300 km^{-2}	Simulation area	R_{sim}	20 km^2
Buildings height Rayleigh parameter	c	20 m	Mean altitude of mobile UAV-UEs	$\mathbb{E}[Z_\infty]$	150 m



(a) UAV-UE coverage probability versus its altitude h_d

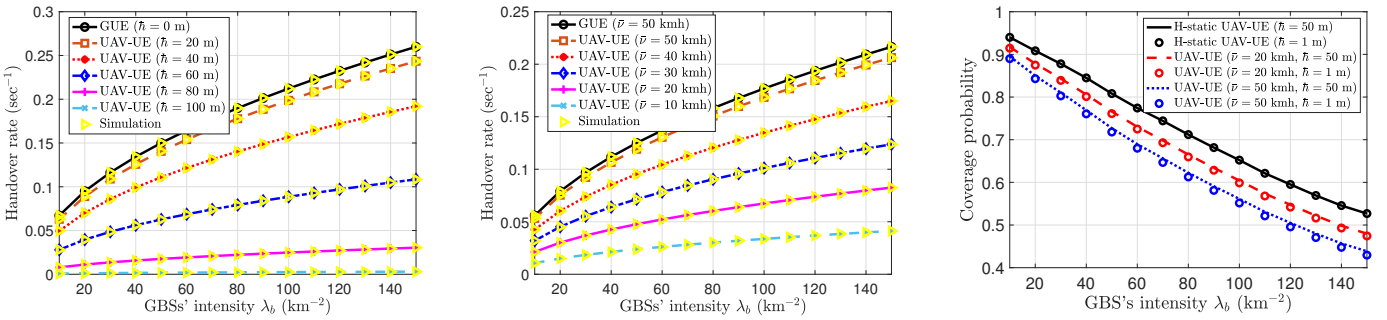


(b) UAV-UE coverage probability versus BS's intensity λ_b

FIGURE 8.8: The derived upper and lower bounds on the static UAV-UE coverage probability are plotted versus the UAV-UE altitude h_d and BS' intensity λ_b.

8.6 Simulation Results and Analysis

For our simulations, we consider a network having the parameter values indicated in Table 8.2. It is worth recalling that we here assume high altitude UAV-UEs where their altitudes are set higher than the height of the ground BSs (for instance, the



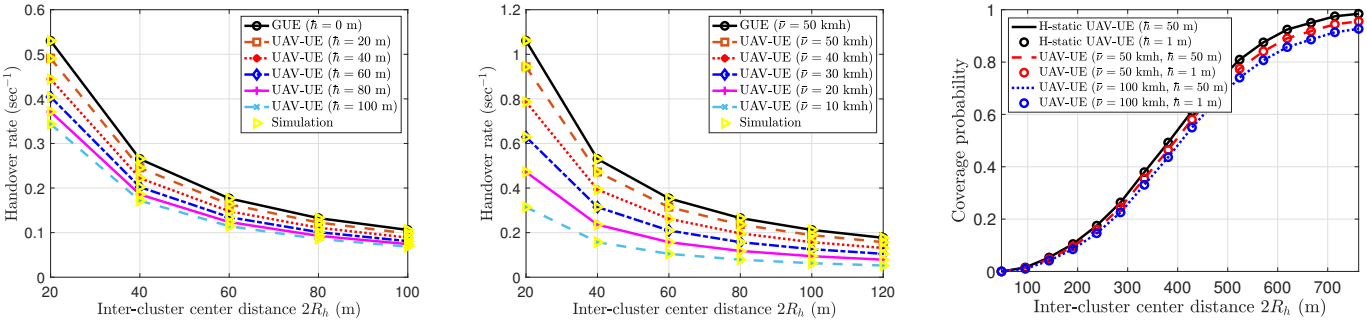
(a) $\bar{\nu} = 50\,\text{kmh}$, $\mu = 300\,\text{km}^{-2}$ (b) $\hbar = 30\,\text{m}$, $\mu = 300\,\text{km}^{-2}$ (c) $\beta = 0.5$, $\mu = 100\,\text{km}^{-2}$, $\vartheta = -10\,\text{dB}$

FIGURE 8.9: Effect of the 3D mobility on the performance of aerial and ground UE when they are associated with their nearest BSs. In (c), H-static refers to a UAV-UE that only moves in the vertical direction within an altitude difference $\hbar$.

UAV-UE altitude h_d is set $120\,\text{m}$ in Table 8.2 while the BS height h_{BS} is $30\,\text{m}$). For this setup, we can quantify the poor cellular connectivity the UAV-UEs undergo and show how this degraded performance can be mitigated by means of cooperative communications from the ground BS.

In Fig. 8.8, we show the effect of the UAV-UE altitude and BSs' intensity on the coverage probability of static UAV-UEs, with that of ground UEs plotted for comparison. Fig. 8.8(a) shows that the coverage probability of high-altitude UAV-UEs monotonically decreases as h_d increases. This is because, as the altitude of UAV-UEs increases, the received interference power substantially increases due to the dominance of LoS links. Fig. 8.8(a) also shows that the derived UB on the coverage probability in (8.13) is considerably tight. Meanwhile, Fig. 8.8(b) illustrates the effect of BSs' intensity λ_b on the performance of UAV-UEs. Except for the nearest association scheme, the coverage probability improves with λ_b since more BSs cooperate to serve the aerial (and ground) UEs. However, when the UAV-UE associates to its nearest BS, the effect of interference increases as the network becomes denser.

Next, we study the impact of 3D mobility on the performance of UAV-UEs. We further compare the performance of UAV-UEs with their ground counterparts moving horizontally with the same speed $\bar{\nu}$. In Fig. 8.9, the handover rate and coverage probability of mobile aerial and ground UEs associated with their nearest BSs are investigated. Fig. 8.9(a) plots the handover rate versus λ_b at different values



(a) $\bar{\nu} = 30\,\text{kmh}$, $\mu = 300\,\text{km}^{-2}$ (b) $\mu = 100\,\text{km}^{-2}$, $\hbar = 50\,\text{m}$ (c) $\mu = 100\,\text{km}^{-2}$, $\beta = 1$

FIGURE 8.10: Effect of the 3D mobility on the performance of aerial and ground UE when they are served via CoMP transmission with the inter-cluster center distance set equal to $2R_h$. In (c), H-static refers to an UAV-UE that only moves in the vertical direction within an altitude difference $\hbar$.

of the altitude difference $\hbar$. Fig. 8.9(a) shows that the analytical result in (8.16) matches the simulation result quite well. As is the case for typical Poisson-Voronoi models, the handover rate grows linearly with the square root of the BS's intensity $\sqrt{\lambda_b}$. Moreover, the handover rate decreases as $\hbar$ increases, which implies that a UAV-UE having frequent up and down motions along its trajectory is susceptible to lower rates of handover. We also note that the handover rate of UAV-UEs is upper bounded by that of ground UEs with $\hbar = 0$. Fig. 8.9(b) shows the effect of the UAV-UE speed $\bar{\nu}$ on its handover rate.[4] Intuitively, the handover rate increases as $\bar{\nu}$ increases since the UAV-UE stays shorter time in the area covered by each BS, i.e., it experiences a shorter sojourn time. Finally, Fig. 8.9(c) investigates the effect of mobility on the UAV-UE coverage probability given an arbitrary handover penalty β. Notice that the coverage probability decreases as $\bar{\nu}$ increases since this leads to higher handover probability (penalized by β). Moreover, the altitude difference $\hbar$ has a marginal effect on the coverage probability of UAV-UEs. This is attributed to the fact that the increase of the altitude difference $\hbar$ for mobile UAV-UEs while keeping the same average flying altitude $\mathbb{E}[Z_\infty]$ relatively yields the same average coverage probability.

[4]Low values of the speed $\bar{\nu}$ suits the motion of UAV-UEs operating surveillance cameras while higher velocities would be suitable for UAV-UEs operating flying taxis.

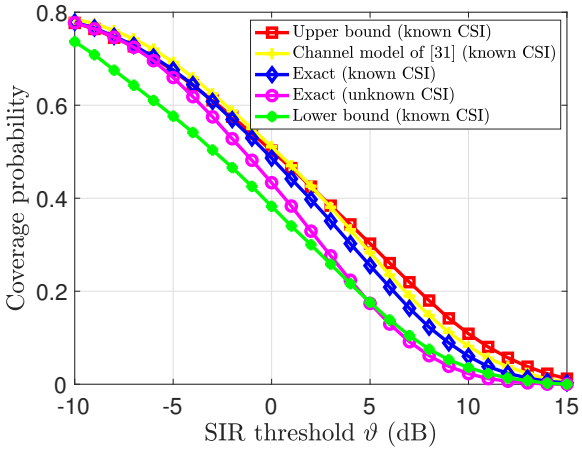


FIGURE 8.11: Comparison of the UAV-UE coverage probability under different setups and assumptions. Specifically, the proposed mathematical model is evaluated versus the channel models of [152] and [142], and also assessed for known and unknown CSI.

In Fig. 8.10, we evaluate the effect of the 3D mobility on the UAV-UE performance under CoMP transmissions. Fig. 8.10(a) shows that the inter-CoMP handover rate monotonically decreases as R_c increases since the UAV-UE would have a longer sojourn time in each cluster. Moreover, the handover rate is shown to decrease as $\hbar$ increases, i.e., when the UAV-UE has frequent up and down motions along its trajectory, which also yields a longer sojourn time in each cell (cluster). We also note that the handover rate of UAV-UEs is upper bounded by that of ground UEs, with $\hbar = 0$. Fig. 8.10(b) shows the effect of the UAV-UE speed $\bar{\nu}$ on the inter-CoMP handover. We note that the handover rate increases as $\bar{\nu}$ increases since the UAV-UE will have a shorter sojourn time in each cluster. Finally, Fig. 8.10(c) shows the effects of the UAV-UE speed and altitude difference on the UAV-UE coverage probability. Fig. 8.10(c) shows that the UB on the coverage probability, characterized in Theorem 8.5.2.1, slightly decreases as $\bar{\nu}$ increases, which corresponds to a higher handover rate (penalized by β). This slight decrease is essentially because as the inter-cluster distance becomes larger, the probability of handover decreases and its effect gradually vanishes. Similar to the nearest association scheme, the altitude difference $\hbar$ has a minor effect on the coverage probability of UAV-UEs. In addition to its impact on the coverage probability, the mobility of UAV-UEs can decrease their throughput, particularly, when accounting for the handover execution time [186].

TABLE 8.3: Channel Model for Urban-Micro with UAV-UEs [152] and [142]

Channel Model	Description
Channel fading	Table B-2 in [152]
Shadowing effect	Log-normal (standard deviation from Table B-3 in [152])
Thermal noise and frequency	$-174\,\mathrm{dBm/Hz}$ power spectral density, $700\,\mathrm{MHz}$ [142]
Probability of LoS	Table B-1 in [152]
System bandwidth	$10\,\mathrm{MHz}$ [142]
Channel estimation	Known CSI

Finally, we compare the obtained upper and lower bounds on the coverage probability with the results based on measurements and system level simulations that are found in [152] and [142]. We consider a system setup whose channel parameter values are indicated in Table 8.3. For a fair comparison, we consider a cooperative transmission scheme from nearest BSs within a collaboration distance R_c according to the proposed deployment of ground BSs, whose parameter values are from Table 8.2. In particular, Fig. 8.11 plots the UAV-UE coverage probability versus the SIR threshold ϑ resulting from the proposed channel models of [152] and [142]. From Fig. 8.11, we can see that the proposed mathematical model gives a performance that is relatively close to the system evaluations based on the channel models of [152] and [142]. Fig. 8.11 also shows that the UAV-UE coverage probability under the assumption of known CSI at the serving BSs surpasses that of the non-coherent transmission (i.e., unknown CSI).

8.7 Conclusion

In this chapter, we have proposed a novel framework for cooperative transmission and probabilistic caching that can be leveraged to provide reliable connectivity and ubiquitous mobility support for UAV-UEs. In order to analytically characterize the performance of UAV-UEs, we have employed Cauchy's inequality and moment approximation of Gamma RVs to derive upper and lower bounds on the UAV-UE coverage probability. Moreover, we have developed a novel 3D RWP model that allowed us to explore the role of UAV-UEs' mobility in cellular networks, particularly, to quantify the handover rate and the impact of their mobility on the achievable performance. For both static and mobile UAV-UEs, we have shown allowing CoMP transmission significantly improves the achievable coverage probability,

e.g., from 28% for the baseline scenario with nearest serving BSs, to 60% for static UAV-UEs (see Fig. 8.3(a)). Furthermore, comparing the performance of UAV-UEs to ground UEs, it is shown that the coverage probability of a UAV-UE is always upper bounded by that of a ground UE owing to the down-tilted antenna pattern and LoS-dominated interference for UAV-UEs. Our results for the case of mobile UAV-UEs have also revealed that their handover rate and handover probability decrease as the altitude difference increases, i.e., in the case of frequent up and down motions of the UAV-UEs along their trajectory. Moreover, while the altitude difference has a minor effect on the coverage probability of mobile UAV-UEs, their speed noticeably degrades their coverage probability.

Chapter 9

Conclusions and Future Directions

There has been a trend in the fifth generation of telecommunication systems towards enabling services at the edge of the networks, as discussed in Chapter 2. These services include edge caching and edge computing. Such key enablers of the 5G technology will help alleviate the heavy burden on the core networks, reduce the overall service delay, and improve the perceived QoE. While prior works in the literature studied the concept of wireless caching and content delivery from various points of views, the work discussed in the thesis stands out with regards to the following aspects. First, we have shown in this thesis that the joint design and optimization of content caching and content delivery (i.e., communication) is vital for improving key application and network KPIs such as average service delay, throughput, offloading gain, energy consumption. Moreover, we have conducted preliminary studies and analysis of content delivery and service communication for contemporary aerial users (i.e., cellular-connected UAVs). By proposing prominent solutions to mitigate the effect of LoS interference and improve the connectivity to the sky, and conducting novel analysis for aerial users under practical antenna and mobility models, the road is paved to envision proper content caching schemes to adequately serve the contemporary sky user.

In the next section, we summarize the key findings of the thesis.

9.1 Summary of the Findings

9.1.1 Inter-cluster Cooperation for D2D Caching Networks

In Chapter 3, we have shown that D2D caching with inter-cluster collaboration as an appealing approach to reduce the network average delay. More specifically, we proposed a novel D2D caching architecture where devices in the same cluster exchange cache content via D2D communication, while the devices in different clusters cooperate by exchanging their cache content via cellular transmission. We formulated the delay minimization problem in terms of the content cache placement and efficiently solved for a local optimal caching solution that is provably showed to be within a factor $(1 - e^{-1})$ of the global optimum. We have also shown that the average throughput can be improved by allowing inter-cluster content sharing. Moreover, by conducting the network throughput scaling analysis, we could explore the asymptotic behavior of the network when the content library size goes asymptotically large. We particularly showed that the network average sum throughput decreases with the library size increase, and the rate of this decrease is controlled by the popularity exponent. We verified the analytical results by simulations and our results proved that the network average delay could be effectively reduced by more than 45% when allowing inter-cluster cooperation.

9.1.2 Joint Caching and Communication for Clustered D2D networks

The study in Chapter 3 is carried out based on the so-called simple protocol model, where the underlying operations of the physical and MAC layers are not considered in the analysis. To provide a more practical study and consider the prominent role of the joint design of caching and communications, we adopted the so-called physical interference analysis in Chapter 4 and Chapter 5. This allowed us to factor in the operation of the MAC and physical layers and jointly design content caching and wireless resource allocation.

First, in Chapter 4, we have conducted a comprehensive analysis of the joint communication and caching for a clustered D2D network undergoing random probabilistic caching. We have first maximized the offloading gain of the proposed network by jointly optimizing the channel access and caching probability. We have

showed that deviating from the optimal access probability yields file sharing more difficult. More precisely, we showed that the underlying system is too conservative for small access probabilities, while the interference is too aggressive for larger access probabilities. Then, we have minimized the overall energy consumption of the proposed network as a function of the caching scheme. The yielded insights have showed that a content with a large size or low popularity has a small probability to be cached.

Lastly, we have proposed a joint spectrum partitioning and content caching optimization framework that is then leveraged to reduce the average service delay for clustered D2D networks. Based on a developed spatiotemporal model, we have characterized the network coverage probability and the content arrival and service rates. We have then formulated the delay joint optimization problem whose decision variables are the content caching and bandwidth allocation. Employing a block coordinate descent (BCD) optimization technique, we have obtained the optimal allocated bandwidth in a closed-form expression, and a suboptimal caching scheme is also provided. Our results have revealed that the joint optimization of spectrum partitioning and caching could substantially reduce the average service delay compared to the legacy caching techniques, e.g., Zipf and uniform caching. Moreover, we have showed that the traffic load and content popularity play an important role in the design of resource allocation and content placement schemes.

Second, in Chapter 5, we have explored the role of cache-assisted CoMP transmissions for clustered D2D networks. We particularly have characterized the coverage probability and offloading gain of the considered clustered D2D network whose devices adopt cooperative transmissions and probabilistic caching scheme. For an improved tractability, we have sought simple yet tight lower bound and approximation of the obtained coverage probability and offloading gain. We have then formulated the offloading gain maximization problem and obtained optimized caching probabilities based on the proposed lower bound and approximation. Our results have showed that allowing CoMP transmission for clustered D2D caching networks can attain up to 300% improvement in the rate coverage probability compared to the single transmission scheme. Finally, we have showed that the proposed PC yields a considerable improvement of the offloading gain over legacy

caching schemes.

9.1.3 The Advantages of MIMO Beamforming for Content Delivery to the Sky

In the second part of the thesis, we have turned our attention to a new architecture of content delivery networks where the ground BSs serve static and mobile aerial UAVs. In order to maintain seamless content transmission and adequate network coverage to such temporary sky users, we have proposed several novel approaches to cancel out the effect of LoS interference that is considered the key limiting factor hindering their performance.

In detail, in Chapter 6, we have proposed a new framework for successful content delivery for multiple ground users co-existing with one aerial user. We have leveraged the independence between the channel to the aerial users w.r.t. of their ground counterparts, so as to efficiently multiplex their transmitted data over the same channel. We have first derived the gain of intended and interfering channel gains. We have then derived an analytical expression for the successful content delivery probability (SCDP). Our results have showed that CB from massive MIMO-enabled BSs can substantially improve the performance of the aerial users in terms of SCDP. For completeness, we have explored the impact of the down-tilt angle of the BS antennas and showed that it yields to a tradeoff between the performance of aerial user and ground users only if the aerial user altitude is below the BS height. The analysis in this chapter is carried out assuming a simple antenna model for simplicity.

9.1.4 The effect of 3D Antenna Pattern on Mobile UAV-UEs

We have continued our study of the content delivery and communication to mobile aerial users in Chapter 7 and Chapter 8. Particularly, motivated by the dominant impact of the antenna patterns on the aerial users, we have studied their performance under practical antenna configurations in Chapter 7. We have particularly characterized the coverage probability of static and mobile aerial users, served from 3D practical antenna patterns as functions of the system parameters, namely, the

number of antenna elements, density of BSs, and the aerial users' altitude. We have also investigated the handover rate of mobile aerial users to reveal the impact of practical antenna patterns on their cell association. The achieved performance under practical antenna patterns have showed to be worse than that attained from a simple antenna model. More importantly, the performance of static and mobile aerial users undergoing nearest association is shown to be slightly impacted by the increase of the number of the antenna elements. In contrast, if mobile aerial users are associated to the ground BS that delivers the highest average signal power, their coverage probability is shown to decrease as the number of antenna elements increases due to the excessive rate of altitude handover that results in frequent handover failure.

9.1.5 Caching and Mobility in the Sky

In Chapter 8, we have proposed a novel framework for cooperative transmission and probabilistic caching that can be applied to provide adequate connectivity and mobility support for aerial users. We have first derived upper and lower bounds on the UAV-UE coverage probability by employing Cauchy's inequality and moment approximation of Gamma RVs. We have showed the impact of several system parameters such as collaboration distance and content availability on the achievable performance of static aerial users. We have further developed a novel 3D mobility model with the aim of exploring the effect of UAV-UEs' mobility in cellular networks and quantifying their handover rate. For both static and mobile aerial users, we have shown that CoMP transmission can significantly improve the achievable coverage probability. By comparing the performance of aerial users to that of their ground counterparts, we have showed that the coverage probability of aerial users is always upper bounded by that of ground users owing to the down-tilted antenna pattern and LoS-dominated interference in the sky. For mobile aerial users, we have also found that their handover rate and handover probability decrease when the altitude difference increases, i.e., in the case of frequent up and down motions of the aerial users. In addition, even though the altitude difference has a slight impact on the coverage probability of mobile UAV-UEs, their speed noticeably deteriorates their coverage probability.

Next, we discuss possible future extensions of the work presented in this thesis. In the discussion below, we detail a possible general and rich research area, targeting joint communication, caching, and computing for terrestrial and aerial (i.e., UAV) mobile networks.

9.2 Joint Caching, Communication, and Computing

With software defined networking (SDN) and network function virtualization (NFV) in place, communication and computing functionalities are converging in 5G networks [187]. Following these developments, jointly optimizing caching and computing capabilities in mobile networks may provide higher efficiency for users' applications with extensive computation demands and continuous content delivery. However, the caching capacity improvement brought by smart computing resource consumption is always neglected. In augmented reality (AR) applications, it is common practice to extract some key features from the originally captured videos in order to save caching and transmission resources [188]. It is therefore crucial to exploit computing resources to alleviate the strain on caching resources for the nodes with poor storage resources. As such, with the virtualized functions of both caching and computing, the coupling of edge caching and edge computing in a joint problem becomes inevitable for 5G networks.

We already discussed in the previous chapters the emergence of edge caching as a promising approach to alleviate the heavy burden on data transmission through caching and forwarding contents to the edge of networks. However, existing studies always treat storage and computing resources separately. Driven by this issue, a new computation-aided edge caching architecture for 5G networks can be proposed where the mobile edge computing (MEC) resources are utilized for enhancing edge caching capability. For instance, files could be compressed into smaller sizes to shorten the delivery time. In such an architecture, we propose to jointly schedule the caching and computing resources at the wireless network edge.

MEC, as a key technology toward 5G, provides cloud computing capabilities and task offloading service at the edge of mobile networks [189]. Due to the proximity of MEC servers to end users, tasks can be offloaded and accomplished with

low latency. Although MEC resources seem different from caching resources, they are closely coupled. For instance, the size of a cached file may be reduced if compressed by MEC resources. Thus, some storage space can be saved. From another perspective, the caching capability of nodes is enhanced. Another example is AR, where the key elements of the captured video can be extracted from the original data through information processing and computing. As the size of key elements is small, they can be cached and distributed easily. Based on the received key elements, end users may reconstruct the original image or video files.

9.2.1 Recent Works

We now discuss few recent works that study both edge caching and computing from different perspectives [190–197], and the stream of more recent works as in [48, 192, 198–201]. For instance, in [190], the MEC is introduced for providing computing capabilities at the edge of networks to improve the latency performance of wireless networks. A MEC-enabled HetNet is composed of the multi-tier networks with access point (AP) (i.e., MEC servers), which have different transmission power and different computing capabilities. In this framework, the authors considered multiple-type mobile users with different sizes of computation tasks, and they offload the tasks to a MEC server, and receive the computation resulting data from the server. The successful edge computing probability considering both the computation and communication performance is introduced as a KPI.

The concept of MEC has been recently introduced to supplement cloud computing by deploying MEC servers at the network edge so as to reduce the network delay and alleviate the load on cloud data centers. However, compared to a resourceful cloud, a MEC server has limited resources. When each MEC server operates independently, it cannot handle all of the computational and big data demands stemming from the users' devices. Consequently, the MEC server cannot provide significant gains in overhead reduction due to data exchange between users' devices and remote cloud. The integration of MEC with a mobile network raises a number of challenges related to the coordination and control of joint communication, computation and caching. Therefore, joint computing, caching, communication, and control (4C) at the edge with MEC server collaboration is strongly needed

for big data applications.[1] In order to address these challenges, in [191], the problem of joint 4C in big data MEC is formulated as an optimization problem whose goal is to maximize the bandwidth saving while minimizing delay, subject to the local computation capability of user devices, computation deadline, and MEC resource constraints.

In [192], the authors envisioned a collaborative joint caching and computing strategy for on-demand video streaming in MEC networks. The design aims at enhancing the widely used Adaptive BitRate (ABR) streaming technology, where multiple bitrate versions of a video can be delivered so as to adapt to the heterogeneity of user capabilities and the varying network condition. This strategy faces two main challenges: (i) not only the videos but their appropriate bitrate video versions have to be effectively selected to store in the caches, and (ii) the transcoding relationships among different versions need to be taken into account to effectively utilize the computing capacity at the MEC servers. Specifically, owing to their real-time computing capability, MEC servers can perform transcoding of a video to different variants to satisfy the user requests. Each variant is a bitrate version of the video and a higher bitrate version can be transcoded to a lower bitrate version. The collaborative joint caching and computing problem is formulated as an Integer Linear Program (ILP) that minimizes the backhaul network cost, subject to the cache storage and computing capacity constraints.

In [193], the authors designed a novel information-centric heterogeneous networks framework aiming at enabling content caching and computing. Due to the virtualization of the whole system, communication, computing, and caching resources can be shared among all users associated with different virtual service providers. A virtual resource allocation optimization problem is formulated, where the gains of not only virtualization but also caching and computing are taken into consideration. In contrast to previous works on MEC, which mainly focuses on computation offloading, the authors in [194] introduced a new concept of task caching. Task caching refers to the caching of completed task application and their

[1]The meaning of control in the 4C context is to coordinate and integrate the communication, computation, and caching models in a distributed manner.

related data in edge cloud. The authors investigated the problem of joint optimization of task caching and offloading on edge cloud with the computing and storage resource constraint.

The key enablers of the caching and computing paradigms are content caching and computation offloading strategies, respectively. In order to jointly tackle these issues in MEC-enabled cellular networks, the authors in [195] formulated the computation offloading decision, resource allocation and content caching strategy as an optimization problem, considering the total revenue of the network. This revenue is formulated in terms of assigning radio resources and allocating computation resources to the users. The authors transformed the original problem into a convex problem and then decomposed it in order to solve it in a distributed and efficient way. Finally, with recent advances in distributed convex optimization, an alternating direction method is used to develop a multipliers-based algorithm to solve the optimization problem.

The problem of distribution and proactive caching of computing tasks in fog networks under latency and reliability constraints is studied in [196]. In such a scenario, computing can be executed either locally at the user device or offloaded to serving edge computing nodes (cloudlets). Moreover, cloudlets exploit both their computing and storage capabilities by proactively caching popular task computation results to minimize computing latency. To this end, a clustering method to group spatially proximate user devices with mutual task popularity interests and their serving cloudlets is proposed. Then, cloudlets can proactively cache the popular tasks' computations of their cluster members to minimize computing latency. Additionally, the problem of distributing tasks to cloudlets is formulated as a matching game in which a cost function of computing delay is minimized under latency and reliability constraints.

Most existing studies focus on cache allocation and coding design for caching [197]. However, grid power supply with fixed power might be costly and possibly infeasible, especially when the load changes dynamically over time. In [197], the authors investigated the energy consumption of the MEC server problem in cellular networks. Given the average download latency constraints, the authors took the MEC servers' energy consumption, backhaul capacities and content popularity

distributions into account and formulate a joint optimization framework to minimize the energy consumption of the system. A genetic algorithm is applied to solve such a complicated joint optimization problem.

9.2.2 Future Directions

Many new challenges will arise when joint caching, communication, and computation are considered. For instance,

1. New KPIs need to be developed to account for the reward/cost from computation-assisted caching, and to balance the trade-off between, caching, serving or computing. For instance, the trade-off between caching and computing arises when the MEC decides either to cache plain files or perform compression to reduce the size of a cached content. Similarly, sending compressed content or plain files entails another trade-off between computing and communication. Typically, the offloading gain is adopted in edge caching research, and recently, successful edge computing probability [190] is used for MEC-enabled heterogeneous networks.

2. Although individual computing and caching resources of the edge nodes can be collaboratively used to improve the content caching performance, the cooperative approach also brings in several issues. The first one comes into effect because of the variability in computing and caching capabilities of different nodes, which make the joint resource scheduling a complicated task. Machine learning is a feasible approach to address the problem. For instance, a reinforcement learning algorithm is adopted in [202] to adaptively arrange serving capabilities of carrier nodes (vehicles) in a vehicular ad-hoc networks. A data carrier node has to handover the data to the next data carrier node before moving out of the interest region. More precisely, a fuzzy logic is used to evaluate the next data carrier node (vehicle), and the reinforcement learning is used to evaluate the possible future reward after selecting the next data carrier node.

3. Coded caching can be used along with coded computing where many (cooperative) MECs can share the computation of the offloaded tasks, and then

the computed data is encoded into segments before being transferred among them. A collaborative distributed computing framework might be an interesting topic for investigation where resource-constrained end-user devices outsource their computation to the upper-layer computing resources at the edge and cloud layers. This framework extends the standard MEC originally formulated by European Telecommunications Standards Institute (ETSI), which focuses only on individual MEC entities.

4. One possible architecture for applying both caching and computing is a network of mobile BSs, e.g., mobile UAVs or moving terrestrial BSs. For instance, it is shown in [203] that significant communication throughput gains can be achieved by mobile UAVs over static UAVs/fixed terrestrial BSs by optimizing the UAV's trajectory. However, adopting the framework of caching and computing for such architectures has not been addressed yet in the literature. Moreover, when dealing with aerial users (i.e., UAV-UEs as shown in the previous chapters), adopting joint content caching and task offloading might pose new challenges. For instance, aerial users have unique channel characteristics that drastically impact their cell association and thereby identifying the content caching computing offloading strategies. Moreover, as aerial users are usually (3D) mobile, the effect of their trajectories on the content caching and computing offloading schemes need to be factored in.[2]

[2]Other potential research directions might also include the use of machine learning for edge caching and adoption of security, bloackchain, and cognitive radio for such caching networks, e.g., see a few relevant works [204–206], [207–212], and [213–227].

Appendices

Appendix A

Appendix A

A.1 Proof of Lemma 3.5.2.1

We define the ground set that describes the cache placement elements in all clusters as

$$\mathcal{S} = \{s_1^1, ..., s_k^f, ..., s_k^m, ..., s_K^1, ..., s_K^m\} \tag{A.1}$$

where s_k^f is an element denoting the placement of file f into the VCC of cluster k. This ground set can be partitioned into K disjoint subsets $\{S_1, S_2, ..., S_K\}$, where $S_k = \{s_k^1, s_k^2, ..., s_k^m\}$ is the set of all files that might be placed in the VCC of cluster k.

Let us express the cache placement by the adjacency matrix $\mathbf{X} = [x_{k,f}]_{K\times m} \in \{0,1\}_{K\times m}$. Moreover, we define the corresponding cache placement set $A \subseteq \mathcal{S}$ such that $s_k^f \in A$ if and only if $x_{k,f} = 1$. Hence, the constraints on the cache capacity of the VCC of cluster $k \in \mathcal{K}$ can be expressed as $A \subseteq \mathcal{S}$, where

$$\mathcal{H} = \{A \subseteq \mathcal{S} : |A \cap S_k| \leq N \text{ for all } k = 1, \ldots, K\} \tag{A.2}$$

The above expression is derived directly from the constraint that the maximum cache size per cluster is N files, i.e., $\sum_{f=1}^{m} x_{k,f} \leq N$. Comparing $\mathcal{H}$ in (A.2) with the definition of partition matroid in (3.17), it is clear that our constraints form a partition matroid with $l = K$ and $k_i = N$. Additionally, since $k_i = N$ for all $i = \{1, 2, \ldots, K\}$, it is easy to see that our constraints also form a uniform partition matroid. This proves Lemma 1.

A.2 Proof of Lemma 3.5.2.2

We consider two cache placement sets A and A', where $A \subset A'$. For a certain cluster $k \in \mathcal{K}$, we consider adding the caching element $s_k^f \in \mathcal{S} \setminus A'$ to both placement sets. This means that a file f is added to cluster k, where the corresponding cache placement element has not been placed in either A or A'. The marginal value of adding an element s_k^f to a set is defined as the change in the file download time after adding this element to the set. The average download time for a file f with mean size $\overline{S}$ is $\frac{\overline{S}}{R_D}$, $\frac{\overline{S}}{R_{WL}/N_a}$, or $\frac{\overline{S}}{R_{BH}/N_b}$ if the file is obtained from the local cluster, a randomly chosen remote cluster, or the backhaul, respectively. As shown in Section 3.5.1, N_a and N_b are random variables that are functions of the cluster cache size, popularity exponent, and the library size. For our work, we assume that $\frac{R_{WL}}{N_a} > \frac{R_{BH}}{N_b}$ always holds. For the sake of simplicity, we replace $\frac{R_{WL}}{N_a}$ and $\frac{R_{BH}}{N_b}$ with their averages, $\overline{R_{WL}}$ and $\overline{R_{BH}}$, respectively. Now, the aggregate transmission rate assumption is $R_D > \overline{R_{WL}} > \overline{R_{BH}}$.

For D_k in (3.6) to be a supermodular function, the difference in the marginal values between the two sets A and A' must be non-positive. For a user u belonging to cluster k and requesting content $f \in \mathcal{F}$, we distinguish between these different cases: as $\overline{R_{WL}}$, similarly, we denote $\frac{R_{BH}}{N_b}$ as $\overline{R_{BH}}$. Now, the aggregate transmission rate assumption is $R_D > \overline{R_{WL}} > \overline{R_{BH}}$.

1. According to placement A', user u obtains file f from a remote cluster j', i.e., $s_{j'}^f \in A'$ and $j' \neq k$. In this case, the marginal value with respect to A' is

$$G(A' \cup \{s_k^f\}) - G(A') = 0 \tag{A.3}$$

   According to placement A, user u obtains file f from a remote cluster j, i.e., $s_j^f \in A$, again the marginal value is zero. However, if $s_j^f \notin A$, the marginal value is given by

$$G(A \cup \{s_k^f\}) - G(A) = P_{k,f}\Big(\frac{\overline{S}}{\overline{R_{WL}}} - \frac{\overline{S}}{\overline{R_{BH}}}\Big) \tag{A.4}$$

2. In this case, we assume that $s_i^f = s_k^m$, i.e., the requested file f is cached in

cluster k. According to placement A', user u obtains file f from the local cluster k. Hence, the marginal value is given by

$$G(A' \cup \{s_i^f\}) - G(A') = P_{k,f}\Big(\frac{\overline{S}}{R_D} - \frac{\overline{S}}{\overline{R_{WL}}}\Big) \tag{A.5}$$

According to placement A, user u obtains file f from a remote cluster j when $s_j^f \in A$, again the marginal value is given by

$$G(A \cup \{s_i^f\}) - G(A) = P_{k,f}\Big(\frac{\overline{S}}{R_D} - \frac{\overline{S}}{\overline{R_{WL}}}\Big) \tag{A.6}$$

However, if $s_j^f \notin A$, the marginal value is written as

$$G(A \cup \{s_i^f\}) - G(A) = P_{k,f}\Big(\frac{\overline{S}}{R_D} - \frac{\overline{S}}{\overline{R_{BH}}}\Big) \tag{A.7}$$

Accordingly, the difference in marginal values between A and A' in all cases is

$$G(A \cup \{s_i^f\}) - G(A) - (G(A' \cup \{s_i^f\}) - G(A')) \leq 0 \tag{A.8}$$

It is clear that $g(A) \leq g(A')$ for $A \subseteq A' \subseteq \mathcal{S}$, or equivalently, $g(A) - g(A') \leq 0$. From the definition of supermodularity, it is clear that the delay per request in the $k-$th cluster, D_k, is a supermodular set function. The weighted sum of supermodular functions is also a supermodular function [95], and so the network average delay D in (3.8) is a supermodular function. For the monotone non-increasing property, it is intuitive to see that the delay will never increase by caching new files. Hence, Lemma 2 proves that problem (3.8) is a monotonically non-increasing supermodular set function minimized under uniform partition matroid constraints.

Appendix B

Appendix B

B.1 Proof of lemma 4.3.0.1

Laplace transform of the inter-cluster aggregate interference $I_{\Phi_p^!}$, evaluated at $s = \frac{\theta r^\alpha}{P_d}$, can be evaluated as

$$
\begin{aligned}
\mathscr{L}_{I_{\Phi_p^!}}(s) &= \mathbb{E}\left[e^{-s\sum_{\Phi_p^!}\sum_{y\in\mathcal{B}^p} g_{yx}\|x+y\|^{-\alpha}}\right] \\
&= \mathbb{E}_{\Phi_p}\left[\prod_{\Phi_p^!}\mathbb{E}_{\mathcal{B}^p,g_{yx}}\prod_{y\in\mathcal{B}^p} e^{-sg_{yx}\|x+y\|^{-\alpha}}\right] \\
&= \mathbb{E}_{\Phi_p}\left[\prod_{\Phi_p^!}\mathbb{E}_{\mathcal{B}^p}\prod_{y\in\mathcal{B}^p}\mathbb{E}_{g_{yx}} e^{-sg_{yx}\|x+y\|^{-\alpha}}\right] \\
&\overset{(a)}{=} \mathbb{E}_{\Phi_p}\left[\prod_{\Phi_p^!}\mathbb{E}_{\mathcal{B}^p}\prod_{y\in\mathcal{B}^p}\frac{1}{1+s\|x+y\|^{-\alpha}}\right] \\
&\overset{(b)}{=} \mathbb{E}_{\Phi_p}\prod_{\Phi_p^!}\exp\Big(-p\overline{n}\int_{\mathbb{R}^2}\Big(1-\frac{1}{1+s\|x+y\|^{-\alpha}}\Big)f_Y(y)\,\mathrm{d}y\Big) \\
&\overset{(c)}{=} \exp\left(-\lambda_p\int_{\mathbb{R}^2}\Big(1-\exp\Big(-p\overline{n}\int_{\mathbb{R}^2}\Big(1-\frac{1}{1+s\|x+y\|^{-\alpha}}\Big)f_Y(y)\,\mathrm{d}y\Big)\,\mathrm{d}x\right) \\
&\overset{(d)}{=} \exp\left(-\lambda_p\int_{\mathbb{R}^2}\Big(1-\exp\Big(-p\overline{n}\int_{\mathbb{R}^2}\Big(1-\frac{1}{1+s\|z\|^{-\alpha}}\Big)f_Y(z-x)\,\mathrm{d}y\Big)\,\mathrm{d}x\right) \\
&\overset{(e)}{=} \exp\left(-2\pi\lambda_p\int_{v=0}^{\infty}\Big(1-\exp\Big(-p\overline{n}\int_{u=0}^{\infty}\Big(1-\frac{1}{1+su^{-\alpha}}\Big)f_U(u|v)\,\mathrm{d}u\Big)v\,\mathrm{d}v\right) \\
&= \exp\left(-2\pi\lambda_p\int_{v=0}^{\infty}\Big(1-\exp\big(-p\overline{n}\int_{u=0}^{\infty}\frac{s}{s+u^\alpha}f_U(u|v)\,\mathrm{d}u\big)v\,\mathrm{d}v\Big)\right) \\
&= \exp\Big(-2\pi\lambda_p\int_{v=0}^{\infty}\Big(1-\mathrm{e}^{-p\overline{n}\varphi(s,v)}\Big)v\,\mathrm{d}v\Big), \qquad \text{(B.1)}
\end{aligned}
$$

where $\varphi(s,v) = \int_{u=0}^{\infty} \frac{s}{s+u^{\alpha}} f_U(u|v)\,\mathrm{d}u$. (a) follows from the Rayleigh fading assumption, (b) follows from the probability generating functional (PGFL) of PPP Φ_c^p, (c) follows from the PGFL of the parent PPP Φ_p, (d) follows from change of variables $z = x + y$, and (e) follows from converting the cartesian coordinates to the polar coordinates. Hence, Lemma 1 is proven.

B.2 Proof of lemma 4.3.0.2

Laplace transform of the intra-cluster aggregate interference I_{Φ_c}, evaluated at $s = \frac{\theta r^{\alpha}}{P_d}$, is written as

$$
\begin{aligned}
\mathscr{L}_{I_{\Phi_c}}(s|v_0) &= \mathbb{E}\left[e^{-s\sum_{y\in\mathcal{A}^p} g_y \|x_0+y\|^{-\alpha}}\right] \\
&= \mathbb{E}_{\mathcal{A}^p, g_y} \prod_{y\in\mathcal{A}^p} e^{-s g_y \|x_0+y\|^{-\alpha}} \\
&= \mathbb{E}_{\mathcal{A}^p} \prod_{y\in\mathcal{A}^p} \mathbb{E}_{g_y} e^{-s g_y \|x_0+y\|^{-\alpha}} \\
&\overset{(a)}{=} \mathbb{E}_{\mathcal{A}^p} \prod_{y\in\mathcal{A}^p} \frac{1}{1+s\|x_0+y\|^{-\alpha}} \\
&\overset{(b)}{=} \exp\Big(-p\overline{n}\int_{\mathbb{R}^2}\Big(1-\frac{1}{1+s\|x_0+y\|^{-\alpha}}\Big) f_Y(y)\,\mathrm{d}y\Big) \\
&\overset{(c)}{=} \exp\Big(-p\overline{n}\int_{\mathbb{R}^2}\Big(1-\frac{1}{1+s\|z_0\|^{-\alpha}}\Big) f_Y(z_0-x_0)\,\mathrm{d}z_0\Big) \\
&\overset{(d)}{=} \exp\Big(-p\overline{n}\int_{h=0}^{\infty}\Big(1-\frac{1}{1+sh^{-\alpha}}\Big) f_H(h|v_0)\,\mathrm{d}h\Big) \\
&= \exp\Big(-p\overline{n}\int_{h=0}^{\infty}\frac{s}{s+h^{\alpha}} f_H(h|v_0)\,\mathrm{d}h\Big) \\
&= \exp\Big(-p\overline{n}\int_{h=0}^{\infty}\frac{s}{s+h^{\alpha}} f_H(h|v_0)\,\mathrm{d}h\Big) \\
\mathscr{L}_{I_{\Phi_c}}(s) &\approx \exp\Big(-p\overline{n}\int_{h=0}^{\infty}\frac{s}{s+h^{\alpha}} f_H(h)\,\mathrm{d}h\Big) \qquad \text{(B.2)}
\end{aligned}
$$

where (a) follows from the Rayleigh fading assumption, (b) follows from the PGFL of the PPP Φ_c^p, (c) follows from changing of variables $z_0 = x_0 + y$, (d) follows from converting the cartesian to polar coordinates, and the approximation comes from neglecting the correlation of the intra-cluster interfering distances, i.e., the common part v_0, as in [76]. Hence, Lemma 2 is proven.

B.3 Proof of lemma 4.3.0.3

First, to prove concavity, we proceed as follows.

$$\begin{aligned}\frac{\partial \mathbb{P}_o}{\partial b_i} &= q_i + q_i\big(\overline{n}(1-b_i)e^{-\overline{n}b_i} - (1-e^{-\overline{n}b_i})\big)\mathbb{P}(R_1 > R_0)\\ \frac{\partial^2 \mathbb{P}_o}{\partial b_i \partial b_j} &= -q_i\big(\overline{n}e^{-\overline{n}b_i} + \overline{n}^2(1-b_i)e^{-\overline{n}b_i} + \overline{n}e^{-\overline{n}b_i}\big)\mathbb{P}(R_1 > R_0)\end{aligned} \tag{B.3}$$

It is clear that the second derivative $\frac{\partial^2 \mathbb{P}_o}{\partial b_i \partial b_j}$ is negative. Hence, the Hessian matrix $\mathbf{H}_{i,j}$ of $\mathbb{P}_o(p^*, b_i)$ w.r.t. b_i is negative semidefinite, and the function $\mathbb{P}_o(p^*, b_i)$ is concave with respect to b_i. Also, the constraints are linear, which imply that the necessity and sufficiency conditions for optimality exist. The dual Lagrangian function and the KKT conditions are then employed to solve **P2**. The KKT Lagrangian function of the energy minimization problem is given by

$$\mathcal{L}(b_i, w_i, \mu_i, v) = \sum_{i=1}^{N_f} q_i b_i + q_i(1-b_i)(1-e^{-b_i\overline{n}})\mathbb{P}(R_1 > R_0) + v(M - \sum_{i=1}^{N_f} b_i) + \sum_{i=1}^{N_f} w_i(b_i - 1) - \sum_{i=1}^{N_f} \mu_i b_i \tag{B.4}$$

where v, w_i, μ_i are the dual equality and two inequality constraints, respectively. Now, the optimality conditions are written as,

$$\boldsymbol{\nabla}_{b_i}\mathcal{L}(b_i^*, w_i^*, \mu_i^*, v^*) = q_i + q_i\big(\overline{n}(1-b_i)e^{-\overline{n}b_i} - (1-e^{-\overline{n}b_i})\big)\mathbb{P}(R_1 > R_0) - v^* + w_i^* - \mu_i^* = 0 \tag{B.5}$$

$$w_i^* \geq 0 \tag{B.6}$$

$$\mu_i^* \leq 0 \tag{B.7}$$

$$w_i^*(b_i^* - 1) = 0 \tag{B.8}$$

$$\mu_i^* b_i^* = 0 \tag{B.9}$$

$$(M - \sum_{i=1}^{N_f} b_i^*) = 0 \tag{B.10}$$

1. $w_i^* > 0$: We have $b_i^* = 1$, $\mu_i^* = 0$, and

$$\begin{aligned} & q_i - q_i(1 - e^{-\overline{n}})\mathbb{P}(R_1 > R_0) = v^* - w_i^* \\ & v^* < q_i - q_i(1 - e^{-\overline{n}})\mathbb{P}(R_1 > R_0) \end{aligned} \tag{B.11}$$

2. $\mu_i^* < 0$: We have $b_i^* = 0$, and $w_i^* = 0$, and

$$\begin{aligned} & q_i + \overline{n} q_i \mathbb{P}(R_1 > R_0) = v^* + \mu_i^* \\ & v^* > q_i + \overline{n} q_i \mathbb{P}(R_1 > R_0) \end{aligned} \tag{B.12}$$

3. $0 < b_i^* < 1$: We have $w_i^* = \mu_i^* = 0$, and

$$v^* = q_i + q_i\big(\overline{n}(1 - b_i^*)e^{-\overline{n}b_i} - (1 - e^{-\overline{n}b_i})\big)\mathbb{P}(R_1 > R_0) \tag{B.13}$$

By combining (B.11), (B.12), and (B.13), with the fact that $\sum_{i=1}^{N_f} b_i^* = M$, Lemma 3 is proven.

B.4 Proof of lemma 4.5.0.1

Under the assumption of one active D2D link within a cluster, there is no intra-cluster interference. Also, the Laplace transform of the inter-cluster interference is similar to that of the PPP [108] whose density is the same as that of the parent PPP. In fact, this is true according to the displacement theory of the PPP [110], where each interferer is a point of a PPP that is displaced randomly and independently of all other points. For the sake of completeness, we prove it here. Starting from the

third line of the proof of Lemma 1, we get

$$\begin{aligned}\mathscr{L}_{I_{\Phi_p^!}}(s) &\overset{(a)}{=} \exp\Bigg(-2\pi\lambda_p \mathbb{E}_{g_u} \int_{v=0}^{\infty} \mathbb{E}_{u|v}\Big[1-e^{-sP_d g_u u^{-\alpha}}\Big] v\,\mathrm{d}v \Bigg), \\ &= \exp\Big(-2\pi\lambda_p \mathbb{E}_{g_u}\big[\int_{v=0}^{\infty}\int_{u=0}^{\infty} \big(1-e^{-sP_d g_u u^{-\alpha}}\big) f_U(u|v)\,\mathrm{d}u\, v\,\mathrm{d}v\,\big]\Big) \\ &\overset{(b)}{=} \exp\Bigg(-2\pi\lambda_p \mathbb{E}_{g_u} \underbrace{\int_{v=0}^{\infty} v\,\mathrm{d}v - \int_{v=0}^{\infty}\int_{u=0}^{\infty} e^{-sP_d g_u u^{-\alpha}} f_U(u|v)\,\mathrm{d}u\, v\,\mathrm{d}v}_{\mathcal{R}(s,\alpha)} \Bigg)\end{aligned} \tag{B.14}$$

where (a) follows from the PGFL of the parent PPP [108], and (b) follows from $\int_{u=0}^{\infty} f_U(u|v)\,\mathrm{d}u = 1$. Now, we proceed by calculating the integrands of $\mathcal{R}(s,\alpha)$ as follows.

$$\begin{aligned}\mathcal{R}(s,\alpha) &\overset{(c)}{=} \int_{v=0}^{\infty} v\,\mathrm{d}v - \int_{u=0}^{\infty} e^{-sP_d g_u u^{-\alpha}} \int_{v=0}^{\infty} f_U(u|v) v\,\mathrm{d}v\,\mathrm{d}u \\ &\overset{(d)}{=} \int_{v=0}^{\infty} v\,\mathrm{d}v - \int_{u=0}^{\infty} e^{-sP_d g_u u^{-\alpha}} u\,\mathrm{d}u \\ &\overset{(e)}{=} \int_{u=0}^{\infty} (1-e^{-sP_d g_u u^{-\alpha}}) u\,\mathrm{d}u \\ &\overset{(f)}{=} \frac{(sP_d g_u)^{2/\alpha}}{\alpha} \int_{u=0}^{\infty} (1-e^{-t}) t^{-1-\frac{2}{\alpha}} du \\ &\overset{(g)}{=} \frac{(sP_d)^{2/\alpha}}{2} g_u^{2/\alpha} \Gamma(1+2/\alpha),\end{aligned} \tag{B.15}$$

where (c) follows from changing the order of integration, (d) follows from $\int_{v=0}^{\infty} f_U(u|v) v\,\mathrm{d}v = u$, (e) follows from changing the dummy variable v to u, (f) follows from changing the variables $t = sg_u u^{-\alpha}$, and (g) follows from solving the integration of (f) by parts. Substituting the obtained value for $\mathcal{R}(s,\alpha)$ into (B.14), and taking the expectation over the exponential random variable g_u, with the fact that $\mathbb{E}_{g_u}[g_u^{2/\alpha}] = \Gamma(1-2/\alpha)$, we get

$$\mathscr{L}_{I_{\Phi_p^!}}(s) = \exp\Big(-\pi\lambda_p (sP_d)^{2/\alpha} \Gamma(1+2/\alpha)\Gamma(1-2/\alpha)\Big), \tag{B.16}$$

Substituting this expression with the distance probability density function $f_R(r)$ into the coverage probability equation yields

$$\begin{aligned}
\mathrm{P_{c_d}} &= \int_{r=0}^{\infty} \mathrm{e}^{-\pi\lambda_p (sP_d)^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)} \frac{r}{2\sigma^2} \mathrm{e}^{\frac{-r^2}{4\sigma^2}} \mathrm{dr}, \\
&\overset{(h)}{=} \int_{r=0}^{\infty} \frac{r}{2\sigma^2} \mathrm{e}^{-\pi\lambda_p \theta^{2/\alpha} r^2 \Gamma(1+2/\alpha)\Gamma(1-2/\alpha)} \mathrm{e}^{\frac{-r^2}{4\sigma^2}} \mathrm{dr}, \\
&\overset{(i)}{=} \int_{r=0}^{\infty} \frac{r}{2\sigma^2} \mathrm{e}^{-r^2 Z(\theta,\sigma,\alpha)} \mathrm{dr}, \\
&= \frac{1}{4\sigma^2 Z(\theta,\alpha,\sigma)}
\end{aligned} \tag{B.17}$$

where (h) comes from the substitution ($s = \frac{\theta r^\alpha}{P_d}$), and (i) from $Z(\theta,\alpha,\sigma) = (\pi\lambda_p\theta^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha) + \frac{1}{4\sigma^2})$.

Appendix C

Appendix C

C.1 Proof of Lemma 5.4.0.1

In the following, by saying $u \in \Phi_{cp}$, we mean that $\boldsymbol{y} \in \Phi_{cp}$, where $u = \|\boldsymbol{x} + \boldsymbol{y}\|$.

$$\begin{aligned}\mathscr{L}_{I_{\text{out}}}(t) &= \mathbb{E}\left[e^{-t\gamma_d \sum_{\Phi_p^!} \sum_{u\in\Phi_{cp}} G_u u^{-\alpha}}\right] \\ &= \mathbb{E}_{\Phi_p}\left[\prod_{\Phi_p^!} \mathbb{E}_{\Phi_{cp},G_u} e^{-t\gamma_d \sum_{u\in\Phi_{cp}} G_u u^{-\alpha}}\right] \qquad \text{(C.1)} \\ &= \mathbb{E}_{\Phi_p}\Big[\prod_{\Phi_p^!} \mathbb{E}_{\Phi_{cp}} \prod_{u\in\Phi_{cp}} \mathbb{E}_{u,G_u} e^{-t\gamma_d G_u u^{-\alpha}}\Big], \qquad \text{(C.2)}\end{aligned}$$

where $G_u = G_{\boldsymbol{y}}$ for ease of exposition; from the Rayleigh fading assumption, we get

$$\begin{aligned}\mathscr{L}_{I_{\text{out}}}(t) &= \mathbb{E}_{\Phi_p}\Big[\prod_{\Phi_p^!} \mathbb{E}_{\Phi_{cp}} \prod_{u\in\Phi_{cp}} \mathbb{E}_u \frac{1}{1+t\gamma_d u^{-\alpha}}\Big] \\ &\overset{(a)}{=} \mathbb{E}_{\Phi_p}\Big[\prod_{\Phi_p^!} \exp\Big(-p\bar{n}\int_{u=0}^{\infty}\Big(1-\frac{1}{1+t\gamma_d u^{-\alpha}}\Big) f_U(u|v)\,\mathrm{d}u\Big],\end{aligned}$$

where (a) follows from the probability generating functional (PGFL) of the Gaussian PPP Φ_{cp}. Notice that, in step (a), the PGFL of the Gaussian PPP Φ_{cp} is adopted for the intensity function given in polar coordinates rather than Cartesian coordinates, i.e., $v = \|\boldsymbol{x}\|$ and $u = \|\boldsymbol{x}+\boldsymbol{y}\|$. Substituting $\int_{u=0}^{\infty}\big(1-\frac{1}{1+t\gamma_d u^{-\alpha}}\big) f_{U|V}(u|v)\,\mathrm{d}u =$

$\int_{u=0}^{\infty} \frac{t\gamma_d}{u^\alpha + t\gamma_d} f_{U|V}(u|v)\,\mathrm{d}u = \zeta(v,t)$, we get

$$\begin{aligned}\mathscr{L}_{I_{\text{out}}}(t) &= \mathbb{E}_{\Phi_p}\Big[\prod_{\Phi_p^!} \exp\big(-p\bar{n}\zeta(v,t)\big)\Big] \\ &\overset{(b)}{=} \exp\bigg(-2\pi\lambda_p \int_{v=0}^{\infty} \Big(1-\exp\big(-p\bar{n}\zeta(v,t)\big)\Big) v\,\mathrm{d}v\bigg), \end{aligned} \tag{C.3}$$

where (b) follows from the PGFL of the PPP Φ_p. Hence, Lemma 5.4.0.1 is proven.

C.2 Proof of Theorem 5.4.1.1

By conditioning on $S_{\Phi_{cpm}} = s_{\Phi_{cpm}} = \sum_{i=1}^{k} h_i^{-\alpha}$, we derive a bound on Laplace transform of inter-cluster interference based on Taylor's series expansion. Starting from equation (C.2) in Appendix C.1, we have

$$\begin{aligned}\mathcal{L}_{I_{\text{out}}}(t|k) &= \mathbb{E}_{\Phi_p}\bigg[\prod_{\Phi_p^!} \mathbb{E}_{\Phi_{cp}} \prod_{u\in\Phi_{cp}} \mathbb{E}_{u,G_u} \exp\big(-tG_u u^{-\alpha}\big)\bigg] \\ &\overset{(a)}{=} \mathbb{E}_{G_u}\exp\Big(-2\pi\lambda_p \int_{v=0}^{\infty} \big(1-\exp\big(-p\bar{n}(1-\zeta'(v,t))\big)\big) v\,\mathrm{d}v\Big) \\ &\overset{(b)}{\approx} \mathbb{E}_{G_u}\exp\Big(-2\pi\lambda_p \int_{v=0}^{\infty} \big(1-(1-p\bar{n}(1-\zeta'(v,t))\big) v\,\mathrm{d}v\Big) \\ &= \exp\Big(-2\pi p\bar{n}\lambda_p \overbrace{\Big(\int_{v=0}^{\infty} v\,\mathrm{d}v - \mathbb{E}_{G_u}\int_{v=0}^{\infty} \zeta'(v,t) v\,\mathrm{d}v\Big)}^{J(t)}\Big), \end{aligned}$$

where $\zeta'(v,t) = \int_{u=0}^{\infty} e^{-t\gamma_d G_u u^{-\alpha}} f_{U|V}(u|v)\,\mathrm{d}u$ and $G_u = G_{\boldsymbol{y}}$ for ease of notation; (a) follows from tracking the proof of Lemma 5.4.0.1 up until equation (C.3), (b) follows from Taylor series expansion for exponential function $e^{-x} \approx 1-x$ when x is small. It is worth mentioning that the obtained $\mathscr{L}_{I_{\text{out}}}(t|k)$ in the above is Laplace transform of an upper bound on the interference. Correspondingly, the resulting rate coverage probability Υ_m and offloading gain $\mathbb{P}_o^{\sim}(\boldsymbol{c})$ are lower bounds on their exact values. We proved in **[amer2018minimizing]** that $J(t) = \frac{(t\gamma_d)^{2/\alpha}}{2}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)$, which proves (5.16). Plugging the result obtained in (5.16) into (5.6) yields the lower bound on the offloading gain in (5.17), which completes the proof.

C.3 Proof of Lemma 5.4.2.1

With reference to Fig. 5.1, the nearest serving distance h_1 is defined as the distance from the typical client at $(0,0)$ to its nearest provider within the same cluster. Following [228], we define the point generating function (PGF) of the number of active clients that cache content m within a ball $\mathbf{b}(o,h_1)$ with radius h_1 and centered around the origin o as:

$$\begin{aligned} G_N(\vartheta) &= \mathbb{E}\left[\vartheta^{\sum_{\boldsymbol{y}_{0i}\in\Phi_{cpm}}\mathbf{1}\{\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|<h_1\}}\right] \\ &= \mathbb{E}_{\Phi_c,\boldsymbol{x}_0}\prod_{\boldsymbol{y}_{0i}\in\Phi_{cpm}}\left[\vartheta^{\mathbf{1}\{\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|<h_1\}}\right] \\ &\overset{(a)}{=} \mathbb{E}_{\boldsymbol{x}_0}\exp\Big(-c_m p\bar{n}\int_{\mathbb{R}^2}(1-\vartheta^{\mathbf{1}\{\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|<h_1\}})f_{\boldsymbol{Y}_{0i}}(\boldsymbol{y}_{0i})d\boldsymbol{y}_{0i}\Big) \\ &\overset{(b)}{=} \mathbb{E}_{\boldsymbol{x}_0}\exp\Big(-c_m p\bar{n}\int_{\mathbb{R}^2}(1-\vartheta^{\mathbf{1}\{\|\boldsymbol{z}_0\|<h_1\}})f_{\boldsymbol{Y}_{0i}}(\boldsymbol{z}_0-\boldsymbol{x}_0)d\boldsymbol{z}_0\Big), \end{aligned}$$

where $\mathbf{1}\{.\}$ is the indicator function, and $\boldsymbol{x}_0\in\mathbb{R}^2$ is a RV modeling the location of representative cluster center relative to the origin o, with a realization $\boldsymbol{X}_0=\boldsymbol{x}_0$; (a) follows from the PGFL of the PPP Φ_{cpm} along with its intensity function $c_m p\bar{n} f_{\boldsymbol{Y}_{0i}}(\boldsymbol{y}_{0i})$, and (b) follows from change of variables $\boldsymbol{z}_0=\boldsymbol{x}_0+\boldsymbol{y}_{0i}$. By converting Cartesian coordinates to polar coordinates with $h=\|\boldsymbol{z}_0\|$, we get

$$\begin{aligned} G_N(\vartheta) &= \mathbb{E}_{V_0}\exp\Big(-c_m p\bar{n}\int_{h=0}^{\infty}(1-\vartheta^{\mathbf{1}\{h<h_1\}})f_{H|V_0}(h|v_0)\,\mathrm{d}h\Big) \\ &\overset{(c)}{=} \mathbb{E}_{V_0}\exp\Big(-c_m p\bar{n}\int_{h=0}^{h_1}(1-\vartheta)f_{H|V_0}(h|v_0)\,\mathrm{d}h\Big) \\ &\overset{(d)}{=} \int_{v_0=0}^{\infty}f_{V_0}(v_0)\exp\Big(-c_m p\bar{n}\int_{h=0}^{h_1}(1-\vartheta)f_{H|V_0}(h|v_0)\,\mathrm{d}h\Big)\,\mathrm{d}v_0\,, \end{aligned}$$

where $V_0\in\mathbb{R}$ is a RV modeling the distance from representative cluster's center to the origin o, with a realization $V_0=v_0=\|\boldsymbol{x}_0\|$; (c) follows from the definition of the indicator function $\mathbf{1}\{h<h_1\}$, and (d) follows from unconditioing over v_0. To clarify how the normal distribution $f_{\boldsymbol{Y}_{0i}}(\boldsymbol{z}_0-\boldsymbol{x}_0)$ is converted to the Rician distribution $f_{H|V_0}(h|v_0)$, consider first the representative cluster centered at $\boldsymbol{x}_0\in\Phi_p$, with a distance $v_0=\|\boldsymbol{x}_0\|$ from the origin. A randomly-selected active provider belonging to the representative cluster has its coordinates in $\mathbb{R}^2$ chosen independently from Gaussian distributions with standard deviation σ. Then, by definition, the distance

h from such an active provider to the origin has Rician PDF denoted as $f_{H|V_0}(h|v_0)$. Recall that $f_{V_0}(v_0) = \text{Rayleigh}(v_0, \sigma)$ from the definition of Gaussian PPP.

Now, the CDF of nearest serving distance $F_{H_1}(h_1)$ can be derived as

$$F_{H_1}(h_1) = 1 - G_N(0) = 1 - \int_{v_0=0}^{\infty} f_{V_0}(v_0) \exp\Big(- c_m p \bar{n} \int_0^{h_1} f_{H|V_0}(h|v_0)\, \mathrm{d}h \Big)\, \mathrm{d}v_0\,. \tag{C.4}$$

Applying Leibniz integral rule, we obtain the nearest distance PDF as

$$\begin{aligned} f_{H_1}(h_1) &= -\frac{\partial}{\partial h_1} \int_{v_0=0}^{\infty} f_{V_0}(v_0) e^{-c_m p \bar{n} \int_0^{h_1} f_{H|V_0}(h|v_0) \mathrm{d}h}\, \mathrm{d}v_0 \\ &= -\int_{v_0=0}^{\infty} f_{V_0}(v_0) \frac{\partial}{\partial h_1} e^{-c_m p \bar{n} \int_0^{h_1} f_{H|V_0}(h|v_0) \mathrm{d}h}\, \mathrm{d}v_0 \\ &= c_m p \bar{n} \int_0^{\infty} f_{V_0}(v_0) \frac{\partial}{\partial h_1} \Big[\int_0^{h_1} f_{H|V_0}(h|v_0)\, \mathrm{d}h \Big] e^{-c_m p \bar{n} \int_0^{h_1} f_{H|V_0}(h|v_0) \mathrm{d}h}\, \mathrm{d}v_0 \\ &= c_m p \bar{n} \int_{v_0=0}^{\infty} f_{V_0}(v_0) f_{H_1|V_0}(h_1|v_0) e^{-c_m p \bar{n} \int_0^{h_1} f_{H|V_0}(h|v_0) \mathrm{d}h}\, \mathrm{d}v_0\,, \end{aligned}$$

The distance PDF $f_{H_1}(h_1)$ can be calculated numerically from (5.19). However, a tractable yet accurate approximation can be obtained using Jensen's inequality as follows:

$$\begin{aligned} f_{H_1}(h_1) &= \frac{\partial}{\partial h_1} F_{H_1}(h_1) \\ &\overset{(a)}{\approx} \frac{\partial}{\partial h_1} \left(1 - e^{-c_m p \bar{n} \int_0^{h_1} \int_0^{\infty} f_{V_0}(v_0) f_{H|V_0}(h|v_0) \mathrm{d}v_0 \mathrm{d}h} \right) \\ &\overset{(b)}{=} \frac{\partial}{\partial h_1} \left(1 - \exp\Big(- c_m p \bar{n} \big(1 - \exp(-\frac{h_1^2}{4\sigma^2}) \big) \Big) \right) \qquad \text{(C.5)} \\ &= \frac{c_m p \bar{n} h_1 e^{-c_m p \bar{n} \left(1 - e^{\frac{-h_1^2}{4\sigma^2}}\right) - \frac{h_1^2}{4\sigma^2}}}{2\sigma^2}, \qquad \text{(C.6)} \end{aligned}$$

where (a) follows from Jensen's inequality applied to the CDF, and (b) follows from

$$F_{H_1}(h_1) = \int_0^{h_1} \int_{v_0=0}^{\infty} f_{V_0}(v_0) f_{H|V_0}(h|v_0)\, \mathrm{d}v_0\, \mathrm{d}h = 1 - \exp(\frac{-h_1^2}{2\sigma^2}). \tag{C.7}$$

This completes the proof.

C.4 Proof of Lemma 5.4.2.2

The conditional variance $\mathrm{Var}\left[S_{\Phi^!_{cpm}}|H_1=h_1\right]$ can be expressed as

$$\begin{aligned}\mathrm{Var}\left[S_{\Phi^!_{cpm}}|H_1=h_1\right] &= \mathrm{Var}\Big[\sum_{\boldsymbol{y}_{0i}\in\Phi^!_{cpm}} \|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|^{-\alpha}\Big] \\ &\overset{(a)}{=} \int_{\mathbb{R}^2} \frac{1}{\|\boldsymbol{x}_0+\boldsymbol{y}_{0i}\|^{2\alpha}} c_m p\bar{n} f_{\boldsymbol{Y}_{0i}}(\boldsymbol{y}_{0i})\,\mathrm{d}\boldsymbol{y}_{0i} \\ &\overset{(b)}{=} c_m p\bar{n}\int_{\mathbb{R}^2} \frac{1}{\|\boldsymbol{z}_0\|^{2\alpha}} f_{\boldsymbol{Y}_{0i}}(\boldsymbol{z}_0-\boldsymbol{x}_0)\,\mathrm{d}\boldsymbol{z}_0\,, \end{aligned} \tag{C.8}$$

where (a) follows from the mean and variance for PPPs [75, Corollary 4.8], along with the Gaussian PPP assumption Φ_{cm}; (b) follows from the substitution $\boldsymbol{z}_0=\boldsymbol{x}_0+\boldsymbol{y}_{0i}$, where $\{\boldsymbol{x}_0,\boldsymbol{y}_{0i},\boldsymbol{z}_0\}\in\mathbb{R}^2$. By converting the Cartesian coordinates coordinates to polar coordinates, where $h=\|\boldsymbol{z}_0\|$, and unconditioning over v_o, we get

$$\begin{aligned}\mathrm{Var}\left[S_{\Phi^!_{cpm}}|H_1=h_1\right] &= c_m p\bar{n}\int_{h_1}^{\infty} h^{-2\alpha} f_{H|V_0}(h|v_0)\,\mathrm{d}h \\ &= c_m p\bar{n}\int_{v_0=0}^{\infty} f_{V_0}(v_0)\int_{h=h_1}^{\infty} h^{-2\alpha} f_{H|V_0}(h|v_0)\,\mathrm{d}h\,\mathrm{d}v_0 \\ &\overset{(c)}{=} c_m p\bar{n}\int_{h=h_1}^{\infty} h^{-2\alpha}\int_{v_0=0}^{\infty} f_{V_0}(v_0) f_{H|V_0}(h|v_0)\,\mathrm{d}v_0\,\mathrm{d}h\,, \end{aligned} \tag{C.9}$$

where (c) follows from changing the order of integration. Finally, we proceed as follows:

$$\begin{aligned}\mathrm{Var}\big[S_{\Phi^!_{cpm}}|H_1=h_1\big] &= c_m p\bar{n}\int_{h=h_1}^{\infty} h^{-2\alpha}\int_0^{\infty} \frac{v_0}{\sigma^2} e^{\frac{-v_0^2}{2\sigma^2}} \quad \frac{h}{\sigma^2} e^{\frac{-(v_0^2+h^2)}{2\sigma^2}} I_0\big(\frac{hv_0}{\sigma^2}\big)\,\mathrm{d}v_0\,\mathrm{d}h \\ &= \frac{c_m p\bar{n}}{\sigma^2}\int_{h=h_1}^{\infty} h^{1-2\alpha}\int_{v_0=0}^{\infty} \frac{v_0}{\sigma^2} e^{\frac{-v_0^2}{2\sigma^2}} e^{\frac{-(v_0^2+h^2)}{2\sigma^2}} I_0\big(\frac{hv_0}{\sigma^2}\big)\,\mathrm{d}v_0\,\mathrm{d}h \\ &= \frac{c_m p\bar{n}}{\sigma^2}\int_{h=h_1}^{\infty} h^{1-2\alpha} e^{\frac{-h^2}{2\sigma^2}}\int_{v_0=0}^{\infty} \frac{v_0}{\sigma^2} e^{\frac{-v_0^2}{\sigma^2}} I_0\big(\frac{hv_0}{\sigma^2}\big)\,\mathrm{d}v_0\,\mathrm{d}h \quad \text{(C.10)} \\ &\overset{(d)}{=} \frac{c_m p\bar{n}}{2\sigma^2}\int_{h_1}^{\infty} h^{1-2\alpha} e^{\frac{-h^2}{4\sigma^2}}\,\mathrm{d}h \overset{(e)}{=} c_m p\bar{n}\int_{\frac{h_1^2}{4\sigma^2}}^{\infty} \tau^{-2\alpha} e^{-\tau}\,\mathrm{d}\tau \quad \text{(C.11)} \\ &\overset{(f)}{=} c_m p\bar{n}\Gamma\left(-2\alpha+1,\frac{h_1^2}{4\sigma^2}\right), \quad \text{(C.12)} \end{aligned}$$

where (d) follows from solving the inner integratal of (C.10), (e) follows from the substitution $\tau=\frac{h^2}{4\sigma^2}$, and (f) follows from solving the integration of (C.11), where $\Gamma(\cdot,\cdot)$ denotes the upper incomplete gamma function. This completes the proof.

Appendix D

Appendix D

D.1 Proof of Theorem 6.3.0.2

The SCDP is defined as the probability of downloading content with a received SIR higher than a target threshold ϑ, i.e.,

$$\begin{aligned}\mathbb{P}^{v}_{\mathrm{c}|r} &= \mathbb{P}\Big(\frac{\frac{P_t}{K}\zeta_v(r)\left|\boldsymbol{w}_{i1}\boldsymbol{h}_{i1}\right|^2}{I} > \vartheta\Big)\\ &= \mathbb{P}\Big(\left|\boldsymbol{w}_{i1}\boldsymbol{h}_{i1}\right|^2 > \frac{\vartheta K}{P_t\zeta_v(r)}I\Big)\\ &\overset{(a)}{=} \mathbb{E}_I\left[\sum_{i=0}^{M_v-1}\frac{s_v^i}{i!}I^i e^{-s_v I}\right]\\ &\overset{(b)}{=} \sum_{i=0}^{M_v-1}\frac{(-s_v)^i}{i!}\mathcal{L}^{(i)}_{I|r}(s_v),\end{aligned} \tag{D.1}$$

where $I = I_{\text{in}} + I_{\text{out}}$, (a) follows from $|\boldsymbol{w}_{i1}\boldsymbol{h}_{i1}|^2 \sim \Gamma(M_v, \frac{\eta}{m_v})$, and (b) follows from the Laplace transform of interference, along with the assumption of independence

between the intra- and inter-cell interference. Next, we derive the Laplace transform of interference from:

$$\begin{aligned}
\mathcal{L}_{I|r}(s_v) &= \mathbb{E}_I\left[e^{-s_v I}\right] \\
&= \mathbb{E}_{h_{iK}} e^{-s_v h_{iK} P(r)^2} \mathbb{E}_{\Phi} \prod_{j\in\Phi^o} \mathbb{E}_{h_{jK}} e^{-s_v h_{jK} P(u_j)^2} \\
&\overset{(a)}{=} \left(1+s_v \eta P_v(r)^2\right)^{-(K-1)} e^{-2\pi\lambda \mathbb{E}_{h_{jK}} \int_{\nu=r}^{\infty} \left(1-\exp\left(-s_v h_{jK} P(\nu)^2\right)\right)\nu \mathrm{d}\nu} \\
&\overset{(b)}{=} e^{-(K-1)\log(1+s_v \eta P_v(r)^2)} e^{-2\pi\lambda \mathbb{E}_{h_{jK}} \int_{\nu=r}^{\infty} \left(1-\exp\left(-s_v h_{jK} P(\nu)^2\right)\right)\nu \mathrm{d}\nu} \\
&\overset{(c)}{=} e^{-(K-1)\log(1+s_v \eta P_v(r)^2)} e^{-2\pi\lambda \int_{\nu=r}^{\infty} \left(1-\mathbb{P}_l(\nu)\delta_l(\nu,s_v)-\mathbb{P}_n(\nu)\delta_n(\nu,s_v)\right)\nu \mathrm{d}\nu} \\
&= e^{\varpi(s_v)},
\end{aligned}$$

where (a) follows from $h_{iK} \sim \Gamma(K-1,\eta)$ and the PGFL of PPP Φ [75]. (b) follows from the fact that $x = e^{\log(x)}$, and (c) follows since $h_{jK} \sim \Gamma(K,\eta)$. In [147], it is proved that $\sum_{i=0}^{M_v-1} \frac{(-s_v)^i}{i!}\mathcal{L}_{I|r}^{(i)}(s_v) = \sum_{i=0}^{M_v-1} p_i$, with $p_i = \frac{(-s_v)^i}{i!}\mathcal{L}_{I|r}^{(i)}(s_v)$ computed from the recursive relation: $p_i = \sum_{l=0}^{i-1} \frac{i-l}{i} p_l t_{i-l}$, where $t_k = \frac{(-s_v)^k}{k!}\varpi^{(k)}(s_v)$. After some algebraic manipulation, $\mathbb{P}_{\mathrm{c}|r}^{v}$ can be expressed in a compact form $\mathbb{P}_{\mathrm{c}|r}^{v} = \|e^{\boldsymbol{T}_{M_v}}\|_1$ as in [147]. In summary, we first derive the conditional log-Laplace transform $\varpi(s_v)$ of the aggregate interference. Then, we calculate the n-th derivative of $\varpi(s_v)$ to populate the entries t_n of the lower triangular Toeplitz matrix $\boldsymbol{T}_{M_v}$. The conditional SCDP can be then computed from $\mathbb{P}_{\mathrm{c}|r}^{v} = \|e^{\boldsymbol{T}_{M_v}}\|_1$.

Appendix E

Appendix E

E.1 Proof of Corollary 7.5.1.1

To study scalability, we consider a simple case with $m_v = A_v = 1$, $v \in \{l, n\}$, hence, $\varpi_l = \frac{\vartheta d_0^{\alpha_l}}{G(r_0,h)}$. We also assume that N_t is small and $\frac{h}{r} \to 0$, which is a reasonable assumption for sparsely-deployed networks. The conditional coverage probability $\mathbb{P}_{c|r_0}$ is then simplified to:

$$\begin{aligned}
\mathbb{P}_{c|r_0} &= e^{-2\pi\lambda_b \int_{r=r_0}^{\infty} \frac{\varpi_l \zeta_v(r) r}{1+\varpi_l \zeta_v(r)} \mathrm{d}r} \\
&\overset{(a)}{\geq} 1 - 2\pi\lambda_b \int_{r_0}^{\infty} \frac{\frac{\vartheta d_0^{\alpha_l} G(r,h) d^{-\alpha_v}}{G(r_0,h)}}{1 + \frac{\vartheta d_0^{\alpha_l} G(r,h) d^{-\alpha_v}}{G(r_0,h)}} r \, \mathrm{d}r \\
&= 1 - 2\pi\lambda_b \int_{r_0}^{\infty} \frac{\vartheta d_0^{\alpha_l} G(r,h) d^{-\alpha_v}}{G(r_0,h) + \vartheta d_0^{\alpha_l} G(r,h) d^{-\alpha_v}} r \, \mathrm{d}r \,, \qquad \text{(E.1)}
\end{aligned}$$

where (a) follows from $e^{-x} \geq 1 - x$. Recall that $G(r,h) = \frac{1}{N_t} \frac{\sin^2 \frac{N_t \pi}{2} (\sin(\arctan(\frac{h}{r})))}{\sin^2 \frac{\pi}{2} (\sin(\arctan(\frac{h}{r})))}$. From [229], we have $\arctan(\frac{h}{r}) \simeq \frac{4h}{\pi r}$ and $\sin(\frac{4h}{\pi r}) \simeq \frac{4h}{\pi r}$ for $h \ll r$, since $\sin(x) \simeq x$ for small x. Also, $\sin^2 \frac{N_t \pi}{2} (\frac{4h}{\pi r}) \simeq (\frac{2N_t h}{r})^2$ for small N_t and $h \ll r$.

$$\begin{aligned}
\mathbb{P}_{c|r_0} &\simeq 1 - 2\pi\lambda_b \int_{r_0}^{\infty} \frac{\vartheta (d_0^{\alpha_l}/d^{\alpha_v}) N_t}{N_t + \vartheta (d_0^{\alpha_l}/d^{\alpha_v}) N_t} r \, \mathrm{d}r \\
&= 1 - 2\pi\lambda_b \int_{r_0}^{\infty} \frac{\vartheta (d_0^{\alpha_l}/d^{\alpha_v})}{1 + \vartheta (d_0^{\alpha_l}/d^{\alpha_v})} r \, \mathrm{d}r \,. \qquad \text{(E.2)}
\end{aligned}$$

Since $\mathbb{P}_{c|r_0}$ in (E.2) is not a function of N_t, the UAV-UE coverage probability does not scale asymptotically with N_t, and the effect of N_t on the coverage probability is minor.

Appendix F

Appendix F

F.1 Proof of Theorem 8.3.0.1

We proceed to obtain an UB on the coverage probability as follows:

$$\begin{aligned}\mathbb{P}\Big(\frac{\kappa P_t J}{I_{\text{out}}} > \vartheta\Big) &= \mathbb{P}\Big(\kappa P_t J > \vartheta I_{\text{out}}\Big)\\ &= \mathbb{E}_{I_{\text{out}}}\Bigg[\mathbb{P}\Big(\kappa P_t J > \vartheta I_{\text{out}}\Big)\Bigg]\\ &\overset{(a)}{\approx} \mathbb{E}_{I_{\text{out}}}\Big[\sum_{i=0}^{K-1}\frac{(\vartheta/\kappa P_t\theta)^i}{i!}I_{\text{out}}^i \exp\Big(-\frac{\vartheta}{\kappa P_t\theta}I_{\text{out}}\Big)\Big]\\ &\overset{(b)}{=} \mathbb{E}_{I_{\text{out}}}\Big[\sum_{i=0}^{K-1}\frac{(-\varpi)^i}{i!}\frac{d^i}{d\varpi^i}\mathcal{L}_{I_{\text{out}}|\boldsymbol{r}_\kappa}(\varpi)\Big], \end{aligned} \tag{F.1}$$

where (a) follows from the PDF of Gamma RV whose shape and scale parameters are θ from (8.10), and $K = m_l\kappa$, respectively. (b) follows from $\varpi = \frac{\vartheta}{\kappa P_t\theta}$, along with the Laplace transform of interference, i.e., the RV I_{out}. Next, we derive the Laplace transform of interference:

$$\begin{aligned}\mathcal{L}_{I_{\text{out}}|\boldsymbol{r}_\kappa}(\varpi) = \mathbb{E}_{I_{\text{out}}}\Big[e^{-\varpi I_{\text{out}}}\Big] &= \mathbb{E}\Bigg[e^{-\sum_{j\in\Phi_b\setminus\mathcal{B}(0,R_c)}\varpi\chi_j P(u_j)^2}\Bigg]\\ &= \mathbb{E}_{\Phi_b,\chi_j}\Bigg[\prod_{j\in\Phi_b\setminus\mathcal{B}(0,R_c)} e^{-\varpi\chi_j P(u_j)^2}\Bigg]\\ &\overset{(a)}{=} \exp\Bigg(-2\pi\lambda_b\int_{\upsilon=R_c}^{\infty}\Big(1-\mathbb{E}_\chi e^{-\varpi\chi P(\upsilon)^2}\Big)\upsilon\,\mathrm{d}\upsilon\Bigg) \qquad \text{(F.2)}\\ &\overset{(b)}{=} \exp\Bigg(-2\pi\lambda_b\int_{\upsilon=R_c}^{\infty}\Big(1-\delta_l\mathbb{P}_l(\upsilon)-\delta_n\mathbb{P}_n(\upsilon)\Big)\upsilon\,\mathrm{d}\upsilon\Bigg) \qquad \text{(F.3)}\\ &\overset{(c)}{=} e^{\Omega(\varpi)_{|\boldsymbol{r}_\kappa}}, \qquad \text{(F.4)}\end{aligned}$$

where $\delta_l = \left(1+\frac{\varpi P_l(\upsilon)^2}{m_l}\right)^{-m_l}$, and $\delta_n = \left(1+\frac{\varpi P_n(\upsilon)^2}{m_n}\right)^{-m_n}$; (a) follows from the PGFL of PPP along with Cartesian to polar coordinates conversion [75], (b) follows from the moments of the Gamma RV $\chi \sim \mathrm{Gamma}(m_\upsilon, 1/m_\upsilon)$ modeling the interfering channel gain, and (c) follows from $\Omega(\varpi)_{|\boldsymbol{r}_\kappa} = -2\pi\lambda_b \int_{\upsilon=R_c}^{\infty} \Big(1-\delta_l \mathbb{P}_l(\upsilon) - \delta_n \mathbb{P}_n(\upsilon)\Big)\upsilon \,\mathrm{d}\upsilon$. In [147], it is proved that $\sum_{i=0}^{K-1} \frac{(-\varpi)^i}{i!}\mathcal{L}^{(i)}_{I|\boldsymbol{r}_\kappa}(\varpi) = \sum_{i=0}^{K-1} p_i$, where $p_i = \frac{(-\varpi)^i}{i!}\mathcal{L}^{(i)}_{I|\boldsymbol{r}_\kappa}(\varpi)$ can be computed from the recursive relation: $p_i = \sum_{l=0}^{i-1} \frac{i-l}{i} p_l t_{i-l}$, with $t_{i-1} = \frac{(-\varpi)^{i-1}}{(i-1)!}\Omega^{(i-1)}(\varpi)$, and $\Omega^{(i-1)}(\varpi) = \frac{d^{i-1}}{d\varpi^{i-1}}\Omega(\varpi)_{|\boldsymbol{r}_\kappa}$. After some algebraic manipulation as in [147], $\mathbb{P}^l_{\mathrm{c}|\boldsymbol{r}}$ can be expressed in a compact form $\mathbb{P}^l_{\mathrm{c}|\boldsymbol{r}} = \|e^{\boldsymbol{T}_k}\|_1$, where $\|.\|_1$ represents the induced ℓ_1 norm, and $\boldsymbol{T}_K$ is the lower triangular Toeplitz matrix whose entries are t_i, $i = |\{1,\ldots,K\}$. This completes the proof.

F.2 Proof of Corollary 8.3.0.1

We first write the exponent power of (F.2) as

$$\begin{aligned}\Omega(\varpi)_{|\boldsymbol{r}_\kappa} &= -2\pi\lambda_b \mathbb{E}_\chi \int_{\upsilon=R_c}^{\infty} \big(1-e^{-\varpi\chi P(\upsilon)^2}\big)\upsilon \,\mathrm{d}\upsilon \\ &\overset{(a)}{=} -2\pi\lambda_b \mathbb{E}_\chi \int_{\upsilon=R_c}^{\infty} \big(1-e^{-\varpi\varsigma\chi(\upsilon^2+h^2)^{-\alpha_l/2}}\big)\upsilon \,\mathrm{d}\upsilon\,,\end{aligned}$$

where (a) follows from $P(\upsilon)^2 = P_t A_l G_s \big(\upsilon^2+h^2\big)^{-\alpha_l/2}$ and substituting $\varsigma = P_t A_l G_s$. Let $g = \upsilon^2 + h^2$, and $\mathrm{d}g = 2\upsilon \,\mathrm{d}\upsilon$, we hence get

$$\Omega(\varpi)_{|\boldsymbol{r}_\kappa} = -\pi\lambda_b \mathbb{E}_\chi \int_{g=R_c^2+h^2}^{\infty} \big(1-e^{-\varpi\varsigma\chi g^{-\alpha_l/2}}\big)\,\mathrm{d}g\,. \tag{F.5}$$

By changing the variables $y = g^{-\alpha_l/2}$, $g = y^{-2/\alpha_l}$, and $\mathrm{d}g = \frac{-2}{\alpha_l} y^{\frac{-2}{\alpha_l}-1}\,\mathrm{d}y$, and solving the reproduced integrals as in [230], we get

$$\begin{aligned}\Omega(\varpi)_{|\boldsymbol{r}_\kappa} &= \pi\lambda_b R_{ch}^2 - \delta_l \pi\lambda_b (\varpi\varsigma)^{\delta_l}\mathbb{E}_\chi\Big[\chi^{\delta_l}\gamma(-\delta_l, \varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big] \\ &\overset{(a)}{=} \pi\lambda_b R_{ch}^2 - \delta_l \pi\lambda_b (\varpi\varsigma)^{\delta_l}\mathbb{E}_\chi\Big[\chi^{\delta_l}\epsilon_1 F_1(-\delta_l; 1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big],\end{aligned} \tag{F.6}$$

where $R_{ch}^2 = R_c^2 + h^2$, $\delta_l = \frac{2}{\alpha_l}$, $\epsilon = \frac{(\varpi\varsigma\chi)^{-\delta_l} R_{ch}^2}{\delta_l}$, and $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}$ is the lower incomplete Gamma function; (a) follows from ${}_1F_1(s; s+1; -x) = \frac{s}{x^s}\gamma(s,x)$, where ${}_1F_1(\cdot;\cdot;\cdot)$ is the confluent hypergeometric function of the first kind. By rearranging

(F.6), we can obtain

$$\Omega(\varpi)_{|\boldsymbol{r}_\kappa} = \pi\lambda_b R_{ch}^2 \left(1 - \mathbb{E}_\chi\Big[{}_1F_1(-\delta_l; 1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big]\right). \tag{F.7}$$

The non-zero terms in $\boldsymbol{T}_k$ can be then determined from:

$$\begin{aligned}
t_k &= \frac{(-\varpi)^k}{k!}\Omega(\varpi)^{(k)}_{|\boldsymbol{r}_\kappa} \\
&= \pi\lambda_b R_{ch}^2 \frac{(-\varpi)^k}{k!}\frac{d^k}{d\varpi^k}\left[1 - \mathbb{E}_\chi\Big[{}_1F_1(-\delta_l; 1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big]\right] \\
&= \pi\lambda_b R_{ch}^2 \mathbb{E}_\chi\left[\frac{(-\varpi)^k}{k!}(-\varsigma\chi R_{ch}^{-\alpha_l/2})^k \frac{d^k}{d(-\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})^k}\Big[1 - {}_1F_1(-\delta_l; 1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big]\right] \\
&= \pi\lambda_b R_{ch}^2 \mathbb{E}_\chi\left[\frac{(\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})^k}{k!} \frac{d^k}{d(-\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})^k}\Big[1 - {}_1F_1(-\delta_l; 1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big]\right] \\
&\overset{(b)}{=} \pi\lambda_b R_{ch}^2 \left(\mathbf{1}\{k=0\} - \frac{(\varpi\varsigma R_{ch}^{-\alpha_l/2})^k}{\Gamma(k+1)}\frac{\delta_l}{(\delta_l - k)}\mathbb{E}_\chi\Big[\chi^k {}_1F_1(k-\delta_l; k+1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big]\right),
\end{aligned}$$

where (b) follows from the derivatives for hypergeometric functions: $\frac{d^k}{dz^k}{}_1F_1(a;b;z) = \frac{\prod_{p=0}^{k-1}(a+p)}{\prod_{p=0}^{k-1}(b+p)} \times {}_1F_1(a+k; b+k; z)$. By letting $a_k = (\varpi\varsigma R_{ch}^{-\alpha_l/2})^k$, we get

$$t_k = \pi\lambda_b R_{ch}^2 \left(\mathbf{1}\{k=0\} - \frac{\delta_l a_k}{(\delta_l - k)\Gamma(k+1)}\mathbb{E}_\chi\Big[\chi^k {}_1F_1(k-\delta_l; k+1-\delta_l; -\varpi\varsigma\chi R_{ch}^{-\alpha_l/2})\Big]\right).$$

Lastly, to get a closed-form expression for t_k, we average over $\chi \sim \text{Gamma}(m_l, 1/m_l)$ as follows:

$$\begin{aligned}
t_k &= \pi\lambda_b R_{ch}^2 \left(\mathbf{1}\{k=0\} - b_k \int_{\chi=0}^{\infty} \chi^{k+m_l-1} e^{-m_l\chi} {}_1F_1(k-\delta_l; k+1-\delta_l; -\varpi\varsigma R_{ch}^{-\alpha_l/2}\chi)\,\mathrm{d}\chi\right) \\
&= \pi\lambda_b R_{ch}^2 \left(\mathbf{1}\{k=0\} - b_k\Gamma(k+m_l)m_l^{-(k+m_l)} {}_2F_1(k+m_l, k-\delta_l; k+1-\delta_l; -\varpi\varsigma R_{ch}^{-\alpha_l/2} m_l)\right) \\
&\overset{(c)}{=} \pi\lambda_b R_{ch}^2 \Big(\mathbf{1}\{k=0\} - c_k {}_2F_1(k+m_l, k-\delta_l; k+1-\delta_l; -\varpi\varsigma R_{ch}^{-\alpha_l/2} m_l)\Big),
\end{aligned} \tag{F.8}$$

where $b_k = \frac{\delta_l a_k m_l^{m_l}}{(\delta_l - k)\Gamma(k+1)\Gamma(m_l)}$, $c_k = \frac{m_l^{m_l}}{\Gamma(m_l)} b_k = \frac{\delta_l a_k \Gamma(k+m_l) m_l^{-k}}{(\delta_l - k)\Gamma(k+1)\Gamma(m_l)}$, and (c) follows from solving the integral in (F.8) [231, Eq. 7.525] and rearranging the right hand side. This completes the proof.

F.3 Proof of Lemma 8.4.1.1

When $\varphi_t = 0$, the conditional probability of handover can be expressed as

$$\mathbb{P}(H|r_0) = 1 - \mathbb{E}_{\rho_t, Z_t, Z_{t-1}}\left[e^{-\pi\lambda_b\left((\bar{\nu}\cos(\varphi_t))^2 + 2r_0\bar{\nu}\cos(\varphi_t)\right)}\right] \tag{F.9}$$

$$\overset{(a)}{\leq} 1 - \underbrace{e^{-\pi\lambda_b\mathbb{E}_{\rho_t, Z_t, Z_{t-1}}\left(\frac{2r_0\bar{\nu}\varrho_t}{\sqrt{\varrho_t^2+(z_t-z_{t-1})^2}} + (\frac{\bar{\nu}\varrho_t}{\sqrt{\varrho_t^2+(z_t-z_{t-1})^2}})^2\right)}}_{\mathbb{P}(\bar{H}|r_0)}, \tag{F.10}$$

where (a) follows from Jensen's inequality, which implies that $\mathbb{E}[e^{-x}] \geq e^{\mathbb{E}[-x]}$ as e^{-x} is a convex function of x. $\mathbb{P}(\bar{H}|r_0)$ is an LB on the probability of no handover conditioned on r_0. We obtain $\mathbb{P}(\bar{H}|r_0)$ in (F.10) as follows:

$$\mathbb{P}(\bar{H}|r_0) = e^{-\pi\lambda_b\mathbb{E}_{\rho_t, Z_t, Z_{t-1}}\left(\frac{2r_0\bar{\nu}\varrho_t}{\sqrt{\varrho_t^2+(z_t-z_{t-1})^2}} + (\frac{\bar{\nu}\varrho_t}{\sqrt{\varrho_t^2+(z_t-z_{t-1})^2}})^2\right)}$$

$$\overset{(b)}{=} e^{-\pi\lambda_b\mathbb{E}_{Z_t, Z_{t-1}}\left[r_0\sqrt{\pi}\bar{\nu}_1F_1\left(\frac{1}{2};0;\pi(z_t-z_{t-1})^2\mu\right) + \pi\mu\bar{\nu}^2\left(\frac{1}{\pi\mu} - (z_t-z_{t-1})^2e^{\pi\mu(z_t-z_{t-1})^2}\Gamma\left(0,\pi(z_t-z_{t-1})^2\mu\right)\right)\right]},$$

where (b) follows from averaging over ρ_t whose PDF is $f_{\rho_t}(\varrho_t)$. By changing the variables: $p = z_t - z_{t-1}$, with $f_P(p) = \frac{\hbar - |p|}{\hbar^2}, \forall -\hbar \leq p \leq \hbar$, we get

$$\mathbb{P}(\bar{H}|r_0) = e^{-\frac{\pi\lambda_b r_0\sqrt{\pi}\bar{\nu}}{\hbar^2}\int_{-\hbar}^{\hbar}(\hbar-|p|)_1F_1\left(\frac{1}{2};0;\pi p^2\mu\right)\mathrm{d}p} e^{-\frac{\pi\lambda_b\pi\mu\bar{\nu}^2}{\hbar^2}\int_{-\hbar}^{\hbar}(\hbar-|p|)\left(\frac{1}{\pi\mu} - p^2e^{\pi\mu p^2}\Gamma\left(0,\pi p^2\mu\right)\right)\mathrm{d}p} \tag{F.11}$$

$$\overset{(c)}{=} e^{-\frac{\pi\lambda_b r_0\sqrt{\pi}\bar{\nu}}{\hbar^2}\left(2\frac{\pi\hbar^2\mu G^{2,2}_{2,3}\left(\hbar^2\pi\mu\middle|\begin{matrix}\frac{1}{2},\frac{1}{2}\\0,1,-\frac{1}{2}\end{matrix}\right) - G^{2,2}_{2,3}\left(\hbar^2\pi\mu\middle|\begin{matrix}1,\frac{3}{2}\\1,2,0\end{matrix}\right)}{\pi^2\mu}\right)} e^{-\pi\lambda_b\zeta(\mu,\hbar)}, \tag{F.12}$$

where (c) follows from solving the left integral of (F.11) [231, Section 7.8], and the substitution

$$\begin{aligned}\zeta(\mu,\hbar) &= \frac{\pi\mu\bar{\nu}^2}{\hbar^2}\int_{-\hbar}^{\hbar}(\hbar-|p|)\left(\frac{1}{\pi\mu} - p^2e^{\pi\mu p^2}\Gamma\left(0,\pi p^2\mu\right)\right)\mathrm{d}p\\ &= \bar{\nu}^2 - \frac{\pi\mu\bar{\nu}^2}{\hbar^2}\int_{-\hbar}^{\hbar}(\hbar-|p|)p^2e^{\pi\mu p^2}\Gamma\left(0,\pi p^2\mu\right)\mathrm{d}p\\ &\overset{(d)}{=} \bar{\nu}^2 - \frac{2\pi\mu\bar{\nu}^2}{\hbar^2}\int_{0}^{\hbar}(\hbar-p)p^2e^{\pi\mu p^2}\Gamma\left(0,\pi p^2\mu\right)\mathrm{d}p\,.\end{aligned} \tag{F.13}$$

where (d) follows from the symmetry of the integrand. From (F.12) and (F.13), with the fact that $\mathbb{P}(H|r_0) = 1 - \mathbb{P}(\bar{H}|r_0)$, the proof is completed.

F.4 Proof of Proposition 8.4.2.1

Following the Buffon's needle approach for hexagonal cells [184], we have

$$\mathbb{E}[N] = \frac{4\sqrt{3}}{3\pi\ell}\mathbb{E}[V_h]\mathbb{E}[T] = \frac{2}{\pi R_h}\mathbb{E}[V_h]\mathbb{E}[T], \tag{F.14}$$

where $\mathbb{E}[V_h]$ represents the average horizontal speed of the UAV-UE. Given the constant speed assumption, $\mathbb{E}[V_h] = \bar{\nu}\mathbb{E}[\cos(\varphi_t)]$, where $\varphi_t = \arccos\big(\frac{\varrho_t}{\sqrt{\varrho_t^2+(z_t-z_{t-1})^2}}\big)$. We hence have

$$\begin{aligned}\mathbb{E}[V_h] &= \mathbb{E}_{\rho_t, Z_t, Z_{t-1}}\Big[\frac{\bar{\nu}\varrho_t}{\sqrt{\varrho_t^2 + (z_t - z_{t-1})^2}}\Big] \\ &\overset{(a)}{=} \frac{\sqrt{\pi}\bar{\nu}}{2}\mathbb{E}_{Z_t, Z_{t-1}}\Big[{}_1F_1\big(\frac{1}{2}; 0; \pi(z_t - z_{t-1})^2\mu\big)\Big]\end{aligned} \tag{F.15}$$

where (a) follows from averaging over the RV ρ_t. By proceeding similar to Appendix F.3 to obtain $\mathbb{E}[V_h]$, the handover rate can be obtained from $H = \frac{\mathbb{E}[N]}{\mathbb{E}[T]}$. This completes the proof.

Bibliography

[1] V. N. I. Cisco, "Global mobile data traffic forecast update, 2016–2021 white paper", *Document ID, 1454457600805266*, 2016.

[2] R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Delay analysis for wireless D2D caching with inter-cluster cooperation", in *Ieee global communications conference (GLOBECOM)*, Singapore, 2017.

[3] R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Inter-cluster cooperation for wireless D2D caching networks", *Ieee transactions on wireless communications*, vol. 17, no. 9, pp. 6108–6121, 2018.

[4] R. Amer, M. M. Butt, and N. Marchetti, "Caching at the edge in low latency wireless networks", *Wireless automation as an enabler for the next industrial revolution*, pp. 209–240, 2020.

[5] R. Amer, M. M. Butt, H. ElSawy, M. Bennis, J. Kibilda, and N. Marchetti, "On minimizing energy consumption for D2D clustered caching networks", in *Ieee global communications conference (GLOBECOM)*, 2018, pp. 1–6.

[6] R. Amer, M. M. Butt, and N. Marchetti, "Optimizing joint probabilistic caching and channel access for clustered D2D networks", *In preparation for submission to journal of communications and networks*, 2020.

[7] R. Amer, H. Elsawy, M. M. Butt, E. A. Jorswieck, M. Bennis, and N. Marchetti, "Optimized caching and spectrum partitioning for D2D enabled cellular systems with clustered devices", *In ieee transactions on communications*, 2020.

[8] R. Amer, H. ElSawy, J. Kibiłda, M. M. Butt, and N. Marchetti, "Cooperative transmission and probabilistic caching for clustered D2D networks", in *Ieee wireless communications and networking conference (WCNC)*, Marrakech, Morocco, 2019.

[9] R. Amer, H. Elsawy, J. Kibiłda, M. M. Butt, and N. Marchetti, "Performance analysis and optimization of cache-assisted CoMP for clustered D2D networks", *Submitted to ieee transactions on mobile computing*, 2019.

[10] R. Amer, W. Saad, and N. Marchetti, "Towards a connected sky: Performance of beamforming with down-tilted antennas for ground and UAV user co-existence", *Ieee communications letters*, pp. 1–1, 2019.

[11] R. Amer, W. Saad, B. Galkin, and N. Marchetti, "Performance analysis of mobile cellular-connected drones under practical antenna configurations", in *In proc. of ieee international conference on communications (ICC)*, Dublin, 2020.

[12] ——, "On the performance of mobile cellular-connected drones under practical antenna configurations", *In preparation for submission to ieee transactions on vehicular technology*, 2020.

[13] R. Amer, W. Saad, H. ElSawy, M. Butt, and N. Marchetti, "Caching to the sky: Performance analysis of cache-assisted CoMP for cellular-connected UAVs", in *Proc. of the ieee wireless communications and networking conference (WCNC)*, Marrakech, Morocco, 2019.

[14] R. Amer, W. Saad, and N. Marchetti, "Mobility in the sky: Performance and mobility analysis for cellular-connected UAVs", *Ieee transactions on communications*, pp. 1–1, 2020.

[15] C. Chaccour, R. Amer, B. Zhou, and W. Saad, "On the reliability of wireless virtual reality at terahertz (THz) frequencies", in *10th ifip international conference on new technologies*, Spain, 2019.

[16] B. Galkin, R. Amer, E. Fonseca, and L. A. DaSilva, "Intelligent uav base station selection in urban environments: A supervised learning approach", in *Ieee 3rd 5g world forum (5gwf)*, 2020.

[17] B. Galkin, E. Fonseca, R. Amer, L. A. DaSilva, and I. Dusparic, "Reqiba: Regression and deep q-learning for intelligent uav cellular user to base station association", *Arxiv preprint arxiv:2010.01126*, 2020.

[18] J. G. Andrews, "Seven ways that hetnets are a cellular paradigm shift", *Ieee communications magazine*, vol. 51, no. 3, pp. 136–144, 2013.

[19] I. S. C. Trial, "Rethinking the small cell business model", 2012.

[20] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution", *Ieee communications magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[21] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks", *Ieee communications magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[22] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching", *Ieee transactions on information theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[23] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance", *Ieee journal on selected areas in communications*, vol. 34, no. 1, pp. 176–189, 2016.

[24] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations", *Ieee journal on selected areas in communications*, vol. 34, no. 4, pp. 907–922, 2016.

[25] E. Bastug, J.-L. Guénégo, and M. Debbah, "Proactive small cell networks", in *20th international conference on telecommunications (ICT)*, IEEE, 2013, pp. 1–5.

[26] V. Etter, M. Kafsi, and E. Kazemi, "Been there, done that: What your mobility traces reveal about your behavior", in *Mobile data challenge by nokia workshop, in conjunction with int. conf. on pervasive computing*, 2012.

[27] J. Robinson, P. Muller, T. Noke, T. L. Lim, W. Glausi, L. Fullerton, and D. Hamar, *Dynamic information management system and method for content delivery and sharing in content-, metadata-and viewer-based, live social networking among users concurrently engaged in the same and/or similar content*, US Patent 8,707,185, 2014.

[28] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions", *Ieee communications magazine*, vol. 54, no. 9, pp. 22–28, 2016.

[29] B. D. Higgins, J. Flinn, T. J. Giuli, B. Noble, C. Peplin, and D. Watson, "Informed mobile prefetching", in *Proceedings of the 10th international conference on mobile systems, applications, and services*, ACM, 2012, pp. 155–168.

[30] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery", *Ieee transactions on wireless communications*, vol. 13, no. 5, pp. 2894–2905, 2014.

[31] M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems", *Ieee communications surveys tutorials*, pp. 1–1, 2019.

[32] K. P. Valavanis and G. J. Vachtsevanos, *Handbook of unmanned aerial vehicles*. Springer, 2015, vol. 1.

[33] R. Austin, *Unmanned aircraft systems: UAVS design, development and deployment*. John Wiley & Sons, 2011, vol. 54.

[34] R. W. Beard and T. W. McLain, *Small unmanned aircraft: Theory and practice*. Princeton university press, 2012.

[35] M. Asadpour, B. Van den Bergh, D. Giustiniano, K. A. Hummel, S. Pollin, and B. Plattner, "Micro aerial vehicle networks: An experimental analysis of challenges and opportunities", *Ieee communications magazine*, vol. 52, no. 7, pp. 141–149, 2014.

[36] R. S. Stansbury, M. A. Vyas, and T. A. Wilson, "A survey of uas technologies for command, control, and communication (c3)", in *Unmanned aircraft systems*, Springer, 2008, pp. 61–78.

[37] I. Bucaille, S. Héthuin, A. Munari, R. Hermenier, T. Rasheed, and S. Allsopp, "Rapidly deployable network for tactical applications: Aerial base station with opportunistic links for unattended and temporary events absolute example", in *Milcom 2013-2013 ieee military communications conference*, IEEE, 2013, pp. 1116–1120.

[38] E. Baştu, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs", *Eurasip journal on wireless communications and networking*, vol. 2015, no. 1, p. 41, 2015.

[39] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers", *Ieee communications magazine*, vol. 54, no. 8, pp. 16–22, 2016.

[40] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers", *Ieee transactions on information theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[41] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks", *Ieee transactions on wireless communications*, vol. 16, no. 5, pp. 3401–3415, 2017.

[42] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks", *Ieee journal on selected areas in communications*, vol. 34, no. 5, pp. 1222–1234, 2016.

[43] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks", in *Ieee 12th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WIOPT)*, 2014, pp. 569–574.

[44] H. Ahlehagh and S. Dey, "Hierarchical video caching in wireless cloud: Approaches and algorithms", in *Communications (ICC), 2012 ieee international conference on*, IEEE, 2012, pp. 7082–7087.

[45] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative hierarchical caching in 5G cloud radio access networks", *Ieee network*, vol. 31, no. 4, pp. 35–41, 2017.

[46] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks?a technology overview", *Ieee communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

[47] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems", *Journal of communications and networks*, vol. 18, no. 2, pp. 135–149, 2016.

[48] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges", *Ieee network*, vol. 30, no. 4, pp. 46–53, 2016.

[49] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data", *Ieee communications magazine*, vol. 53, no. 10, pp. 190–199, 2015.

[50] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures", *Journal of optical communications and networking*, vol. 7, no. 11, B38–B45, 2015.

[51] M. Chiang and T. Zhang, "Fog and iot: An overview of research opportunities", *Ieee internet of things journal*, vol. 3, no. 6, pp. 854–864, 2016.

[52] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks", in *Global communications conference (GLOBECOM), 2015 ieee*, IEEE, 2015, pp. 1–6.

[53] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud ran", *Ieee transactions on wireless communications*, vol. 15, no. 9, pp. 6118–6131, 2016.

[54] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching", in *Smart antennas (WSA 2016); proceedings of the 20th international itg workshop on*, VDE, 2016, pp. 1–5.

[55] B. Azari, O. Simeone, U. Spagnolini, and A. M. Tulino, "Hypergraph-based analysis of clustered co-operative beamforming with application to edge caching", *Ieee wireless communications letters*, vol. 5, no. 1, pp. 84–87, 2016.

[56] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks", in *Information theory (ISIT), 2016 ieee international symposium on*, IEEE, 2016, pp. 2029–2033.

[57] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels", in *Information theory (ISIT), 2015 ieee international symposium on*, IEEE, 2015, pp. 809–813.

[58] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency", in *Information science and systems (CISS), 2016 annual conference on*, IEEE, 2016, pp. 320–325.

[59] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks", *Ieee transactions on information theory*, vol. 62, no. 2, pp. 849–869, 2016.

[60] S. Hosny, F. Alotaibi, H. E. Gamal, and A. Eryilmaz, "Towards a mobile content marketplace", in *2015 ieee 16th international workshop on signal processing advances in wireless communications (spawc)*, 2015, pp. 675–679.

[61] B. Bai, L. Wang, Z. Han, W. Chen, and T. Svensson, "Caching based socially-aware D2D communications in wireless content delivery networks: A hypergraph framework", *Ieee wireless communications*, vol. 23, no. 4, pp. 74–81, 2016.

[62] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks", in *Communications (ICC), 2016 ieee international conference on*, IEEE, 2016, pp. 1–6.

[63] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching", *Ieee transactions on information theory*, vol. 60, no. 7, pp. 4286–4298, 2014.

[64] N. Golrezaei *et al.*, "Base-station assisted device-to-device communications for high-throughput wireless video networks", *Ieee transactions on wireless communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[65] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience", *Ieee journal on selected areas in communications*, vol. 35, no. 5, pp. 1046–1061, 2017.

[66] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-d placement of an aerial base station in next generation cellular networks", in *2016 ieee international conference on communications (icc)*, IEEE, 2016, pp. 1–5.

[67] L. Wang, S. Chen, and M. Pedram, "Power management of cache-enabled cooperative base stations towards zero grid energy", in *Proc. of ieee international conference on communications (ICC)*, 2018, pp. 1–6.

[68] D. Bamburry, "Drones: Designed for product delivery", *Design management review*, vol. 26, no. 1, pp. 40–48, 2015.

[69] M. Mozaffari, A. Taleb Zadeh Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5G with UAVs: Foundations of a 3D wireless cellular network", *Ieee transactions on wireless communications*, vol. 18, no. 1, pp. 357–372, 2019.

[70] F. Song, J. Li, M. Ding, L. Shi, F. Shu, M. Tao, W. Chen, and H. V. Poor, "Probabilistic caching for small-cell networks with terrestrial and aerial users", *Ieee transactions on vehicular technology*, vol. 68, no. 9, pp. 9162–9177, 2019.

[71] S. Andreev *et al.*, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands", *Ieee journal on selected areas in communications*, vol. 33, no. 1, pp. 67–80, 2015.

[72] J. Chen and D. Gesbert, "Optimal positioning of flying relays for wireless networks: A los map approach", in *Proc. of ieee international conference on communications (ICC)*, Paris, France, 2017, pp. 1–6.

[73] C. Yang *et al.*, "Energy efficiency in wireless cooperative caching networks", in *Ieee international conference on communications (ICC)*, Sydney, Australia, 2014.

[74] S. E. Hajri and M. Assaad, "Energy efficiency in cache-enabled small cell networks with adaptive user clustering", *Ieee transactions on wireless communications*, vol. 17, no. 2, pp. 955–968, 2018.

[75] M. Haenggi, *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.

[76] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks", *Ieee transactions on wireless communications*, vol. 15, no. 7, pp. 4957–4972, 2016.

[77] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Fundamentals of cluster-centric content placement in cache-enabled device-to-device networks", *Ieee transactions on communications*, vol. 64, no. 6, pp. 2511–2526, 2016.

[78] D. Malak, M. Al-Shalash, and J. G. Andrews, "Spatially correlated content caching for device-to-device communications", *Ieee transactions on wireless communications*, vol. 17, no. 1, pp. 56–70, 2018.

[79] H. Chen and Y. Xiao, "Cache access and replacement for future wireless internet", *Ieee communications magazine*, vol. 44, no. 5, pp. 113–123, 2006.

[80] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks", *Ieee transactions on information theory*, vol. 62, no. 2, pp. 849–869, 2016.

[81] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks", *Ieee transactions on wireless communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[82] N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Wireless video content delivery through distributed caching and peer-to-peer gossiping", in *Asilomar conference on signals, systems and computers*, CA, USA, 2011.

[83] B. Chen, C. Y. Yang, and G. Wang, "High throughput opportunistic cooperative device-to-device communications with caching", *Ieee transactions on vehicular technology*, vol. PP, no. 99, pp. 1–1, 2017.

[84] S. Shalmashi and S. B. Slimane, "Cooperative device-to-device communications in the downlink of cellular networks", in *2014 ieee wireless communications and networking conference (WCNC)*, 2014, pp. 2265–2270.

[85] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks", *Ieee transactions on communications*, vol. 64, no. 6, pp. 2438–2452, 2016.

[86] G. Caire and A. F. Molisch, "Femtocaching and D2D communications: A new paradigm for video-aware wireless networks.", *Intel technology journal*, vol. 19, no. 1, 2015.

[87] S. W. Jeon, S. N. Hong, M. Ji, G. Caire, and A. F. Molisch, "Wireless multihop device-to-device caching networks", *Ieee transactions on information theory*, vol. 63, no. 3, pp. 1662–1676, 2017.

[88] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane", *Arxiv preprint arxiv:1309.0604*, 2013.

[89] K. Sundaresan, E. Chai, A. Chakraborty, and S. Rangarajan, "Skylite: End-to-end design of low-altitutde uav networks for providing lte connectivity", *Arxiv preprint arxiv:1802.06042*, 2018.

[90] M. K. Karray and M. Jovanovic, "A queueing theoretic approach to the dimensioning of wireless cellular networks serving variable-bit-rate calls", *Ieee transactions on vehicular technology*, vol. 62, no. 6, pp. 2713–2723, 2013.

[91] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications", in *Ieee conference on computer communications (INFOCOM)*, NY, USA, 1999.

[92] T. W. R. Collings, "A queueing problem in which customers have different service distributions", *Appl. statist.*, vol. 34, no. 1, pp. 75–82, 1974.

[93] C. Wang and M.-S. Chen, "On the complexity of distributed query optimization", *Ieee transactions on knowledge and data engineering*, vol. 8, no. 4, pp. 650–662, 1996.

[94] Y. Sun, Z. Chen, and H. Liu, "Delay analysis and optimization in cache-enabled multi-cell cooperative networks", in *Ieee global communications conference (GLOBECOM)*, Wash. DC, USA, 2016, pp. 1–7.

[95] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrak, "Maximizing a supermodular set function subject to a matroid constraint", in *12th international IPCO conference*, NY, USA, 2007.

[96] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks", in *Ieee conference on computer communications (INFOCOM)*, Hong Kong, 2015.

[97] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks", *Ieee transactions on wireless communications*, vol. 15, no. 1, pp. 131–145, 2016.

[98] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks", *Ieee transactions on information theory*, vol. 61, no. 12, pp. 6833–6859, 2015.

[99] M. Naldi, "Approximation of the truncated zeta distribution and zipf's law", *Arxiv preprint arxiv:1511.01480*, 2015.

[100] Y. Zhong, W. Zhang, and M. Haenggi, "Stability analysis of static poisson networks", in *Ieee international symposium on information theory (ISIT)*, 2015, pp. 2812–2816.

[101] K. Stamatiou and M. Haenggi, "Random-access poisson networks: Stability and delay", *Ieee communications letters*, vol. 14, no. 11, pp. 1035–1037, 2010.

[102] Y. Zhong, T. Q. Quek, and X. Ge, "Heterogeneous cellular networks with spatio-temporal traffic: Delay analysis and scheduling", *Ieee journal on selected areas in communications*, vol. 35, no. 6, pp. 1373–1386, 2017.

[103] M. Gharbieh, H. ElSawy, A. Bader, and M. S. Alouini, "Spatiotemporal stochastic modeling of iot enabled cellular networks: Scalability and stability analysis", *Ieee transactions on communications*, vol. 65, no. 8, pp. 3585–3600, 2017.

[104] H. Wu and H. Lu, "Energy and delay optimization for cache-enabled dense small cell networks", *Arxiv preprint arxiv:1803.03780*, 2018.

[105] W. Huang, W. Chen, and H. V. Poor, "Energy efficient pushing in AWGN channels based on content request delay information", *Ieee transactions on communications*, vol. 66, no. 8, pp. 3667–3682, 2018.

[106] W. Jiang, Y. Gong, Y. Cao, X. Wu, and Q. Xiao, "Energy-delay-cost tradeoff for task offloading in imbalanced edge cloud based computing", *Arxiv preprint arxiv:1805.02006*, 2018.

[107] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks", in *Ieee international conference on communications (ICC)*, London, UK, 2015, pp. 3358–3363.

[108] J. G. Andrews *et al.*, "A tractable approach to coverage and rate in cellular networks", *Ieee transactions on communications*, vol. 59, no. 11, pp. 3122–3134, 2011.

[109] Ericsson, "Radio waves and health", 2006. [Online]. Available: `http://www.ericsson.com/health`.

[110] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: Volume ii: General theory and structure*. Springer Science & Business Media, 2007.

[111] W. Szpankowski, "Stability conditions for some distributed systems: Buffered random access systems", *Advances in applied probability*, vol. 26, no. 2, pp. 498–515, 1994.

[112] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times", in *Mathematical proceedings of the cambridge philosophical society*, vol. 58, 1962, pp. 497–520.

[113] L. Kleinrock, *Queueing systems, vol. 1*. NY, USA: Wiley, 1975.

[114] B. Chen, C. Yang, and Z. Xiong, "Optimal caching and scheduling for cache-enabled D2D communications", *Ieee communications letters*, vol. 21, no. 5, pp. 1155–1158, 2017.

[115] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[116] N. Deng and M. Haenggi, "The benefits of hybrid caching in gauss-poisson D2D networks", *Ieee journal on selected areas in communications*, vol. 36, no. 6, pp. 1217–1230, 2018.

[117] S. H. Chae, T. Q. S. Quek, and W. Choi, "Content placement for wireless cooperative caching helpers: A tradeoff between cooperative gain and content diversity gain", *Ieee transactions on wireless communications*, vol. 16, no. 10, pp. 6795–6807, 2017.

[118] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks", *Ieee transactions on wireless communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[119] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Edge computing meets millimeter-wave enabled VR: Paving the way to cutting the cord", in *Ieee wireless communications and networking conference (wcnc)*, 2018, pp. 1–6.

[120] K. E. Hunter, S. A. Sprigg, N. J. Meijers, C. S. Wurster, and P. E. Jacobs, *Retail proximity marketing*, US Patent App. 13/833,110, 2013.

[121] H. Elsawy, H. Dahrouj, T. Y. Al-naffouri, and M. Alouini, "Virtualized cognitive network architecture for 5G cellular networks", *Ieee communications magazine*, vol. 53, no. 7, pp. 78–85, 2015.

[122] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks", *Ieee transactions on wireless communications*, vol. 16, no. 5, pp. 3401–3415, 2017.

[123] S. H. Chae, T. Q. S. Quek, and W. Choi, "Content placement for wireless cooperative caching helpers: A tradeoff between cooperative gain and content diversity gain", *Ieee transactions on wireless communications*, vol. 16, no. 10, pp. 6795–6807, 2017.

[124] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery", in *Proceedings of the 16th acm international symposium on mobile ad hoc networking and computing*, ACM, 2015, pp. 127–136.

[125] G. Zheng, H. A. Suraweera, and I. Krikidis, "Optimization of hybrid cache placement for collaborative relaying", *Ieee communications letters*, vol. 21, no. 2, pp. 442–445, 2017.

[126] A. Daghal, H. Zhu, and J. Wang, "Content delivery analysis in multiple devices to single device communications", *Ieee transactions on vehicular technology*, 2018.

[127] B. Chen, C. Yang, and G. Wang, "High throughput opportunistic cooperative device-to-device communications with caching", *Ieee transactions on vehicular technology*, vol. 66, no. 8, pp. 7527–7539, 2017.

[128] G. Kim, B. Hong, W. Choi, and H. Park, "Mds coded caching leveraged by coordinated multi-point transmission", *Ieee communications letters*, 2018.

[129] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, "Social network aware device-to-device communication in wireless networks", *Ieee transactions on wireless communications*, vol. 14, no. 1, pp. 177–190, 2015.

[130] X. Hu, L. Meng, and A. D. Striegel, "Evaluating the raw potential for device-to-device caching via co-location", *Procedia computer science*, vol. 34, pp. 376–383, 2014.

[131] M. Lee, A. F. Molisch, N. Sastry, and A. Raman, "Individual preference probability modeling and parameterization for video content in wireless caching networks", *Ieee/acm transactions on networking*, vol. 27, no. 2, pp. 676–690, 2019.

[132] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks", in *Proc. of ieee international conference on communications (ICC)*, London, UK, 2015.

[133] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal", *Ieee communications letters*, vol. 21, no. 3, pp. 584–587, 2017.

[134] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching.", *Ieee trans. wireless communications*, vol. 15, no. 10, pp. 6626–6637, 2016.

[135] 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, no. 11, pp. 7500–7514, 2018.

[136] M. Naslcheraghi, M. Afshang, and H. S. Dhillon, "Modeling and performance analysis of full-duplex communications in cache-enabled D2D networks", in *2018 ieee international conference on communications (ICC)*, 2018, pp. 1–6.

[137] P. Parida, H. S. Dhillon, and A. F. Molisch, "Downlink performance analysis of cell-free massive MIMO with finite fronthaul capacity", in *Ieee 88th vehicular technology conference (vtc-fall)*, 2018, pp. 1–6.

[138] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs", *Ieee transactions on wireless communications*, vol. 15, no. 6, pp. 3949–3963, 2016.

[139] M. M. Azari, F. Rosas, A. Chiumento, and S. Pollin, "Coexistence of terrestrial and aerial users in cellular networks", in *Proc. of ieee globecom workshops (gc wkshps)*, Singapore, 2017, pp. 1–6.

[140] X. Lin, V. Yajnanarayana, S. D. Muruganathan, S. Gao, H. Asplund, H.-L. Maattanen, M. Bergstrom, S. Euler, and Y.-P. E. Wang, "The sky is not the limit: LTE for unmanned aerial vehicles", *Ieee communications magazine*, vol. 56, no. 4, pp. 204–210, 2018.

[141] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective", *Ieee transactions on communications*, vol. 67, no. 3, pp. 2580–2604, 2019.

[142] G. Geraci, A. Garcia-Rodriguez, L. G. Giordano, D. López-Pérez, and E. Björnson, "Understanding UAV cellular communications: From existing networks to massive MIMO", *Ieee access*, vol. 6, pp. 67 853–67 865, 2018.

[143] B. Galkin, J. Kibilda, and L. Da Silva, "A stochastic model for UAV networks positioned above demand hotspots in urban environments", *Ieee transactions on vehicular technology*, pp. 1–1, 2019.

[144] V. V. Chetlur and H. S. Dhillon, "Downlink coverage analysis for a finite 3-D wireless network of unmanned aerial vehicles", *Ieee transactions on communications*, vol. 65, no. 10, pp. 4543–4558, 2017.

[145] R. Bhagavatula, R. W. Heath Jr, and K. Linehan, "Performance evaluation of MIMO base station antenna designs", *Antenna systems and technology magazine*, vol. 11, no. 6, pp. 14–17, 2008.

[146] K. Hosseini, W. Yu, and R. S. Adve, "Large-scale MIMO versus network MIMO for multicell interference mitigation", *Ieee journal of selected topics in signal processing*, vol. 8, no. 5, pp. 930–941, 2014.

[147] X. Yu, C. Li, J. Zhang, M. Haenggi, and K. B. Letaief, "A unified framework for the tractable analysis of multi-antenna wireless networks", *Ieee transactions on wireless communications*, vol. 17, no. 12, pp. 7965–7980, 2018.

[148] M. M. Azari, F. Rosas, and S. Pollin, "Cellular connectivity for UAVs: Network modeling, performance analysis and design guidelines", *Ieee transactions on wireless communications*, pp. 1–1, 2019.

[149] X. Xu and Y. Zeng, "Cellular-connected UAV: Performance analysis with 3D antenna modelling", in *Ieee international conference on communications workshops (icc workshops)*, 2019, pp. 1–6.

[150] S. Euler, H. Maattanen, X. Lin, Z. Zou, M. Bergström, and J. Sedin, "Mobility support for cellular connected unmanned aerial vehicles: Performance and analysis", in *Ieee wireless communications and networking conference (wcnc)*, 2019, pp. 1–6.

[151] A. Fakhreddine, C. Bettstetter, S. Hayat, R. Muzaffar, and D. Emini, "Handover challenges for cellular-connected drones", in *Proc. of acm workshop on micro aerial vehicle networks, systems, and applications*, Seoul, Republic of Korea, 2019, pp. 9–14.

[152] G. T. 36.777, "Technical specification group radio access network; study on enhanced LTE support for aerial vehicles", *Tech. rep., 5g americas*, Dec. 2017.

[153] T.-T. Vu, L. Decreusefond, and P. Martins, "An analytical model for evaluating outage and handover probability of cellular wireless networks", *Wireless personal communications*, vol. 74, no. 4, pp. 1117–1127, 2014.

[154] M. Banagar and H. S. Dhillon, "Performance characterization of canonical mobility models in drone cellular networks", *Ieee transactions on wireless communications, to appear*, 2020.

[155] P. K. Sharma and D. I. Kim, "Random 3D mobile UAV networks: Mobility modeling and coverage probability", *Ieee transactions on wireless communications*, vol. 18, no. 5, pp. 2527–2538, 2019.

[156] M. Vondra, M. Ozger, D. Schupke, and C. Cavdar, "Integration of satellite and aerial communications for heterogeneous flying vehicles", *Ieee network*, vol. 32, no. 5, pp. 62–69, 2018.

[157] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems", *Ieee network, to appear*, 2019.

[158] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on uav communications for 5g and beyond", *Arxiv preprint arxiv:1903.05289*, 2019.

[159] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies", *Ieee wireless communications*, vol. 26, no. 1, pp. 120–127, 2019.

[160] C. D'Andrea, A. Garcia-Rodriguez, G. Geraci, L. G. Giordano, and S. Buzzi, "Cell-free massive MIMO for UAV communications", *Arxiv preprint arxiv:1902.03578*, 2019.

[161] A. Rahmati, Y. Yapıcı, N. Rupasinghe, I. Guvenc, H. Dai, and A. Bhuyany, "Energy efficiency of RSMA and NOMA in cellular-connected mmwave UAV networks", *Arxiv preprint arxiv:1902.04721*, 2019.

[162] N. Cherif, M. Alzenad, H. Yanikomeroglu, and A. Yongacoglu, "Downlink coverage and rate analysis of an aerial user in integrated aerial and terrestrial networks", *Arxiv preprint arxiv:1905.11934*, 2019.

[163] L. Liu, S. Zhang, and R. Zhang, "Multi-beam UAV communication in cellular uplink: Cooperative interference cancellation and sum-rate maximization", *Ieee transactions on wireless communications*, vol. 18, no. 10, pp. 4679–4691, 2019.

[164] W. Mei, Q. Wu, and R. Zhang, "Cellular-connected UAV: Uplink association, power control and interference coordination", *Ieee transactions on wireless communications*, vol. 18, no. 11, pp. 5380–5393, 2019.

[165] L. Qualcomm, *Unmanned aircraft systems' trial report*, 2017.

[166] B. Van der Bergh, A. Chiumento, and S. Pollin, "LTE in the sky: Trading off propagation benefits with interference costs for aerial nodes", *Ieee communications magazine*, vol. 54, no. 5, pp. 44–50, 2016.

[167] L. Liu, S. Zhang, and R. Zhang, "CoMP in the sky: UAV placement and movement optimization for multi-user communications", *Ieee transactions on communications*, pp. 1–1, 2019.

[168] S. Zhang and R. Zhang, "Trajectory optimization for cellular-connected UAV under outage duration constraint", *Arxiv preprint arxiv:1901.04286*, 2019.

[169] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach", *Ieee transactions on wireless communications*, vol. 18, no. 4, pp. 2125–2140, 2019.

[170] P. K. Sharma and D. I. Kim, "Coverage probability of 3-D mobile UAV networks", *Ieee wireless communications letters*, vol. 8, no. 1, pp. 97–100, 2019.

[171] S. Enayati, H. Saeedi, H. Pishro-Nik, and H. Yanikomeroglu, "Moving aerial base station networks: A stochastic geometry analysis and design perspective", *Ieee transactions on wireless communications*, vol. 18, no. 6, pp. 2977–2988, 2019.

[172] C. Zhu and W. Yu, "Stochastic modeling and analysis of user-centric network MIMO systems", *Ieee transactions on communications*, vol. 66, no. 12, pp. 6176–6189, 2018.

[173] M. Ding, P. Wang, D. López-Pérez, G. Mao, and Z. Lin, "Performance impact of LoS and NLoS transmissions in dense cellular networks", *Ieee transactions on wireless communications*, vol. 15, no. 3, pp. 2365–2380, 2016.

[174] M. M. Azari, F. Rosas, and S. Pollin, "Reshaping cellular networks for the sky: Major factors and feasibility", in *Proc. of ieee international conference on communications (ICC)*, Kansas City, USA, 2018, pp. 1–7.

[175] ITU-R, "Recommendation p.1410-5: Propagation data and prediction methods required for the design of terrestrial broadband radio access systems operating in a frequency range from 3 to 60 GHz", 2012.

[176] R. W. Heath Jr, T. Wu, Y. H. Kwon, and A. C. Soong, "Multiuser MIMO in distributed antenna systems with out-of-cell interference", *Ieee transactions on signal processing*, vol. 59, no. 10, pp. 4885–4899, 2011.

[177] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks", *Ieee transactions on mobile computing*, vol. 2, no. 3, pp. 257–269, 2003.

[178] C. Bettstetter, H. Hartenstein, and X. Pérez-Costa, "Stochastic properties of the random waypoint mobility model", *Wireless networks*, vol. 10, no. 5, pp. 555–567, 2004.

[179] E. Hyytia, P. Lassila, and J. Virtamo, "Spatial node distribution of the random waypoint mobility model with applications", *Ieee transactions on mobile computing*, vol. 5, no. 6, pp. 680–694, 2006.

[180] X. Lin, R. K. Ganti, P. J. Fleming, and J. G. Andrews, "Towards understanding the fundamentals of mobility in cellular networks", *Ieee transactions on wireless communications*, vol. 12, no. 4, pp. 1686–1698, 2013.

[181] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad hoc mobile networks", *Ieee transactions on information theory*, vol. 51, no. 6, pp. 1917–1937, 2005.

[182] P. K. Sharma and D. I. Kim, "Coverage probability of 3D UAV networks with RWP mobility-based altitude control", in *Ieee international conference on communications workshops (icc workshops)*, 2018, pp. 1–6.

[183] X. Xu, Z. Sun, X. Dai, T. Svensson, and X. Tao, "Modeling and analyzing the cross-tier handover in heterogeneous networks", *Ieee transactions on wireless communications*, vol. 16, no. 12, pp. 7859–7869, 2017.

[184] H. Tabassum, M. Salehi, and E. Hossain, "Mobility-aware analysis of 5G and B5G cellular networks: A tutorial", *Arxiv preprint arxiv:1805.02719*, 2018.

[185] S. Sadr and R. S. Adve, "Handoff rate and coverage analysis in multi-tier heterogeneous networks", *Ieee transactions on wireless communications*, vol. 14, no. 5, pp. 2626–2638, 2015.

[186] R. Arshad, H. ElSawy, S. Sorour, T. Y. Al-Naffouri, and M. Alouini, "Velocity-aware handover management in two-tier cellular networks", *Ieee transactions on wireless communications*, vol. 16, no. 3, pp. 1851–1867, 2017.

[187] S. Andreev, O. Galinina, A. Pyattaev, J. Hosek, P. Masek, H. Yanikomeroglu, and Y. Koucheryavy, "Exploring synergy between communications, caching, and computing in 5G-grade deployments", *Ieee communications magazine*, vol. 54, no. 8, pp. 60–69, 2016.

[188] X. Wang, T. Kwon, Y. Choi, H. Wang, and J. Liu, "Cloud-assisted adaptive video streaming and social-aware video prefetching for mobile users", *Ieee wireless communications*, vol. 20, no. 3, pp. 72–79, 2013.

[189] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading", *Arxiv preprint arxiv:1702.05309*, 2017.

[190] C. Park and J. Lee, "Mobile edge computing-enabled heterogeneous networks", *Arxiv preprint arxiv:1804.07756*, 2018.

[191] A. Ndikumana, N. H. Tran, T. M. Ho, Z. Han, W. Saad, D. Niyato, and C. S. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing", *Arxiv preprint arxiv:1803.11512*, 2018.

[192] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili, "Collaborative multi-bitrate video caching and processing in mobile-edge computing networks", in *Wireless on-demand network systems and services (WONS), 2017 13th annual conference on*, IEEE, 2017, pp. 165–172.

[193] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. Leung, "Virtual resource allocation for heterogeneous services in full duplex-enabled scns with mobile edge computing and caching", *Ieee transactions on vehicular technology*, vol. 67, no. 2, pp. 1794–1808, 2018.

[194] Y. Hao, M. Chen, L. Hu, M. S. Hossain, and A. Ghoneim, "Energy efficient task caching and offloading for mobile edge computing", *Ieee access*, vol. 6, pp. 11 365–11 373, 2018.

[195] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing", *Ieee transactions on wireless communications*, vol. 16, no. 8, pp. 4924–4938, 2017.

[196] M. S. Elbamby, M. Bennis, and W. Saad, "Proactive edge computing in latency-constrained fog networks", in *Networks and communications (EuCNC), 2017 european conference on*, IEEE, 2017, pp. 1–6.

[197] Z. Luo, M. LiWang, Z. Lin, L. Huang, X. Du, and M. Guizani, "Energy-efficient caching for mobile edge computing in 5G networks", *Applied sciences*, vol. 7, no. 6, p. 557, 2017.

[198] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing", *Ieee wireless communications*, vol. 25, no. 3, 2018.

[199] R. Wang, J. Yan, D. Wu, H. Wang, and Q. Yang, "Knowledge-centric edge computing based on virtualized D2D communication systems", *Ieee communications magazine*, vol. 56, no. 5, pp. 32–38, 2018.

[200] M. Chen, Y. Qian, Y. Hao, Y. Li, and J. Song, "Data-driven computing and caching in 5G networks: Architecture and delay analysis", *Ieee wireless communications*, vol. 25, no. 1, pp. 70–75, 2018.

[201] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications", *Ieee access*, vol. 5, pp. 6757–6779, 2017.

[202] C. Wu, T. Yoshinaga, Y. Ji, T. Murase, and Y. Zhang, "A reinforcement learning-based data storage scheme for vehicular ad hoc networks", *Ieee transactions on vehicular technology*, vol. 66, no. 7, pp. 6336–6348, 2017.

[203] J. Lyu, Y. Zeng, and R. Zhang, "Cyclical multiple access in uav-aided communications: A throughput-delay tradeoff", *Ieee wireless communications letters*, vol. 5, no. 6, pp. 600–603, 2016.

[204] R. Amer, A. A. El-Sherif, H. Ebrahim, and A. Mokhtar, "Stability analysis for multi-user cooperative cognitive radio network with energy harvesting", in *Ieee international conference on computer and communications (iccc)*, 2016, pp. 2369–2375.

[205] R. Amer, A. A. El-sherif, H. Ebrahim, and A. Mokhtar, "Cooperation and underlay mode selection in cognitive radio network", in *International conference on future generation communication technologies (fgct)*, 2016, pp. 36–41.

[206] R. Amer, A. A. El-Sherif, H. Ebrahim, and A. Mokhtar, "Cooperative cognitive radio network with energy harvesting: Stability analysis", in *International conference on computing, networking and communications (icnc)*, 2016, pp. 1–7.

[207] T. A. Odetola, H. R. Mohammed, and S. R. Hasan, "A stealthy hardware trojan exploiting the architectural vulnerability of deep learning architectures: Input interception attack (iia)", *Arxiv preprint arxiv:1911.00783*, 2019.

[208] T. A. Odetola, O. Oderhohwo, and S. R. Hasan, "A scalable multilabel classification to deploy deep learning architectures for edge devices", *Arxiv preprint arxiv:1911.02098*, 2019.

[209] H. Mohammed, T. A. Odetola, S. R. Hasan, S. Stissi, I. Garlin, and F. Awwad, "(hiadiot): Hardware intrinsic attack detection in internet of things; leveraging power profiling", in *2019 ieee 62nd international midwest symposium on circuits and systems (mwscas)*, IEEE, 2019, pp. 852–855.

[210] O. Oderhohwo, H. Mohammed, T. Odetola, T. N. Guo, S. Hasan, and F. Dogbe, "An edge intelligence framework for resource constrained community area network", in *2020 ieee 63rd international midwest symposium on circuits and systems (mwscas)*, IEEE, 2020, pp. 97–100.

[211] O. Oderhohwo, T. A. Odetola, H. Mohammed, and S. R. Hasan, "Deployment of object detection enhanced with multi-label multi-classification on

edge device", in *2020 ieee 63rd international midwest symposium on circuits and systems (mwscas)*, IEEE, 2020, pp. 986–989.

[212] H. Mohammed, T. A. Odetola, and S. R. Hasan, "How secure is distributed convolutional neural network on iot edge devices?", *Arxiv preprint arxiv:2006.09276*, 2020.

[213] M. Baza, N. Lasla, M. Mahmoud, G. Srivastava, and M. Abdallah, "B-ride: Ride sharing with privacy-preservation, trust and fair payment atop public blockchain", *Ieee transactions on network science and engineering*, 2019.

[214] M. Baza, A. Sherif, M. Mahmoud, S. Bakiras, X. Lin, and M. Abdallah, "Privacy preserving blockchain-based energy trading schemes for electric vehicles", *Ieee transactions of vehicular technology*, 2021.

[215] M. Baza, J. Baxter, N. Lasla, M. Mahmoud, M. Abdallah, and M. Younis, "Incentivized and secure blockchain-based firmware update and dissemination for autonomous vehicles", in *Connected and autonomous vehicles in smart cities*, CRC press, 2020.

[216] M. Baza, M. Mahmoud, G. Srivastava, W. Alasmary, and M. Younis, "A light blockchain-powered privacy-preserving organization scheme for ride sharing services", *Proc. of the ieee 91th vehicular technology conference (vtc-spring), antwerp, belgium*, May 2020.

[217] M. Baza, "Blockchain-based secure and privacy-preserving schemes for connected vehicles", PhD thesis, Tennessee Technological University, 2020.

[218] W. Al Amiri, M. Baza, K. Banawan, M. Mahmoud, W. Alasmary, and K. Akkaya, "Towards secure smart parking system using blockchain technology", in *2020 ieee 17th annual consumer communications and networking conference (ccnc)*, 2020, pp. 1–2.

[219] M. Baza, R. Amer, M. Mahmoud, G. Srivastava, and W. Alasmary, "A privacy-preserving energy trading scheme for electric vehicles", in *2021 ieee 18th annual consumer communications and networking conference (ccnc)*, 2021, pp. 1–6.

[220] M. Baza, A. Salazar, M. Mahmoud, M. Abdallah, and K. Akkaya, "On sharing models instead of data using mimic learning for smart health applications", in *2020 ieee international conference on informatics, iot, and enabling technologies (iciot)*, 2020, pp. 231–236.

[221] M. Baza, R. Amer, G. Srivastava, M. Mahmoud, W. Alasmary, and M. Younis, "Efficient privacy-preserving charging coordination with linkabilty-resistance in the future smart grid", in *Ieee international conference on communications workshops (icc workshops)*, 2021, pp. 1–6.

[222] W. Al Amiri, M. Baza, M. Mahmoud, b. K. Banawan, W. Alasmary, and K. Akkaya, "Privacy-preserving smart parking system using blockchain and private information retrieval", *Proc. of the ieee international conference on smart applications, communications and networking (smartnets 2019)*, 2020.

[223] R. Amer, M. Baza, M. M. Butt, and N. Marchetti, "Optimizing joint probabilistic caching and channel access for clustered d2d networks", *Arxiv preprint arxiv:2003.02676*, 2020.

[224] M. Baza, R. Amer, G. Srivastava, M. Mahmoud, W. Alasmary, and M. Younis, "Efficient privacy-preserving charging coordination with linkability-resistance in the future smart grid", *Proc. of ieee international conference on communications workshops (icc workshops), montreal, canada*, 2021.

[225] F. Amsaad, A. Razaque, M. Baza, and G. Srivastava, "An efficient and reliable lightweight puf for iot-based applications", *Proc. of ieee international conference on communications workshops (icc workshops), montreal, canada*, 2021.

[226] A. Sherif, M. Baza, M. Mahmoud, and G. Srivastava, "Efficient search over encrypted medical cloud data with known-plaintext/background and unlinkability security", *Ieee transactions on industrial informatics*, 2021.

[227] A. Sherif, M. Baza, and M. Mahmoud, "Cses: Customized searchable encryption scheme with efficient key management over medical cloud data", *2021 international symposium on networks, computers and communications (isncc)*, 2021.

[228] M. Afshang, C. Saha, and H. S. Dhillon, "Nearest-neighbor and contact distance distributions for thomas cluster process", *Ieee wireless communications letters*, vol. 6, no. 1, pp. 130–133, 2017.

[229] S. Rajan, S. Wang, R. Inkol, and A. Joyal, "Efficient approximations for the arctangent function", *Ieee signal processing magazine*, vol. 23, no. 3, pp. 108–111, 2006.

[230] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks", *Ieee transactions on communications*, vol. 59, no. 11, pp. 3122–3134, 2011.

[231] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.