An Energy Efficient D2D Communication Model with Guaranteed QoS for Cloud Radio Access Networks

Isuru Janith Ranawaka

(169229L)

Degree of Master of Science

Department of Electronic and Telecommunication

University of Moratuwa
Moratuwa, Sri Lanka

Abstract

This work proposes a spectrum selection scheme and a transmit power minimization scheme for a device-to-device (D2D) network cross-laid with a cloud radio access network (CRAN). The D2D communications are allowed as an overlay to the CRAN as well as in the unlicensed industrial, scientific and medical radio (ISM) band. A link distance based spectrum selection scheme is proposed and closed-form approximations are derived for the link distance thresholds to select the operating band of the D2D users. Furthermore, analytical expressions are derived to calculate the minimum required transmit power to achieve a guaranteed level of quality of service in each operating band. The results demonstrate that the proposed scheme achieves nearly 50% power saving compared to a monolithic (purely overlay or purely ISM band) D2D network. Moreover, this work creates an immense space for communication technologies to be wisely managed and utilized by application layer requirements through CRAN architecture. Caching strategies for content replication across end user devices and effective content delivery strategies can be implemented for forthcoming video streaming applications.

Keywords:
cloud radio access networks, device-to-device communication, , overlay communication, proactive caching, underlay communication.

Declaration

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Asuru Ranawaka — Signature

2020-11-27 — Date

The above candidate has carried out research for the Masters thesis under my supervision.

Signature of the supervisor

Date

Acknowledgement

This work was carried out during the years 2017-2020 at the Department of Electronic and Telecommunication Engineering, University of Moratuwa.

I owe my deepest gratitude to my supervisors Dr. Tharaka Samarasinghe and Dr. Kasun Hemachandra for their incredible guidance to successfully complete the research. A conference paper have been published based on the research outcomes and I would like to thank all of the coauthors for their immense support. I would like to thank the Senate Research Council, University of Moratuwa for supporting this work under grant SRC/LT/2018/2.

I would like to thank all the authors and owners of the materials that have been referred in this thesis and finally, I would like to thank all the people who supported to successfully complete the research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The telecommunication industry has experienced an exponential growth in the number of mobile subscribers [10,30], and a rapid increase in data usage, over the last decade. Although, there is a large growth in the number of subscribers, the net revenue generated by the service providers has remained almost constant over time. However, to maintain quality and uninterrupted services, the network operators have to increase their network capacity proportionately to the data demand. The network capacity primarily depends on the channel capacity and the number of channels. The channel capacity cannot be improved beyond the Shannon limit. Hence, to improve the network capacity, operators need to investigate novel network deployment methods such as small cells, HetNets [10], and massive MIMO. However, these legacy techniques have led to complex interference management issues that result in increase in the total cost of ownership (TCO). The TCO comprises of capital expenditure (CAPEX) and operating expenditure (OPEX). The CAPEX is a one-time investment, that includes costs incurred from land acquisition to initiation of the operations. The OPEX is a recurring cost that includes costs incurred from site maintenance to labor costs. The main challenge for network architects is to support high network capacity while keeping a low TCO, and this motivates the modification of existing network architectures to build low cost, highly scalable networks.

The network designers have considered cloud based network architectures as a feasible solution to support high data demand and a lower TCO. To this end, cloud radio access networks (CRANs) have been introduced [10], and the CRANs operate on top of generic cloud infrastructures. CRANs consist of remote radio heads (RRHs), baseband units (BBUs), and hardware pools such as general purpose processors (GPUs) [10]. RRHs send and receive the modulated signals to and from the cloud processing units deployed on the cloud infrastructure. The BBUs are deployed on the cloud, and are responsible for signal processing, data extraction, and data transfers. The introduction of the cloud architectures into telecommunications has created a unique space for new technological innovations and energy efficient,

low cost networks.

Although CRANs help to reduce the TCO of mobile networks, network capacity is not adequate to satisfy proliferating data demand. Hence, caching schemes are proposed at network edges to reduce the backhaul traffic, and real implementations of cache-enabled edge CRANs (E-CRANs) are given in [30, 43]. This technology has reduced backhaul traffic significantly but still network edges need to handle a large volume of traffic. Hence, traffic offloading techniques [1, 12, 21] are proposed to reduce the fronthaul traffic. However, these approaches require separate access points (APs) for operation, which results in additional costs for the network operators.

To this end, D2D networks are proposed where end devices directly communicate using different technologies. Literally, it moves some heavy computational functions such as calculating user destination, selecting transmit power, deciding communication bandwidth into mobile devices. Due to limitations of power, processing capacity, and memory of mobile devices fully autonomous D2D networks are not pragmatic. Hence, controlled networks are proposed where BSs do heavy calculations and command mobile devices to transmit data. However, due to the lack of global view of the network, high processing capacity at traditional BSs lot of communication pitfalls and significant degradation of QoS occur in D2D networks.

This has motivated research on power-efficient and QoS guaranteed D2D communication protocols and user association schemes. We have identified the key benefits of CRAN architecture to support D2D communication. CRAN provides the global overview of mobile network that enables accurate calculations of network parameters. Moreover, it provides high processing power, memory capacity, and high availability, Hence, we have used CRAN to develop our system model and algorithms.

This research proposes a spectrum selection scheme for a D2D network cross laid with an E-CRAN. The proposed scheme provides guaranteed QoS while minimizing power consumption. In contrast to a monolithic D2D network, a hybrid D2D network is proposed, where D2D communications are allowed in the unlicensed ISM band as well as an overlay to the E-CRAN. The contributions of this paper can be summarized as follows.

- Link distance based spectrum selection scheme is proposed for paired D2D users.
- Link length thresholds for user allocation in each band are analytically obtained.
- The minimum transmit power required to provide a guaranteed QoS level in each band are analytically derived.

The remainder of this thesis is organized as follows. Chapter 2 presents the literature survey and related work. It articulates the evolution of network architectures from 1G

to CRANs, communication technologies of cellular and D2D networks, and content caching strategies at network edges and end-user devices. Furthermore, network performance metrics for cellular and D2D networks that have been used as analytical tools in the research are discussed. Chapter 3 presents the problem statement and chapter 4 presents analytical results that derived in the research. Furthermore, Chapter 5 presents numerical results obtained through extensive simulations to validate the analytical results. Chapter 6 presents applications of the derived models and pending work. Chapter 7 concludes the thesis.

Chapter 2

Literature Survey

2.1 Network Architecture Evolution

Mobile networks are continuously evolving thanks to enhancing technological and business requirements. Primarily, the key elements of a mobile network are base stations (BSs) and mobile core networks [22]. The coverage area of a BS is called as a cell and a cell-based network is named as a cellular network. A cellular network consists of both baseband functions and radio functions. The key functions of baseband processing are interference handling, modulation, demodulation, and fast Fourier transforms (FFTs). In addition to that, digital-to-analog conversion, analog-to-digital conversion, and power amplification [30] are grouped as radio functions. The placement of these components differ depending on the network architecture.

2.1.1 Traditional Networks

First generation (1)G and second generation (2)G mobile networks are identified as traditional networks [10]. The proximity of the antenna module and the radio module is a key characteristic of the traditional networks, with Coaxial cables being the key communication medium used for the communication between antenna modules and radio modules. X2 interface and S1 interface are primary interfaces of the traditional networks as shown in Figure 2.1. X2 interface is used for the communication between BSs, and the S1 interface connects the BSs and the mobile core network. Moreover, the BBUs and the antennas are co-located, hence, expanding of a cell requires the deployment of BBUs and antennas for each cell. This leads to high TCO, high power consumption, and under utilization of resources.

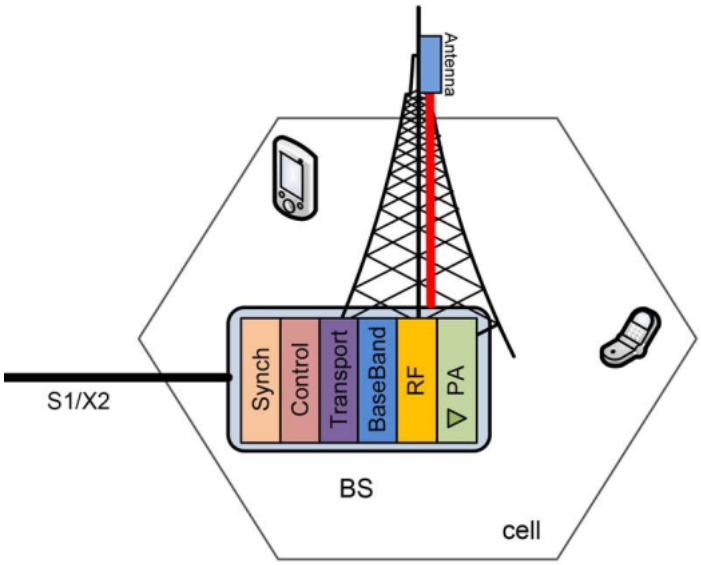

Figure 2.1: 1G 2G Base Station Architecture

Separation of the BBU and the RRH

The main components of the BS are divided into the BBUs and the remote radio heads (RRHs) as shown in Figure 2.2 in 3G networks. The BBU is responsible for signal processing related activities, while the RRH is responsible for radio functions. The distance between the RRH and its connected BBU can be extended up to 40 km and might be varied based on the communication media. Mainly, network designers use optical fibers and microwave links to establish the connection between the BBU and RRHs [10]. The BBU and the RRH placement greatly reduces the maintenance and cooling costs of a network. The primary bottleneck for the placement arises from processing and propagation delays. Most geographically separated RRHs can be statically connected to a single BBU, which reduces the CAPEX. THe RRHs are connected by Daisy chained architecture and the Ir interface is used to connect with the RRHs and the BBUs. Common public radio interface (CPRI) is used for IQ data transmission between RRHs and BBUs on the Ir interface [10].

2.1.2 CRAN Networks

Due to the continuous enhancement of network architectures, BBUs are grouped into BBU pools [16]. The BBU pools are implemented using general-purpose processors on top of the cloud infrastructure and deployed on different computing clusters. This architecture was firstly introduced by IBM [17]. A BBU pool is a cluster which composes of general-purpose processors to perform baseband processing. X2 interface is in a new form and often referred to as X2+, which is organized for inter-cluster communication. There might be dedicated

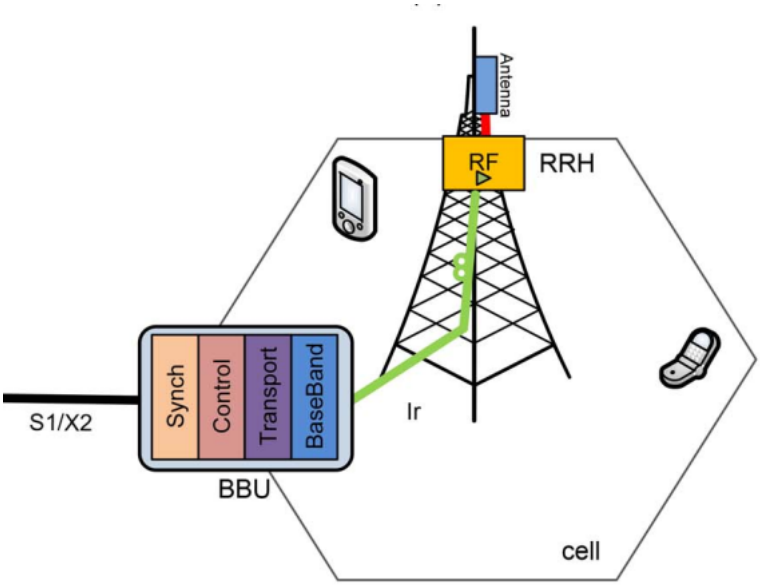

Figure 2.2: 3G Base Station Architecture

hardware pools for latency intensive processing such as FFT calculations to reduce the computational latency compared to general-purpose computations.

The fronthaul network spans from the RRHs up to the BBU Pools. The backhaul does connect the BBU pool to the mobile core network. RRHs are co-located with the antennas and they are connected with the high-performance processors of the BBU Pool via low latency, high bandwidth optical transport links. According to Figure 2.3, there are densely deployed RRHs with a shorter coverage area. Furthermore, interference management is handled at the BBUs, that enables the utilization of cloud computing capabilities for running compute-intensive and latency intensive algorithms. Moreover, it eases the implementation of interference management techniques such as inter-cell interference coordination (ICIC), and coordination multi point (CoMP) [10].

L1 functionality of the physical layer is implemented at the BBU, but it does generate the high bandwidth IQ data transmission in between the BBU and the RRH. Dark fiber, WDM/OTN, unified fixed and mobile access (for indoor coverage deployment) and carrier Ethernet are the main transport networks used in the CRANs. Furthermore, network equipment such as CPRI2EthernetGateway, IQ data routing switch, CPRI Mux, and X2 OTN gateway are used based on the transport medium. The high bandwidth links between RRHs and BBUs can be further fine tuned by using techniques such as reducing sampling rate, non-linear quantization, IQ data compression, and sub carrier compression [10].

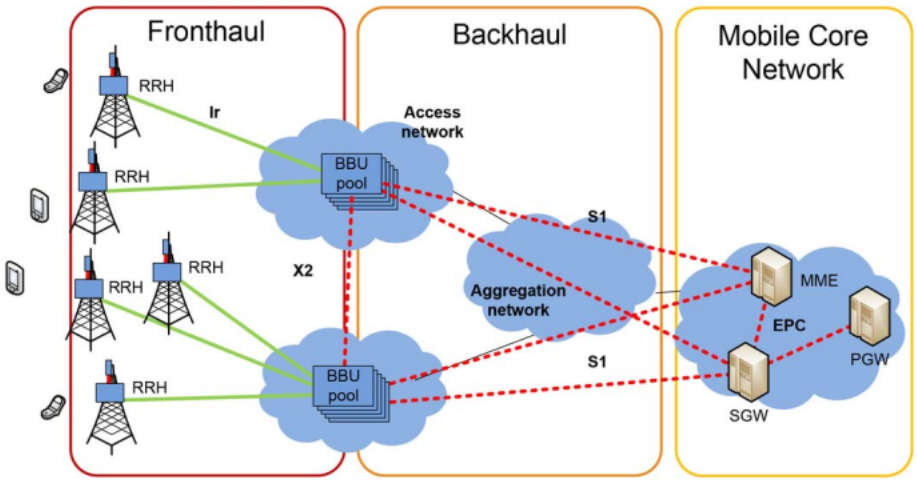

Figure 2.3: CRAN Architecture

Advantages of CRANs

Primarily, the TCO of network capacity expansion is reduced with the introduction of CRANs, as compared to the legacy networks. Since BSs only consist of RRHs and high bandwidth links, the rate of failure and the maintenance cost is lower as well. Most of the BBU bound functions are executed on top of the general-purpose processors, hence, the capital expenditure of establishing new hardware is reduced. As the current data demand is increasing exponentially, high computational power is required for data processing as well. To this end, CRANs provide huge space for data processing and data analysis at a lower cost, thanks to cloud processing. This makes future networks more intelligent and autonomous.

According to [10], RAN accounts for 80% of the power consumption of a mobile network. The consumption consists of power used for power amplification, RRHs, BBU pools, and air conditioning. Out of this, 41% is used for electrification of the equipment and 46% is used for cooling. Hence, with the introduction of CRANs, RRHs can be cooled using natural cooling resources that leads to a power-saving of 67% to 80% compared to traditional RANs.

According to the data usage patterns, BBUs can cater to irregular traffic patterns easily. Thus, on-demand processing capabilities can be easily implemented on top of CRANs. For an example, during office hours, most of the people travel to urban areas, that increases the data usage around the city. Similarly, during the off-peak times, the data traffic of the city is comparatively lower compared to suburban areas. This emphasizes that we can optimize the processing power based on the traffic patterns. Cloud processing capabilities and dynamic booting of BBUs facilitate this process [29, 30].

Furthermore, users have heterogeneous data requirements, thus mobile operators provide service level agreements for the users at signup. Implementation of service level agreements is infeasible for traditional networks because of the requirement of data analysis services.

However, it has become a reality with the introduction of CRANS, with the help of general-purpose processors. Upgrading and repairing network elements are comparatively easier with CRANs as well. Moreover, high concurrent data processing capabilities have helped for high throughput and low latency communication.

Long term evolution-advanced (LTE-A) technology, that 4G introduced, does use orthogonal frequency division multiple access (OFDMA). It uses dense orthogonal careers to carry data and eNB is used as a scheduler for resource allocation [7]. Moreover, LTE-A does use inter-cell interference coordination (ICIC), and coordination multi-point (CoMP) techniques to control inter-cell interference. These technologies are facilitated by the CRAN architecture.

The centralized co-located BBUs alleviate network maintenance challenges, as they are automatically reconfigured to recover from errors. It supports a vast reduction in the human intervention compared to traditional networks. A CRAN with a virtualized BBU pool enables smooth migration for technology upgrades and it requires less hardware changes. Co-locating BBUs in a BBU pool enables more frequent CPU updates. It is, therefore, possible to benefit from IT improvements via merging telecommunication with the latest advancements in IT technologies.

Challenges of CRANs

One of the key problems of CRANs is deploying a larger number of RRHs over a smaller area. This may lead to environmental and social issues. In addition to this, it requires optical fiber networks and expertise on networks. The lack of expertise might cause unknown technical problems and fault tolerance issues. The BBU cooperation and cluster formation have a different set of challenges. In a clustered environment, selecting a leader and identifying workers to process data will be an issue to handle. Although existing algorithms can be used, adapting them to the CRAN architecture being investigated [10].

Virtualization is a very important aspect of CRANs, as BBU resources are shared among different network operators and the network functions are executed on top of generic processors. Hence, clear separation between hardware and software is required for generic APIs. Application-layer virtualization might not be directly applied to CRANs and it might need careful design and implementation. Network function virtualization (NFV) and software-defined networks (SDNs) are currently evolving techniques in such related virtualization. Since CRAN uses general-purpose processors (GPPs), latency intensive processing such as FFTs might not be ideal. This might lead to inaccurate data interpretations and may cause unexpected errors in mobile networks. Hence, latency intensive processing should be carried out using a dedicated hardware pool. However, the integration of a dedicated hardware pool and BBUs is not yet standardized, thus it may lead to numerous integration problems.

Furthermore, upgrading existing traditional networks into CRAN supported features create several problems such as device incompatibility, communication interface incompatibility, and message and data format incompatibility. Thus, upgrading requires complete removal of existing infrastructure, which leads to heavy losses for network operators. This may result in a partial integration of CRANs into mobile networks or a very high integration time [22].

There will be synchronization issues at BSs as well. Since CRAN is an open platform base station solution, many operators may share the resources and resource allocation and synchronisation through shared media. This may lead to many open ended and unknown problems. Moreover, incorrect synchronization may cause bad user experience and data losses.

The transport network should support strict jitter and latency requirements, and it should support high bandwidth transmissions. It should also be cost-effective. The most advanced CoMP scheme, which is joint transmission, requires timing accuracy in collaboration process between the BSs. This is hard to achieve in practice. To this end, the subframe processing delay on a link between the RRHs and the BBUs should be kept below 1 ms to meet HARQ requirements. Due to the delay requirements of the HARQ mechanism, generally, the maximum distance between the RRH and the BBU must not exceed 20–40 km. Special security and resilience mechanisms need to be implemented due to the co-location of many BBUs as well. Solutions enabling connection of BBUs should be reliable, should support high bandwidth, low latency, and low cost, with flexibility in topology interconnections. Thus, C-RAN must be more reliable than traditional optical networks such as synchronous digital hierarchy (SDH).

Implementation of CRANs

Physical layer components of CRANs are deployed in a centralized or a partially centralized fashion. In centralized solutions, L1, L2 and L3 functionalities are kept within the BBU pool, and the RRH only relays high bandwidth signals to the BBU. This strategy creates a huge burden on the transport network, but it simplifies resource sharing and interference mitigation techniques such as CoMP. For partially centralized solutions, L1 functionality is kept within the RRH and demodulated signals are sent over the network to the BBU pool. This, creates more complexity in the sharing of resources, thus makes CoMP support harder.

The physical medium plays a vital role in deploying CRANs. Mostly, fiber transport networks are used in the CRANs. Dark fiber is used in many networks as it is relatively cheap compared to other solutions. However, Dark fibers are recommended for BBU pools with lesser than 10 macro BSs. The process consumes considerable fiber resources, thus scalability will be a challenging issue. Therefore, wavelength-division multiplexing (WDM) and optical transport network (OTN) solutions are used within the base stations having

limited fiber resources. This improves the bandwidth of fronthaul links by 40% to 80%. However, despite improving bandwidth, the cost associated with the upgrade is significantly high.

In addition to this, systems based on coarse WDM like UniPON, provides both PON services and CPRI transmission. These systems are suitable for indoor coverage deployment, offer 14 different wavelengths per optical cable, and reduce overall cost as a result of sharing. Career Ethernet is the other medium used for the transport networks. Carrier Ethernet ensures 99.99% service availability. The main challenge of using carrier Ethernet is synchronization. Synchronization refers to the synchronization of the phase and the frequency alignment, which is mandatory for a BS.

Network components that are used in building the transport network of the CRAN architecture are presented next. CPRI2Ethernet gateway is used in the Ethernet medium and used to map CPRI data to the Ethernet packets. China Mobile research institute has developed an IQ data routing switch to support more than 1000 carriers for large scale BBU hotels. It uses the dynamic circuit network (DCN) technology, and it can be used to easily implement load balancing between BBUs. CPRI mux is a device that aggregates traffic from various radios and encapsulates them to transport over a minimum number of optical interfaces. It can also implement IQ compression/decompression and has optical interfaces. If OTN is chosen as a transport network solution, then a CPRI to OTN gateway is required as well.

Next, we look into IQ compression schemes and solutions for optimized bandwidth utilization in links between the RRH and the BBU Pool. In CRANs, the expected data rate is more than 40% above the radio interface, depending on the modulation scheme. RRHs transmit raw IQ samples towards the BBU cloud, thus efficient compression schemes are required. Currently, there are techniques in the literature such as reducing signal sampling rate, non-linear quantization, frequency sub-carrier compression, or IQ data compression [10]. Techniques can be mixed and a chosen scheme exhibits a trade-off between achievable compression ratio, algorithm and design complexity, computational delay, signal distortion, and power consumption. Reducing the sampling rate is a low complex solution that improves compression up to 66%. On the other hand, non-linear quantization improves the quantization SNR along with the increased Ir interface complexity and IQ data compression normalizes power levels to average symbol power reference and improves the compression ratio. Lack of well-designed algorithms is the main issue with these implementations. Sub career compression is achieved by implementing the FFT/Inverse FFT (IFFT) blocks at the RRH itself, which allows a 40% reduction in the Ir interface load.

While considering the RRH implementation, the existing RRHs are expected to work in a fully centralized CRAN architecture in a plug-and-play manner. In the case of a partially

centralized CRAN architecture, L1 needs to be incorporated into RRH. The major concerns that need to be monitored are the delay caused by transmitting over large distances, and achieving higher bit rates compared to previous deployments. With regards to BBU implementation, supporting multi-standard platforms for BBUs, and supportive processor interfaces, are considered to be the main concerns. Multi-mode base stations are used with reconfigurable software interfaces. In current deployments, field programmable gate arrays (FPGAs) and digital signal processors (DSPs) are used to implement the signal processing. Due to improvement in IT, general-purpose processors are used for BBU implementation with the help of virtualization.

Virtualization is a technology used to create abstract and isolated environments over common hardware platforms. Network function virtualization (NFV) and the software-defined networks(SDN) are the main technologies supported by the CRANs. The network virtualization contains a group of virtual nodes and virtual links. Multiple virtual networks coexist on the same physical substrate. This enables many virtual operators to perform on the same hardware pool. Virtualization is mainly supported by the operating system and it handles job scheduling, resource sharing, and management of virtual BBU pools [7,10,43]. Mainly, virtualization is two-fold with regards to CRANs, namely, network resource virtualization, and computational resource virtualization. Realizing the virtualization of computational resources includes ensuring massive parallelism for real-time applications, minimizing the computation latency within the OS, reducing the communication latency among VBS entities, and keeping the clocks synchronized among BSs. Moreover, there are virtual network interfaces operated on shared media in the virtualization of network resources.

Finally, there is the heterogeneous implementation of CRANs in the form of Hetnets, super hotspots, APs in railways, and highways. Hetnets are deployed by adding BBU pools to macro base stations and deploying additional RRHs into cells. Cell splitting is used to increase the system capacity by splitting macrocells into smaller cells and deploying RRHs. Providing additional frequency bands for RRH's operation and deploying more RRHs are considered by overlay networks. Railways and highways are equipped with RRHs to handle smooth handovers over traditional RANs.

2.1.3 D2D Networks

A high density of mobile users and the booming data demand have pushed the RRHs and BBUs to their capacity limits. Also, due to new mobile applications that require high data rates, low latency and high throughput network protocols should be facilitated. This motivates direct communication between mobile terminals (MTs). To this end, D2D communication is first proposed in FlashLinQ networks, where D2D networks focus on applications

such as content distribution and location-aware advertisements [6]. There are two types of D2D networks called controlled networks and autonomous networks as stated in [20].

Controlled D2D Networks

The control plane is operated by the BSs and they have a simple implementation compared to autonomous networks. The BS controls the resource allocation and mode selection at the mobile devices. The control plane focuses on white space detection, collision avoidance, and synchronization [7]. BS assisted communication is established through LTE-A network architecture. Moreover, mobile devices may use the cellular spectrum for communication, underlay to the cellular network, or overlay to the cellular network. In underlay communications, the spectral efficiency is higher, at the expense of complex interference management compared to overlay communications. Figure 2.4 illustrates the deployment of a controlled D2D Network.

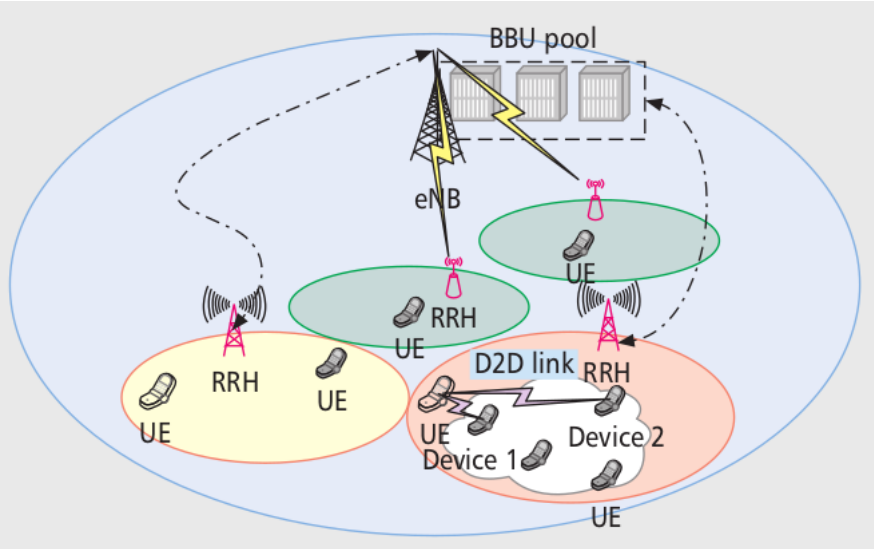

Figure 2.4: CRAN Assisted Controlled D2D Network

Autonomous D2D Networks

Autonomous D2D networks are clustered deployments based on some indices. The indices include physical layer parameters such as proximity and application layer parameters such as user interest etc. Mainly, the control plane and the data plane exist within mobile devices. The implementation of these networks is not straightforward and complex due to the self-operating nature of the devices.

There are semi-autonomous networks as well. In semi-autonomous networks, the mobile devices are geographically clustered, and a selected leader is responsible for the control plane

operations. The slave nodes are used for data transmission. On the other hand, every node is responsible for the control plane and data plane operations in a fully autonomous network. Mainly, these kind of networks are implemented using out band communication, where communication occurs in the industrial scientific and medical radio band (ISM). However, power level monitoring of the operating nodes and achieving quality of service (QoS) requirements of the communication, resource allocation, and interference handling are considered complex due to limited processing power and limited memory in mobile devices [7, 37].

There is a common set of open issues in D2D networks concerning interference management, power management, modulation format handling, resource allocation, and channel measurement. To analyze these issues and to find theoretical solutions, D2D networks are modeled using mathematical tools. Next, we introduce some mathematical tools that are used in the literature.

2.1.4 D2D Network Modeling

Mainly, ad-hoc networks, sensor networks, mobile D2D networks consist of transmitters (TXs) and receivers (RXs). The spatial distribution of the TXs and RXs directly affects the communication. Following the literature [2, 11], the received signal power mainly depends on the distance between the TX and the RX, and the fading environment. Most papers consider the fading distribution to be Rayleigh for mathematical tractability. Stochastic geometry is the main tool used for location modeling. Homogeneous Poisson point processes (PPP) are used in R^2 infinite plane and independent and uniform distribution (iud) is considered when the region is finite.

Important properties of a homogeneous PPP are listed below [5, 14, 15, 25]. If the point process ϕ is a homogeneous PPP, then

- for any two disjoint Borel sets $A1$ and $A2$, the random variables $P(n, A1)$ and $P(n, A2)$ describing the number of points of a process falling in these sets are independent,
- the number of points $P(n, A)$ falling in a bounded Borel set A is distributed according to the Poisson distribution with parameter $\lambda L(A)$, where $L(A)$ is the area of A, and λ is the mean density of points,
- $P(n, A)$ and $E[P(n, A)]$ follows

$$P(n, A) = \frac{[\lambda L(A)]^n}{n!} e^{-\lambda L(A)} \tag{2.1}$$

and

$$E[P(n, A)] = \lambda L(A), \tag{2.2}$$

- void and contact probabilities $C(r, A)$ are given by

$$P(0, A) = e^{-\lambda L(A)} \tag{2.3}$$

and

$$C(r, A) = 1 - P(0, A) = 1 - e^{-\lambda L(A)}. \tag{2.4}$$

There are two methods when considering independent and uniform distribution of points in a finite area $A \subsetneq R^2$. First method is by making approximations and the second is resorting to exact mathematical modeling. Most of the papers resort to approximation method given in [9, 19], that gives the void distribution as follows:

- For a circle of radius r, we have the void distribution

$$V(0, r) = \left(1 - \left(\frac{r}{R}\right)^2\right)^N. \tag{2.5}$$

To model clustered networks, most of the work in the literature use Poisson hole processes (PHP). There are D2D networks where D2D communication is only allowed within the given geographic region. Hence, D2D transmitters can transmit only within the bounded region as shown in Figure 2.5. D2D devices are also modeled using a Poisson dipole process (PDP), where TXs are modeled using a homogeneous PPP, and D2D RXs are located at a fixed distance from the TX within the given region. Independent thinning of a PDP results in a PHP. The distribution of the distance between the TX and the RX follows the Rayleigh PDF and the distribution of the distance between the TXs follows the Rician PDF.

2.1.5 Network Performance Metrics

Network designing requires baseline estimations of key performance metrics to analyze the efficiency of network architectures. Physical layer metrics such as data rate and channel capacity do not provide sufficient insights for the current requirements of networks. Current applications require a high QoS level to operate smoothly, hence, delay, jitter, and delay violation probabilities need to be considered. Currently, channel models such as Rayleigh

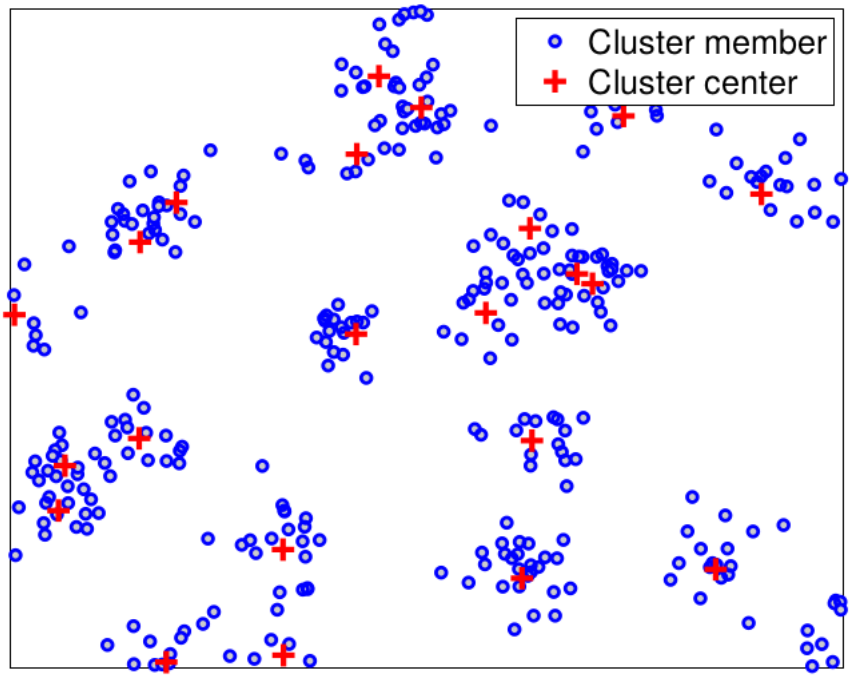

Figure 2.5: D2D Clustered Network

fading models with specified Doppler spectrum are inadequate to capture these parameters. Thus, a link-level channel model called effective capacity (EC) is proposed in [38].

Effective Capacity

As stated above, extracting QoS measures out of physical layer channel models is not a straight forward task and might lead to inaccurate results. Hence, channel modeling of the link layer with queue analyzing techniques is used to derive expressions for the effective capacity. Let's consider a data source that transmits data with a constant data rate of μ and the receiver receives data via a time-varying channel. The delay violation probability (DVP) for a delay $D(t)$ associated with a link, is defined as

$$\mathsf{Pr}\left\{D(t) > D_{max}\right\} \leq \epsilon, \tag{2.6}$$

where D_{max} and ϵ are upper delay threshold and upper bound for DVP, respectively. Furthermore, let $r(t)$ be the instantaneous channel capacity. The effective capacity is defined as

$$\alpha(u) = -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)d_\tau}], \tag{2.7}$$

where $u \geq 0$. If $\alpha(u)$ exists for a large queue,

$$\mathsf{Pr}\left\{D(t) > D_{max}\right\} \approx e^{-\theta(\mu)D_{max}} \tag{2.8}$$

for large D_{max}, and the QoS component is denoted by $\theta(\mu)$ such that

$$\theta(\mu) = \alpha^{-1}(\mu), \tag{2.9}$$

where $\alpha^{-1}(.)$ indicates the inverse function of the EC.

Lower the QoS indicates higher DVP and higher QoS indicates lower DVP, hence, $\theta(\mu)$ directly indicates the QoS gain of the channel. If $\theta(\mu)$ of a channel is known, then the probability of achieving the QoS can be calculated using r_s, D_{max}, ϵ. If $\theta(r_s) \geq \rho$ for $\rho = \frac{-\log \epsilon}{D_{max}}$ then the respective channel is considered to satisfying the required QoS level. Moreover, effective capacities of the network with parallel wireless links, network with tandem wireless links, and network with variable sources can be found in [38].

Transmission Capacity

Transmission capacity is defined as the product of the user density and the successful transmission probability of a network. A higher transmission capacity implies a higher activated D2D link density. Moreover, transmission capacity is dependent on the user density, interference, distance between TX and RX, and the fading environment. Transmission capacities for different kinds of networks are analyzed and proposed in [40]. The mathematical expressions for successful transmission probability depend on the considered network model. For instance, consider an underlay network model having cellular and D2D users that share the same spectrum. We denote the parameters of the cellular network by subscript 0 and the parameters of the D2D network by subscript 1. The signal to interference ratio (SIR) of the system, assuming an interference-limited network is given by

$$SIR_n = \frac{P_n \delta_{n0} R_{n0}^{-\alpha}}{\sum_{j \in 0,1} I_j} \tag{2.10}$$

where $n \in \{0, 1\}$,

$$I_j = (\frac{P_j}{P_n}) \sum_{k \in \Pi_j} \delta_{j,k} |X_{jk}|^{-\alpha}, \tag{2.11}$$

and P_n, δ_{n0} and R_{n0} are transmission power, fading coefficient of the channel between TX and RX and the distance between TX and RX, respectively. Π represents the transmitters of each network and subscript n0 represents the channel between transmitter and typical receiver. Hence, the successful transmission probability with respect to a given threshold can be written as

$$\mathsf{Pr}\left\{SIR_n \geq T_n\right\} = e^{-C_\alpha T_n^{\frac{2}{\alpha}} R_{n0}^2 \sum_{j \in (0,1)} \lambda_j (\frac{P_j}{P_n})^{\frac{2}{\alpha}}} \tag{2.12}$$

where λ_j denotes the user density, $C_\alpha = \pi T(1 + \frac{2}{\alpha}) T(1 - \frac{2}{\alpha})$ and $T(.)$ is the standard gamma function. Mainly, transmission capacity is used for power estimations in networks and for

statistical estimation.

2.2 D2D Communication

Communication means exchanging information or messages between two endpoints. These endpoints can be MS and BS, BS and BS, MS and MS. Communication between a MS and a MS is called D2D communication, and it is standardized in LTE-A architecture. D2D communication relies on three main functional blocks, which are mode selection, power allocation, and resource allocation. In this section, we consider D2D communication technologies, spectrum sharing techniques, interference management, and their impact on transmission power of the network. D2D communication can be classified as inband or out-band communication. Furthermore, inband communication can be classified as underlay communication and the overlay communication.

2.2.1 Underlay Communication

A cross-tier communication in the same cellular spectrum is identified as underlay communication. It increases spectral efficiency, but interference management complexity is an issue. Let us consider an underlay network where the cellular channel is divided into sub-channels and a fraction of $\beta \in [0, 1]$ is used for D2D communication. An expression for the successful transmission probability can be formulated to provide important insights related to the network as in [18]. Consider that the D2D users form a homogeneous PPP ϕ_d with intensity λ_d and that the cellular users form a homogeneous PPP ϕ_c with intensity λ_b. The interference at a typical receiver can be written as

$$I_d = \sum_{X_i \in k\beta\phi_{d,0} n\{0\}} P_{d,i} G_i ||X_i||^{-\alpha} + \sum_{X_i \in \phi_c} P_{c,i} G_i ||X_i||^{-\alpha}, \tag{2.13}$$

where G and X denote the fading coefficients and the distance distributions, respectively. The successful transmission probability for a given threshold can be written as

$$\mathsf{Pr}\left\{SINR \geq x\right\} = exp\left(-N_0 x - c\beta x^{\frac{2}{\alpha}} - \frac{1}{2sinc(\frac{2}{\alpha})}(\beta x)^{\frac{2}{\alpha}}\right), x \geq 0. \tag{2.14}$$

The spectral efficiency R_d is given as

$$R_d = k \int_0^\infty \frac{e^{-N_0 x}}{1+x} exp\left(-c\beta x^{\frac{2}{\alpha}} - \frac{1}{2sinc(\frac{2}{\alpha})}(\beta x)^{\frac{2}{\alpha}}\right) \tag{2.15}$$

According to the above expressions, decreasing β increases spectral efficiency, hence, we

can interpret that inband D2D users should access small bandwidth with higher power density compared to spreading the power over a large bandwidth. However, small β compromises the spectrum resource available to the D2D transmissions, which affects the D2D throughput. Moreover, to minimize the interference of the underlay communication, frequency hopping techniques can be used.

2.2.2 Overlay Communication

Overlay communication uses orthogonal frequency resources for D2D communication and cellular communication, hence, it has zero cross-tier interference. Moreover, it increases throughput and rate of the network, but spectral efficiency is reduced. In most cases, the cellular uplink is divided into two orthogonal portions and a fraction η is used for D2D communication, and $1-\eta$ is used for cellular communication. Let's consider a network where D2D users are modeled with homogeneous PPP ϕ with intensity λ. The successful transmission probability is given by,

$$\mathsf{Pr}\left\{SINR \geq x\right\} = \exp\left(-N_0 x - c x^{\frac{2}{\alpha}}\right), x \geq 0 \tag{2.16}$$

where

$$c = \frac{kq\left(\frac{\lambda}{\epsilon} - \left(\frac{\lambda}{\epsilon} + \lambda\pi\mu^2\right)e^{-\epsilon\pi\mu^2}\right)}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}, c \geq 0 \tag{2.17}$$

and $k, q, \epsilon, \mu, \alpha$ are aloha access probability, potential D2D UEs, D2D distance parameter, mode selection threshold and path loss exponent, respectively. Furthermore, spectral efficiency R_d of D2D links is given by

$$R_d = k\int_0^\infty \frac{e^{-N_0 x}}{1+x} e^{-c x^{\frac{2}{\alpha}}} dx. \tag{2.18}$$

Apart from that, cellular interference needs to be considered separately. In this case, additive white Gaussian noise has a significant impact on successful transmission probability. Moreover, for sparse deployment, the network is noise limited, and the dense network would be interference limited.

2.2.3 Outband Communication

Outband communication mainly occurs in the ISM band. ISM band operates in two frequency bands at 2.4 GHz and 5 GHz. Due to its high frequency, the 5 GHz band is used for short-range communication. Some papers have assumed narrowband channels [24, 32], while others have assumed wideband channels for the ISM band. If we consider wideband tech-

niques, the main advantages are from the frequency hopping and the limited interference, but we cannot obtain closed-form solutions for analytical results. Besides, outband D2D may suffer from the uncontrolled nature of the unlicensed spectrum. It should be noted that only cellular devices with two wireless interfaces (e.g., LTE and Wi-Fi) can use outband D2D, and thus users can have simultaneous D2D and cellular communications. Furthermore, Bluetooth and Wi-Fi direct operate in the outband frequency range, and currently, millimeter-wave communication is getting popular as well.

Millimeter-wave communication happens in the frequency range of $30-300$ GHz, and the power decay is proportional to the square of the career frequency [18]. Since propagation loss is high in mmWave communication, highly directional antennas are deployed. Line of sight signals with short wavelengths are used, hence, it gives high throughput and low multi-user interference.

2.2.4 Wi-Fi Offloading

Wi-Fi offloading has been proposed as prominent solution to data explosion in cellular networks. Mainly, there are two types of solutions called opportunistic offloading and delayed offloading [27]. When a MS has reached a Wi-Fi zone, it automatically starts to transmit data via the Wi-Fi network, otherwise, it transmits data via the cellular network. Estimation of handover thresholds and delaying in handover incidents are the main issues related to opportunistic offloading. Balasubaramanium *et al* [36] has proposed delayed Wi-Fi offloading. It predicts the time of accessing the Wi-Fi network by the MS and buffers data for later delivery. Delayed Wi-Fi networks inherently requires to implement a data session with the access point (AP) and the session should be persisted. Moreover, network coverage, data rate, and cost of delay bound are the main factors when deciding the amount of data to be transferred. The efficiency of delay tolerance and fast switching directly affects the performance of Wi-Fi offloading. Geo *et al* [1, 21] have proposed a cost bearing mechanism for cellular operators and AP operators using the Nash bargaining theory.

Also, the IEEE 802.11 standard provides Wi-Fi direct to directly communicate between mobile devices instead of going through the APs. However, the peer discovery process utilizes more radio resources and power of the mobile devices, which leads to quicker battery drain. Hence, BS assisted Wi-Fi direct protocols are proposed and used with the LTE-A architecture. Cooperative streaming and social gaming have hugely benefited from this architecture, and it has been found out that 30% offload increases cell throughput by 50%. In addition to this, there are Integrated Femto-Wi-Fi networks that use cellular frequency for providing Wi-Fi capabilities instead of the ISM band. Those APs are called F-APs and used in the industry. With the help of Wi-Fi offloading, we can also shut down some BSs that are based on cellular traffic, which reduces TCO significantly.

The paper [41] has implemented a quality-aware traffic offloading (QATO) framework for offload traffic using Wi-Fi direct. They have developed a web browser for uploading videos and photos and they measure the QoS for decision making and for communication with mobile devices. According to their results, they have shown that this saves energy by 38% for downloading and 70% for uploading. Hence, we can conclude that Wi-Fi offloading has a significant impact on reducing the load on BSs. However, still there are open ended problems in the areas of peer discovery and handover handling in Wi-Fi traffic offloading.

2.2.5 Mode Selection and Admission Controlling

Mode selection is crucial in a heterogeneous D2D networks. Mainly, there are three communication modes, namely, inband underlay, inband overlay, and outband communication. Selection of the communication mode depends on the parameters of the network. Some networks might prefer QoS compared to power minimization. In contrast to that, some networks may be power sensitive compared to the reliability of the communication. Hence, mode selection depends on the network type. Furthermore, there is a selection decision between time division duplexing (TDD) and frequency division duplexing (FDD) modes . For D2D networks, TDD is recommended due to scarce frequency resources. Duplexing allows simultaneous communication as well [22, 35].

Admission control is mainly decided based on QoS fluctuations and the interference level in the network. Admission control can be classified in to two categories as pure analytical approaches and simulation based approaches. Pure analytical results rely heavily on the assumptions, hence, they can be used as a baseline for the network. However, in a realistic setting, the results may deviate from the results. Simulations have resulted in more accurate results according to the literature, but it is hard to utilize pure simulation results to predict the future. Joint optimization algorithms can also be seen in the literature. Joint optimization is a mix of analytical and simulation results, thus it gives superior results, and we can use them to effectively predict different states of the network.

2.2.6 Power Controlling and Energy Harvesting

Power controlling is a hot topic in D2D networks because mobile devices are power constrained. Therefore, estimating the required power to satisfy a QoS level is a necessary in D2D communication. Currently, many networks use real-time channel state information (CSI) to estimate the channel condition, and the required power is based on that. However, the dynamic environment of mobile devices due to movements and inaccurate channel estimations at the receiver causes inaccurate power estimation. Hence, [32] proposes statistical feature based power controlling (SFPC) using D2D success likelihood and opportunistic

access control. Although opportunistic access control reduces interference and maximizes area spectral efficiency, power can be varied only at zero or peak power. To this end, they have proposed a power variation between zero and maximum power using their proposed algorithm. The main advantages of SFPC compared to other schemes are reduced latency, resilience to user's movement, improved link reliability and the low processing overhead.

Joint coordination scheduling and power controlling are used in many networks as well. The paper [13] proposes a scheme in which the users attached to a given BS is divided into power zones. The power zones are divided into time and frequency blocks and users allocated to the power zones are considered for power controlling under specific conditions of the power zones. They have formulated graph problems and have used graph-based theories to solve the power calculation problems. Furthermore, they have used the maximum weight clique problem to obtain the analytical results.

Energy harvesting is another field of interest in D2D networks [8]. Mainly, this is used in sensor network deployments, where autonomous operations are mandatory. They store the power received from signals and use them to charge their batteries. Energy harvesting can be done primarily using environmental energy sources, or RF energy harvesting networks (RF-EHNs). Energy gain from environmental sources is not steady and varies due to weather and environmental conditions, hence, it is not an ideal solution, RF-EHNs are proposed as a solution. RF-EHNs are implemented in APs and they have energy transmission zones and information transmission zones. Dedicated ambient RF sources are deployed to transmit a steady energy flow in RF-EHNs. Moreover, energy harvesting zones contain a low power microcontroller and a low power RF transceiver. The received energy can be written as

$$E_h(d) = \epsilon f(d, \alpha) \tag{2.19}$$

where $f(d, \alpha) = P_r(d) \times 10^L |r|^2$, $L = -\alpha \log_{10}(\frac{d}{d_0})$ and $P_r(d), \epsilon, \alpha, d_0, r$ are received power, conversion efficiency, path loss exponent, reference distance and random number following the complex Gaussian distribution, respectively.

2.3 D2D Caching Techniques

Nowadays, location-based networks have been transformed into user-centric content-based networks. Recently, it has been shown that more than 70% of the traffic in the internet is due to internet videos. This emphasizes that low latency and high data rates are necessities. As a facilitator, caching techniques are introduced, and caches are deployed at different caching layers. Primarily, content caching is categorized into two main functions as content placement and content delivery.

2.3.1 Content Placement

Content placement is placing data from content servers in storage of intermediate servers for easy access. Content placement strategies differ based on network type as well. Mainly, we consider caching at content delivery networks (CDNs), mobile network edges and D2D networks. The efficiency of content placement is measured using the cache hit rate, and it depends on the content popularity prediction and cache loading and replacement algorithms. Firstly, we will focus on the content placement strategies used in content distribution networks (CDNs)

Content Placement at Content Distribution Networks

According to the literature [3, 33, 34], least recently used (LRU) and least frequently used (LFU) cache replacement strategies provide a high hit ratio for static content. However, these methods provide poor performance for time-varying content. Moreover, LRU and LFU both assume that heavily used latest content is highly in demanded compared to newly generated recent data. This assumption does not hold for large user bases. To this end, online and offline linear regression is used to predict popularity. The offline linear regression uses the access history of the content and the online model uses real-time data within a time frame to predict the content popularity. The paper [39] proposes a hybrid popularity prediction scheme using a hybrid model of online and offline linear regression.

Furthermore, [39, 42] say that major popularity prediction algorithms such as linear regression, logistic regression, temporal evolution regression, decision trees and hierarchical clustering, are not idealfor content-centric networks as they do not focus on the network topology and the flow distribution. Hence, they have proposed in between content popularity prediction framework called BEACON, that uses the network topology and the flow distribution to calculate the popularity. For instance, let us consider a network having intermediate nodes B, C, D and the content server is E and the content requester is A. When A requests new content in legacy approaches, content is cached at every intermediate node B, C, D. However, according to BEACON, caching happens only at the nearest node A.

According to [28], information-centric networks (ICNs), content-centric networks (CCNs), content delivery networks (CDNs), and web caching are heavily used in LRU and LFU schemes, although they lack a time-dependent analysis. Moreover, the formulation of an optimization problems and its solutions can be used only for limited scenarios. In many practical situations, it lacks reliability. To this end, Zipf's law does not account for the fluctuation in the time series, thus prediction models such as autoregression (AR) and neural network models have been proposed. Although they lead to reliable solutions, computational overhead and memory overhead is very high. Hence, [31] proposes a model combining the

AR model with time windows, that they have named as prediction cache. The following equation depicts the model they have proposed.

$$x_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \epsilon_t, \tag{2.20}$$

where $x_t, \alpha_i, \epsilon_t$ denote time series data, AR coefficient using Burg method and the residual, respectively.

Furthermore, with the introduction of SDNs, new innovations have been introduced for caching technologies of content-centric networks, and Openflow has been used for implementation. The SDN controller, the SDN cache, and the SDN Proxy are primarily the three main components of a SDN network. The SDN controller has the view of the network and its topology, hence, it can manage all flow decisions. The SDN proxy and the SDN cache are used to implement the caching layer at SDN networks. The SDN controller directs the traffic of interest (TOI) to the relevant SDN proxy. Caching policies of the SDN networks are categorized as caching everything, caching based on regular expression, the cache only size matching data, cache data of a target domain, and cache data with a given type.

A cooperative caching strategy is proposed in [4] which is applicable for clusters. A network can be partitioned into a set of nodes and these nodes are categorized into clusters. Each cluster has its own caching decision, cache replacement algorithm, and forwarding strategy. According to the literature, a cache is divided into on path caching and off-path caching. On path caching is done at the data path and off-path caching is done at a specific location. However, both types require content storage (CS), pending interest table (PIT), and forwarding interest base (FIB). If the content is present in the CS, it can be served. Otherwise, the content is checked at the PIT. If the PIT does not contain the requested data, the request is forwarded to the FIB. If the FIB is also unable to find the data, the data is flooded to the network to be cached by relevant nodes. They primarily use leave copy every (LCE) and leave copy down (LCD) methods for content placement.

It is stated in [34] that 20% of the most popular content account for 80% of the network traffic. Hence, identifying the most popular content is paramount. The above-stated paper has proposed lifetime popularity score (LPS) for content prediction using a stochastic model. Historic view counts and lifetime view counts are taken into account for the LPS. They have modeled users using a Poisson point process and have worked on identifying more important features of popular content using Bayesian theory.

Content Placement at Cloud Radio Access Networks

Caching at the CRANs is more complicated than caching at CDNs. Since CRANs are composed of RRHs and BBUs, many external factors such as user mobility, signaling envi-

ronment, and geographical clustering of RRHs should be considered while designing a cache. Moreover, limitations of dynamic fronthaul and backhaul conditions and highly volatile user density and preference will further complicate the process. According to the literature [29,30], hierarchical caching for CRANs using asymptotic limits of caching and the mean filed theory is used to improve the reliability of caching implementation. Moreover, multicast beamforming and dynamic clustering are used to minimize power consumption. Cooperative caching and clustering based on the Akaike information criterion have been used to formulate CRAN cluster and predict content popularity based on geographic locations.

Low complexity search algorithms are used to reduce the average caching failure rate in CRANs. However, the main challenge of the network is user wise content prediction and predicting the mobility patterns of users. Hence, proactive caching is introduced with the utilization of cloud infrastructure. Since cloud computing capabilities enable the use of machine learning algorithms to predict content, the paper [42] proposes an echo state network with a sublinear algorithm to implement caching techniques. It captures content request distribution for each user, mobility pattern of the user, and the users' preference over the content to formulate the caching layer at the edge of the network. Moreover, the Hadoop framework is used for data extraction and social information-based future access patterns are recognized to further improve the caching efficiency.

Caching at End User Devices

D2D caching is primarily used with the 5G D2D networks with the standardization of D2D communication. The main challenges with the end-user MS caching are limited power and limited storage. Hence, controlled D2D networks are used to formulate the caches that have the global view of MSs at the BSs. Moreover, an opportunistic multihop transmission has been considered to offload the network by exploiting the MS capabilities in the cellular network. Large virtual unified cache spaces are formulated using MSs. The MSs are clustered based on geographical regions or based on interest matches. Storing multiple content over a virtually unified distributed memory is the main challenge faced with the MS cluster formation. Channel state information, the popularity of the content, and available bandwidth resources are the main facts that have been considered for D2D caching. Some literature has shown that backpack theory is also used for clustering [28].

2.3.2 Content Delivery

The content delivery phase is of utmost importance in cache design. Minimum data download delay and reliable data delivery are essential for new D2D networks. Mainly, content delivery networks use different techniques and algorithms to transfer data. Primarily, we

are looking to wireless data transmission of D2D networks. Literature has shown two types of content delivery network formulation types known as self-organizing distributed networks and centrally coordinated distributed networks. Self-organizing distributed networks do not have a central coordinator and they take decisions based on the local information. Coordinated networks have a central manager that instructs the nodes to transfer data.

Message passing is investigated in the literature and algorithms and techniques such as sum-product algorithm, forward and backward algorithm, Pearl's belief propagation, Bayesian networks, Kalman filtering, marginalization, Tanner graphs, and the Viterbi algorithm are used. According to [23], they have developed low cost high efficient Tanner graph inherited factor graph-based sum-product algorithm for message passing without involving a central coordinator.

Furthermore, [42] proposes an energy-efficient cluster oriented solution for multimedia delivery using LTE and Wi-Fi direct called ECO-M. Although Wi-Fi offloading is popular, opportunistic communication through D2D have become a promising technology. However, Wi-Fi direct discovery algorithm is not time optimized and can cause delays in the range of seconds.

The energy consumption of nodes is not taken into consideration while designing many content delivery protocols. To this end [42] proposes an energy-efficient and quality-aware protocol for content delivery. They have grouped users into clusters based on user preference, and the cluster head is selected based on battery level and the communication channel quality. The cluster head communicates with the BS using the cellular spectrum, and cluster members communicate with the cluster head using Wi-Fi direct.

The paper [26] proposes a non-centralized distributed content delivery protocol using belief propagation. It uses local information available to MSs and minimum communication with BSs to decide the content delivery protocol. The simulation results have shown that it reduces average download delay significantly compared to centralized versions. They have stated that minimizing average download delay subject to BSs storage capacities is NP-hard. Hence, collaborative caching with belief propagation has created a sub-optimal solution for content delivery problems.

Chapter 3

System Model and Problem Formulation

3.1 Problem Formulation

3.1.1 State of the art

Based on the literature, mobile networks have been evolved from the 1G network to CRANs. 1G networks mainly focus on the successful transmission of signals and with the introduction of 2G networks, factors such as resource utilization, QoS, and operational costs are considered. However, those technologies are not adequate to satisfy the high data demand due to the advancement of current mobile applications. To this end, 3G architecture is introduced with a higher bit rate by significantly improving network capacity. Later, LTE and LTE-A architectures are introduced to further maximize the network capacity.

However, those network architectures are not scalable with the cost of the expanding of networks. The TCO is increased proportionately to the network capacity. Hence, researches are carried out to investigate novel architectures and strategies to maintain flat TCO while placating rapidly increasing network capacity. To this end, CRANs are proposed and IT technological advancements are leveraged into mobile networks. Moreover, D2D networks are designed to further assist the cellular networks for traffic offloading. There are autonomous and semi-autonomous D2D networks and they are an integral part of CRANs. Integrating D2D networks and CRANs is not a straightforward methodology. Integrating D2D and CRAN networks has open-ended unknown complexities. According to the literature, the most prominent areas can be categorized into protocol management, power management, spectrum management, and user admission control. However, those areas are not independent of each other, but, have a high correlation among them. For instance, spectrum allocation directly decides the interference level of the mobile communication and

interference is a factor in deciding transmission power. Hence, we can elaborate that power management and spectrum management has interconnected behavior. These correlations further exacerbate the integration problem.

3.1.2 Current challenges and solutions

Primary concerns of D2D networks are power efficiency and QoS requirement satisfaction. The literature has shown hybrid networks, that D2D networks and cellar networks coexist and hybrid networks play a prominent role in reducing network traffic. However, there are still some unaddressed challenges such as, mobile cluster formulation, user allocation into clusters, switching between legacy and D2D networks, smooth handovers of signals between networks, operational strategies for the telecommunication industry. Many kinds of research highlighted that D2D networks are suffering from limited power capacity of mobile terminals, hence, it won't be a good supporter for traffic offloading. To this end, we investigated on power-efficient protocols for D2D networks with the help of CRANs. Moreover, achieving good QoS via D2D networks is another primary challenge that researches have faced. Hence, we extensively investigated on minimizing power efficiency while maintaining required QoS for D2D networks, and findings are presented in this thesis. we propose a hybrid D2D network with power-efficient communication protocols, where mobile users are operated, overlay to the cellular network, and outband to the cellular network. Furthermore, we guarantee that QoS requirements are satisfied with the D2D network admission controlling algorithm proposed in our solution.

3.2 System Model

We consider an E-CRAN cross laid with a D2D network, which comprises of geographically clustered RRHs, a content cache and a BBU pool. The RRHs are spatially distributed in the entire 2-D plane according to a homogeneous PPP Φ_{bs} of intensity λ_{bs}. Each RRH uses a fixed transmit power P_{bs}. Three types of users, namely, data consumers (DCs), data producers (DPs), and external users (EUs) are considered in our model. The DCs are connected to their nearest RRH, and they request content from their connected RRH. The DPs cache the most popular content files from the edge cloud cache, such that the cache hit probability (CHP) for a given file is p. Moreover, we assume that a typical DC is at the origin and hereafter, we refer to it as the DC. The spatial distributions of the DCs and the DPs are modeled using homogeneous PPPs Φ_{dc} and Φ_{dp} with intensities of λ_{dc} and λ_{dp}, respectively. The EUs, that operate in the ISM band, are modeled using a PPP Φ_{ext} of intensity λ_{ext}, and each EU transmits with a fixed power P_{ext}.

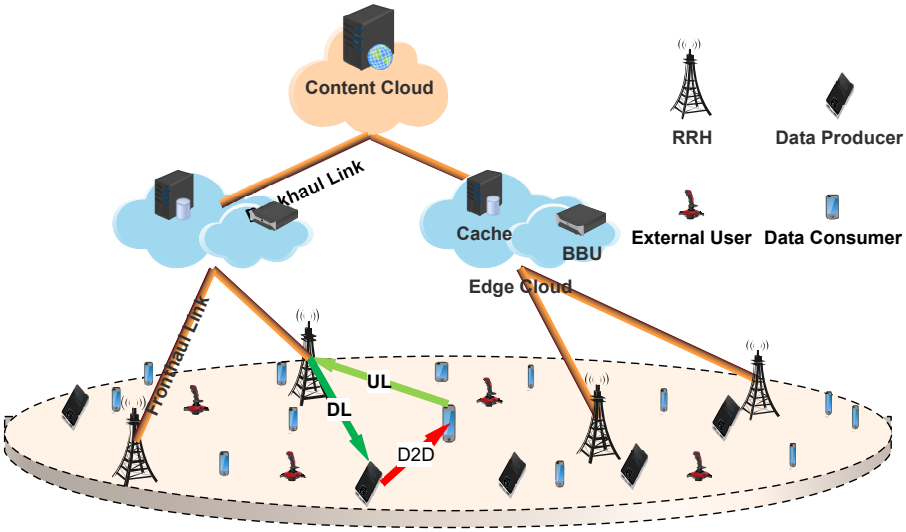

Figure 3.1: Communication Modes and System Model

An interference limited network is assumed where the additive noise is negligible compared to interference. For all links, Rayleigh fading is assumed where the channel power coefficients are independently and identically distributed exponential random variables (RVs) of unit mean. A distance dependent path loss model with exponent $\alpha > 2$ is also used to model large-scale fading, while the effects of shadowing are neglected due to being short range D2D links.

We assume that all DCs will make their requests simultaneously. A request of a DC will generally be served by the RRH, but if there is a DP in the vicinity that has the requested file in its cache, the DC may get served by this DP. Delivering the file directly from the edge cloud is referred to as cellular mode and the latter case is termed D2D mode. The DC will get served by the DP only if the content delivery through D2D communication can provide equal or higher QoS compared to content delivery through the RRH. The transmission delay violation probability (DVP) with respect to a given delay threshold $D_{\max}$, i.e., for link delay D, $\mathsf{Pr}\{D > D_{\max}\}$, is used to measure the QoS. Intuitively, lower the DVP, higher the QoS experienced by the user.

For D2D links, the distance between the DC and the serving DP is used to determine whether the communication happens in the ISM band or in the cellular band with overlay spectrum access. These two modes are referred to as outband mode and overlay mode, respectively. Under similar network conditions, outband users, who are assumed to operate at a higher frequency band, have a small coverage area compared to overlay users, who operate at a lower frequency band. Therefore, the DC and DP pairs with short links are allocated into the outband mode, pairs having moderately long links are allocated into the overlay mode, and pairs with long links will not transmit in the D2D mode as they fail to satisfy the QoS requirements. The content delivery procedure for our system model is

summarized in Algorithm 1, where $d^{\star}_{\text{ou}}$ and $d^{\star}_{\text{ol}}$ are the distance thresholds for outband and overlay modes, respectively.

Algorithm 1 Admission and Transmit Power Control

for each request
    d ← calculate the distance between DC and DP
    if ($d \leq d^{\star}_{\text{ou}}$) **then**
        $P \leftarrow$ calculate the outband power
        if ($P \leq P_{\max}$) **then**
            transmit in outband network using power P
        else
            transmit using cellular communication
    else if ($d^{\star}_{\text{ou}} \leq d \leq d^{\star}_{\text{ol}}$) **then**
        $P \leftarrow$ calculate the overlay power
        if ($P \leq P_{\max}$) **then**
            transmit in overlay network using power P
        else
            transmit using cellular communication
    else
        transmit using cellular communication
do

Obtaining analytical expressions for the optimal values of $d^{\star}_{\text{ou}}$ and $d^{\star}_{\text{ol}}$, and the minimum required transit powers of the DPs are the main contributions of this thesis, which are presented in proceeding sections. The notations used in this book are given in Table 3.1.

Table 3.1: Notation Description

Description	Notation
Bandwidth of a cellular channel	B_{bs}
Bandwidth of an outband channel	B_{ou}
Bandwidth of an overlay channel	B_{ol}
Application level processing delay	c
Distance from the DC to the nearest RRH	$d_{\text{bs},0}$
Distance from the DC to the k^{th} DP operating in outband	$d_{\text{ou},k}$
Distance from the DC to the k^{th} DP in overlay	$d_{\text{ol},k}$
SIR of channel between DC to RRH	γ_{bs}
SIR of channel between DC to k^{th} DP in outband	$\gamma_{\text{ou},k}$
SIR of channel between DC to k^{th} DP in overlay	$\gamma_{\text{ol},k}$
Delay of the channel between DC and RRH	$d_{\text{bs},0}$
Delay of the channel between DC to the k^{th} DP in outband	$D_{\text{ou},k}$
Delay of the channel between DC to the k^{th} DP in overlay	$D_{\text{ol},k}$
Fading coefficient of the channel between DC and RRH	$h_{\text{bs},0}$
Fading coefficient of the channel between DC to the k^{th} DP in outband	$h_{\text{ou},k}$
Fading coefficient of the channel between DC to the k^{th} DP in overlay	$h_{\text{ol},k}$

Chapter 4

Analytical Results

4.1 Distance Threshold Computation

The computation of distance thresholds requires several intermediate results namely, the DVPs for each communication mode and the spatial intensities of the DPs in each D2D band. We begin by assuming that the DC requests a file of size M from the edge cloud. The edge cloud may deliver the file directly through the RRH or via a DP. The DVPs of each mode are used to make this decision. The DVP of a link between the DC and its nearest RRH is given by the following lemma.

4.1.1 Delay violation probability of a cellular communication

Lemma 1. *The DVP of the link between the DC and the nearest RRH is given by*

$$\mathrm{T}_{\mathrm{bs}}(D_{\max}) = \frac{(\gamma_{\mathrm{bs}}^{\star})^{\frac{2}{\alpha}}}{(\gamma_{\mathrm{bs}}^{\star})^{\frac{2}{\alpha}} + \operatorname{sinc}\left(\frac{2}{\alpha}\right)}, \tag{4.1}$$

where $\gamma_{\mathrm{bs}}^{\star} = \frac{M}{2^{B_{\mathrm{bs}}(D_{\max}-c)}} - 1$.

Proof. Considering D to be the sum of the propagation and processing delays, the DVP conditioned on $d_{\mathrm{bs},0}$, which is the distance between the DC and the RRH, is given by

$$\begin{aligned}
\mathsf{Pr}\left\{D > D_{\max} \mid d_{\mathrm{bs},0}\right\} &= \mathsf{Pr}\left\{\frac{M}{B_{\mathrm{bs}}\log(1+\gamma_{\mathrm{bs}})} + c > D_{\max} \mid d_{\mathrm{bs},0}\right\}, \\
&= \mathsf{Pr}\left\{\gamma_{\mathrm{bs}} < 2^{\frac{M}{B_{\mathrm{bs}}(D_{\max}-c)}} - 1 \mid d_{\mathrm{bs},0}\right\}, \\
&= \mathsf{Pr}\left\{\gamma_{\mathrm{bs}} < \gamma_{\mathrm{bs}}^{\star} \mid d_{\mathrm{bs},0}\right\}. \qquad (4.2)
\end{aligned}$$

The signal-to-interference-ratio (SIR) at the receiver for the link of interest is given by

$$\gamma_{\text{bs}} = \frac{P_{\text{bs}} h_{\text{bs},0} d_{\text{bs},0}^{-\alpha}}{\sum_{j \in \Phi'_{\text{bs}}} P_{\text{bs}} h_{\text{bs},j} d_{\text{bs},j}^{-\alpha}}, \tag{4.3}$$

where Φ'_{bs} represents the point process governing the locations of the interfering RRHs. Evaluating (4.2) is well studied in the literature [32], and the DVP conditioned on $d_{\text{bs},0}$ is given by

$$\mathsf{Pr}\left\{D > D_{\max} \mid d_{\text{bs},0}\right\} = 1 - \exp\left(\frac{-\pi\lambda_{\text{bs}}(\gamma^{\star}_{\text{bs}})^{\frac{2}{\alpha}} d_{\text{bs},0}^2}{\text{sinc}\left(\frac{2}{\alpha}\right)}\right). \tag{4.4}$$

Then,

$$\text{T}_{\text{bs}}(D_{\max}) = \int_0^{\infty} \mathsf{Pr}\left\{D > D_{\max} \mid r\right\} f_{d_{\text{bs},0}}(r) dr, \tag{4.5}$$

where probability density function (pdf) of the distance to nearest RRH is given by $f_{d_{\text{bs},0}}(r) = 2\pi\lambda_{\text{bs}} r \exp\left(-\pi\lambda_{\text{bs}} r^2\right)$.

$$\begin{aligned}
\text{T}_{\text{bs}}(D_{\max}) &= \int_0^{\infty} \mathsf{Pr}\left\{D > D_{\max} \mid r\right\} 2\pi\lambda_{\text{bs}} r \exp\left(-\pi\lambda_{\text{bs}} r^2\right) dr, \\
&= \int_0^{\infty} \left(1 - \exp\left(\frac{-\pi\lambda_{\text{bs}}(\gamma^{\star}_{\text{bs}})^{\frac{2}{\alpha}} r^2}{\text{sinc}\left(\frac{2}{\alpha}\right)}\right)\right) 2\pi\lambda_{\text{bs}} r \exp\left(-\pi\lambda_{\text{bs}} r^2\right) dr, \\
&= 1 - \frac{sinc(\frac{2}{\alpha})}{(\gamma^{\star}_{\text{bs}})^{\frac{2}{\alpha}} + sinc(\frac{2}{\alpha})}, \\
&= \frac{(\gamma^{\star}_{\text{bs}})^{\frac{2}{\alpha}}}{(\gamma^{\star}_{\text{bs}})^{\frac{2}{\alpha}} + \text{sinc}\left(\frac{2}{\alpha}\right)}
\end{aligned} \tag{4.6}$$

□

Next, $\text{T}_{\text{bs}}(D_{\max})$ is compared with the two (outband and overlay) DVP values achieved in the D2D mode, assuming that the DP with the requested content is located at a distance equal to the threshold distance. Note that (4.4) can be used to make this comparison. However, this leads to decision thresholds which are functions of $d_{\text{bs},0}$ as well. Physically, this means each DC has its own decision thresholds, that depend on its distance from the RRH. This makes it prohibitively hard for us to obtain the intensities of the point processes governing the locations of the overlay and outband DPs, *i.e.*, λ_{ou} and λ_{ol}, respectively, which we require to calculate the DVP values in the D2D mode. Therefore, we have averaged out the effect of $d_{\text{bs},0}$, and obtain a threshold valid for the entire network. With the idea of this common threshold, next we derive the λ_{ou} and λ_{ol}.

To this end, we thin PPP Φ_{dc} into three point processes to represent DCs served by RRHs, by a DP as outband and by a DP as an overlay. Moreover, we assume that outband DCs and

the overlay DCs form homogeneous PPPs Φ_{ou} and Φ_{ol}, respectively. Since the segmentation of DCs into outband and overlay depends on the distance between DCs and DPs, the thinning of Φ_{dc} will not result in homogeneous PPPs. However, similar approximations are used in [5,14,15,25] with sufficient accuracy. With this assumption, we formally obtain expressions for λ_{ou} and λ_{ol} through the following lemma.

4.1.2 Outband and overlay user intensities

Lemma 2. *The intensities of the two point processes* Φ_{ou} *and* Φ_{ol} *are given by* $\lambda_{\text{ou}} = \lambda_{\text{dc}}\left[1 - e^{-p\pi\lambda_{\text{dp}}(d^{\star}_{\text{ou}})^2}\right]$ *and* $\lambda_{\text{ol}} = \lambda_{\text{dc}}\left[e^{-p\pi\lambda_{\text{dp}}(d^{\star}_{\text{ou}})^2} - e^{-p\pi\lambda_{\text{dp}}\left(d^{\star}_{\text{ol}}\right)^2}\right]$ *respectively.*

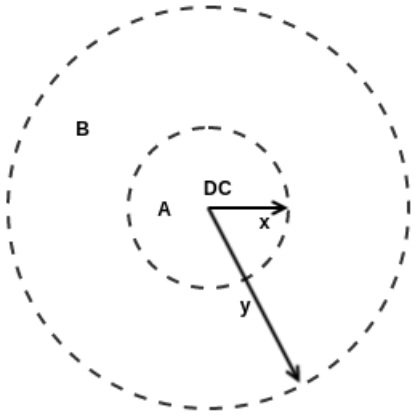

Figure 4.1: Communication Mode Selection

Proof. According to the Figure 4.1, the region A is used for outband communication and B is used for overlay communication. Moreover, x indicates the outband communication distance and it satisfies $x \leq d^{\star}_{\text{ou}}$ and y indicates the total communication range of DC to DP and it should satisfies $y \leq d^{\star}_{\text{ol}}$. The probability of existence of a DP having the requested file within the distance of $d^{\star}_{\text{ou}}$ from the DC is given by $\left[1 - e^{-p\pi\lambda_{\text{dp}}(d^{\star}_{\text{ou}})^2}\right]$ using the null probability of PPP Φ_{dp}, where we have assumed the intensity of the DPs containing the requested file is $p\lambda_{\text{dp}}$. The edge cloud randomly selects a DP within the distance of $d^{\star}_{\text{ou}}$ from the DC, which will transmit in outband to the DC. Hence, multiplying the probability by λ_{dc} gives the intensity of DPs, who are eligible to transmit in outband. Assuming a one to one mapping of DCs to DPs, this intensity is equal to λ_{ou}.

Similarly, the probability of existence of a DP having the requested content between the distance of $d^{\star}_{\text{ou}}$ and $d^{\star}_{\text{ol}}$ is given by $\left[1 - e^{-p\pi\lambda_{\text{dp}}(d^{\star}_{\text{ol}})^2}\right] - \left[1 - e^{-p\pi\lambda_{\text{dp}}(d^{\star}_{\text{ou}})^2}\right]$. The randomly selected DP will transmit to the DC in the overlay band, and hence, multiplying this probability by λ_{dc} gives λ_{ol}. □

Next, we present the DVPs of the links between the DC and the serving DP (k^{th}) in outband, as well as the serving DP in the overlay band.

4.1.3 Delay violation probabilities of outband and overlay communication

Lemma 3. *The DVPs of the link between the DC and the k^{th} outband DP, and the k^{th} overlay DP are given by* $\mathsf{Pr}\{D_{\mathrm{ou},k} > D_{\max} \mid d_{\mathrm{ou},k}\} = \mathrm{T}_{\mathrm{ou},k}(D_{\max}, d_{\mathrm{ou},k})$ *and* $\mathsf{Pr}\{D_{\mathrm{ol},k} > D_{\max} \mid d_{\mathrm{ol},k}\} = \mathrm{T}_{\mathrm{ol},k}(D_{\max}, d_{\mathrm{ol},k})$, *respectively, where*

$$\mathrm{T}_{\mathrm{ou},k}(D_{\max}, d_{\mathrm{ou},k}) = 1 - \exp\left(\frac{-\pi\left[\lambda_{\mathrm{ou}} E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) + P_{\mathrm{ext}}^{\frac{2}{\alpha}}\lambda_{\mathrm{ext}}\right](\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}} d_{\mathrm{ou},k}^{2}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\mathrm{ou},k}^{\frac{2}{\alpha}}}\right) \tag{4.7}$$

and

$$\mathrm{T}_{\mathrm{ol},k}\left(D_{\max}, d_{\mathrm{ol},k}\right) = 1 - \exp\left(\frac{-\pi\lambda_{\mathrm{ol}} E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right)(\gamma_{\mathrm{ol}}^{\star})^{\frac{2}{\alpha}} d_{\mathrm{ol},k}^{2}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\mathrm{ol},k}^{\frac{2}{\alpha}}}\right). \tag{4.8}$$

Proof. Following the proof of the Lemma 1, We have

$$\mathrm{T}_{\mathrm{ou},k}\left(D_{\max}, d_{\mathrm{ou},k}\right) = \mathsf{Pr}\left\{\gamma_{\mathrm{ou},k} < \gamma_{\mathrm{ou}}^{\star} \mid d_{\mathrm{ou},k}\right\}, \tag{4.9}$$

$$\mathrm{T}_{\mathrm{ol},k}\left(D_{\max}, d_{\mathrm{ol},k}\right) = \mathsf{Pr}\left\{\gamma_{\mathrm{ol},k} < \gamma_{\mathrm{ol}}^{\star} \mid d_{\mathrm{ol},k}\right\}, \tag{4.10}$$

where $\gamma_{\mathrm{ou}}^{\star} = \frac{M}{2^{B_{\mathrm{ou},k}(D_{\max}-c)}} - 1$ and $\gamma_{\mathrm{ol}}^{\star} = \frac{M}{2^{B_{\mathrm{ol},k}(D_{\max}-c)}} - 1$. To evaluate the (4.9) and (4.10), we derive the SIRs of the links as follows

$$\gamma_{\mathrm{ou},k} = \frac{P_{\mathrm{ou},k} h_{\mathrm{ou},k} d_{\mathrm{ou},k}^{-\alpha}}{\sum_{j\in\Phi'_{\mathrm{ou}}} P_{\mathrm{ou},j} h_{\mathrm{ou},j} d_{\mathrm{ou},j}^{-\alpha} + \sum_{j\in\Phi_{\mathrm{ext}}} P_{\mathrm{ext},j} h_{\mathrm{ext},j} d_{\mathrm{ext},j}^{-\alpha}}, \tag{4.11}$$

$$\gamma_{\mathrm{ol},k} = \frac{P_{\mathrm{ol},k} h_{\mathrm{ol},k} d_{\mathrm{ol},k}^{-\alpha}}{\sum_{j\in\Phi'_{\mathrm{ol}}} P_{\mathrm{ol},j} h_{\mathrm{ol},j} d_{\mathrm{ol},j}^{-\alpha}}, \tag{4.12}$$

Where Φ'_{ou} and Φ'_{ol} are the PPPs governing the locations of interfering transmitters in the Φ_{ou} and Φ_{ol}. Evaluating (4.9) and (4.10) using (4.11) and (4.12) as in the proof of Lemma 1 concludes the proof. □

Using the DP intensities in each band, the distance thresholds of the outband overlay networks can be evaluated by considering the DVP values in each band, and we formally state this in the following Lemma.

4.1.4 Outband and overlay distance threshold calculation

Lemma 4. *The outband and overlay distance thresholds are given by*

$$d_{\mathrm{ou}}^{\star} = \left(\frac{\sqrt{B^2 + 4AC} - B}{2A} \right)^{\frac{1}{2}} \tag{4.13}$$

and

$$d_{\mathrm{ol}}^{\star} = \left(\frac{\sqrt{E^2 + 4DC} - E}{2D} \right)^{\frac{1}{2}}, \tag{4.14}$$

where $A = \frac{\pi^2 p (\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}} \lambda_{\mathrm{dp}} \lambda_{\mathrm{dc}} E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right)}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}}$, $B = \frac{\pi (\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}} P_{\mathrm{ext}}^{\frac{2}{\alpha}} \lambda_{\mathrm{ext}}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}}$, $C = \mid \ln\left(1 - \mathrm{T}_{\mathrm{bs}}\left(D_{\max}\right)\right) \mid$, $D = \frac{\pi^2 p \lambda_{\mathrm{dc}} \lambda_{\mathrm{dp}} E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right) (\gamma_{\mathrm{ol}}^{\star})^{\frac{2}{\alpha}}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}}$ *and* $E = \frac{\pi p \lambda_{\mathrm{dc}} E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right) (\gamma_{\mathrm{ol}}^{\star})^{\frac{2}{\alpha}} \left[e^{-\pi \lambda_{\mathrm{dp}} (d_{\mathrm{ou}}^{\star})^2} - 1\right]}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}}$.

Proof. We assume selected DP is at the distance threshold $d_{\mathrm{ou}}^{\star}$ from the DC. Using maximum transmit power, it should satisfy the following condition

$$\begin{aligned} \mathrm{T}_{\mathrm{ou},k}\left(D_{\max}, d_{\mathrm{ou}}^{\star}\right) &\leq T_{\mathrm{bs}}(D_{\max}), \\ \exp\left(\frac{-\pi \left[\lambda_{\mathrm{ou}} E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) + P_{\mathrm{ext}}^{\frac{2}{\alpha}} \lambda_{\mathrm{ext}}\right] (\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}} (d_{\mathrm{ou}}^{\star})^2}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}} \right) &\geq \left(1 - T_{\mathrm{bs}}(D_{\max})\right), \\ \frac{\pi \left[\lambda_{\mathrm{ou}} E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) + P_{\mathrm{ext}}^{\frac{2}{\alpha}} \lambda_{\mathrm{ext}}\right] (\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}} (d_{\mathrm{ou}}^{\star})^2}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}} &\leq \mid \ln\left(1 - T_{\mathrm{bs}}(D_{\max})\right) \mid \end{aligned} \tag{4.15}$$

to be eligible for an outband D2D link.

Substituting λ_{ou} and using first order Taylor series approximation $e^{-ax} = 1 - ax$, (4.15) can be simplified to

$$\frac{\pi \left[\pi p \lambda_{\mathrm{dp}} \lambda_{\mathrm{dc}} (d_{\mathrm{ou}}^{\star})^2 E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) + P_{\mathrm{ext}}^{\frac{2}{\alpha}} \lambda_{\mathrm{ext}}\right] (\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}} (d_{\mathrm{ou}}^{\star})^2}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\max}^{\frac{2}{\alpha}}} \leq \mid \ln\left(1 - \mathrm{T}_{\mathrm{bs}}\left(D_{\max}\right)\right) \mid . \tag{4.16}$$

Solving (4.16), we obtain $d_{\mathrm{ou}}^{\star}$. Same procedure can be used to obtain an expression for $d_{\mathrm{ol}}^{\star}$. □

Clearly, $d_{\mathrm{ou}}^{\star}$ and $d_{\mathrm{ol}}^{\star}$ depend on $E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right)$ and $E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right)$, respectively. Obtaining analytical expressions for these expectations appears to be intractable since the probability distribution of the transmit powers of the DPs is not known. Therefore, assuming worst case conditions, the interferes are allowed to transmit at their maximum power, hence $E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) = P_{\max}^{\frac{2}{\alpha}}$

and $E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right) = P_{\max}^{\frac{2}{\alpha}}$. This simplifies $d_{\mathrm{ou}}^{\star}$ and $d_{\mathrm{ol}}^{\star}$ such that $A = \frac{\pi^2 p(\gamma_{\mathrm{ou}}^{\star})^{\frac{2}{\alpha}}\lambda_{\mathrm{dp}}\lambda_{\mathrm{dc}}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}$,$D = \frac{\pi^2 p(\gamma_{\mathrm{ol}}^{\star})^{\frac{2}{\alpha}}\lambda_{\mathrm{dp}}\lambda_{\mathrm{dc}}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}$, and $E = \frac{\pi p\lambda_{\mathrm{dc}}\left(\gamma_{\mathrm{ol}}^{\star}\right)^{\frac{2}{\alpha}}\left[e^{-\pi\lambda_{\mathrm{dp}}\left(d_{\mathrm{ou}}^{\star}\right)^2}-1\right]}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}$, which provides a lower bounds for $d_{\mathrm{ou}}^{\star}$ and $d_{\mathrm{ol}}^{\star}$. Thresholds can be further refined in a system setting using an iterative computation scheme. Initially, the distance thresholds and the transmit power of each DP are calculated under the worst case conditions. In the next iteration, $E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right)$ and $E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right)$ are evaluated using the transmit powers of the previous iteration, and the distance thresholds and the transmit power of each DP are recalculated. This procedure is repeated until the distance thresholds are converged to a fixed value.

According to analytical results, one can see that when $d_{\mathrm{ou}}^{\star} \to 0$, all DPs will be allocated to overlay band. Since reducing the outband threshold will allocate more DPs into the overlay network, the interference in the overlay band will increase. Therefore, when $d_{\mathrm{ou}}^{\star} \to 0$, $d_{\mathrm{ol}}^{\star}$ also decays exponentially. Furthermore, when $d_{\mathrm{ou}}^{\star}$ increases, $d_{\mathrm{ol}}^{\star}$ also increases. When the outband region expands more users are allocated to the outband. Hence, the interference in the overlay region will be reduced, providing more communication opportunities in the overlay band.

4.2 Transmit Power Computation

Next, we calculate the parameter P in Algorithm 1, the required minimum power of each DP to transmit data. Assume DC requests a file from the edge cloud, the edge cloud selects the k^{th} DP with the requested file to serve the DC. We first decide the operating band of the DP by comparing the link length with the distance thresholds. Next, the required minimum powers of each DP can be computed such that DVP with a D2D link is at most equal to the DVP of content delivery through an RRH.

The following Lemma formally states the required minimum power for a selected DP in each band to achieve a DVP equal to the cellular mode.

4.2.1 Minimum transmit power calculation

Lemma 5. *The minimum transmit power of the k^{th} DP allocated to the outband network or the overlay network can be given as*

$$P'_{\mathrm{ou},k} = \left[\frac{\lambda_{\mathrm{ou}}E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) + P_{\mathrm{ext}}^{\frac{2}{\alpha}}\lambda_{\mathrm{ext}}}{\lambda_{\mathrm{bs}}}\right]^{\frac{\alpha}{2}} \left(\frac{\gamma_{\mathrm{ou}}^{\star}}{\gamma_{\mathrm{bs}}^{\star}}\right)\left(\frac{d_{\mathrm{ou},k}}{d_{\mathrm{bs},0}}\right)^{\alpha} \tag{4.17a}$$

and

$$P'_{\mathrm{ol},k} = \left[\frac{\lambda_{\mathrm{ol}} E\left(P_{\mathrm{ol},j}^{\frac{2}{\alpha}}\right)}{\lambda_{\mathrm{bs}}}\right]^{\frac{\alpha}{2}} \left(\frac{\gamma^{\star}_{\mathrm{ol}}}{\gamma^{\star}_{\mathrm{bs}}}\right) \left(\frac{d_{\mathrm{ol},k}}{d_{\mathrm{bs},0}}\right)^{\alpha}, \tag{4.17b}$$

respectively.

Proof. First, we consider an outband DP. We have

$$\begin{aligned} T_{\mathrm{ou},k}\left(D_{\max}, d_{\mathrm{ou},k}\right) &\leq T_{\mathrm{bs}}\left(D_{\max}, d_{\mathrm{bs},0}\right) \\ \exp\left(\frac{-\pi\left[\lambda_{\mathrm{ou}} E\left(P_{\mathrm{ou},j}^{\frac{2}{\alpha}}\right) + P_{\mathrm{ext}}^{\frac{2}{\alpha}}\lambda_{\mathrm{ext}}\right]\left(\gamma^{\star}_{\mathrm{ou}}\right)^{\frac{2}{\alpha}} d_{\mathrm{ou},k}^{2}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right) P_{\mathrm{ou},k}^{\frac{2}{\alpha}}}\right) &\geq \\ \exp\left(\frac{-\pi\lambda_{\mathrm{bs}}(\gamma^{\star}_{\mathrm{bs}})^{\frac{2}{\alpha}} d_{\mathrm{bs},0}^{2}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}\right). \end{aligned} \tag{4.18}$$

Solving (4.18), we obtain the $P_{\mathrm{ou},k}$ as in Lemma 5. Similarly, an expression for $P_{\mathrm{ol},k}$can be obtained. □

One can observe that $P_{\mathrm{ou},k}$ and $P_{\mathrm{ol},k}$ depend on the ratio $\frac{d_{\mathrm{ou},k}}{d_{\mathrm{bs},0}}$ and the mean interference power of in band.

The distance thresholds find the feasible set of outband DPs and the overlay DPs. However, since we used an averaged DVP for cellular mode, all DPs in the feasible set may not be able to satisfy the maximum transmit power constraint for individual links. Therefore, the DPs in feasible set are individually checked for maximum power constraint violation. The DCs with selected DPs who are not capable of satisfying the power constraint are served using their connected RRHs. The DP intensities in outband and in the overlay band after this refining are given in the following Lemma.

4.2.2 Fine tuning intensities of overlay and outband users

Lemma 6. *Refined intensities of the outband and the overlay band are given by*

$$\lambda_{\mathrm{ou}}^{th} = \left[\frac{12\lambda_{\mathrm{ou}}}{\pi^3\lambda_{\mathrm{bs}}^3\left(d^{\star}_{\mathrm{ou}}\right)^2}\right]\left(\frac{P_{\max}}{\beta}\right)^{\frac{2}{\alpha}} \tag{4.19}$$

and

$$\lambda_{\mathrm{ol}}^{th} = \frac{\lambda_{\mathrm{ol}}}{\left(\left(d^{\star}_{\mathrm{ol}}\right)^2 - \left(d^{\star}_{\mathrm{ou}}\right)^2\right)}\left[\frac{12}{\pi^3\lambda_{\mathrm{bs}}^3}\left(\frac{P_{\max}}{\beta}\right)^{\frac{2}{\alpha}} - \left(d^{\star}_{\mathrm{ou}}\right)^2\right], \tag{4.20}$$

respectively, where $\beta = \left[\frac{\left(\lambda_{\mathrm{ou}} P_{\max}^{\frac{2}{\alpha}} + P_{\mathrm{ext}}^{\frac{2}{\alpha}}\lambda_{\mathrm{ext}}\right)}{\lambda_{\mathrm{bs}}}\right]^{\frac{\alpha}{2}}\left(\frac{\gamma^{\star}_{\mathrm{ou}}}{\gamma^{\star}_{\mathrm{bs}}}\right)$ *and* $\eta = \left[\frac{\lambda_{\mathrm{ol}} P_{\max}^{\frac{2}{\alpha}}}{\lambda_{\mathrm{bs}}}\right]^{\frac{\alpha}{2}}\left(\frac{\gamma^{\star}_{\mathrm{ol}}}{\gamma^{\star}_{\mathrm{bs}}}\right)$.

Proof. Considering the outband communication,

$$\begin{aligned}
\lambda_{\text{ou}}^{th} &= \lambda_{\text{ou}} \mathsf{Pr}\left\{P_{\text{ou},k} \leq P_{\max}\right\}, \\
&= \int_0^{\infty} \lambda_{\text{ou}} \mathsf{Pr}\left\{P_{\text{ou},k} \leq P_{\max} \mid r\right\} f_{d_{\text{bs},0}}(r) dr, \\
&= \int_0^{\infty} \lambda_{\text{ou}} \mathsf{Pr}\left\{\beta \left(\frac{d_{\text{ou},k}}{r}\right)^{\alpha} \leq P_{\max}\right\} f_{d_{\text{bs},0}}(r) dr, \\
&= \int_0^{\infty} \lambda_{\text{ou}} \mathsf{Pr}\left\{d_{\text{ou},k} \leq \left(\frac{P_{\max}}{\beta}\right)^{\frac{1}{\alpha}} r\right\} f_{d_{\text{bs},0}}(r) dr, \\
&= \int_0^{\infty} \lambda_{\text{ou}} \left[\frac{\left(\frac{P_{\max}}{\beta}\right)^{\frac{2}{\alpha}} r^2}{\left(d_{\text{ou}}^{\star}\right)^2}\right] 2\pi\lambda_{\text{bs}} r exp\left(-\pi\lambda_{\text{bs}} r^2\right) dr. \qquad (4.21)
\end{aligned}$$

Evaluating (4.21), we obtain the λ_{ou}^{th}. Following a similar approach and $\mathsf{Pr}\left\{d_{\text{ol},k} \leq r\right\} = \frac{r^2 - (d_{\text{ou}}^{\star})^2}{\left(d_{\text{ol}}^{\star}\right)^2 - (d_{\text{ou}}^{\star})^2}$, one can obtain the expression for λ_{ol}^{th}. □

We have assumed maximum interference in each band when calculating the exact intensities.

Chapter 5

Simulation Results and Discussion

In this section, we provide numerical and simulation results to validate our assumptions and to identify the benefits of our proposed algorithm. The parameters used in the simulations are given in Table II. Note that the simulation results are obtained by relaxing the assumptions used in the analysis.

Table 5.1: Simulation Parameters

Parameter	Value
RRH power (P_{bs})	$100mW$
Maximum power of an end device ($P_{\max}$)	$2.5mW$
Power of an external user (P_{ext})	$2mW$
Radius of the simulated area (R)	$3000m$
DP intensity(λ_{dp})	10^{-4}
DC intensity(λ_{dc})	10^{-3}
EU intensity (λ_{ext})	$10^{-3.5}$
RRH intensity (λ_{bs})	$10^{-5.5}$
Path loss exponent (α)	3.5
File size (M)	$80kB$
Channel bandwith ($B_{\text{bs}}, B_{\text{ou}}, B_{\text{ol}}$)	$5MHz$
Application level delay threshold ($D_{\max}$)	$0.5ms$
Processing delay (c)	$0.1ms$

5.1 Validation of approximations

We first validate the assumption Φ_{ou} and Φ_{ol} are homogeneous PPPs. For this we use DPs and DCs that are spatially distributed following homogeneous PPPs. Then the DPs and DPs randomly paired and based on their link lengths. We split them into outband and overlay using a threshold distance d. The coverage probability of a typical DC in each band is evaluated using simulation and compare with the theoretical coverage probability

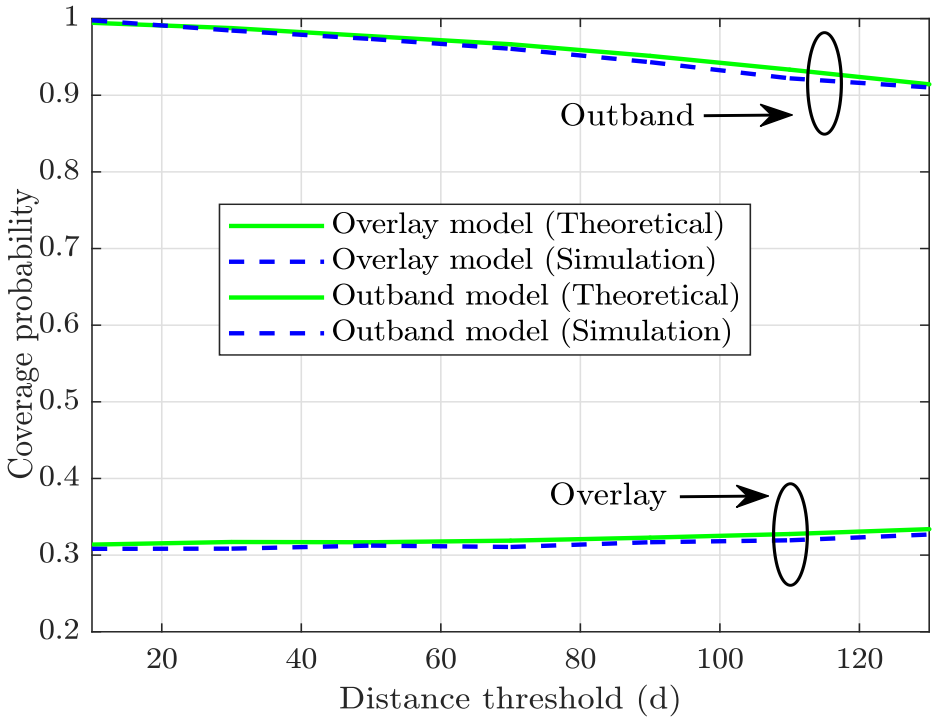

Figure 5.1: Validation of the independent thinning approximation

obtained by assuming that Φ_{ou} and Φ_{ol} are homogeneous PPP. Fig. 5.1 demonstrates that the simulation results closely match with theoretical results, validating our approximation.

Figure 5.2 shows variation of the D2D link intensity in each band with λ_{ext}. The subscripts *ou* and *ol* are used to denote outband and overlay, respectively. The superscripts *ap*, *th*, *al*, and *f* are used to denote maximum interference approximation, theoretical value, iterative optimization and monolithic (fully overlay or underlay) schemes, respectively. Increasing λ_{ext} will result in a reduction in D2D links in both outband and overlay networks. The reduction rate is faster for the outband network. As λ_{ext} increases, the threshold distance $d^{\star}_{\text{ou}}$ is reduced such that intensity of the network does not violates the required QoS of the network, allowing less D2D users in outband. Moreover, this allocates more users to the overlay network, resulting in higher interference. Therefore, $d^{\star}_{\text{ol}}$ is also reduced with a slower rate compared to $d^{\star}_{\text{ou}}$, to maintain QoS. Furthermore, one can observe that iterative optimization provides more D2D communication opportunities compared to our approximate solution. However, it requires higher computational time. Therefore, the user allocation scheme can be selected based on the available resources. The theoretical D2D intensities closely follow the results obtained through simulation. Also it can be seen D2D opportunities have increased 4-5 times with the hybrid model compared to pure overlay or outband D2D networks.

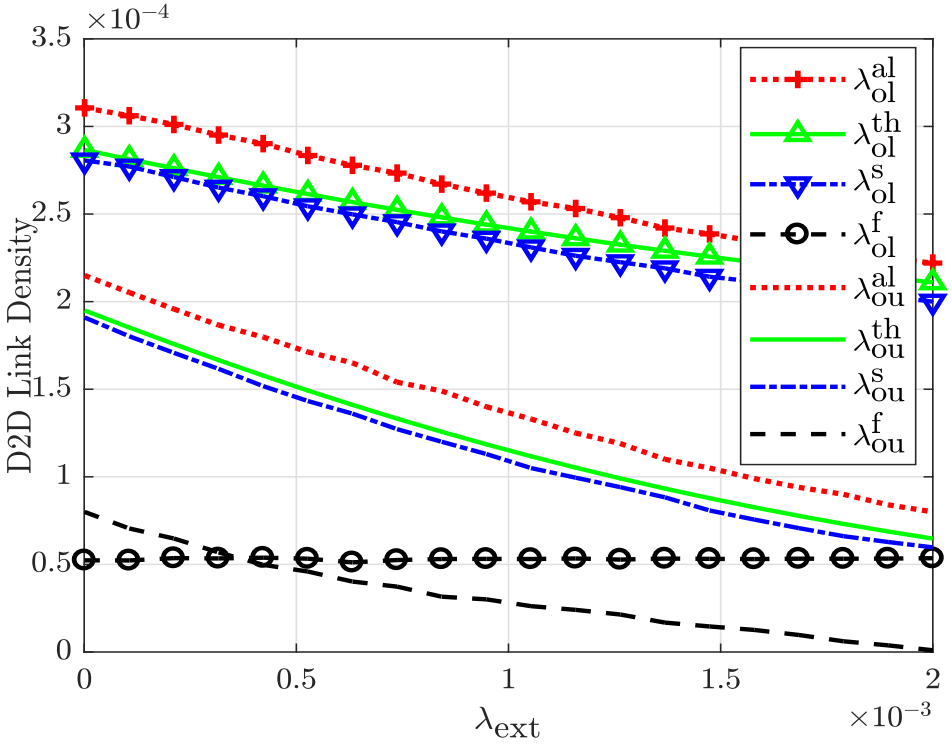

Figure 5.2: Intensity of each D2D network against the external user intensity

5.2 Variation of D2D user intensities based on external interference

Figs. 5.3 presents the user intensities in each band with varying DP intensities. At first, increasing λ_{dp} results in a linear increase in D2D intensities in each band. As λ_{dp} is further increased, the user intensities begin to saturate. The saturation occurs mainly due to the increment of the interference power in each band such that additional DC-DP pairs will not satisfy the QoS requirement with D2D communications. Initially, overlay intensity is higher than the outband intensity since the sparse network in low λ_{dp} regime results in a low probability of finding close proximity DP-DC pairs. Therefore, more D2D links are eligible for the overlay network with lower λ_{dp}. However, increasing λ_{dp} will results in a dense network, where the probability of finding DC-DP pairs with shorter communication distance is higher. Therefore, the number of links satisfying the outband threshold will be higher and the outband intensity overtakes the overlay intensity beyond a certain value of λ_{dp}.

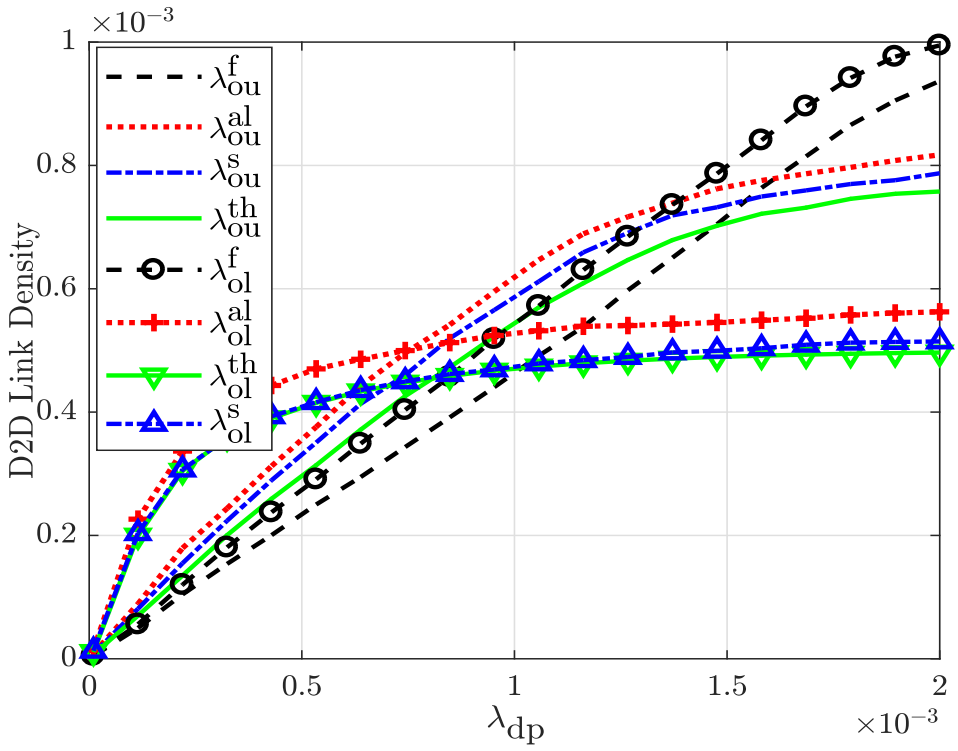

Figure 5.3: Intensity of each D2D network against the DP intensity

5.3 Analysis of power consumption of data consumers based on data producer intensity

Fig. 5.4 compares the average power consumption of a user in the hybrid network, fully overlay network and the fully outband network under three different λ_{ext} values. As expected, increasing λ_{ext} increases the power consumption of the outband networks since higher transmit power is required to maintain the QoS. Also, the power consumption of the fully overlay network is unaffected by λ_{ext} as expected. One can see that the hybrid network saves nearly 50% of the power compared to monolithic networks, indicating the energy efficiency of our proposed model. Again, it can be seen that the iterative optimization results in lower power consumption at the devices. However, it may result in higher power consumption at the infrastructure nodes due to the increased complexity.

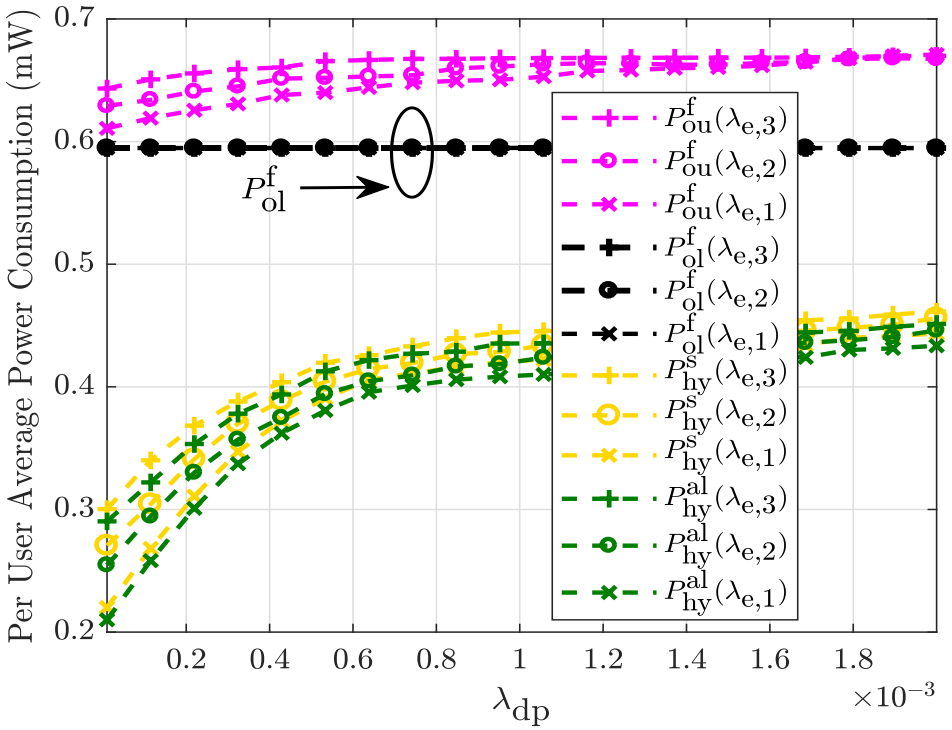

Figure 5.4: Power consumption of the D2D network against the DP intensity

Chapter 6

Applications and Future Work

6.1 Proximity based video streaming

Our proposed scheme placates D2D networks to satisfy the required QoS, while serving content requests of proximity data consumers. Content streaming applications are becoming popular and they will be heavily used in 5G networks. The proposed solution can assist such applications in different aspects. Firstly, it can be used to formulate D2D clusters such that D2D communication guarantees required QoS. Hence, uninterrupted video streaming can be supported. The limited battery power of mobile devices affects the achievable QoS, resulting D2D communications unpopular in telecommunication industry. To this end, our solution allocates D2D users such that the average power consumption of the network is minimized. This will contribute towards significant power savings at mobile stations.

In addition, the proposed scheme provides economic benefits for the telecommunication industry by supporting D2D networks. Since this is a semi-autonomous network, the legacy cellular network is responsible for the management of the D2D network. Therefore, the operators can charge for D2D services. Moreover, data producers that cache content on their mobile terminals can get lower priced service packages from the operators, since they contribute to better QoS in the network. The network operators can offload storage requirements to users and users can earn economic benefits throughout that.

6.2 Intelligent content filtering

CRANS inherently utilizes cloud computation infrastructures to run computationally-intensive algorithms. Hence, we can use novel machine learning and artificial intelligence techniques to support data caching strategies. For instance, for communities such as universities, schools are well suited to deploy D2D networks, because those communities have common interests in educational materials. Hence, AI technologies can be used to proactively cache content

among data producers of such communities to deliver content. Moreover, per-user based interest tracking and caching strategies can be easily introduced with the help of cloud technologies.

6.3 Future work

The key outcome of our research is developing D2D network formation algorithms based on QoS and power constraints on top of CRAN architecture. This model was developed based on several assumptions that may not fully characterize a practical networks. Therefore, there will be a gap between theoretical estimates that has been articulated throughout this research and practical applications of it. To this end, we need to further investigate applications of the proposed scheme and identify the shortfalls and possible solutions. Furthermore, our work creates a new pitch for interconnecting application layer technologies such as caching techniques, AI, and ML models with the network layer. We can introduce our solution as a cross-layer optimization for D2D networks. Hence, we need to further research on how to utilize CRANs computational capacity to further optimize D2D networks.

Chapter 7

Conclusion

We have proposed a spectrum selection and transmit power minimization scheme for a D2D network cross-laid with a CRAN, where D2D communications are allowed as an overlay to the CRAN and in the ISM band. Analytical approximations were derived to calculate spectrum selection thresholds and the required minimum transmit power to achieve an assured QoS level. Moreover, theoretical approximations were derived for the D2D user intensity in each band, which can be used to derive key performance metrics such as coverage probability and transmission capacity. The proposed scheme achieves nearly 50% power savings compared to a monolithic D2D network, where D2D communication occurs only at overlay to the network or at the ISM band.

Moreover, the work presented in this thesis can be used as a starting point to future work on energy efficient D2D models. Our research has reduced the application layer and network layer complexities. This research motivates us to explore more on CRAN architecture and utilize advanced IT infrastructure and capabilities to solve telecommunication related bottlenecks and issues. Furthermore, the presented work proposes revenue generation methods not only for the network operators, but for the subscribers.

Bibliography

[1] A. Kliks and N. Dimitriou and A. Zalonis and O. Holland. WiFi traffic offloading for energy saving. In *ICT 2013*, pages 1–5, May 2013.

[2] A. Abdallah, M. M. Mansour, and A. Chehab. A distance-based power control scheme for D2D communications using stochastic geometry. In *Proc. IEEE Vehicular Technology Conference*, Sep. 2017.

[3] E. Ben Abdelkrim, M. A. Salahuddin, H. Elbiaze, and R. Glitho. A hybrid regression model for video popularity-based cache replacement in content delivery networks. In *IEEE Global Communications Conference*, pages 1–7, Dec 2016.

[4] R. M. Abuteir, A. Fladenmuller, and O. Fourmaur. Sdn based architecture to improve video streaming in home networks. In *IEEE 30th International Conference on Advanced Information Networking and Applications*, pages 220–226, March 2016.

[5] K. S. Ali, H. ElSawy, and M. Alouini. Modeling cellular networks with full-duplex D2D communication: A stochastic geometry approach. *IEEE Trans. Commun.*, 64(10):4409–4424, Oct 2016.

[6] R. I. Ansari, C. Chrysostomou, S. A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, and J. J. P. C. Rodrigues. 5g d2d networks: Techniques, challenges, and future prospects. *IEEE Systems Journal*, 12(4):3970–3984, Dec 2018.

[7] A. Asadi, Q. Wang, and V. Mancuso. A survey on device-to-device communication in cellular networks. *Communications Surveys and Tutorials,IEEE*, 16(4):1801–1819, Fourthquarter. 2014.

[8] R. Atat, L. Liu, N. Mastronarde, and Y. Yi. Energy harvesting-based d2d-assisted machine-type communications. *IEEE Transactions on Communications*, 65(3):1289–1302, March 2017.

[9] F. Baccelli and B. Blaszczyszyn. *Stochastic Geometry and Wireless Networks, Vol.1. Delft, the Netherlands.* NOW, 2010.

[10] 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. *Communications Surveys and Tutorials,IEEE*, 17(1):405–426, Firstquarter. 2015.

[11] H. Chen, L. Liu, H. S. Dhillon, and Y. Yi. QoS-aware D2D cellular networks with spatial spectrum sensing: A stochastic geometry view. *IEEE Trans. Commun.*, 67(5):3651–3664, May 2019.

[12] D. Suh and H. Ko and S. Pack. Efficiency Analysis of WiFi Offloading Techniques. *IEEE Trans. Veh. Technol.*, 65(5):3813–3817, May 2016.

[13] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M. Alouini. Coordinated scheduling and power control in cloud-radio access networks. *IEEE Transactions on Wireless Communications*, 15(4):2523–2536, April 2016.

[14] H. ElSawy and E. Hossain. On stochastic geometry modeling of cellular uplink transmission with truncated channel inversion power control. *IEEE Trans. Wireless Commun.*, 13(8):4454–4469, Aug 2014.

[15] H. ElSawy, E. Hossain, and M. Alouini. Analytical modeling of mode selection and power control for underlay D2D communication in cellular networks. *IEEE Trans. Commun.*, 62(11):4147–4161, Nov 2014.

[16] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis. Distributed wireless communication system: A new architecture for future public wireless access. *IEEE Commun. Mag.*, 41(3):108–113, Mar. 2003.

[17] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis. Wireless network cloud: Architecture and system requirements. *IBM Journal of Research and Development*, 54(1):4:1–4:12, Jan. 2010.

[18] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis. D2d-aware device caching in mmwave-cellular networks. *IEEE J. Sel. Areas Commun.*, 35(9):2025–2037, Sep. 2017.

[19] M. Haenggi. *Stochastic Geometry for Wireless Networks*. Cambridge University Press, 2013.

[20] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed. A survey of device-to-device communications: Research issues and challenges. *IEEE Communications Surveys Tutorials*, 20(3):2133–2168, thirdquarter 2018.

[21] K. Lee and J. Lee and Y. Yi and I. Rhee and S. Chong. Mobile Data Offloading: How Much Can WiFi Deliver? *IEEE/ACM Transactions on Networking*, 21(2):536–550, April 2013.

[22] K. Katsalis, N. Nikaein, E. Schiller, R. Favraud, and T. I. Braun. 5g architectural design patterns. In *2016 IEEE International Conference on Communications Workshops*, pages 32–37, May 2016.

[23] F. R. Kschischang, B. J. Frey, and H. . Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, Feb 2001.

[24] P. Kumar, C. M. Shamrao, and M. Mukherjee. Sum-rate maximization with joint power allocation and mode selection in D2D-enabled 5G cellular networks. In *IEEE International Conference on Communications Workshops*, pages 1–6, May 2019.

[25] X. Lin, J. G. Andrews, and A. Ghosh. Spectrum sharing for device-to-device communication in cellular networks. 13(12):6727–6740, Dec 2014.

[26] J. Liu, B. Bai, J. Zhang, and K. B. Letaief. Content caching at the wireless network edge: A distributed algorithm via belief propagation. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2016.

[27] M. A. Abana and M. Peng and Z. Zhao and L. A. Olawoyin. Coverage and Rate Analysis in Heterogeneous Cloud Radio Access Networks With Device-to-Device Communication. *IEEE Access*, 4:2357–2370, May 2016.

[28] H. Nakayama, S. Ata, and I. Oka. Caching algorithm for content-oriented networks using prediction of popularity of contents. In *IEEE International Symposium on Integrated Network Management*, pages 1171–1176, May 2015.

[29] M. Peng, Y. Li, Z. Zhao, and C. Wang. System architecture and key technologies for 5G heterogeneous cloud radio access networks. *Network,IEEE*, 29(2):6–14, March. 2015.

[30] M. Peng, S. Yan, K. Zhang, and C. Wang. Fog-computing-based radio access networks: issues and challenges. *Network,IEEE*, 30(4):46–53, July. 2016.

[31] C. Richier, R. Elazouzi, T. Jimenez, E. Altman, and G. Linares. Predicting popularity dynamics of online contents using data filtering methods. In *54th Annual Allerton Conference on Communication, Control, and Computing*, pages 31–38, Sep 2016.

[32] P. Sun, K. G. Shin, H. Zhang, and L. He. Transmit power control for D2D-underlaid cellular networks based on statistical features. *IEEE Trans. Veh. Technol.*, 66(5):4110–4119, May. 2017.

[33] K. Thar, Thant Zin Oo, Chuan Pham, S. Ullah, Doo Ho Lee, and C. S. Hong. Efficient forwarding and popularity based caching for content centric network. In *International Conference on Information Networking*, pages 330–335, Jan 2015.

[34] A. F. R. Trajano and M. P. Fernandez. Contentsdn: A content-based transparent proxy architecture in software-defined networking. In *IEEE 30th International Conference on Advanced Information Networking and Applications*, pages 532–539, March 2016.

[35] Antonio Virdis, Giovanni Nardini, and Giovanni Stea. Modeling unicast device-to-device communications with simulte. In *2016 1st International Workshop on Link and System Level Simulations*, pages 1–6, July. 2016.

[36] W. Hu and G. Cao. Quality-Aware Traffic Offloading in Wireless Networks. *IEEE Trans. Mobile Comput.*, 16(11):3182–3195, Nov 2017.

[37] L. Wei, Z. Zheng, J. Corander, and G. Taricco. On the outage capacity of orthogonal space-time block codes over multi-cluster scattering MIMO channels. *IEEE Trans. Commun.*, 63(5):1700–1711, May. 2015.

[38] Dapeng Wu and R. Negi. Effective capacity: A wireless link model for support of quality of service. *IEEE Trans. Commun.*, 2(4):630–643, July. 2003.

[39] Z. Xiaoqiang, Z. Min, and W. Muqing. An in-network caching scheme based on betweenness and content popularity prediction in content-centric networking. In *IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications*, pages 1–6, Sep 2016.

[40] Y. Yang, Y. Zhang, L. Dai, J. Li, S. Mumtaz, and J. Rodriguez. Transmission capacity analysis of relay-assisted device-to-device overlay/underlay communication. 13(1):380–389, Feb. 2017.

[41] H. Yao, D. Zeng, H. Huang, S. Guo, A. Barnawi, and I. Stojmenovic. Opportunistic offloading of deadline-constrained bulk cellular traffic in vehicular dtns. *IEEE Transactions on Computers*, 64(12):3515–3527, Dec 2015.

[42] M. Chen W. Saad C. Yin and M. Debbah. Echo state networks for proactive caching and content prediction in cloud radio access networks. In *Proc. IEEE Global Telecommunications Conference*, pages 1–6, Dec. 2016.

[43] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor. Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks. *IEEE J. Sel. Areas Commun.*, 34(5):1207–1221, May. 2016.