Efficient Data Delivery in 5G Mobile Communication Networks

vorgelegt von M. Sc.

Yilin Li

geb. in Gansu, VR China

von der Fakultät IV – Elektrotechnik und Informatik

der Technischen Universität Berlin

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Rafael Schaefer (Technische Universität Berlin)

Gutachter: Prof. Giuseppe Caire, Ph.D. (Technische Universität Berlin)

Gutachter: Prof. Dr.-Ing. Gerhard Wunder (Freie Universität Berlin)

Gutachter: Prof. Dr. Joerg Widmer (IMDEA Networks Institute)

Tag der wissenschaftlichen Aussprache: 22. Feb. 2019

Berlin 2019

To my family and friends.

Acknowledgements

The research work for this dissertation was carried out during the years 2015–2018 at the German Research Center, Huawei Technologies Duesseldorf GmbH, Munich, Germany, and at Technische Universität Berlin (TU Berlin), Berin, Germany. The years of the Ph.D. program at Huawei and TU Berlin have been an incredible journey in my life, and I would like to express my gratitude to many people for their support during this journey.

First and foremost, I would like to thank my advisors Dr. Jian Luo and Prof. Giuseppe Caire. I owe my deepest gratitude to Dr. Luo, who has guided me with his solid engineering knowledge, high standards of rigorous works, and strong work ethic and dedications to research. I was fortunate to be admitted into Prof. Caire's research group, and his deep mathematical knowledge, great senses for research directions, and strong writing and presentation skills have significantly improved my research capabilities. I am extremely grateful for all the encouragement, inspiration, and support from my advisors, which have shaped me into an independent researcher and professional engineer.

I would also like to thank the dissertation committee members Prof. Gerhard Wunder and Prof. Joerg Widmer for providing insightful comments to my research and dissertation. In addition, I would like to express my gratitude and appreciation to my colleagues, Dr. Mario H. Castañeda Garcia, Dr. Nikola Vučić, Dr. Emmanouil Pateromichelakis, Dr. Ronald Böhnke, and Zhongfeng Li, for invaluable comments and discussions,

which have inspired me a lot for my research work. I would also like to express my special thanks to the department head Dr. Wen Xu and Dr. Richard A. Stirling-Gallacher for their guidance and support of my research work in Huawei.

I am sincerely grateful to our "PhD student squad" members, Anastasios Kakkavas, Marcin Iwanow, Marcin Pikus, Naveed Iqbal, Taylan Şahin, and Tobias Laas, for various valuable technical discussions and a lot of amazing moments. Moreover, I would also like to thank my colleagues Hanqian Yu, Yunyan Chang, Dr. Xitao Gong, Dr. Qi Wang, Dr. Chan Zhou, Dr. Hanwen Cao, Dr. Zhao Zhao, Dr. Apostolos Kousaridas, Dr. Konstantinos Manolakis, Dr. Martin Schubert, Dr. Mate Boban, Dr. Onurcan İşcan, Dr. Ömer Bulakçi, Dr. Panagiotis Spapis, Dr. Samer Bazzi, Dr. Sandip Gangakhedkar, Xiaosha Tang, Wei Zhao, Jingna Fu, Geer Li, Jianjun Wu, Jian Li, Xinyue Niu, Shiyu Lin, Xueying Chen, as well as all other colleagues in the whole advanced wireless technology Laboratory and the German Research Center of Huawei, for the support and collaboration in research projects and daily work.

For my wonderful and memorable life in Germany, I want to show my warm thanks all my friends: Li Zhong, Jinjian Wang, Yaqi Xu, Lei Zhou, Dr. Licong Zhang, and many others.

Last but certainly not least, my deepest thanks to my parents, for their unconditional support and love, and my girlfriend Wanqiu Luo, for sharing this unbelievable Ph.D. journey with me.

Munich, December 2nd, 2018.

Yilin Li

Zusammenfassung

Die ständig zunehmende Verbreitung intelligenter Geräten, die Einführung neuer benutzerorientierter unternehmenskritischer Anwendungen, sowie ein exponentieller Anstieg der Nachfrage und Nutzung mobiler Daten, belasten die bestehenden Mobilfunknetze erheblich. Tatsächlich muss der Mobilfunk der fünften Generation (5G) ein Paradigmenwechsel sein, der sehr hohe Trägerfrequenzen mit großen Bandbreiten, extrem dicht verteilten Basisstationen und Geräten, und eine große Anzahl an Antennen umfasst. Das Ziel dieser Dissertation ist es, einen präzisen Rahmen für die Datenbereitstellung für 5G-Mobilfunknetze zu entwickeln, und die Herausforderungen bei Design und Analyse der aufkommenden Kommunikationsnetze zu bewältigen, die eine flächendeckende Abdeckung mit hoher Übertragungsgeschwindigkeit und eine nahtlose Benutzererfahrung ermöglichen.

Zunächst werden Kommunikationsnetzwerke betrachtet, die in Millimeterwellen (mm-Wellen) Bändern operieren, wobei mm-Wellen aufgrund der Verfügbarkeit von reichlich Spektrum eine besonders vielversprechende Technologie für den 5G Mobilfunk sind. Dennoch erfordert eine stark gerichtete Übertragung, die für die Nutzung von mm-Wellenbändern zur Kompensation hoher Ausbreitungsverluste unerlässlich ist, ein spezifisches Design der ersten Zellerkennung bei mm-Wellen, da eine herkömmliche omnidirektionale Übertragung bei der Bereitstellung von Zellenerkennungsinformationen möglicherweise versagt. Um dieses Problem zu lösen, wird ein nachvollziehbarer analytischer Rahmen für die Zellerkennung bei mm-Wellen mit Strahlformung vorgeschlagen, der auf einem in-

formationstheoretischen Ansatz basiert und mehrere repräsentative Übertragungsschemata berücksichtigt, um die Leistung der Zellerkennung in Bezug auf die Erkennungslatenz und den Ressourcenaufwand zu charakterisieren. Diese Dissertation, die die Analyse der Zellerkennungsperformance nutzt, liefert wichtige Erkenntnisse für das Design der mm-Wellen Zellerkennung: eine vollständige Suche mit einem Einzelstrahl optimiert die Erkennungslatenz, führt aber einem erhöhten Ressourcenaufwand, während eine Mehrstrahlsuche den Ressourcenaufwand erheblich reduziert und die Flexibilität bietet, einen Kompromiss zwischen Latenz und Ressourcenaufwand zu erreichen.

Ein weiterer Schwerpunkt dieser Dissertation sind neue Technologien, die Raummultiplex für den Mobilfunk nutzen. Um eine große Anzahl von Verbindungen gleichzeitig zu übertragen, wurde eine Richtantennenanordnung mit Strahlformung als vielversprechender Kandidat zum Erreichen einer beispiellosen räumlichen Isolation konzipiert. Um eine hohe Effizienz der räumlichen Wiederbenutzung und damit eine Verbesserung der Systemleistung zu erreichen, wird ein Optimierungsproblem formuliert, das die erreichbare Datenrate eines heterogenen Multihop-Netzwerks maximiert, wobei sowohl Downlink- als auch Uplink-Übertragungen über Backhaul- und Zugangsverbindungen berücksichtigt werden. Das Optimierungsproblem wird systematisch zerlegt und es wird gezeigt, dass es NP-schwer ist. Außerdem wird ein heuristisches gemeinsames Planungs- und Ressourcenzuweisungsschema einschließlich Verbindungsplanung, Übertragungsdauer- und Leistungszuweisung vorgeschlagen, um die erreichbare Datenrate zu maximieren. In Verbindung mit einem effizienten Pfadauswahlverfahren wird gezeigt, dass die Datenrate dem theoretischen Optimum sehr nahe kommt, jedoch mit deutlich geringerer Latenz.

Schließlich wird in dieser Dissertation die Datenübermittlung in Mobil-

funknetzen betrachtet, mit besonderem Schwerpunkt auf Fahrzeugnetzwerken (V2X: vehicle-to-everything), bei denen die Datenübermittlung auf der Carry-and-Forward Strategie der Fahrzeug-zu-Fahrzeug-Kommunikation beruht, die durch die Verfügbarkeit von Fahrzeug-zu-Infrastruktur-Kommunikation unterstützt wird, um weite Strecken zwischen Fahrzeugen zu überbrücken. Trotzdem erfordert die V2X-Kommunikation ein spezifisches Design des Multihop-Routings, um die Leistung der Datenübermittlung zu verbessern. Um das Problem zu lösen, bietet diese Dissertation einen nachvollziehbaren analytischen Rahmen, um die Leistung der Datenübermittlung in V2X-Netzwerken zu untersuchen, wobei Übermittlungslatenz und Datenrate als Leistungskennzahlen berücksichtigt werden. Basierend auf theoretischen Analysen werden sowohl globale als auch verteilte Optimierungsprobleme formuliert, um die Übermittlungslatenz zu minimieren und gleichzeitig die Datenrate zu maximieren. Die vorgeschlagenen Optimierungsprobleme werden als konvex verifiziert und dann mit Methoden der konvexen Optimierungstheorie gelöst, gefolgt von entsprechenden globalen und verteilten Multihop-Routingalgorithmen, um den optimalen Pfad für die Datenübermittlung auszuwählen. Insbesondere bieten die vorgeschlagenen Routingalgorithmen eine erhebliche Verbesserung gegenüber klassischen Routingalgorithmen für Fahrzeuge im Sinne einer Minimierung der Latenzzeit bei gleichzeitiger Maximierung der Datenrate und sie geben Einblicke in das Design der Routingalgorithmen für Datenübermittlung in Mobilfunknetzen.

Abstract

The ever-increasing proliferation of smart devices, the introduction of new user-oriented mission-critical applications, together with an exponential rise in mobile data demand and usage, have been creating a significant burden on the existing cellular networks. Indeed, fifth generation (5G) mobile communications will need to be a paradigm shift that includes very high carrier frequencies with massive bandwidths, extreme base station and device densities, and unprecedented numbers of antennas. The goal of this dissertation is to develop an accurate data delivery framework and address design and analysis challenges, including initial access, link scheduling, resource allocation, and routing, for 5G mobile communication networks that are envisioned to provide universal high-rate coverage and seamless user experience.

First, this dissertation considers communication networks operating in millimeter wave (mm-wave) bands, where the availability of abundant spectrum makes mm-wave a prominent candidate technology for 5G mobile communications. Highly directional transmission, which is essential for the exploitation of mm-wave bands to compensate for high propagation loss, necessitates a specific design of mm-wave initial cell discovery, as conventional omnidirectional broadcast may fail in delivering cell discovery information. To address this issue, a tractable analytical framework is proposed to characterize beamformed cell discovery, where several representative broadcast schemes are studied to investigate discovery performance including latency and overhead. Leveraging the analysis of the discovery performance, this dissertation provides key

insights for the design of mm-wave beamformed cell discovery: Single beam exhaustive scan optimizes latency but leads to overhead penalty, and multiple beam simultaneous scan significantly reduces overhead and provides the flexibility to achieve a trade-off between latency and overhead.

Then, the focus of this dissertation shifts to emerging technologies that exploit spatial multiplexing for mobile communications. By enabling a large number of links to be simultaneously transmitted, directional antenna arrays with beamforming are promising to reach unprecedented levels of spatial isolation. To achieve the high efficiency of spatial reuse in improving system performance, an optimization problem that maximizes the achievable data rate of a multihop heterogeneous network, considering both downlink and uplink transmissions on backhaul and access links, is formulated. The optimization problem is systematically decomposed and demonstrated as NP-hard, and a heuristic joint scheduling and resource allocation scheme, including link scheduling, transmission duration allocation, and power allocation, is proposed to maximize the achievable data rate. In conjunction with an efficient path selection algorithm, it is demonstrated that the data rate closely approaches the theoretical optimum, yet with significantly lower latency.

Finally, this dissertation considers the data delivery in mobile communication networks, with a special focus on vehicle-to-everything (V2X) networks, in which the data delivery relies on the carry-and-forward strategy of vehicle-to-vehicle communications assisted by vehicle-to-infrastructure communications that are capable of bridging long-range vehicular connectivity. Nevertheless, V2X communications necessitate a specific design of multihop routing to enhance data delivery performance. To address the issue, this dissertation provides a tractable analytical framework to investigate the data delivery performance in V2X networks tak-

ing into account latency and data rate as performance metrics. Based on a theoretical analysis, both global and distributed optimization problems that minimize latency while maximizing data rate are formulated. The proposed optimization problems are verified to be convex and then solved using convex optimization theory, followed by corresponding global and distributed multihop routing algorithms to select the optimal route for data delivery. In particular, the proposed routing algorithms provide considerable improvement over classical vehicular routing algorithms, in the sense of minimizing latency while maximizing data rate, and shed design insights into the multihop routing algorithm for data delivery in 5G mobile communication networks.

Table of Contents

List of Figures

List of Tables

Acronyms

3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
ABS	Almost Blank Subframe
ADC	Analog-to-Digital Converters
AP	Access Point
AWGN	Additive White Gaussian Noise
BS	Base Station
BT	Bidirectional Transmission
CD	Code-Division
CDF	Cumulative Distribution Function
CDL	Cell Discovery Latency
CDMA	Code-Division Multiple Access
CDO	Cell Discovery Overhead
CG-MIS	Conflict Graph Maximum Independent Set
CMOS	Complementary Metal-Oxide-Semiconductor
CP	Cyclic Prefix
CS	Complementary Slackness
D2D	Device-to-Device
DAC	Digital-to-Analog Converters
DF	Dual Feasibility
DPP	Drift Plus Penalty

DR	Dynamic Routing
DSRC	Dedicated Short-Range Communication
eICIC	enhanced Inter-Cell Interference Coordination
eNB	evolved Node Base station
E2E	End-to-End
FD	Frequency-Division
FDD	Frequency-Division Duplex
FDMA	Frequency-Division Multiple Access
FTP	File Transfer Protocol
gNB	next generation Node Base station
Gbps	Gigabits per second
GHz	GigaHertz
GI	Guard Interval
GPSR	Greedy Perimeter Stateless Routing
HetNets	Heterogeneous Networks
i.i.d	independent and identically distributed
IAB	Integrated Access and Backhaul
IoT	Internet of Things
IoV	Internet of Vehicles
IP	Internet Protocol
ITS	Intelligent Transportation System
JSRA	Joint Scheduling and Resource Allocation
KKT	Karush-Kuhn-Tucker
LOS	Line-Of-Sight
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution Advanced
mm-wave	millimeter wave
MAC	Medium Access Control
MHz	MegaHertz

MINLP	Mixed Integer Nonlinear Programming
MIS	Maximum Independent Set
MWST	Minimum Weight Spanning Tree
NLOS	Non-Line-Of-Sight
NP	Non-deterministic Polynomial-time
NR	New Radio
NUM	Network Utility Maximization
OFDM	Orthogonal Frequency-Division Multiplexing
P2M	Point-to-Multipoint
P2P	Point-to-Point
PDF	Probability Density Function
PF	Primal Feasibility
PHY	PHYsical layer
PMF	Probability Mass Function
PPP	Poisson Point Process
PSS	Primary Synchronization Signal
QoS	Quality of Service
RAN	Radio Access Network
RF	Radio Frequency
RR	Round-Robin
RSRP	Reference Signal Received Power
RSU	Road Side Unit
Rx	Receiver
SD	Space-Division
SDMA	Space-Division Multiple Access
SF	Shadowing Factor
SG	Spreading Gain
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio

SPR	Shortest Path Routing
SSS	Secondary Synchronization Signal
TD	Time-Division
TDD	Time-Division Duplex
TDMA	Time-Division Multiple Access
TRP	Transmission Reception Point
Tx	Transmitter
UE	User Equipment
V2G	Vehicle-to-Grid
V2I	Vehicle-to-Infrastructure
V2P	Vehicle-to-Pedestrian
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VANETs	Vehicular Ad-hoc NETworks
WCDMA	Wideband Code-Division Multiple Access

Chapter 1

Introduction

1.1 Background and Motivation

The gradual, yet steady evolution of mobile communications, initiated with the first generation, and towards the second generation, the third generation, and the fourth generation (4G), has been witnessed by the world over the last few decades. Originated from the voice-only system, the introduction of advanced physical layer (PHY) technologies, realistic channel measurement and modeling, well-defined radio access network (RAN) architectures, and the penetration of packet-based Internet, have significantly contributed towards the gradual evolution [1]. The packet-based long-term evolution (LTE) system, embodying 4G, has now been widely deployed and is reaching its maturity. Considering the incremental improvement of current mobile communication systems and the increasing popularity of smart devices [2], it is natural for researchers to ponder "what is next?"

Driven by emerging user-oriented mobile multimedia applications, like video streaming, remote surgery, and online gaming, mobile data explosion is upcoming and will continue. The quantitative evidence released by the latest Cisco visual networking index [3] reveals that annual global IP traffic will reach 3.3 zettabytes (1 trillion gigabytes) by 2021, and will have increased 127-fold from 2005 to 2021.

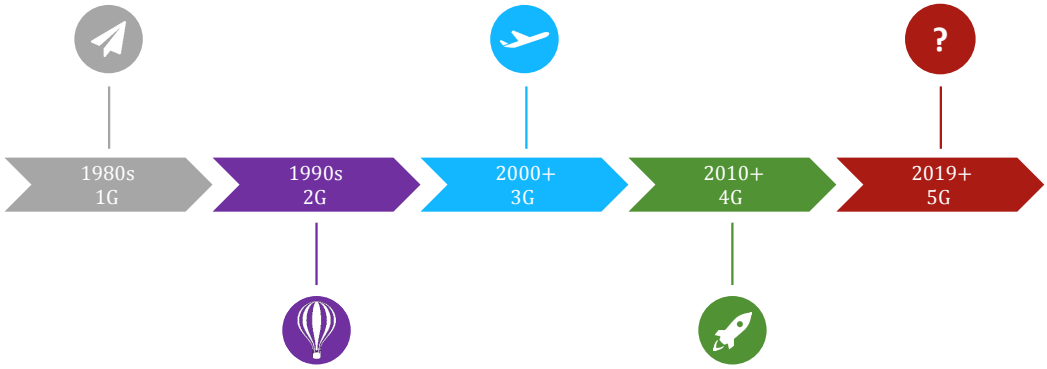


Figure 1.1: Evolution of mobile communications.

In addition to the deluge of data, the number of mobile devices could reach three times as high as global population in 2021 [3], due to new applications in directions of augmented reality, Internet of things (IoT), and Internet of vehicles (IoV), etc. Supporting such sheer volume of data usage and rapid increase in connectivity is an extremely daunting task in present 4G LTE cellular systems, where an incremental approach of LTE will not come close to meeting the demands that mobile communication networks will face by 2020 [4].

The combined effect of the highly visible demand for more network capacity, the possibility of "Internet of everything" composed of billions of miscellaneous devices, and the increasing integration of past and current cellular and Wi-Fi standards, trigger the next major evolution in mobile communications — the 5G (fifth generation). Fig 1.1 provides an illustration of the evolution trends from 1G to 5G. In just the past years, preliminary interest and discussions about a possible 5G standard have evolved into a full-fledged conversation that has captured the attention and imagination of researchers and engineers around the world. Industry and Academia are engaging in collaborative projects such as METIS [5] and mmMAGIC [6], while the 3rd Generation Partnership Project (3GPP) is driving 5G standardization activities as New Radio (NR) [7]. 5G mobile communications envision orders of magnitude higher data rates, ubiquitous coverage and connectivity, and a massive reduction in round-trip latency and energy consumption, which would be supported by the paradigm shift that includes the "big three" 5G technologies:

very high carrier frequencies with enormous bandwidths, *unprecedented numbers of antennas*, and *extreme base station densities*.

Nowadays, the vast majority of commercial radio communications, including AM/FM, GPS, cellular, and Wi-Fi, have crowded in a narrow band of the radio frequency (RF) spectrum from 300 megahertz (MHz) to 3 gigahertz (GHz), often termed as the "sweet spot" [8], due to its encouraging propagation features for radio communication applications. Regardless of the efficacy of densification and offloading [9], there is only one way to put large amounts of new bandwidth into play: up in frequency [10]. Fortunately, immense amounts of relatively idle spectrum do exist in the millimeter wave (mm-wave) range of 6–300 GHz, where the available bandwidths in these bands, for example in Ka-band (26.5–40 GHz), V-band (57–71 GHz), and E-band (71–76 GHz and 81–86 GHz), can significantly exceed all allocations in contemporary cellular networks [11]. Fig. 1.2 illustrates the availability of mm-wave spectrum. The main reason that mm-wave spectrum has lay idle, until recently, is that it had been deemed unsuitable for mobile communications because of rather hostile propagation qualities, including strong path loss, atmospheric absorption, and low penetration through obstacles, which confine the application of mm-wave spectrum to short-range transmission [12, 13].

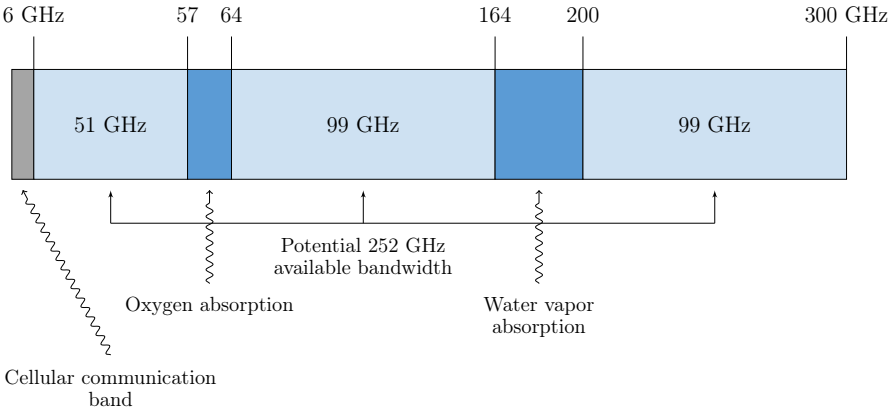


Figure 1.2: Mm-wave spectrum availability in 6–300 GHz.

Thanks to the very small wavelengths of mm-wave signals, a large number of miniaturized antennas, combined with advanced low power complementary metal-

oxide-semiconductor (CMOS) RF circuits, are exploited for deploying mm-wave communication networks [14]. This brings the concept of sectorized and directional beams, almost like flashlights, as opposed to the age-old omnidirectional beams. The narrow beam, however, leads to the difficulty in establishing an association between a user equipment (UE) and a base station (BS), particularly for initial access. To find each other, a UE and a BS may need to scan lots of angular positions where a narrow beam could possibly be found. In current LTE systems, initial access is performed using omnidirectional signals, whereas beamforming or other directional transmissions can only be performed after a physical link is established [15]. More importantly, the omnidirectional transmission at mm-wave frequencies, indeed, may generate a mismatch between the relatively short range at which a cell can be discovered, and the much longer range at which a UE can receive or transmit data using beamforming [16, 17]. These issues are particularly important to 5G mobile communications and motivate a thorough performance analysis of beamformed initial access with an accurate framework.

Another straightforward but extremely effective way to increase the network capacity, addressed as one of the "big three" 5G technologies above, is making the cells smaller. Cell shrinking has numerous benefits, such as efficient spectrum reuse, the compensation of penetration loss, and the deployment of a large number of small cells giving rise to heterogeneous networks (HetNets) [18, 19]. HetNets are typically composed of small cells, which are deployed for improving coverage and augmenting overall capacity, and a legacy macrocell that seamlessly cooperates with the small cells. A major cost consideration for HetNets is the backhaul from cell edge to core network [20]. In contrast to the legacy wired backhaul, HetNets are likely to be connected via wireless backhaul infrastructure [21]. Characteristics of capacity, latency, and deployment cost are expected to vary from case to case, which enables wireless backhaul a viable and economical alternative to fully exploit the heterogeneity of HetNets. To make sure that backhaul does not become a bottleneck for 5G HetNets, the problem of joint backhaul and access optimization is becoming

a pressing concern, which drives the need of designing innovative scheduling and resource management schemes.

Besides, cell densification, by its nature, allows remote users to establish relatively long radio hops via small cell access point (AP) with macrocell BS. The new-found interest in multihopping for 5G mobile communications is reflected in recent academic works such as [22, 23, 24, 25, 26], as well as in 3GPP Release-15 standardization activities [27]. In fact, the noise-limitedness of mm-wave transmissions greatly simplifies multihop routing problems for a generic network deployment with fixed topology and can help to close the gap between theoretically optimal solutions and practical implementations. Many high data rate 5G applications, particularly the ones for vehicular communications, such as the innovative operation of vehicle, safe and eco-friendly driving, and ubiquitous infotainment services, have yet high mobility at the same time, which brings new challenges in providing a full-fledged solution to capture the connectivity, efficiency, and scalability of vehicular networks. Hence, an accurate path selection model needs to be properly designed for optimizing the data delivery performance in 5G mobile multihop networks, with the specific interest in vehicle-to-everything (V2X) networks.

1.2 Contributions and Outline

Motivated by the issues addressed in Section 1.1, this dissertation mainly aims to tackle the challenges of 5G mm-wave mobile communication systems, including initial access, scheduling and radio resource management, and multihop routing, and design algorithms and methods to optimize the overall performance of the considered systems. The main technical contributions of this dissertation are covered in Chapter 2 to Chapter 4, which are summarized as follows.

Beamformed Cell Discovery in 5G Millimeter Wave Communication Networks. An essential design challenge for mm-wave communication networks is cell discovery, which refers to the preliminary process that allows a user to first

establish links to a communication network. As mentioned in Section 1.1, mm-wave communication systems generally rely on directional transmission in order to maintain a viable connection. The directional transmission, therefore, entails a special design of mm-wave cell discovery, as the conventional omnidirectional broadcast may fail in delivering cell discovery information to the users of mm-wave communication networks. In Chapter 2, we propose an analytical framework for mm-wave beamformed cell discovery, based on an information-theoretical approach, to characterize the key performance metrics for the cell discovery under various network configurations, beamforming architectures, and beam scan models. By leveraging four fundamental but representative broadcast schemes, we demonstrate that the mm-wave beamformed cell discovery is analytically tractable. Specifically, we show that analog beamforming and hybrid beamforming perform as well as digital beamforming in terms of cell discovery latency. Moreover, single beam exhaustive scan is shown to provide the lowest discovery latency but lead to the highest burden for the network to perform cell discovery, referred to as cell discovery overhead. By contrast, multiple beam simultaneous scan can significantly reduce the overhead and achieve a trade-off between latency and overhead. In addition, the performance analysis is extended by considering miscellaneous use-case-dependent frame structures and block error rates. As a result, we demonstrate that different frame structure options provide the flexibility in reaching desired latency and overhead, which are also shown to be relatively insensitive to extreme low block error rates.

Joint Scheduling and Resource Allocation Optimization in 5G Millimeter Wave Heterogeneous Networks. The very high dimensional antenna arrays, benefited from the very small wavelengths of mm-wave signals and low power 1-bit analog-to-digital/digital-to-analog converters (ADC/DAC), enable the utilization of high directional transmission in order to fully exploit spatial resource reuse. In Chapter 3, we leverage link scheduling, radio resource allocation, and routing algorithms to enhance the overall performance of a mm-wave HetNet, taking into account both backhaul and access links and a combination of both downlink and

uplink transmissions. The HetNet, composed of a macrocell BS and several small cell APs, is modeled as graph nodes, and the transmission links among these nodes are abstracted as graph edges. Based on this, an optimization problem, in terms of maximizing long-term average data rate, is formulated as a mixed integer nonlinear programming (MINLP) problem, considering constraints such as interference, transmission time and power allocation, and duplex mode. As MINLP problems are in general non-deterministic polynomial-time (NP)-hard, we propose a heuristic joint scheduling and resource allocation algorithm, which allows non-interfered links to be simultaneously transmitted with prolonged transmission duration, to solve the optimization problem. Moreover, we propose a dynamic routing algorithm that selects paths consisting of low-load hops to further enhance the end-to-end (E2E) data rate achieved by the proposed joint scheduling and resource allocation algorithm with predefined static paths. Compared to the baseline multiple access and interference coordination schemes, such as time-division multiple access (TDMA) and enhanced inter-cell interference coordination (eICIC), the results show that the proposed algorithms offer considerable improvements in data rate, and approach the theoretical optimum acquired by the drift plus penalty (DPP) with backpressure scheme with a small degradation. The optimal solution causes yet intolerable latency in delivering data to users, while the proposed algorithms achieve near-optimal performance in terms of data rate but with much lower latency. The performance of the proposed algorithms for different frame structures, duplex modes, and network configurations are also demonstrated.

Multihop Routing for Data Delivery in 5G Millimeter Wave V2X Networks. 5G is not only a new access technology but also a user-centric network concept that aims to address the application requirements of all the stakeholders in the connected world. Among the stakeholders, vehicles, in particular, will benefit from 5G at both system and application levels. In Chapter 4, the focus is shifted up from traditional personal users to vehicular users, which will be an integral part of 5G mobile communications that promise to provide ubiquitous connectivity and

ultra-reliable transmissions with low latency. While the idea of streamlining the innovative operation of vehicle, facilitating safe and eco-friendly driving, and offering ubiquitous infotainment services for commuting passengers seems yet far-fetched, efforts are already underway to achieve efficient data delivery with bulky throughput, high reliability, and low latency for diverse requirements of different V2X services and applications. Nevertheless, the data delivery in V2X networks is particularly challenging, as its performance depends highly on the efficiency of data routing. On the one hand, inter-vehicle communications may introduce non-negligible latency due to frequent disconnections. On the other hand, inter-RSU (road side unit) communications can suffer from limited coverage and insufficient backhaul. Hence, the development of an accurate analytical framework for the data delivery in V2X networks is proposed in Chapter 4, where closed-form mathematical expressions for both hop-wise and E2E latency and data rate have been derived. Based on the expressions, optimization problems aiming to maximize the weighted sum of latency and data rate are formulated considering both global and distributed scenarios. The formulated problems are then solved by convex optimization theory, and multihop routing algorithms, based on the solutions, are proposed for both global and distributed data delivery to select the optimal route. In the end, the rigorousness of the theoretical analysis, together with the efficiency of the proposed routing algorithms, are demonstrated by numerical evaluations.

Finally, Chapter 5 concludes this dissertation by summarizing the key contributions and discussing potential future research directions.

1.3 Previous Publications and New Contributions

This dissertation is partly based on the author's previous publications [11, 16, 28, 29, 30]. The relations between this dissertation and the previous publications are listed below:

- The design of broadcast schemes and the analysis of 5G mm-wave beamformed

cell discovery for standalone networks have been partly published in [16].

- The analysis of the 5G mm-wave beamformed cell discovery for non-standalone networks, for different frame structure designs, and for different block error rates has been published in [28] with detailed derivations and extensive simulations.
- The considerations for the radio resource allocation in 5G mm-wave mobile communication networks have been published in [11].
- The joint scheduling and resource allocation optimization for 5G mm-wave HetNets has been partly published in [29].
- The multihop routing for the data delivery in 5G mm-wave V2X networks has been partly published in [30].

In all the previous publications, co-authors[1] have provided many valuable discussions and insightful comments to derive solutions and improve the presentation of results.

Except for the contributions in the previous publications, this dissertation also includes the following new (unpublished) contributions:

- In Section 2.4.5, the work in [28] is extended by considering the mm-wave beamformed cell discovery with beam-scanning reception. Moreover, design insights for the cell discovery of 5G mobile communication systems are addressed in Section 2.6.
- In Section 3.4, the design of multihop dynamic routing algorithm for path selection of 5G mm-wave HetNets is proposed in addition to the joint scheduling and resource allocation algorithm addressed in Section 3.3 and in [29].
- In Section 3.5, the performance of data rate and latency achieved by the proposed joint scheduling and resource allocation algorithm and the dynamic routing algorithm for different number of users, frame structures, as well as duplex modes are evaluated.

[1]Dr. Jian Luo, Prof. Giuseppe Caire, Dr. Wen Xu, Dr. Mario H. Castañeda Garcia, Dr. Nikola Vučić, Dr. Emmanouil Pateromichelakis, Dr. Richard A. Stirling-Gallacher, Dr. Ronald Böhnke, and Zhongfeng Li.

- In Section 3.5.6, the application of the proposed joint scheduling and resource allocation algorithm and the dynamic routing algorithm to the data delivery in a platoon of vehicles are addressed.
- In Section 4.3.4, the distributed data delivery problem is formulated in addition to the global data delivery problem addressed in Section 4.3.3 and in [30]. Besides, more thorough derivations of the latency and data rate of both the global and distributed data delivery problems are presented in Section 4.4.1. Based on these, the distributed multihop routing algorithm is proposed in Section 4.4.3 in addition to the global multihop routing algorithm addressed in Section 4.4.3 and in [30].
- In Section 4.5, the performance of latency and data rate achieved by the proposed global routing algorithm and the distributed routing algorithm for different broadcast schemes, vehicle arrival rates, as well as the support of wireless backhaul on data delivery are evaluated.

Chapter 2

Beamformed Cell Discovery in 5G Millimeter Wave Communication Networks

2.1 Introduction

As mentioned in the previous chapter, the spectral efficiency of the current sub-6 GHz communication systems is already approaching theoretical limits [31]. With the wide coverage of frequencies ranging from 6 and 300 GHz, mm-wave bands have drawn significant attention for the next generation mobile communication systems [10, 32], where the available bandwidths are much wider than today's cellular allocations [11, 33]. Despite its promising characteristics in offering vast spectrum opportunities, mm-wave bands suffer from increased isotropic free space loss, higher penetration loss, and propagation attenuation, which also pose significant challenges

This chapter has been published in [16, 28]. I am the primary author of these works. Coauthors Dr. Mario H. Castañeda Garcia, Dr. Ronald Böhnke, Dr. Nikola Vučić, Dr. Richard A. Stirling-Gallacher, and Dr. Wen Xu have provided many valuable discussions and insights to these works, and Dr. Jian Luo and Prof. Giuseppe Caire are my supervisors.

for cellular communications [34]. In order to combat these issues, one distinctive feature of mm-wave communications is acquiring high antenna gain at both transceiver sides for enhanced data transmissions [35, 36]. Fortunately, advances in CMOS RF circuits combined with the small wavelengths of mm-wave bands enable a large number of electrically steerable antenna elements to be placed at both AP and UE sides. These high-dimensional antenna arrays can provide further gains via adaptive beamforming and spatial multiplexing [37].

Nevertheless, the requirement for narrow beam communication renders the design of cell discovery a central and novel challenge for mm-wave communication systems [38]. Specifically, cell discovery refers to the procedures by which a UE discovers a potential mm-wave cell and establishes a link-layer connection [39]. The cell discovery is the critical prerequisite for any subsequent communications, as UE and the AP of a mm-wave cell have no information about transmission and reception beam directions upon cell discovery, and they must search over a potentially large angular space to find each other. Besides, the cell discovery for mm-wave communication systems will be more time and resource consuming compared to the conventional systems, as mm-wave signals are vulnerable to blockages and thus are prone to frequent beam misalignment.

In addition, the reliance on highly directional transmission and reception considerably complicates the cell discovery in mm-wave bands [40]. While conventional cellular systems, such as 3GPP LTE/LTE-A [41, 42, 43], support multi-antenna diversity techniques and spatial multiplexing with beamforming, the underlying design assumption is that cell discovery can be conducted entirely with omnidirectional transmissions or transmissions using fixed antenna patterns [44]. An LTE BS, as an example, generally does not apply beamforming when transmitting synchronization/broadcast signals. Directional transmissions are typically exploited only after initial access has been established.

Last but not least, for mm-wave communications, omnidirectional transmission may fail in cell discovery procedure, as the utilization of highly directional antenna

would create a mismatch between discoverable range and achievable range [45, 46]. Specifically, for systems operating at mm-wave bands, applying conventional cell discovery technique would lead to a smaller discoverable area (blue) than the achievable area (gray) in which reasonable data rates can be achieved, as illustrated in Fig. 2.1. Therefore, the mm-wave cell discovery is expected to be designed properly to establish communication links via directional transmission and to exploit resources in the spatial dimension. In this chapter, we develop an analytical framework and detailed performance analysis for the cell discovery in mm-wave communication system.

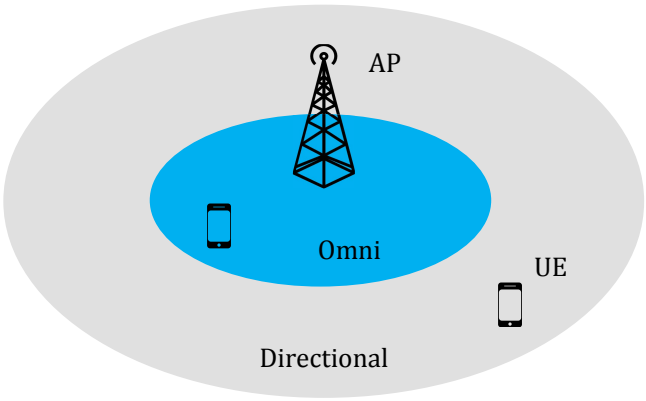


Figure 2.1: Illustration of the mismatch between discoverable area and achievable area of mm-wave communication networks.

2.1.1 Related Works

Cell discovery, particularly at mm-wave bands, has been investigated by a few standard organizations in recent years [47, 48, 49, 50, 51]. Physical broadcast channel is defined in 5G NR for delivering cell discovery information and supports the time-division multiplexing of synchronization signals for both single beam and multi-beam scenario [47, 48]. The IEEE 802.11ad and 802.15.3c standard adopted a two-level initial beamforming training protocol for the 60 GHz unlicensed band, where a coarse-grained sector level sweep phase is followed by an optional beam refinement phase [49, 50]. However, these standards are mainly designed for indoor communications within an ad hoc type network without significant mobility or range.

In addition to the aforementioned standardization activities, several recent research efforts have investigated the cell discovery for mm-wave communication networks [16, 28, 44, 52, 53, 54, 55, 56, 57, 58]. Deafness, as an explicit consequence of directional transmission and reception, emerges when the main beam of a transmitter (Tx) and the intended beam of a receiver (Rx) are not aligned [52]. To address this issue, communication links are established by a beam scan procedure [53, 54], in which an exhaustive scan over all possible combinations of transmission and reception directions is performed through a sequence of broadcast signals. Leveraging the omnidirectional synchronization from macrocell BS, followed by a sequential spatial search, authors in [55] showed that cell discovery efficiency can be enhanced with a two-step sequential spatial search procedure.

Incorporating the concept of beam scan facilitates beamforming procedure, however introduces alignment latency that is the time of matching beam pairs to complete cell discovery. A hierarchical search procedure was proposed in [56], where BS first performs an exhaustive search over wide beams, then refines to narrow beams. Several cell discovery options with different modifications to LTE cell discovery procedure were proposed in [44], which has observed that cell discovery latency can be reduced by omnidirectional transmission from BS during cell search, and digital beamforming can even further reduce the latency. By adopting limited feedback-type directional codebooks, a low-complexity beamforming approach for cell discovery was proposed in [57]. The exhaustive and hierarchical strategies were compared in [58], which shows that hierarchical search generally achieves smaller cell discovery latency, but exhaustive search gives better coverage to cell-edge users.

In addition to the beam scan aspect, the beam alignment problem of cell discovery design for mm-wave communication system has also been investigated in several literatures [59, 60, 61, 62, 63]. By exploiting the information from sub-6 GHz channels, authors in [59] proposed a beam alignment algorithm with low training overhead for mm-wave system. In [60], an efficient beam alignment scheme has been proposed to achieve a competitive performance in terms of scalability and

robustness. Authors in [61] proposed an initial access framework including several beam paring protocols to investigate the trade-off between initial access latency and user-perceived downlink throughput. In [62], a bi-section search algorithm for beam alignment in mm-wave system was shown to achieve higher overall throughput than exhaustive and iterative algorithms. Authors of [63] proposed an enhanced beam codebook design for a better beam alignment.

Furthermore, geometry information has been recently recognized as a promising candidate to improve the efficiency of mm-wave cell discovery [64, 65, 66, 67, 68]. Leveraging the context information related to user positions, authors in [64] proposed a geographically-located context-based approach to improve directional cell discovery process. The fundamental trade-offs of cell discovery process with directional antennas were investigated in [65]. By incorporating a joint search of UE by both mm-wave macrocell and small cells, authors in [66] showed that the performance of cell discovery can be improved. Similarly, authors [67] argued that analog beamforming can still be a viable choice when the context information of mm-wave BS is available at UE. The exploitation of statistical blockage models to achieve a considerable data rate gain was discussed in [68].

All the aforementioned works are either a point-to-point analysis ([58, 64, 65]) or only consider one user with a few nearby BSs ([55, 67, 68]) and, a system-level analysis of cell discovery in mm-wave networks in terms of not only cell discovery latency (CDL), but also cell discovery overhead (CDO), has not been considered ([44, 53]). Some works have utilized stochastic blockage model, such as a line-of-sight (LOS) ball blockage model [69], an exponential decreasing LOS probability function with respect to link length [70], or a blockage model which also incorporates an outage state [71], as a powerful mathematical tool to analyze key performance metrics in mm-wave cell discovery. However, [69, 70, 71] all assume that the association between a user and its serving BS has already been established, while in fact the cell discovery is a key design challenge and a performance limiting factor for mm-wave communications.

2.1.2 Contributions

In this chapter, we consider the fact that in initial access phase, rough beam alignment would be sufficient, and aligned beam refinement can be done in later stages via scheduled resources. The most important task of broadcast during initial access is conveying the necessary information for cell discovery to UE in the most efficient way. Therefore, we investigate four fundamental and representative broadcast schemes that enable various multiplexing schemes of radio resources, such as time, frequency, code, and space. In particular, the broadcast schemes are compliant with the basic cell discovery procedures of LTE, IEEE 802.11ad and IEEE 802.15.3c, including a baseline single beam exhaustive search scheme wherein an AP scans through all coverage spaces during cell discovery, and three multiple beam simultaneous scan schemes that exploit different multiplexing gains in frequency, code, and space.

Then, we develop a framework to analyze the performance of mm-wave cell discovery under these broadcast schemes. The key metrics including CDL and CDO are derived considering different broadcast schemes, network deployments (standalone and non-standalone), as well as frame structure designs. The analysis is then validated against detailed system-level simulations, and the evaluation results allow us to reveal some insights into the design of mm-wave cell discovery. The main contributions of this chapter are summarized as follows:

- Development of an accurate analytical framework for mm-wave cell discovery under various broadcast schemes: With respect to different network deployments, the cell discovery procedure is distinguished by standalone and non-standalone network. Specifically, in the standalone scenario, the beam scan for the broadcast of cell discovery signal is divided into different scenarios depending on various combinations of beam scan area and UE location. These scenarios are shown to result in different CDL and CDO. In particular, it is demonstrated that analog and hybrid beamforming perform as well as digital

beamforming in terms of CDL.

- Comparison of broadcast schemes in CDL and CDO: The baseline scheme, i.e., single beam exhaustive scan, leads to the best CDL performance, under certain conditions, which is the opposite to an intuitive consideration that multiple beam simultaneous scan can accelerate the discovery. This scheme, however, causes high CDO due to the impact of guard interval (GI) reserved for beam switching. By contrast, the schemes wherein AP applies multiple beam simultaneous scan, generally give better CDO performance. However, due to a longer scanning duration at each beam direction compared to the single beam exhaustive scan, these schemes also cause high CDL. For the three considered multiple beam simultaneous scan schemes, the best trade-off between CDL and CDO is achieved by configuring the number of simultaneous beams.

- A detailed system-level performance evaluation for mm-wave cell discovery: Different from the link-level analysis in [44, 58, 64, 68], we apply system-level performance metrics, including CDL and CDO, to capture the cell discovery efficiency of mm-wave communication network. Our analytical results are validated against detailed system-level simulations.

- Leveraging analysis and simulation results to provide insights into the answers of the key questions: (i) How wide should a beam be? (ii) Is it beneficial to exploit multiple beam simultaneous scan? (iii) If so, how many simultaneous beams should be exploited? (iv) Does cell discovery performance vary with different frame structures? (v) What is the impact of block error rate?

The remainder of this chapter is organized as follows: Section 2.2 presents the system model and Section 2.3 compares and analyzes the different broadcast schemes. Section 2.4 presents the analysis of the cell discovery procedure. The analysis is then verified by extensive simulations in Section 2.5, followed by a summary concluding this chapter in Section 2.6.

2.2 System Model

In this section, we introduce the network, propagation, antenna, and beam scan models. These models are the fundamentals of the analytical framework for mm-wave cell discovery mentioned in Section 2.1.2 and are exploited as the preliminaries of the design of broadcast schemes and the analysis of cell discovery procedure, which will be elaborated in Section 2.3 and in Section 2.4, respectively. Specifically, these models are applied to the calculation of achievable signal-to-noise ratio (SNR)/signal-to-interference-plus-noise ratio (SINR) and *beam duration* (a key metric for performance evaluation detailed in Section 2.2.4) of the broadcast schemes, and also to the analysis of CDL and CDO. The important notations and system parameters defined in this section are summarized in Table 2.1 and will be used in the rest of this chapter.

Table 2.1: System Model Parameters

Notation	Description	Value
r	Cell radius	$100\,\mathrm{m}$
$p(d)$	Probability of a link with distance d to be LOS	See (2.1)
d_1	Parameters in d_1/d_2 model	20
d_2	Parameters in d_1/d_2 model	39
$L(f,d)$	Pathloss of a link at carrier frequency f and distance d	See (2.2)
f	Carrier frequency	$28\,\mathrm{GHz}$
c	Speed of light	$3\times10^8\,\mathrm{m/s}$
n_L	Pathloss exponent	LOS, NLOS: $2.1, 3.17$
SF	Shadowing factor	LOS, NLOS: $3.76, 8.09$
N_Tx	Number of antennas at AP	128
M_Tx	Number of RF chains at AP	128
h_AP	Height of AP antennas	$15\,\mathrm{m}$
h_UE	Height of UE access plane	$1.5\,\mathrm{m}$
N	Number of beam scan areas	$1, 2, 4, \ldots, 128$
N_H	Number of beam scan areas in horizontal dimension	$1, 2, 4, 8, 16$
N_V	Number of beam scan areas in vertical dimension	$1, 2, 4, 8$
θ	Horizontal Tx beamwidth	$2\pi/N_\mathrm{H}$
ϕ	Vertical Tx beamwidth	$\arctan\left(\frac{r}{h_\mathrm{AP}-h_\mathrm{UE}}\right)/N_\mathrm{V}$
G	Antenna beamforming gain	See (2.3)
d_H	Height of approximated rectangular beam pattern (see Fig. 2.3)	$d_\mathrm{H} = d\sin\theta$
d_V	Width of approximated rectangular beam pattern (see Fig. 2.3)	$d_\mathrm{V} = d\sin\phi$
T	Frame length	$200\,\mu\mathrm{s}$
M	Number of simultaneous beams	Divisor of N
S	Number of beam slots	N/M
M_H	Number of simultaneous beams in horizontal dimension	Divisor of N_H
M_V	Number of simultaneous beams in vertical dimension	Divisor of N_V

2.2.1 Network Model

Mm-wave communication networks are going to be densely deployed with more random nature than conventional macro network, and currently there is no real mm-wave AP[1] location data available to extract a better model. Poisson point process (PPP) has been recognized as a powerful mathematical tool to idealize of AP locations. However, the PPP assumption for AP locations is not ubiquitously applicable, as [72] has proved that the SNR value under PPP assumption has a constant gap compared to the stationary AP location model. Besides, the convergence of modeling AP locations to a Poisson network depends significantly on the statistics of the propagation loss of communication link, which requires sufficiently large shadowing variance to achieve reasonable performance under PPP assumption [73]. Therefore, in this chapter we focus on stationary AP location assumption and envision a single-cell downlink communication network with one mm-wave AP located at the center of the cell. The considered AP broadcasts signals via beam scan to UEs in its coverage for cell discovery, where the cell radius r is assumed to be 100 meters.

In general, an AP can be located indoor or outdoor. Without loss of generality, in this chapter we focus on the performance of mm-wave communication networks with outdoor AP. UEs are also assumed to be outdoor and randomly "dropped" in the network. Fig. 2.2 shows an example of the considered network. Some recent papers such as [66, 74] have discussed legacy bands (sub-6 GHz) assisted cell discovery, proposing the joint search of UE between mm-wave small cell and macrocell. In this regard, we also provide a performance analysis of cell discovery for a non-standalone network (see Section 2.4.2), where the existence of legacy band refers to the support of synchronization between AP and UE. Besides this, throughout this chapter we generally assume a standalone mm-wave network.

[1]AP can refer to evolved node base station (eNB) for 4G and/or a next Generation Node Base station (gNB) for 5G in 3GPP terminology, or simply BS and/or AP for general use term.

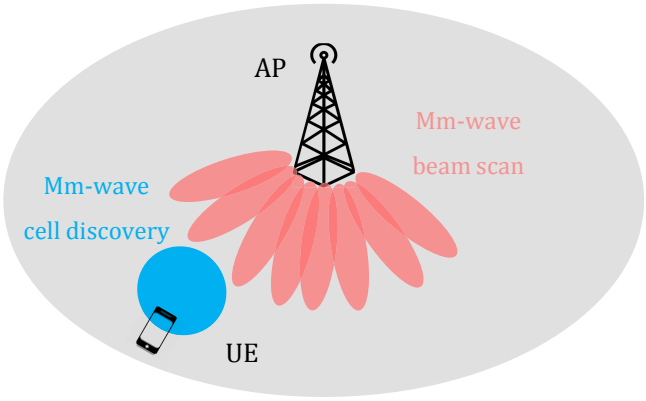


Figure 2.2: Illustration of a single cell standalone mm-wave network with one AP. Only one UE is shown.

2.2.2 Blockage and Propagation Model

Incorporating the blockage model to differentiate the LOS and non-line-of-sight (NLOS) link is a distinctive characteristic for analyzing the mm-wave network performance, compared to the analysis in traditional sub-6 GHz networks. Authors in [38, 65, 67] take advantage of "external localization service" to get the position information of UE provided by e.g. using GPS. Deviated from most of these works, where cell discovery is only applied to LOS paths, we also take into account establishing NLOS links.

We assume that buildings act as propagation blockages for the broadcast of cell discovery signals from the AP. Based on that, the link state between AP and UE is determined to be either LOS or NLOS by considering whether any buildings intersect the direct path between Tx (AP) and Rx (UE). To incorporate the LOS/NLOS state into our system model, we adopt the d_1/d_2 model in [75, 76, 77], where the link between Tx and Rx at distance d in meters is determined to be LOS or NLOS according to the LOS probability $p(d)$:

$$p(d) = \min\left(\frac{d_1}{d}, 1\right)(1 - e^{-d/d_2}) + e^{-d/d_2}. \tag{2.1}$$

The pathloss for the beam formed by the AP to broadcast cell discovery signals

is given by [78, 79]:

$$L(f,d) = 20\log_{10}\frac{4\pi f}{c} + 10n_{\mathrm{L}}\log_{10} d + \mathrm{SF}, \tag{2.2}$$

where f is the carrier frequency in Hz, n_{L} is the pathloss exponent, and d is the distance between Tx and Rx in meters. The first term of (2.2) indicates the free space pathloss at 1 m, where c is the speed of light. The impact of objects such as trees, cars, etc. is not represented in the blockage model and is modeled separately using the shadowing factor (SF) in dB. Note that n_{L} and SF are different for LOS and NLOS links as addressed in Table 2.1. Further, an additive white Gaussian noise (AWGN) channel is assumed within each beam.

The small-scale fading effect is assumed to be Rayleigh fading, where each link is subject to an independent and identically distributed (i.i.d.) exponentially distributed fading power with unit mean. Compared to more realistic small-scale fading models such as Nakagami-m fading, Rayleigh fading leads to much more tractable results with very similar design insights [80, 81].

2.2.3 Antenna and Beamforming Model

We assume that the AP employs steerable antenna arrays of N_{Tx} antennas to support directional communications and can perform both 2D and 3D beamforming. The AP is equipped with M_{Tx} RF chains, such that multiple simultaneous beams can be transmitted. In case of 3D beamforming, Tx beams are scanning over both horizontal and vertical directions with ranges in $[0, 2\pi]$ and $[0, \arctan\left(\frac{r}{h_{\mathrm{AP}}-h_{\mathrm{UE}}}\right)]$, respectively. Here, h_{AP} and h_{UE} represent the height of AP antennas and the height of the access plane of UE, respectively, and r refers to the cell radius introduced in Section 2.2.1. The UEs are assumed to be able to synthesize a quasi-omnidirectional antenna pattern (e.g. as in 802.11ad [49]) for signal reception. 5G NR has discussed the time-division multiplexing of synchronization signals, where UE also exploits directional antenna pattern for the reception of cell discovery signal [47, 48]. In this regard, we also provide a performance analysis of cell discovery for beam-scanning reception at

UE (see Section 2.4.5). Besides this, throughout this chapter we generally assume an omnidirectional signal reception.

Similar to [44, 68], we assume each AP has a codebook of $N = N_{\mathrm{H}} \times N_{\mathrm{V}}$ possible beamforming vectors, which corresponds to N sectorized beam patterns that have non-overlapping main lobes. Here, N_{H} and N_{V} indicates the number of beamforming vectors in horizontal and vertical dimension, respectively. Specifically, the n_{H}-th horizontal beam ($1 \leq n_{\mathrm{H}} \leq N_{\mathrm{H}}$) and n_{V}-th vertical beam ($1 \leq n_{\mathrm{V}} \leq N_{\mathrm{V}}$) cover a sector area centered at the AP, whose angle is within $[(n_{\mathrm{H}}-1)\theta, n_{\mathrm{H}}\theta]$ and $[(n_{\mathrm{V}}-1)\phi, n_{\mathrm{V}}\phi]$, respectively, where θ and ϕ represent the horizontal and the vertical beamwidth, respectively, and are calculated as $\theta = 2\pi/N_{\mathrm{H}}$ and $\phi = \arctan\left(\frac{r}{h_{\mathrm{AP}}-h_{\mathrm{UE}}}\right)/N_{\mathrm{V}}$, respectively.

In case multiple RF chains are equipped, the AP is enabled to perform a spatial search for the broadcast of cell discovery signal with multiple simultaneous beams. In this situation, the main lobe of the desired beam could interfere with the side lobe of other beams transmitted in parallel, and correspondingly the main lobe suffers from inter-beam interference. The side lobe gain is acquired from realistic 16×8 uniform rectangular transmit antenna arrays and will be used for calculating the inter-beam interference (see details in Section 2.3.3) in the system-level simulations in Section 2.5.

For analytical tractability, we assume that the actual antenna pattern is approximated by a sectorized beam pattern ([13, 22, 23, 55, 68, 69, 71, 81]). As depicted in Fig. 2.3, The Tx antenna pattern is approximated by a rectangular sectorized pattern [82], and the beamforming gain G can be calculated as

$$G = \frac{\text{Area of isotropic sphere}}{\text{Area of rectangular}} = \frac{4\pi d^2}{d_{\mathrm{H}} d_{\mathrm{V}}} = \frac{4\pi}{\sin\theta \sin\phi}, \tag{2.3}$$

where the approximated sectorized area (the rectangular) of height d_{H} and width d_{V} can be obtained as $d_{\mathrm{H}} = d\sin\theta$ and $d_{\mathrm{V}} = d\sin\phi$, wherein d is the radius of isotropic sphere of the cell ($d = r$ at cell edge).

The antenna gain is essential for the analysis of cell discovery. Specifically, narrow

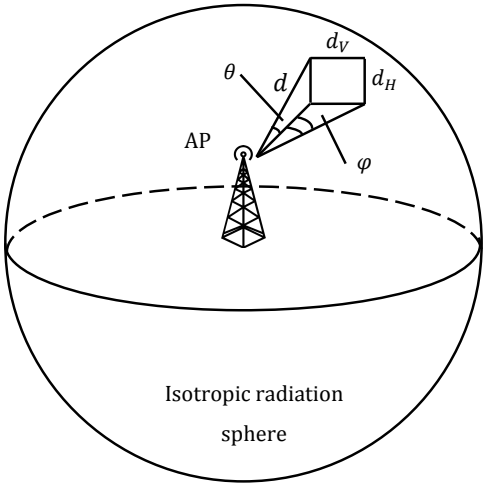


Figure 2.3: Tx beamforming gain is calculated by a rectangular sectorized pattern.

beams, where the values of θ and ϕ decrease, reduce the value of the denominator in (2.3) and lead to relatively higher beamforming gain that eventually increases the experienced SNR at Rx. Correspondingly, less transmission duration is required for delivering the same amount of information from the AP to UEs. This duration, which we refer to as *beam duration*, is a key metric for the performance analysis of cell discovery. More details are provided in Section 2.2.4.

2.2.4 Frame Structure and Beam Scan Model

In this chapter, a time-division duplex (TDD) mm-wave frame structure in Fig. 2.4 is considered, where the system time is divided into consecutive frames with period T [49, 51]. Each cycle begins with a beacon phase followed by a data transmission phase. In the beacon phase, the AP broadcasts cell discovery information via beam scan over different beam slots. The entire cell is covered by N beam scan areas ($1 < N \leq N_{\mathrm{Tx}}$) that correspond to the N non-overlapping beam directions of the sectorized beam patterns. Assuming multiple RF chains are enabled, the AP forms M simultaneous beams ($1 < M \leq M_{\mathrm{Tx}}$, i.e., limited by the number of RF chains at the Tx) to successively scan these areas. Then, the number of slots, where M out of total N non-overlapping beam scan areas are scanned by M simultaneous beams

in each slot, can be denoted and obtained as $S = \frac{N}{M}$ under the assumption that M is a divisor of N.

To broadcast the cell discovery signal to UEs in each scan area, the corresponding beam that scans this area maintains a certain duration for delivering the cell discovery information. We assumed that in each slot, the simultaneously formed beams will maintain identical duration, which is the beam duration mentioned in the previous section, to deliver the information to UEs located in the corresponding areas. Note that these slots are separated by GIs reserved for beam switching in the case of hybrid or analog beamforming [83]. These GIs are specifically used for beam switching and other GIs reserved for e.g., (orthogonal frequency-division multiplexing) OFDM symbol, physical channels, and downlink/uplink transmission, are also considered and not addressed here.The AP periodically scans the cell via angular probing in the beacon phase within each frame, and maintains the order of the beams among different frames, namely the beam scans the same area of the cell during the same slot of each frame.

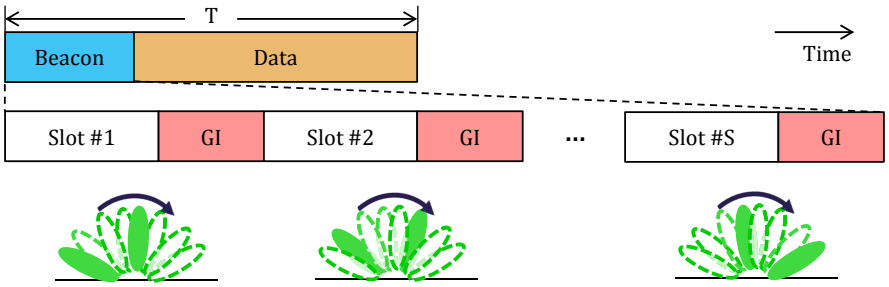


Figure 2.4: Frame structure including a beam scan phase.

As shown in Fig. 2.4 as an example, the entire cell area is covered by eight beam scan areas, where within slot #1, two Tx beams are scanning the two flat oval-shaped green-filled areas (first and fifth beam), and the duration of the scan is the beam duration. Then, the beam rotates and scans the subsequent areas clockwise, two-by-two, with the same duration in each scan area pairs. Obviously in the illustrated example, the beacon phase is partitioned into four slots to cover the entire cell (namely $S = 4$), and within each slot, two beams are formed to convey

the cell discovery information to UEs in the corresponding areas[2]. It is worth noting that two simultaneous beams are explicitly exploited in the example. In case single beam exhaustive scan is applied, the beacon phase would be partitioned into eight slots wherein UEs located in each scan area are covered by a single beam during the corresponding slot.

2.3 Broadcast Schemes Design for Cell Discovery

In this section, we will briefly review the cell discovery in LTE and then investigate the design of broadcast schemes for mm-wave beamformed cell discovery. Apart from single beam exhaustive scan, multiple beam simultaneous scan is applied to exploit the potential benefit of spatial multiplexing. The notations and parameters used in this section are given in Table 2.2 and will be used in the rest of this chapter.

Table 2.2: Broadcast Schemes Design Parameters

Notation	Description	Value
SG	Spreading gain of CD in dB scale	$10\log_{10} M$
I_{B}	Inter-beam interference	See (2.8)
$\mathrm{SNR}_{\mathrm{TD}}$, $\mathrm{SNR}_{\mathrm{FD}}$, $\mathrm{SNR}_{\mathrm{CD}}$, $\mathrm{SNR}_{\mathrm{SD}}$	Receive SNR of scheme TD, FD, CD, and SD	See (2.4)–(2.8)
P	Total Tx power	30 dBm
η	Thermal noise power	-77 dBm
M_{dB}	Multiplexing gain in dB scale	$10\log_{10} M$
B	Bandwidth	1 GHz
U	Amount of cell discovery information	128 bit
t	Beam duration	See (2.9)
n	Finite blocklength	See (2.9)
ϵ	Block error rate	10^{-3}
C	Capacity of AWGN channel	See (2.10)
V	Channel dispersion	See (2.11)
$Q(.)$	Complementary Gaussian cumulative distribution function	See (2.9)
t_{TD}, t_{FD}, t_{CD}, t_{SD}	Beam duration of scheme TD, FD, CD, and SD	See (2.17)–(2.18)
t_{GI}	Guard interval duration	See (2.20)

[2]In the illustrated example, the cell is partitioned in $N = N_{\mathrm{H}} \times N_{\mathrm{V}}$ ($8 = 8 \times 1$) scan areas where the beams scan over these areas only in horizontal direction. However, the beam scan areas can be partitioned as e.g. $8 = 4 \times 2$ where four areas are close to the AP and the remaining four are farther away from the AP. In this case, the beams scan over both vertical and horizontal directions (e.g., near to far in vertical direction under the same vertical beamwidth, then clockwise in horizontal direction under the same horizontal beamwidth).

2.3.1 Cell Discovery in LTE

The main steps for initial access in LTE include scan, measurement, evaluation, and detection [84]. Specifically, by tuning to a specific frequency and measuring the reference signal received power (RSRP), a UE is able to use downlink synchronization channels to find the primary synchronization signal (PSS) and detect slot timing information and PHY ID. Afterwards, UE processes the secondary synchronization signal (SSS) to acquire radio frame timing information, as well as other identities such as cyclic prefix (CP) length and TDD/frequency-division duplex (FDD) ID. However, the cell discovery of LTE is performed omnidirectionally, which cannot be directly applied to mm-wave networks due to the mismatch of discoverable range and achievable range in mm-wave frequencies as addressed in Section 2.1.

2.3.2 Schemes for Broadcast of Cell Discovery Information

To harvest multiplexing capability, different broadcast schemes are investigated to enable AP to broadcast cell discovery information. The schemes considered in this chapter are described as follows:

1. Time-division (TD): AP scans through all possible beamforming directions in its cell area with a single beam at a time (i.e., $M = 1$). We refer to this scheme as the baseline design, which is especially suitable for Tx with a single RF chain.
2. Frequency-division (FD): If multiple RF chains are available, more than one beam can be simultaneously formed ($1 < M \leq N$). In contrast to the baseline, this scheme exploits multiple simultaneous beams multiplexed in frequency domain. Here, we assume that the total available bandwidth and transmission power are equally allocated on each beam.
3. Code-division (CD): Similar to FD, in this scheme multiple simultaneous beams are exploited, yet multiplexed in codes ($1 < M \leq N$). We apply orthogonal codes on the multiplexed beams where a spreading gain (SG) is expected.

Here, CD achieves the SG at the expense of more time (M times more). The performance of CD with orthogonal codes are discussed in Section 2.3.3 and the reason of selecting orthogonal codes are addressed in Appendix 2.7.1. In addition, the total transmission power is also equally allocated on each beam.

4. Space-division (SD): A general scheme, or an extension to TD, is proposed here as SD ($1 < M \leq N$). In this scheme, multiple simultaneous beams are formed at the same time. As mentioned in Section 2.2.3, the main lobe of the desired beam suffers from the inter-beam interference referred to as I_{B}.

A summary of these schemes is provided in Table 2.3.

Table 2.3: Broadcasting Schemes

Scheme	Multiplexing	N (Number of beam scan areas)	M (Number of simultaneous beams)
TD	Time	$1, 2, 4, \ldots, 128$	1
FD	Space, time, frequency	$1, 2, 4, \ldots, 128$	Divisor of N
CD	Space, time, code	$1, 2, 4, \ldots, 128$	Divisor of N
SD	Space, time	$1, 2, 4, \ldots, 128$	Divisor of N

2.3.3 SNR and Beam Duration Analysis for Broadcast Schemes

The receive SNR at UE for the broadcast schemes, based on their characteristics summarized in Section 2.3.2 and Table 2.3, are derived as follows.

When applying the baseline broadcast scheme, namely TD, the AP forms a single beam to transmit cell discovery signal. Therefore, the total transmission power P will be allocated to the beam. Then, the receive SNR of TD at UE side, denoted as $\mathrm{SNR}_{\mathrm{TD}}$, satisfies

$$\mathrm{SNR}_{\mathrm{TD}} = P + G - L - \eta, \tag{2.4}$$

where P, G, L, and η represent transmission power, antenna gain, pathloss, and thermal noise power of the scanning beam, respectively, all in dB scale.

For the other three broadcast schemes, multiple beams are simultaneously formed at the AP to transmit cell discovery signal to UEs located at different scan areas (beamforming directions) of the cell. In this way, the total transmission power P is

averaged at each simultaneously formed beam, as addressed in 2.3.2. More advanced power allocation schemes aiming to improve the system performance in terms of e.g. data rate and energy efficiency, are also possible and left as future work.

Assuming M simultaneous beams are employed at the AP, the receive SNR of FD at UE side, denoted as $\mathrm{SNR}_{\mathrm{FD}}$, is derived as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{FD}} &= P - M_{\mathrm{dB}} + G - L - (\eta - M_{\mathrm{dB}}) \\ &= P + G - L - \eta, \end{aligned} \tag{2.5}$$

Here, $P - M_{\mathrm{dB}}$ indicates the power split on each of the M simultaneously transmitted beams in dB scale. Due to frequency multiplexing, the total system bandwidth B is also averaged at each beam, which reduces the experienced thermal noise at UE side by also M_{dB} in dB scale. Therefore, the term $\eta - M_{\mathrm{dB}}$ indicates the noise power in $\frac{1}{M}$-th bandwidth (frequency band per beam in FD).

Similar to FD, the total transmission power P is averaged on M simultaneous beams for CD. However, instead of multiplexing in frequency, CD exploits multiple access by different codes, where the total bandwidth B is allocated to each beam. When applying orthogonal codes to CD, the spreading gain SG equals to the number of orthogonal codes [85], namely $\mathrm{SG} = M_{\mathrm{dB}}$ in dB scale. Hence, the receive SNR of CD at UE side, denoted as $\mathrm{SNR}_{\mathrm{CD}}$, can be written as

$$\begin{aligned}\mathrm{SNR}_{\mathrm{CD}} &= P - M_{\mathrm{dB}} + \mathrm{SG} + G - L - \eta \\ &= P + G - L - \eta. \end{aligned} \tag{2.6}$$

It is obvious from (2.5) and (2.6) that in terms of SNR, FD and CD perform exactly the same as TD, i.e.,

$$\mathrm{SNR}_{\mathrm{TD}} = \mathrm{SNR}_{\mathrm{FD}} = \mathrm{SNR}_{\mathrm{CD}}. \tag{2.7}$$

Actually, for a frequency flat channel, achieving orthogonality in time, frequency, or in any other multiplexing dimension, is "exactly" the same in terms of spectral efficiency and SNR, therefore frequency-division multiple access (FDMA) and orthog-

onal code-division multiple access (CDMA) perform exactly the same as TDMA.

The analysis of SNR for SD is a bit tricky. In addition to the transmission power split, which is similar to FD and CD, no multiplexing technique is explicitly applied to SD. As the cell discovery information conveyed in simultaneously transmitted beams are partially identical (e.g. cell ID, slot and frame timing indication, etc.), these beams, which lose orthogonality after transmission due to multi-path channel, will suffer from inter-beam interference. Therefore, the receive SNR of SD at UE side, denoted as $\mathrm{SNR}_{\mathrm{SD}}$, is calculated as

$$\begin{aligned} \mathrm{SNR}_{\mathrm{SD}} &= P - M_{\mathrm{dB}} + G - I_{\mathrm{B}} - L - \eta \\ &\leq \mathrm{SNR}_{\mathrm{TD}}, \end{aligned} \tag{2.8}$$

where I_{B} refers to the inter-beam interference.

As mentioned in Section 2.2.3 and 2.2.4, the beam duration is defined as the duration required for the AP to deliver cell discovery information U in bits to a UE (e.g., cell ID, beam ID, etc.[3]). The beam duration also represents the length of each beam slot in the frame structure illustrated in Fig. 2.4. Within each beam duration, one or multiple beams are formed by the AP to broadcast the cell discovery information, where each beam corresponds to one beam scan area (see Fig. 2.4). In the next slots, the next group of beams (one or multiple) are formed by the AP to broadcast the information to UEs in their corresponding beam scan areas, and also have the same duration.

The beam duration, denoted as t, is derived according to the achievable rate of AWGN channel with bandwidth B, finite blocklength n, and block error rate ϵ as follows [86]:

$$t = \frac{U}{B\left(C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + \frac{\log_2 n}{n}\right)}, \tag{2.9}$$

where C indicates the channel capacity. The term V is referred to as channel

[3]Frequency and time synchronization requirements has not been treated in this analysis and the incorporation of the requirements can be put on top of our result and remains as future work.

dispersion in [86] that characterizes the backoff from channel capacity, in the finite blocklength regime, and measures the stochastic variability of the channel relative to a deterministic channel with the same capacity. These two parameters satisfy

$$C = \log_2(1 + \mathrm{SNR}), \tag{2.10}$$

and

$$V = \frac{\mathrm{SNR}(\mathrm{SNR}+2)}{(\mathrm{SNR}+1)^2} \log_2^2 e, \tag{2.11}$$

where e is the natural logarithm constant. $Q(.)$ is the complementary Gaussian cumulative distribution function. Note that here the derivation of the beam duration t from the blocklength n is implicitly included in (2.9), where given the certain amount of data U, the channel capacity C, and the channel dispersion V, we can obtain the blocklength n [86] and further the beam duration t.

Taking (2.4) into (2.9), we can get the beam duration of TD, denoted as t_{TD}, as follows:

$$t_{\mathrm{TD}} = \frac{U}{B\Big(C_{\mathrm{TD}} - \sqrt{\frac{V_{\mathrm{TD}}}{n}} Q^{-1}(\epsilon) + \frac{\log_2 n}{n}\Big)}, \tag{2.12}$$

where

$$C_{\mathrm{TD}} = \log_2(1 + \mathrm{SNR}_{\mathrm{TD}}), \tag{2.13}$$

and

$$V_{\mathrm{TD}} = \frac{\mathrm{SNR}_{\mathrm{TD}}(\mathrm{SNR}_{\mathrm{TD}}+2)}{(\mathrm{SNR}_{\mathrm{TD}}+1)^2} \log_2^2 e. \tag{2.14}$$

As TD, SD, and CD perform exactly the same in achievable SNR, the beam duration of FD and CD, denoted as t_{FD} and t_{CD}, respectively, can be derived as

$$\begin{aligned} t_{\mathrm{FD}} &= \frac{U}{\frac{B}{M}\Big(C_{\mathrm{FD}} - \sqrt{\frac{V_{\mathrm{FD}}}{n}} Q^{-1}(\epsilon) + \frac{1}{2n}\log_2 n\Big)} \\ &= M \cdot \frac{U}{B\Big(C_{\mathrm{TD}} - \sqrt{\frac{V_{\mathrm{TD}}}{n}} Q^{-1}(\epsilon) + \frac{1}{2n}\log_2 n\Big)} \\ &= M t_{\mathrm{TD}}, \end{aligned} \tag{2.15}$$

and

$$\begin{aligned} t_{\mathrm{CD}} &= \frac{U}{\frac{B}{M}\Big(C_{\mathrm{CD}} - \sqrt{\frac{V_{\mathrm{CD}}}{n}}Q^{-1}(\epsilon) + \frac{1}{2n}\log_2 n\Big)} \\ &= M \cdot \frac{U}{B\Big(C_{\mathrm{TD}} - \sqrt{\frac{V_{\mathrm{TD}}}{n}}Q^{-1}(\epsilon) + \frac{1}{2n}\log_2 n\Big)} \\ &= M t_{\mathrm{TD}}. \end{aligned} \tag{2.16}$$

Here, the term $\frac{B}{M}$ in (2.15) and (2.16) indicates the allocated $\frac{1}{M}$-th bandwidth on each simultaneously formed beam of FD and the M-fold increased code length of CD, respectively.

In summary, the beam durations of FD and CD are identical and are M fold of the beam duration of TD, i.e.,

$$t_{\mathrm{FD}} = t_{\mathrm{CD}} = M t_{\mathrm{TD}}. \tag{2.17}$$

For SD, similar derivation can be conducted, where the beam duration of SD, denoted as t_{SD}, can be written as

$$\begin{aligned} t_{\mathrm{SD}} &= \frac{U}{B\Big(C_{\mathrm{SD}} - \sqrt{\frac{V_{\mathrm{SD}}}{n}}Q^{-1}(\epsilon) + \frac{1}{2n}\log_2 n\Big)} \\ &\overset{(a)}{\geq} \frac{U}{B\Big(C_{\mathrm{TD}} - \sqrt{\frac{V_{\mathrm{TD}}}{n}}Q^{-1}(\epsilon) + \frac{1}{2n}\log_2 n\Big)} \\ &= t_{\mathrm{TD}}. \end{aligned} \tag{2.18}$$

The proof of (a) is provided in Appendix 2.7.2.

It can be observed from (2.17) and (2.18) that the four schemes in Section 2.3.2 could lead to different beam durations. In the next section, we will develop and verify a general analytical framework that can quantify the impact of various broadcast schemes and their corresponding beam durations on the mm-wave cell discovery performance.

2.4 Cell Discovery Analysis

The objective of this section is to characterize the performance of mm-wave cell discovery for the four broadcast schemes introduced in Section 2.3. The performance metrics, namely the CDL and CDO, are introduced in Section 2.4.1. In Section 2.4.2, we develop an analytical framework for the cell discovery of non-standalone networks to characterize the introduced performance metrics. In addition, we analyze the framework for the cell discovery of standalone networks and consider various frame structure designs, in Section 2.4.3 and Section 2.4.4, respectively. The performance analysis of cell discovery with beam-scanning reception is addressed Section 2.4.5. Under the framework for the standalone network, the analysis reveals succinct characterizations of CDL and CDO.

2.4.1 Performance Metrics

As introduced in Section 2.3.1, cell discovery is a basic prerequisite to any communications and is an essential component of the initial access procedure. For a UE to detect the presence of an AP, the cell periodically broadcasts discovery information via angular probing in the beacon phase of a frame as introduced in Section 2.2.4 and illustrated in Fig. 2.4.

On the one hand, operating at extremely high frequencies and with wide bandwidths may quickly drain a UE's battery, thus the UE experiences a transition from connected state (active) to idle mode (inactive). On the other hand, a UE may fail in cell discovery when its connected state and a corresponding scanning beam are mismatched. In other words, a UE should be active before a beam scans its located area. Otherwise, it misses the beam scan phase within the current frame (say frame 1) and has to wait until a beam reaches its located area in the beacon phase of next frame (frame 2). Then, in frame 2, the UE tries to "catch" the beam and decodes the cell discovery information, and if there is a decoding error, with the probability ϵ, the discovery procedure in frame 2 fails, and the UE must wait for another round

of beam scan phase (a complete frame) and try to receive discovery information in frame 3. Obviously, the cell discovery process is under geometric distribution.

Based on this, we define the CDL as the duration between the time when UE is active (z_{active}), and the time when it successfully decodes the complete cell discovery information. The CDO is referred to as the portion of the beacon phase in one frame, which depicts the burden of cell discovery. Without loss of generality, we assume z_{active} of a UE is uniformly distributed in a random frame[4], which corresponds to the realistic scenario that UE could be active at any time. Then, the UE keeps active until the end of a successful cell discovery.

It is intuitively that the smallest CDL refers to the case that a UE is active exactly at the time when a beam starts to scan its located area, where the CDL is just the beam duration t. By contrast, the largest CDL refers to the situation that a UE fails to detect the presence of a cell within K frames (with the probability $\epsilon^K \to 0$, when $K \to \infty$). Here, $K \to \infty$ indicates the fact that all UEs eventually succeed in cell discovery.

In the rest of the section, we provide a precise analytical frameworks to study a more general case, in which UEs are arbitrarily active at a frame and there is no specific relation between z_{active} and the time that a beam starts to scan the located area of UEs. We are interested in which factors are key to characterize the average CDL and CDO. It is worth noting that in Section 2.4.2 and 2.4.3, the normal frame structure illustrated in Fig. 2.4 is applied to the cell discovery analysis. The other design options of frame structure, such as a long frame containing multiple beacon phases, CDO-sensitive case where a normal frame containing a beacon phase is followed by several consecutive data-only frames, and the case where a complete beacon is equally separated into several consecutive frames, are studied in Section 2.4.4.

[4]Note that not all UEs are necessarily active at the same frame.

2.4.2 Cell Discovery for Non-Standalone Networks

Synchronization between AP and UE for a non-standalone network could be achieved by the well-known control plane and user plane split, where the timing is obtained via a macrocell operating at legacy bands. The split allows e.g. reduced latency on application service through selecting user plane nodes that are closer to RAN and enables software defined networking to deliver user plane data more efficiently.

In this case, all UEs are instructed to be active aligning to the beginning of a frame. Without loss of generality, we denote the unique beam slot, corresponding to the scan area in which UE is located, as slot j, which means that in slot j, AP forms a beam to transmit cell discovery information to UEs in this area. As UE is uniformly "dropped" in the cell that is covered by beams scan with totally S slots (each slot corresponds to a scan area depicted in Fig. 2.4, and S is the number of beam slots as summarized in Table 2.1), the probability of the event "UE is located in the area corresponding to slot j", denoted as $p_{\mathrm{O}}(j)$, can be written as

$$p_{\mathrm{O}}(j) = \frac{1}{S}. \tag{2.19}$$

For simplicity, in the rest of the chapter we represent the event "UE is located in the area corresponding to slot j" as a short-term "located in slot j".

For this event, UE attempts to decode the delivered information with the error rate ϵ in K frames. In this case, the average latency of the event, denoted as $t_{\mathrm{O}}(j)$, is given as

$$t_{\mathrm{O}}(j) = \sum_{k=0}^{K-1} \Big(\big(kT + (j-1)t' + t\big)\epsilon^k(1-\epsilon) + KT\epsilon^K \Big), \tag{2.20}$$

where $t' = t + t_{\mathrm{GI}}$. K indicates the number of frames used for cell discovery of the UE. The term $(j-1)t' + t$ indicates the latency in the successful frame, which includes $j-1$ beam slots waited by the UE from the beginning of the frame until a beam starts to scan its location, and the decoding duration t. Then, the average CDL for the non-standalone network is given by the following theorem.

Theorem 2.4.1. *Given the probability and the average latency of the event "UE*

located in slot j discovers the cell within K frames" as $p_{\mathrm{O}}(j)$ and $t_{\mathrm{O}}(j)$, respectively, the average CDL for a non-standalone network, denoted as $\overline{T}_{\textit{non-standalone}}$, over the uniform distribution of j from 1 to S, is given by

$$\overline{T}_{\textit{non-standalone}} = \big(1-\epsilon^K\big)\Big(\frac{S+1}{2}t + \frac{S-1}{2}t_{\mathrm{GI}} + \frac{\epsilon}{1-\epsilon}T\Big). \tag{2.21}$$

When $K \to \infty$, which indicates the fact that UE eventually succeeds in cell discovery, the average CDL can be further simplified to

$$\overline{T}_{\textit{non-standalone}} = \frac{S+1}{2}t + \frac{S-1}{2}t_{\mathrm{GI}} + \frac{\epsilon}{1-\epsilon}T. \tag{2.22}$$

Proof. The proof is provided in Appendix 2.7.3. □

In short, (2.22) is explained by noticing that if there is a decoding error, with the probability ϵ, the discovery procedure fails, and the UE must wait for a whole round and try again, assuming that all errors can be revealed.

2.4.3 Cell Discovery for Standalone Networks

In fact, it is not viable to get a very accurate time and frequency synchronization from legacy bands for mm-wave, as the time domain and frequency domain experience very different granularities [87]. Accordingly, only rough synchronization is expected to be achieved in non-standalone networks. In addition, the request of cell discovery for a UE emerges randomly in practice, which motivates the CDL to be counted exactly from the emergence of the request to the moment of successful discovery. Therefore, in this section we address the cell discovery for standalone network taking into account the UE active time stamp z_{active}.

As a UE is arbitrarily active within a frame, it could be active at the beacon phase or at the data phase. For the former case, there exist three scenarios, denoted as scenario A to C that are elaborated in the remaining of this section, in which a UE could be active before, when, and after a beam scans its location, respectively. Without loss of generality, we assume that the UE is active at slot i and located in slot j (the scan area corresponding to slot j). The latter case is referred to as

scenario D. These four scenarios are illustrated in Fig. 2.5. We further refer to the frame that UE is active at as "the current frame".

It is worth noting that a UE has a chance to finish cell discovery in the current frame only in scenario A, in case it successfully decodes the conveyed information. The reason behind this lies in the fact that only if UE is active before a beam scans its located area, it has the opportunity to receive the complete information. For scenario B to D, UE has to wait for at least one round of scan and tries to detect the cell in the next frames. In the following, we will describe the four scenarios and analyze their specific average CDLs in details. The corresponding CDOs of are derived afterwards. It is obvious that the CDL of scenario A, B, C, and D are different, as the number of slots between the slot a UE is active at (slot i) and the slot a UE is located in (slot j) varies across different scenarios.

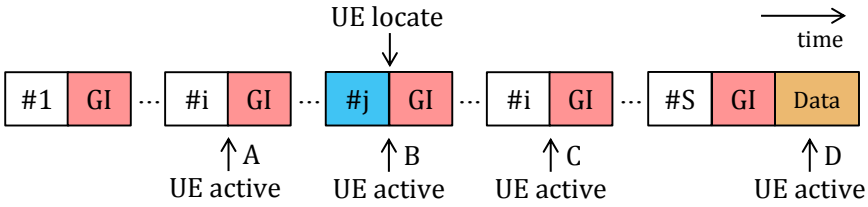


Figure 2.5: Illustration of cell discovery scenarios for standalone networks.

2.4.3.1 Scenario A – UE Becomes Active Before Beam Scans Its Location

In this scenario we have $j > i$. Therefore, UE can "catch" the beam to decode when the beam is scanning its location in the current frame. The probability of scenario A, denoted as $p_{\mathrm{A}}(i)$, which is the event "UE is active at slot i and located in slot j from $i+1$ to S", is written as

$$p_{\mathrm{A}}(i) = p(\text{active at slot } i) \cdot p(\text{located in slot } j) = \frac{t'}{T} \cdot \frac{S-i}{S}. \tag{2.23}$$

Here, as z_{active} is uniformly distributed in the entire frame, the probability of UE active at slot i is $\frac{t'}{T}$. Similarly, due to the uniform distribution of UE location in the

entire cell, the probability of UE located in beam slot j is $\frac{S-i}{S}$. As the event "UE is active at slot i" and the event "UE is located in slot j from $i+1$ to S" are i.i.d, the probability of the joint distribution of event "UE is active at slot i and located in slot j from $i+1$ to S" equals to the product of the probability of the two individual events. For scenarios B to D, the analysis of the joint probability distribution are similar and omitted in Section 2.4.3.2, 2.4.3.3, and 2.4.3.4, respectively.

For this event, UE attempts to decode the conveyed information with the error rate ϵ in the current frame. Then, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot j from $i+1$ to S", denoted as $t_{\mathrm{A}}(i)$, is given by the following theorem.

Theorem 2.4.2. *Averaging over the uniform probability of j from $i+1$ to S, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot j from $i+1$ to S", denoted as $t_{\mathrm{A}}(i)$, is given by*

$$t_{\mathrm{A}}(i) = \left(\frac{t'}{2} + \frac{S-i-1}{2}t' + t\right)(1-\epsilon) + \left(\frac{t'}{2} + T - it' + (K-1)T\right)\epsilon^K + \sum_{k=1}^{K-1}\left(\frac{t'}{2} + T - it' + (k-1)T + \frac{S+i-1}{2}t' + t\right)\epsilon^k(1-\epsilon). \tag{2.24}$$

Proof. The proof is provided in Appendix 2.7.4. □

2.4.3.2 Scenario B – UE Is Active when Beam Is Scanning Its Location

In this scenario we have $j = i$. Therefore, UE is not able to "catch" the beam scanning its location to decode in the current frame. The probability of scenario B, denoted as $p_{\mathrm{B}}(i)$, which is the event "UE is active at slot i and located in slot $j = i$", is written as

$$p_{\mathrm{B}}(i) = p(\text{active at slot } i) \cdot p(\text{located in slot } j) = \frac{t'}{T} \cdot \frac{1}{S}. \tag{2.25}$$

Here, as z_{active} is uniformly distributed at the entire frame, the probability of UE active at slot i is $\frac{t'}{T}$. Similarly, due to the uniform distribution of the UE location

in the entire cell, the probability of UE located in beam slot j is $\frac{1}{S}$.

For this event, UE fails in cell discovery within the current frame, as it is not able to decode the conveyed information within a complete beam duration (we assume this can only be done when UE is active at exactly the beginning of the scanning slot[5]). Therefore, UE attempts to decode the conveyed information with the error rate ϵ in the next frames. Then, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot $j = i$", denoted as $t_{\mathrm{B}}(i)$, is given by the following theorem.

Theorem 2.4.3. *Averaging over the uniform probability of j from 1 to S, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot $j = i$", denoted as $t_{\mathrm{B}}(i)$, is given by*

$$t_{\mathrm{B}}(i) = \frac{t'}{2} + T - it' + \sum_{k=0}^{K-2} \big(kT + (i-1)t' + t\big)\epsilon^k(1-\epsilon) + (K-1)T\epsilon^{K-1}. \quad (2.26)$$

Proof. The proof is provided in Appendix 2.7.5. □

2.4.3.3 Scenario C – UE Is Active After Beam Has Scanned Its Location

In this scenario we have $j < i$. Therefore, similar to scenario B introduced in Section 2.4.3.2, UE is also not able to "catch" the beam scanning its location in the current frame. The probability of scenario C, $p_{\mathrm{C}}(i)$, which is the event "UE is active at slot i and located in slot j from 1 to $i-1$", is written as

$$p_{\mathrm{C}}(i) = p(\text{active at slot } i) \cdot p(\text{located in slot } j) = \frac{t'}{T} \cdot \frac{i-1}{S}. \quad (2.27)$$

Here, as z_{active} is uniformly distributed at the entire frame, the probability of UE active at slot i is $\frac{t'}{T}$. Similarly, due to the uniform distribution of the UE location in the entire cell, the probability of UE located in beam slot j is $\frac{i-1}{S}$.

For this event, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot j from 1 to $i-1$",

[5]Just for the case that the UE experiences high SNR and can decode even from just a part of the receive signal.

denoted as $t_{\mathrm{C}}(i)$, can be obtained similarly to scenario B is given by the following theorem.

Theorem 2.4.4. *Averaging over the uniform probability of j from 1 to $i-1$, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot j from 1 to $i-1$", denoted as $t_{\mathrm{C}}(i)$, is given by*

$$t_{\mathrm{C}}(i) = \frac{t'}{2} + T - it' + \sum_{k=0}^{K-2} \left(kT + \frac{i-2}{2}t' + t \right) \epsilon^k (1-\epsilon) + (K-1)T\epsilon^{K-1}. \quad (2.28)$$

Proof. The proof is provided in Appendix 2.7.6. □

2.4.3.4 Scenario D – UE Is Active at Data Phase

It is clear that in this scenario, UE is also not able to "catch" the beam to decode when the beam is scanning its location in the current frame, as it is active at the data phase. The probability of scenario D, p_{D}, which is the event "UE is active at data phase", is written as

$$p_{\mathrm{D}} = \frac{T - St'}{T}. \quad (2.29)$$

Here, as z_{active} is uniformly distributed in the entire frame, the probability of UE active at the data phase is $\frac{T-St'}{T}$.

For this event, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at data phase", denoted as t_{D}, can be obtained similarly to scenario B and C and is given by the following theorem.

Theorem 2.4.5. *Averaging over the uniform probability of j from 1 to S, the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at data phase", denoted as t_{D}, is given by*

$$t_{\mathrm{D}} = \frac{T - St'}{2} + \sum_{k=0}^{K-2} \left(kT + \frac{(S-1)}{2}t' + t \right) \epsilon^k (1-\epsilon) + (K-1)T\epsilon^{K-1}. \quad (2.30)$$

Proof. The proof is provided in Appendix 2.7.7. □

2.4.3.5 The Overall CDL

Combining the previous four scenarios, namely scenario A to D that covers all the relationships between the slot that UE is active at and the slot that UE is located in, the average CDL for standalone networks, denoted as $\overline{T}_{\text{standalone}}$, is given by the following theorem.

Theorem 2.4.6. *Denote the probability of the event "UE is active at slot i and located in slot j from $i+1$ to S", "UE is active at slot i and located in slot $j=i$", "UE is active at slot i and located in slot j from 1 to $i-1$", and "UE is active at data phase" as $p_{\mathrm{A}}(i)$, $p_{\mathrm{B}}(i)$, $p_{\mathrm{C}}(i)$, and p_{D}, respectively. Denote the average latency of the event "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot j from $i+1$ to S", "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot $j=i$", "UE makes discovery trials within K frames conditioned to UE is active at slot i and located in slot j from 1 to $i-1$", and "UE makes discovery trials within K frames conditioned to UE is active at the data phase" as $t_{\mathrm{A}}(i)$, $t_{\mathrm{B}}(i)$, $t_{\mathrm{C}}(i)$, and t_{D}, respectively. Then, the average CDL for a standalone network, denoted as $\overline{T}_{standalone}$, is given by*

$$\overline{T}_{standalone} = \sum_{i=1}^{S-1} p_{\mathrm{A}}(i)t_{\mathrm{A}}(i) + \sum_{i=1}^{S} p_{\mathrm{B}}(i)t_{\mathrm{B}}(i) + \sum_{i=2}^{S} p_{\mathrm{C}}(i)t_{\mathrm{C}}(i) + p_{\mathrm{D}}t_{\mathrm{D}}. \tag{2.31}$$

When $K \to \infty$, we have

$$\overline{T}_{standalone} = t + \frac{1+\epsilon}{2(1-\epsilon)}T. \tag{2.32}$$

Proof. The proof is provided in Appendix 2.7.8. □

This result is quite inspiring as it shows that the average CDL depends only on the beam duration t, the error probability ϵ, and the frame length T. If the frame length T is fixed, then the dependency of GI on the average CDL is "hidden", where different lengths of GI will not change the average CDL. We consequently argue that analog and hybrid beamforming perform as well as digital beamforming in the context of CDL. However, longer GI of hybrid beamforming would definitely lead to

higher CDO, which will be elaborated in Section 2.4.3.6.

In fact, we have considered the case in [16] that if a UE is not able to "catch" the beam in the current frame, it can successfully decode the information in the next frame. Specifically, the error probability ϵ for this case equals to 0. It is demonstrated in [16] that the average CDL, which is denoted as $\overline{T}'$ and describes the cell discovery for standalone networks with $\epsilon = 0$, can be written as

$$\overline{T}' = t + \frac{T}{2}. \tag{2.33}$$

When $\epsilon \neq 0$, the cell discovery for standalone networks turns to be geometrically distributed as addressed in 2.4.1, compared to a successful cell discovery in the second frame (the next frame of the current frame) when $\epsilon = 0$. In this regard, the average CDL of the cell discovery for the standalone network, $\overline{T}_{\text{standalone}}$, can be derived from $\overline{T}'$ as

$$\begin{aligned}
\overline{T}_{\text{standalone}} &= \sum_{k=0}^{\infty} (\overline{T}' + kT) \cdot \epsilon^k (1-\epsilon) \\
&= \sum_{k=0}^{\infty} \left(t + \frac{T}{2} + kT \right) \epsilon^k (1-\epsilon) \\
&= t + \frac{1+\epsilon}{2(1-\epsilon)} T,
\end{aligned} \tag{2.34}$$

which means that the trial of cell discovery, with the average latency $\overline{T}'$ and the error probability ϵ, will be kept k times until the $k+1$ trial is successful.

2.4.3.6 CDL and CDO for Different Broadcast Schemes

Considering the analysis of different broadcast schemes in Section 2.3, the average CDL for standalone networks of TD, FD, CD, and SD, denoted as $\overline{T}_{\text{TD}}$, $\overline{T}_{\text{FD}}$, $\overline{T}_{\text{CD}}$, and $\overline{T}_{\text{SD}}$, respectively, can be written as

$$\overline{T}_{\text{TD}} = t_{\text{TD}} + \frac{1+\epsilon}{2(1-\epsilon)} T, \tag{2.35}$$

and

$$\begin{aligned}\overline{T}_{\mathrm{FD}} = \overline{T}_{\mathrm{CD}} &= M\mathrm{t}_{\mathrm{TD}} + \frac{1+\epsilon}{2(1-\epsilon)}T \\ &\geq \overline{T}_{\mathrm{SD}} = \mathrm{t}_{\mathrm{SD}} + \frac{1+\epsilon}{2(1-\epsilon)}T \\ &\geq \overline{T}_{\mathrm{TD}}. \end{aligned} \tag{2.36}$$

Similarly, the CDO for standalone networks of TD, FD, CD, and SD, denoted as $\mathrm{CDO}_{\mathrm{TD}}$, $\mathrm{CDO}_{\mathrm{FD}}$, $\mathrm{CDO}_{\mathrm{CD}}$, and $\mathrm{CDO}_{\mathrm{SD}}$, respectively, are provided by

$$\mathrm{CDO}_{\mathrm{TD}} = \frac{(t_{\mathrm{TD}} + t_{\mathrm{GI}})S_{\mathrm{TD}}}{T} = \frac{N(t_{\mathrm{TD}} + t_{\mathrm{GI}})}{T}, \tag{2.37}$$

and

$$\begin{aligned}\mathrm{CDO}_{\mathrm{SD}} &= \frac{(t_{\mathrm{SD}} + t_{\mathrm{GI}})S_{\mathrm{SD}}}{T} \\ &= \frac{\frac{N}{M}(t_{\mathrm{SD}} + t_{\mathrm{GI}})}{T} \\ &\leq \frac{Nt_{\mathrm{TD}} + \frac{N}{M}t_{\mathrm{GI}}}{T} = \mathrm{CDO}_{\mathrm{FD}} = \mathrm{CDO}_{\mathrm{CD}} \\ &\leq \frac{N(t_{\mathrm{TD}} + t_{\mathrm{GI}})}{T} = \mathrm{CDO}_{\mathrm{TD}}. \end{aligned} \tag{2.38}$$

To summarize, we argue that:

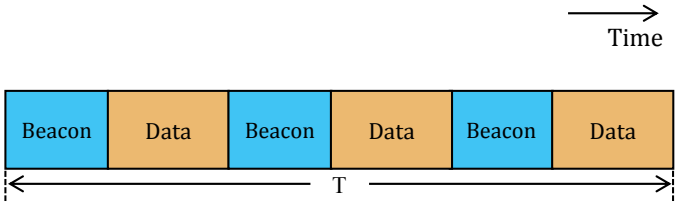
- The average CDL depends only on beam duration, error rate, and frame length.
- Single beam exhaustive scan (TD) outperforms all broadcast schemes in terms of CDL but results in higher CDO, in case the frame length is fixed and a complete beam scan can be finished within one frame.
- Multiple beam simultaneous scan (FD/CD/SD) can significantly reduce CDO and provide the flexibility to achieve a trade-off between CDL and CDO.

2.4.4 Cell Discovery for Different Frame Structures

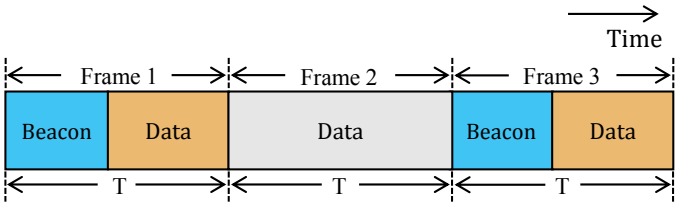
Inspired by the frame structure designs addressed in [88, 89], in this section we provide three additional frame structure designs to address the diverse requirements of CDL and CDO for mm-wave cell discovery.

2.4.4.1 Frame Containing Multiple Beacon Phases

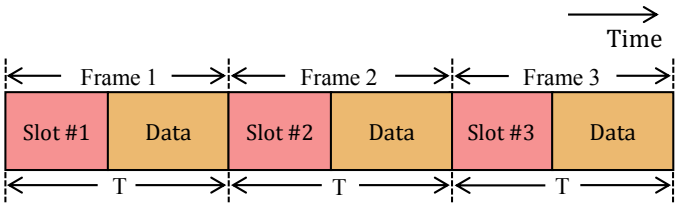
In the previous sections, only one complete beacon phase is accommodated in each frame. However, considering the diversity of frame structure design, e.g., dynamic TDD, multiple beacon phases could be incorporated in one frame to fulfill the requirement of flexible downlink/uplink data adjustment. In this case, we enable the accommodation of W uniformly distributed and separated beacon phases to one single frame. An example of the considered frame structure with $W = 3$ beacon phases is illustrated in Fig. 2.6(a). Other distribution design of multiple beacon phases can be put on top of this case.



(a) Frame with $W = 3$ beacon phases.



(b) Insert $V = 1$ data-only frame between two normal frames.



(c) A beacon phase is equally separated into $X = 3$ frames.

Figure 2.6: Frame structure design examples considering (a) frame with $W = 3$ beacon phases, (b) inserting $V = 1$ data-only frame between two normal frames, and (c) separating one beacon phase equally into $X = 3$ frames.

Intuitively, this frame can be "partitioned" into W consecutive subframes where each subframe is mapping to the normal frame studied in Section 2.4.2 and Section 2.4.3. Specifically, the original cell discovery that is supposed to be done in K

normal frames, where a single complete beacon phase is incorporated in each frame, is now "squeezed" into $\frac{K}{W}$ subframes. Note that here the length of each subframe is $\frac{T}{W}$, and the length of each beacon phase is the same as in the normal frame, which means the data phase is shrunken. Correspondingly, a decreased latency $\frac{T}{W}$, which is a span of the frame length T, can be expected in the CDL of this case. Moreover, from the mathematical point of view, the CDL of this case can be derived by substituting K in (2.24), (2.26), (2.28), and (2.30) with $\frac{K}{W}$, and then the average CDL of this case, denoted as $\overline{T}_{\text{multi-beacon}}$, is written as

$$\overline{T}_{\text{multi-beacon}} = t + \frac{1+\epsilon}{2W(1-\epsilon)}T. \tag{2.39}$$

Obviously, the corresponding CDO is W times of the normal frame.

2.4.4.2 Data-Only Frame

In some scenarios, a large GI, which is reserved for beam switching in the case of hybrid or analog beamforming, would lead to high CDO. In such case, incorporating a beacon phase in every frame may not fulfill the requirement of CDO-fixed or -limited application. Consequently, schemes that alleviate high CDO are proposed in 802.11ad [49], such as inserting data-only frames between consecutive normal frames. An example of inserting $V = 1$ data-only frame between two normal frames is illustrated in Fig. 2.6(b).

Similar to the case that multiple beacon phase are incorporated in one frame, in this case each normal frame can be treated as "prolonged" to a superframe with V data-only frames. Therefore, the original cell discovery that is supposed to be done in K normal frames, where a single complete beacon phase is incorporated in each frame, is now "equivalent" to a superframe with the length $(V+1)T$. Correspondingly, an increased latency $(V+1)T$, which is also a span of the frame length T, can be expected in the CDL of this case. Moreover, from the mathematical point of view, the CDL of this case can be derived by substituting K in (2.24), (2.26), (2.28),

and (2.30) with $(V+1)K$, and then the CDL, denoted as $\overline{T}_{\text{data-only}}$, is written as

$$\overline{T}_{\text{data-only}} = t + \frac{1+\epsilon}{2(1-\epsilon)}(V+1)T. \tag{2.40}$$

In the context of frame-scale, the CDO of the data-only frame is 0. However, when considering the CDO of a superframe including both normal and data-only frames, the CDO is $\frac{1}{V+1}$ times of the original one.

2.4.4.3 Beacon Phase Separation

Another scheme proposed in 802.11ad [49] to alleviate CDO burden is separating a complete beacon phase into consecutive frames. Specifically, only partial angular areas (corresponding to beam slots) in a cell are covered by the beam scan in each frame. This frame structure is expected to be the most general case for cell discovery in 5G NR [47, 48], in which no sufficient long frame length T is supported for incorporating a complete beacon phase to broadcast cell discovery information to all UEs in the cell coverage. An example of separating one beacon phase equally into $X = 3$ frames is illustrated in Fig. 2.6(c).

In this case, each frame can be also treated as "prolonged". However, for a UE in a scanning area, if it misses the beams in its active frame (the current frame), it needs to wait for X frames and then try to decode the discovery information, compared to the case of decoding in the next frame described in Section 2.4.2 and Section 2.4.3. Consequently, the term $\frac{1+\epsilon}{2(1-\epsilon)}T$ in (2.32), is X times as before, as now between two cell discovery trials there are X frames compared to a single frame applied to the situation of Section IV-B and IV-C. Then, the CDL, denoted as $\overline{T}_{\text{separation}}$, is written as

$$\overline{T}_{\text{separation}} = t + \frac{1+\epsilon}{2(1-\epsilon)}XT. \tag{2.41}$$

In the context of frame-scale, the CDO of this case is X times less as the normal frame, however when considering the CDO of a superframe including a complete beacon phase, the CDO remains the same as the normal frame.

To summarize, we argue that:

- The trade-off between CDL and CDO can be extended to different frame structure designs, where incorporating more beacon phases in a frame improves CDL at the price of increased CDO. By contrast, introducing data-only frames and separating a complete beacon phase into different frames save the resource for cell discovery due to reduced CDO, yet suffer from high CDL.
- In particular, the frame structure of separating a beacon phase into consecutive frames is recommended for general 5G mobile communication systems with fixed frame length and potentially short beacon phase. Specifically, separating a beacon phase into as many frames as possible is suitable for CDO-sensitive services but leads to intolerable CDL, while keeping the beacon phase as long as possible in each frame under certain CDO limit is suggested to fulfill the requirement of low-CDL applications.

2.4.5 Cell Discovery for Beam-Scanning Reception

As addressed in Section 2.2.3, UEs are assumed to be able to synthesize a quasi-omnidirectional antenna pattern for signal reception. In order to be compatible with the current standardization activity such as 5G NR, in this section we provide the performance analysis of cell discovery with beam-scanning reception at UE, i.e., UE also applied directional antenna pattern to receive the cell discovery signal transmitted from AP. When applying directional antenna pattern at both Tx and Rx, the angular speed of beam scan at Tx and Rx should be carefully designed in order to avoid the misalignment of transmission and reception beam. Specifically, when single beam exhaustive scan is employed at both Tx and Rx and the angular speeds of transmission and reception beam are equal, it may happen that the transmission and reception beam will never be overlapped, provided that there exists an initial phase difference between these two beams. In this case, no signals can be detected by UE which leads to cell discovery failure. Therefore, in this section we consider two general cases, where the angular speed of either the transmission or reception

beam is large enough compared to the other, such that the fast beam scan over all angular directions can be finished within the duration of one beam direction of the slow beam scan[6].

2.4.5.1 Fast Beam Scan at Transmitter

In this case, within the duration of one UE beam direction, AP can scan over all its angular directions for cell discovery signal transmission. In this way, when the transmission and reception beams are aligned (UE's beam points towards AP), the cell discovery procedure is exactly the same as that studied in the previous sections with omnidirectional reception. Otherwise, the UE should wait until it tunes the reception beam to the position of AP after scanning other angular directions. The issue beam alignment has been investigated in previous works such as [60] as addressed in Section 2.1.1 and is beyond the scope of this dissertation. Therefore in this section, we assume that the transmission and reception beams are able to be aligned and focus on the performance analysis of cell discovery on top of robust beam alignment.

Assume that the "beam duration" of UE, which is similar to the beam duration studied for AP to transmit cell discovery information, equals to the time of AP to scan all its angular directions, namely St. Further assume that the frame length of UE, which includes the beacon phase for discovery information reception and data phase, equals to T_{UE}. Then, the original cell discovery with omnidirectional reception can be imagined similar to the case of data-only frame addressed in Section 2.4.4.2, where a normal frame for cell discovery can be treated as "prolonged" to a superframe with T_{UE}. Therefore, the corresponding CDL for fast beam scan at Tx, denoted as $\overline{T}_{\text{fast-TX}}$, satisfies

$$\overline{T}_{\text{fast-TX}} = t + \frac{1+\epsilon}{2(1-\epsilon)} T_{\text{UE}}. \tag{2.42}$$

[6]Extension to the performance study of cell discovery with more complicated beam-scanning transmission and reception with the issue of beam alignment and mismatch is possible and left as future work.

2.4.5.2 Fast Beam Scan at Receiver

In contrast to the case of fast beam scan at the transmitter, in this case UE can scan over all its angular directions for cell discovery signal reception within one AP beam duration. In this way, when the transmission and reception beams are aligned (UE's beam points towards AP), the cell discovery procedure is also exactly the same as that studied in the previous section of this chapter with omnidirectional reception. However, UE in this case can finish complete beam scan for discovery information reception within the duration of beam alignment as there will be always a chance for it to tunes the reception beam to the position of AP within the beam duration of AP.

Assume that the frame length of UE equals to the beam duration of AP, namely t. Denote the number of slots for UE scan all its angular directions, similar to S for AP, as S_{UE}. Then, the original cell discovery with omnidirectional reception can be imagined similar to the case of the frame containing multiple beacon phases addressed in Section 2.4.4.1, where the beacon phase of a normal frame for cell discovery can be treated as "squeezed" into $\frac{t}{S_{\text{UE}}}$ subframes. Therefore, the corresponding CDL for fast beam scan at Tx, denoted as $\overline{T}_{\text{fast-RX}}$, satisfies

$$\overline{T}_{\text{fast-RX}} = \frac{t}{S_{\text{UE}}} + \frac{1+\epsilon}{2(1-\epsilon)}T. \tag{2.43}$$

2.5 Numerical Evaluation and Discussion

In this section, we evaluate the performance of the proposed broadcast schemes for the mm-wave beamformed cell discovery for standalone networks. We also provide some insights into the questions raised in Section 2.1.2. We first investigate the impact of the baseline scheme, i.e., single beam exhaustive scan, on the cell discovery as well as the beamforming architecture design in Section 2.5.1, as the baseline scheme is the most straightforward broadcast design which has been implemented by 5G NR [47, 48]. Then, based on the analysis in Section 2.4.3 and Section 2.4.4,

we compare the CDL and CDO for all the four broadcast schemes and for different frame structure designs in Section 2.5.2 and Section 2.5.3, respectively. Finally, we compare the cell discovery performance versus block error rate in Section 2.5.4, and show how selecting the error rate could be beneficial.

The simulation results in this section adopt the system model and broadcast schemes design described in Section 2.2 and Section 2.3. In particular, the system carrier frequency and bandwidth are $f = 28\,\text{GHz}$ and $B = 1\,\text{GHz}$, respectively. One AP is located at the center of the cell with cell radius $r = 100\,\text{m}$, and 100 UEs are uniformly dropped in the angular domain of the cell. The adopted propagation and antenna model are explained in Section 2.2.2 and 2.2.3. It is worth noting that UEs are assumed to be almost stationary so the pathloss and shadowing values are fixed during the simulation. Simulation samples are averaged over 1000 independent snapshots. The default system parameter values are summarized in Table 2.1 and Table 2.2.

2.5.1 Impact of Baseline Broadcast Scheme on Cell Discovery and Beamforming Architecture Design

2.5.1.1 How Wide Should a Beam Be (How to Select N)?

In Fig. 2.7 and Fig. 2.8, we plot the analytical and simulation results of the average CDL and CDO of TD for the cell discovery for standalone networks, defined in (2.35) and (2.37), respectively, versus the beam width ($= \frac{360^\circ}{N}$, where N is the number of beam scan areas) for different GIs. Here, GI = 0 indicates the digital beamforming architecture without beam switching time, and the other two values refer to analog/hybrid beamforming with different beam switching times GI = 0.1 µs and GI = 1 µs, respectively. Results show that the simulation results match the analytical result in (2.32), where with the fixed frame length T, the average CDL is "independent" of the selection of GI (tiny fluctuations of the curves refers to non-ideal averaging in the simulations). With wider beam (smaller N), the CDL

increases mainly because of the reduction in the beamforming gain G in (2.3), which correspondingly increases the beam duration t. Further, we note that the CDO degrades when GI becomes larger, which corresponds to our theoretical analysis that larger GI leads to higher CDO.

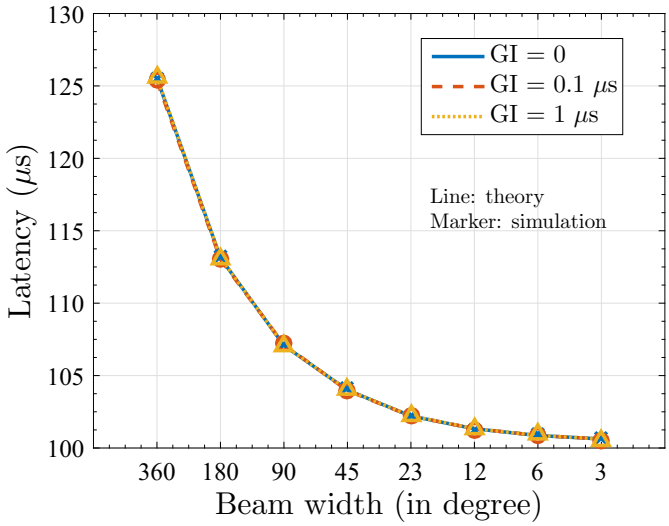


Figure 2.7: Performance of TD for cell discovery are compared in average CDL.

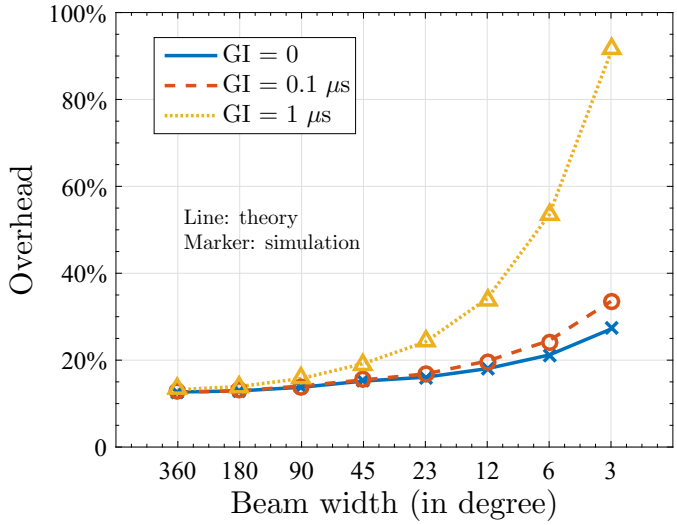


Figure 2.8: Performance of TD for cell discovery are compared in CDO.

In summary, these results indicate that analog and hybrid beamforming perform as well as digital beamforming in terms of CDL. Furthermore, thinner beams (larger

N) significantly decrease CDL. These thinner beams, however, lead to higher CDO.

2.5.2 Cell Discovery Performance Comparison for Different Broadcast Schemes

In this section, based on (2.35)–(2.38), we compare the CDL and CDO of the four broadcast schemes proposed in Section 2.3. In making the comparison, we only consider digital beamforming architecture, i.e., GI = 0, where the performance of the schemes with other GI values can be validated similar to the results. Here, the performance evaluation of the cell discovery for standalone networks addressed in Section 2.4.3 is demonstrated.

2.5.2.1 Is It Beneficial to Exploit Multiple Beam Simultaneous Scan?

To get some insights into the answer of this question, we plot the analytical and simulation results of the average CDL and CDO of different broadcast schemes, defined in (2.35)–(2.38), versus the number of simultaneous beams M in Fig. 2.9 and Fig. 2.10, respectively. The number of beam scan areas N is set as 128. Note that for the baseline scheme TD, M equals to one and thus we plot a single triangle instead of a curve to represent the CDL and CDO of TD in Fig. 2.9 and Fig. 2.10, respectively. A similar representation can be found in Fig. 2.11 and Fig. 2.12.

The results in Fig. 2.9 show that TD achieves the lowest average CDL when the number of beam scan areas N is fixed. FD and CD perform exactly the same, and the curve of SD locates in-between of TD and FD/CD, as given in (2.35)–(2.36). Note that here we plot the average CDL of FD and CD in one curve as their values are exactly the same. Different behaviors are experienced in Fig. 2.10 where in terms of CDO, SD outperforms all schemes as demonstrated in (2.38). It is worth noting that in Fig. 2.10, TD, FD and CD perform the same in CDO because of the

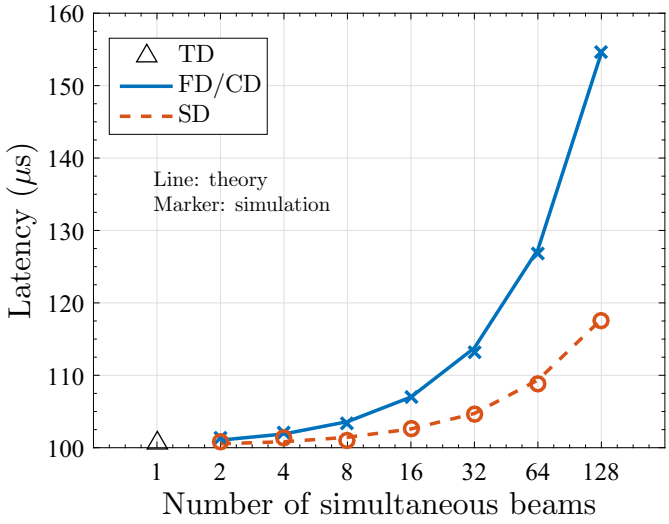


Figure 2.9: Performance of different broadcast schemes for cell discovery are compared in average CDL.

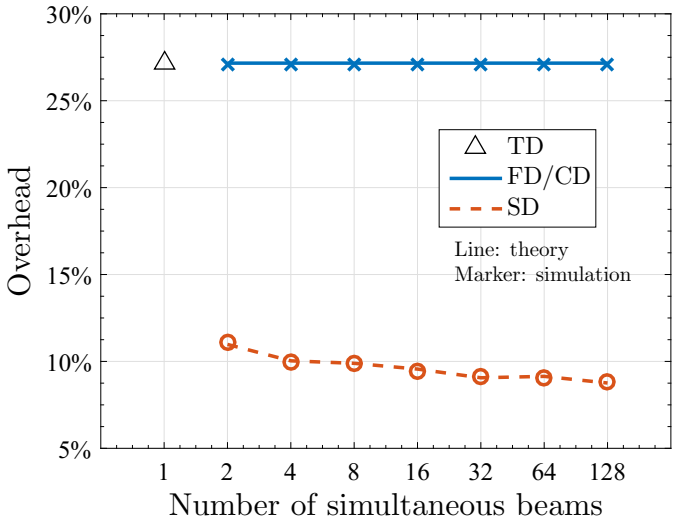


Figure 2.10: Performance of different broadcast schemes for cell discovery are compared in CDO.

assumption of GI = 0, where (2.38) can be reformed as

$$\begin{aligned}\mathrm{CDO_{FD}} = \mathrm{CDO_{CD}} &= \frac{t_{\mathrm{FD/CD}} S_{\mathrm{FD/CD}}}{T} = \frac{\frac{N}{M} \cdot M t_{\mathrm{TD}}}{T} \\ &= \frac{N t_{\mathrm{TD}}}{T} = \mathrm{CDO_{TD}}. \end{aligned} \tag{2.44}$$

In case GI $\neq 0$, the highest CDO can be expected from TD as indicated in (2.38).

2.5.2.2 If So, How Many Simultaneous Beams Should Be Exploited (How to Select M)?

It is clear from Fig. 2.9 that the CDL increases with the number of simultaneous beams M (the CDL keeps fixed for TD as only one beam is exploited). By contrast, the CDO degrades with the increase of M as indicated in Fig. 2.10. Therefore, the results in Fig. 2.9 and Fig. 2.10 do not recommend any optimal value of M unless a targeted performance metric is explicitly stated. Nevertheless, if both CDL and CDO are to be considered, SD provides the flexibility to achieve a trade-off between both metrics. In other words, by configuring the number of simultaneous beams, CDL can be traded with CDO or vice versa.

2.5.3 Cell Discovery Performance Comparison for Different Frame Structures

In this section, based on (2.39)–(2.41), we compare the CDL and CDO of the three frame structures proposed in Section 2.4.4 and the normal frame structure proposed in Section 2.2.4. Similar to Section 2.5.2, we only consider digital beamforming architecture, i.e., GI = 0, where the performance of the schemes with other GI values can be similarly validated.

2.5.3.1 Does Cell Discovery Performance Vary with Different Frame Structures?

Fig. 2.11 and Fig. 2.12 plot the analytical and simulation results of the average CDL and CDO of different broadcast schemes with different frame structure designs including multi-beacon ($W = 3$), data-only frame insertion ($V = 1$), and beacon phase separation ($X = 3$), defined in (2.39)–(2.41), respectively, as well as normal frame structure, versus the number of simultaneous beams M. The number of beam scan areas N is also set as 128.

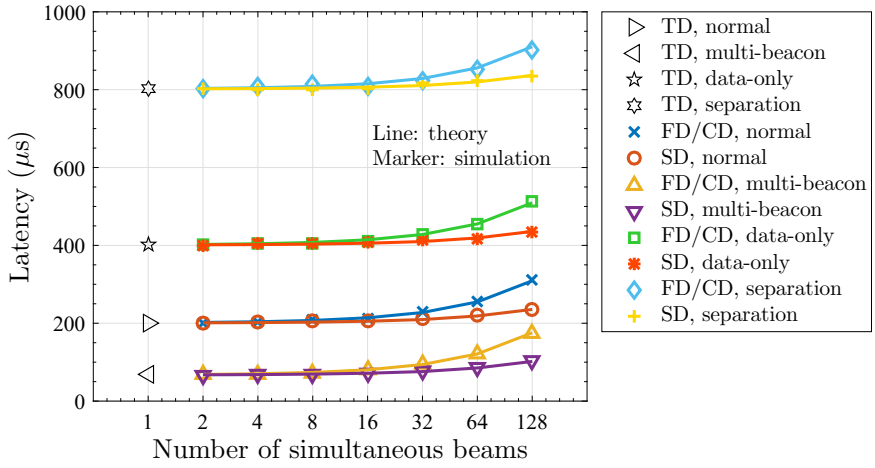


Figure 2.11: Performance of different broadcast schemes for cell discovery with different frame structure designs including normal frame, multi-beacon ($W = 3$), data-only frame insertion ($V = 1$), and beacon phase separation ($X = 3$) are compared in average CDL.

On the one hand, the results in Fig. 2.11 and Fig. 2.12 suggest that the average CDL declines when multiple beacons are incorporated in each frame, while the decrease in CDL leads to triple CDO compared to the normal frame. On the other hand, for data-only frame insertion and beacon separation cases, increased CDLs are observed at the cost of half and identical CDOs as the normal frame. Note that here the CDOs of the data-only frame insertion case and of the beacon separation case refer to superframe-scale. In the context of frame-scale, the CDO of the data-only frame is zero, and the CDO of the frame in beacon phase separation case is $\frac{1}{X}$ times

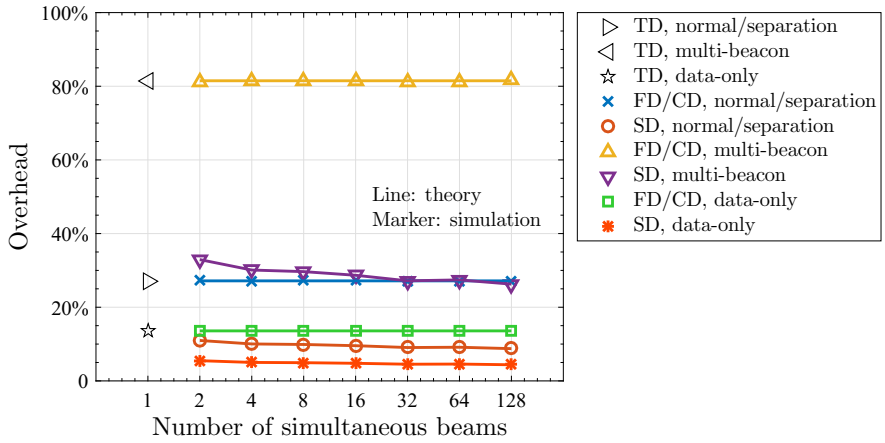


Figure 2.12: Performance of different broadcast schemes for cell discovery with different frame structure designs including normal frame, multi-beacon ($W = 3$), data-only frame insertion ($V = 1$), and beacon phase separation ($X = 3$) are compared in CDO.

of a normal frame.

Based on the results in Fig. 2.9 to Fig. 2.12, we conclude the trade-off between CDL and CDO can be extended to specific frame structure design, in which frame with multiple beacon phases results in lower CDL at the price of higher CDO. On the contrary, for the case of CDO-sensitive application scenario, superframe with embedded data-only frame and/or with equally separated beacon phases are recommended to achieve lower CDO with relatively high CDL. In particular, separating a beacon phase into as many frames as possible is recommended for CDO-sensitive services, while keeping the beacon phase as long as possible in each frame under certain CDO limit is suggested for low-CDL applications.

2.5.4 Cell Discovery Performance Comparison for Different Block Error Rates

2.5.4.1 What Is the Impact of Block Error Rate (How to Select ϵ)?

In Fig. 2.13 and Fig. 2.14, the analytical and simulation results of the average CDL and CDO of different broadcast schemes with $N = 128$ and $M = 128$ versus the

block error rates ϵ are compared.

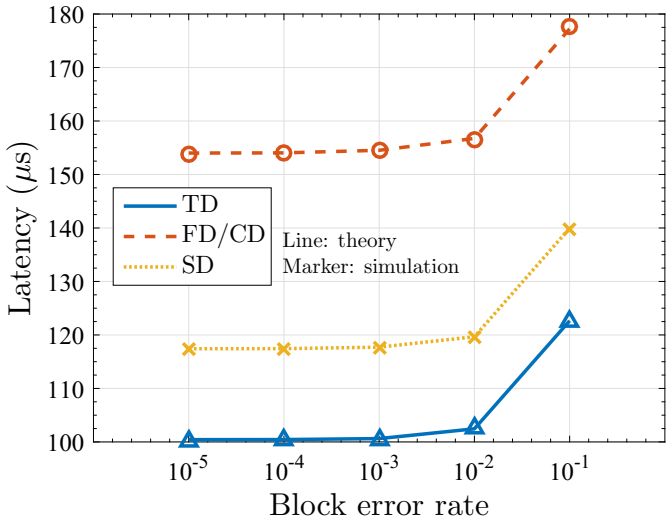


Figure 2.13: Performance of different broadcast schemes for cell discovery with different block error rates are compared in average CDL.

In the figure, we notice that applying codes with lower block error rate yields both lower CDL and CDO. The main reason behind this behavior lies in the fact that a lower block error rate increases the value of the denominator in (2.9), under the fixed SNR and information U, and eventually decreases the beam duration t. However, these results show that the CDL and CDO are relatively insensitive to extremely low block error rates (10^{-5}–10^{-3}). Therefore, a relatively high block error rate (10^{-3}) would be sufficient for initial cell discovery, unless an extreme coding scheme is desired to achieve a better performance.

2.6 Summary

In this chapter, we proposed an analytical framework to investigate the performance of mm-wave beamformed cell discovery. Specifically, we analyzed four broadcast schemes where an AP delivers information to UEs for cell discovery. Based on this, the cell discovery analysis was distinguished by various broadcast schemes, network deployments, frame structure designs, and beamforming architectures. By evaluat-

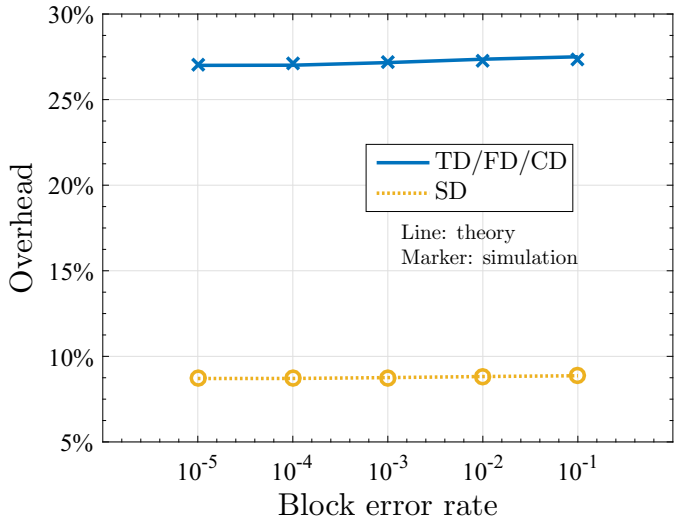


Figure 2.14: Performance of different broadcast schemes for cell discovery with different block error rates are compared in CDO.

ing the performance metrics including the cell discovery latency and overhead, this chapter allows us to gain the following design insights:

- *How wide should a beam be?* Cell discovery latency is optimized when the thinnest beam is formed. Interestingly, the beamforming architecture has no impact on the latency, which makes the performance of analog/hybrid beamforming the same as digital beamforming in terms of the latency. By contrast, thinner beam results in higher overhead.
- *Is it beneficial to exploit multiple beam simultaneous scan?* Multiple beam simultaneous scan leads to a latency penalty on cell discovery. Single beam exhaustive scan is found to be optimal in terms of the latency, in case the frame length is fixed and a complete beam scan can be finished within one frame. This is reversed when considering overhead, where the single beam exhaustive scan, as well as frequency-division/code-division multiple beam scan, suffer from high overhead. The spatial-division multiple beam scan, however, achieves the lowest overhead.
- *If so, how many simultaneous beams should be exploited?* On the one hand, cell discovery latency gets worse as the number of simultaneous formed beams

increases. On the other hand, overhead degrades with more simultaneously formed beams. Therefore, the optimal number of simultaneous beams depends on targeted performance metric. Nevertheless, the best trade-off between latency and overhead can be achieved by spatial-division multiple beam scan. In other words, by configuring the number of simultaneous beams, latency can be traded with overhead or vice versa.

- *Does cell discovery performance vary with different frame structures?* The trade-off between latency and overhead can be further extended to frame structure design, where specific frame structure results in lower latency at the price of higher overhead or vice versa. In particular, the frame structure of separating a beacon phase for cell discovery into consecutive frames is recommended for general 5G mobile communication systems with a fixed frame length and a potentially short beacon. Specifically, separating a beacon into as many frames as possible is suitable for overhead-sensitive services yet with intolerable latency, while keeping the beacon as long as possible in each frame under certain overhead limit is suggested for low-latency applications.
- *What is the impact of block error rate?* It has been demonstrated that the cell discovery latency and overhead are relatively insensitive to extreme low block error rates (10^{-5}–10^{-3}). Therefore, a relatively high block error rate (10^{-3}) would be sufficient for the initial cell discovery, unless an extreme coding scheme is desired to achieve a better performance.

2.7 Appendix

2.7.1 Discussion of Orthogonal Codes for Code-Division Broadcast Scheme

From an information theoretical point of view, code-division is completely equivalent to time-frequency division when orthogonal codes, which do not lose orthogonality

after transmission due to multi-path channel, are utilized. In other words, for a frequency flat channel, achieving orthogonality in time, frequency, or in any other multiplexing dimension, is "exactly" the same in terms of spectral efficiency and SNR, therefore CDMA performs exactly the same as TDMA or FDMA.

A standard direct spreading CDMA system typically correlates the signal with respect to desired spreading code and treats the rest as additive noise. In this case, CDMA is a simplistic way to achieve low rate coding [90]. In short, it consists of concatenating a given channel code with a repetition code of rate 1/SG, where SG is the spreading gain.

Non-orthogonal codes may slightly benefit from special sets of sequences with particularly good correlation coefficients. However in this case, it makes more sense to use orthogonal codes. When considering the robustness to multi-path for non-orthogonality, it has been shown in Qualcomm IS-95 and Wideband CDMA (WCDMA) that random spreading is as good as any other family of sequences. In this case, the receive power after de-spreading is attenuated by a factor 1/SG, and the approach for the Gaussian channel in (2.9) is still valid, taking into account the overhead of SG dimensions per symbol due to spreading. Moreover, the utilization of directional antenna leads to relatively low multi-user interference where more simultaneous transmissions can be supported to exploit spatial multiplexing gain. When the interference is weak, imposing decoding or cancellation of all signals, instead of just the useful one, is not a good strategy.

In conclusion, insisting on treating interference as noise, it is beneficial to consider the orthogonal codes, which is strictly better than any form of random spreading, then we have the family of orthogonal access, namely TDMA, FDMA, and orthogonal CDMA, which are all equivalent and yield exactly the same SNR and channel capacity performance.

2.7.2 Proof of Equation (2.18)

As we have $\mathrm{SNR}_{\mathrm{SD}} \leq \mathrm{SNR}_{\mathrm{TD}}$ according to (2.8), then

$$\begin{aligned} C_{\mathrm{SD}} &= \log_2\left(1+\mathrm{SNR}_{\mathrm{SD}}\right) \leq \log_2\left(1+\mathrm{SNR}_{\mathrm{TD}}\right) \\ &= C_{\mathrm{TD}}, \end{aligned} \tag{2.45}$$

which comes from the fact that the logarithm function log(.) is monotonically increasing.

Moreover, the channel dispersion of SD, V_{SD}, can be expanded as

$$\begin{aligned} V_{\mathrm{SD}} &= \frac{\mathrm{SNR}_{\mathrm{SD}}(\mathrm{SNR}_{\mathrm{SD}}+2)}{(\mathrm{SNR}_{\mathrm{SD}}+1)^2}\log_2^2 e = \frac{\mathrm{SNR}_{\mathrm{SD}}^2+2\mathrm{SNR}_{\mathrm{SD}}}{\mathrm{SNR}_{\mathrm{SD}}^2+2\mathrm{SNR}_{\mathrm{SD}}+1}\log_2^2 e \\ &\overset{(b)}{\leq} \frac{\mathrm{SNR}_{\mathrm{TD}}^2+2\mathrm{SNR}_{\mathrm{TD}}}{\mathrm{SNR}_{\mathrm{TD}}^2+2\mathrm{SNR}_{\mathrm{TD}}+1}\log_2^2 e \\ &= V_{\mathrm{TD}}. \end{aligned} \tag{2.46}$$

Here, (b) in (2.46) comes from the fact that the denominator $\mathrm{SNR}_{\mathrm{TD}}^2+2\mathrm{SNR}_{\mathrm{TD}}+1$ grows faster than the numerator $\mathrm{SNR}_{\mathrm{SD}}^2+2\mathrm{SNR}_{\mathrm{SD}}$.

Therefore, we have

$$\begin{aligned} t_{\mathrm{SD}} &= \frac{U}{B\Big(C_{\mathrm{SD}}-\sqrt{\frac{V_{\mathrm{SD}}}{n}}Q^{-1}(\epsilon)+\frac{\log_2 n}{n}\Big)} \geq \frac{U}{B\Big(C_{\mathrm{TD}}-\sqrt{\frac{V_{\mathrm{TD}}}{n}}Q^{-1}(\epsilon)+\frac{\log_2 n}{n}\Big)} \\ &= t_{\mathrm{TD}}. \end{aligned} \tag{2.47}$$

2.7.3 Proof of Theorem 2.4.1

Given the uniform distribution of slot j from 1 to S, we have

$$\begin{aligned} \overline{T}_{\text{non-standalone}} &= \sum_{j=1}^{S} p_{\mathrm{O}}(j)t_{\mathrm{O}}(j) = \sum_{j=1}^{S}\left(\frac{1}{S}\sum_{k=0}^{K-1}\Big(\big(kT+(j-1)t'+t\big)\epsilon^k(1-\epsilon)+KT\epsilon^K\Big)\right) \\ &= \frac{\epsilon+\epsilon^{K+1}(K-1)-K\epsilon^K}{1-\epsilon}T + KT\epsilon^K + \frac{1}{2}\big(1-\epsilon^K\big)\big((S-1)t'+2t\big) \\ &= \big(1-\epsilon^K\big)\Big(\frac{S+1}{2}t+\frac{S-1}{2}t_{\mathrm{GI}}+\frac{\epsilon}{1-\epsilon}T\Big). \end{aligned} \tag{2.48}$$

When $K \to \infty$, (2.48) is further simplified as

$$\overline{T}_{\text{non-standalone}} = \frac{S+1}{2}t + \frac{S-1}{2}t_{\text{GI}} + \frac{\epsilon}{1-\epsilon}T. \tag{2.49}$$

2.7.4 Proof of Theorem 2.4.2

The average latency $t_{\text{A}}(i)$ can be classified into the following cases:

(a) Successfully decoding in the current frame, which consists of three parts:

- Average latency in slot i: As z_{active} is arbitrarily distributed in slot i, which means that UE can be active at any time in the slot, the average latency is calculated by taking integral of z_{active} from $(i-1)t'$ to it' within slot i, namely

$$\int_{(i-1)t'}^{it'} \frac{it' - z_{\text{active}}}{it' - (i-1)t'} \, \mathrm{d}z_{\text{active}} = \frac{t'}{2}. \tag{2.50}$$

- Average latency between the end of slot i and the beginning of slot j: This part of latency counts for the remaining time from the end of UE active slot to the beginning of UE located slot, which is simply calculated as $(j-i-1)t'$.
- Beam decoding latency: t for the UE to decode the cell discovery information.

(b) Successfully decoding in the $(k+1)$-th frame. Similar to the previous case, the average latency of this case also consists of three parts:

- Average latency in the current frame: As UE fails in decoding cell discovery information in the current frame, the average latency in the current frame is calculated by adding the latency in UE active slot (slot i) and the remaining time in the current frame, namely $\frac{t'}{2} + T - it'$.
- Latency from the end of the current frame to the end of the k-th frame: simply $(k-1)T$.
- Latency from the beginning of the $(k+1)$-th frame to the time of successfully decoding: In the $(k+1)$-th frame, UE has to wait from the beginning of the frame to a beam starts to scan its located slot, which is calculated as $(j-1)t'$, and decodes the cell discovery information within decoding time t. Hence, the

summarized latency of this part is $(j-1)t' + t$.

(c) Failed in all the K frames. For this case, UE is not able to finish cell discovery within K frames, where two parts of average latency are considered as follows:

- Average latency in the current frame: $\frac{t'}{2} + T - it'$, which is the same as the first part of case (b).
- Latency from the beginning of the $(k+1)$-th frame to the end of the K-th frame: simply $(K-1)T$.

Considering the fact that the probability of the above three cases are $1-\epsilon$, $\epsilon^k(1-\epsilon)$, and ϵ^K, respectively, the average latency $t_{\mathrm{A}}(i)$, averaging over the uniform probability of j from $i+1$ to S, is given by

$$\begin{aligned} t_{\mathrm{A}}(i) &= \left(\frac{t'}{2} + \frac{\sum_{j=i+1}^{S}(j-i-1)t'}{S-i} + t\right)(1-\epsilon) + \left(\frac{t'}{2} + T - it' + (K-1)T\right)\epsilon^K \\ &\quad + \sum_{k=1}^{K-1}\left(\frac{t'}{2} + T - it' + (k-1)T + \frac{\sum_{j=i+1}^{S}(j-1)t' + t}{S-i}\right)\epsilon^k(1-\epsilon) \\ &= \left(\frac{t'}{2} + \frac{S-i-1}{2}t' + t\right)(1-\epsilon) + \left(\frac{t'}{2} + T - it' + (K-1)T\right)\epsilon^K \\ &\quad + \sum_{k=1}^{K-1}\left(\frac{t'}{2} + T - it' + (k-1)T + \frac{S+i-1}{2}t' + t\right)\epsilon^k(1-\epsilon). \end{aligned} \tag{2.51}$$

2.7.5 Proof of Theorem 2.4.3

The average latency $t_{\mathrm{B}}(i)$ can be classified into the following cases:

(a) Failed in the current frame. In this case, the average latency is calculated as same as the first part of case (b) and case (c) for the proof of Theorem 2.4.2, namely $\frac{t'}{2} + T - it'$.

(b) Successfully decoding in the $(k+2)$-th frame. Note that, here the index $k+2$ indicates that UE failed in cell discovery within the first frame (the current frame) and the following consecutive k frames. The average latency of this case consists of two parts:

- Latency from the end of the current frame to the end of the $(k+1)$-th frame: simply kT.
- Latency from the beginning of the $(k+2)$-th frame to the time of successfully decoding: $(i-1)t+t$, which is the same as the last part of case (b) for the proof of Theorem 2.4.2 ($j=i$).

(c) Failed in all the K frames. The latency of this case is simply $(K-1)T$.

Considering the fact that the probability of the above three cases are 1, $\epsilon^k(1-\epsilon)$, and ϵ^{K-1}, respectively, the average latency $t_{\mathrm{B}}(i)$, averaging over the uniform probability of j from 1 to S, is given by

$$t_{\mathrm{B}}(i) = \frac{t'}{2} + T - it' + \sum_{k=0}^{K-2} \big(kT + (i-1)t' + t\big)\epsilon^k(1-\epsilon) + (K-1)T\epsilon^{K-1}. \quad (2.52)$$

2.7.6 Proof of Theorem 2.4.4

The proof of Theorem 2.4.4 is very close to the proof of Theorem 2.4.3, as in both scenario B and C, UE is not able to finish the cell discovery in the current frame. The only different point of Theorem 2.4.4 from Theorem 2.4.3 is that the slot j is averaging from 1 to $i-1$. Therefore, the average latency $t_{\mathrm{C}}(i)$, averaging over the uniform probability of j from 1 to $i-1$, is given by

$$\begin{aligned} t_{\mathrm{C}}(i) &= \frac{t'}{2} + T - it' + \sum_{k=0}^{K-2} \left(kT + \frac{\sum_{j=1}^{i-1}(j-1)}{i-1}t' + t\right)\epsilon^k(1-\epsilon) + (K-1)T\epsilon^{K-1} \\ &= \frac{t'}{2} + T - it' + \sum_{k=0}^{K-2} \left(kT + \frac{i-2}{2}t' + t\right)\epsilon^k(1-\epsilon) + (K-1)T\epsilon^{K-1}. \quad (2.53) \end{aligned}$$

2.7.7 Proof of Theorem 2.4.5

The average latency t_{D} consists of the following cases:

(a) Failed in the current frame. With the uniform distribution of z_{active} at the entire frame, i.e., UE can be active at any time in the data phase, the average latency

is calculated by taking integral of z_{active} from St' to T within the data phase, namely

$$\int_{St'}^{T} \frac{z_{\text{active}}}{T - St'} \, \mathrm{d}z_{\text{active}} = \frac{T - St'}{2}. \tag{2.54}$$

(b) Successfully decoding in the $(k+2)$-th frame. In this case, the average latency is calculated as same as case (b) for the proof of Theorem 2.4.3, namely $kT + (j-1)t' + t$.

(c) Failed in all the K frames. The latency of this case is simply $(K-1)T$.

Considering the fact that the probability of the above three cases are 1, $\epsilon^k(1-\epsilon)$, and ϵ^{K-1}, respectively, the average latency t_{D}, averaging over the uniform probability of j from 1 to S, is given by

$$\begin{aligned} t_{\text{D}} &= \frac{T - St'}{2} + \sum_{k=0}^{K-2} \left(kT + \frac{\sum_{j=1}^{S}(j-1)}{S} t' + t \right) \epsilon^k (1-\epsilon) + (K-1)T\epsilon^{K-1} \\ &= \frac{T - St'}{2} + \sum_{k=0}^{K-2} \left(kT + \frac{(S-1)}{2} t' + t \right) \epsilon^k (1-\epsilon) + (K-1)T\epsilon^{K-1}. \end{aligned} \tag{2.55}$$

2.7.8 Proof of Theorem 2.4.6

It is assumed that in scenario A, UE is active before its located slot, which means the last active slot should be slot $S-1$. Thus, the probability of scenario A is distributed over 1 to $S-1$. For scenario B, UE is active at the same slot which it is located in, then the probability of scenario B is distributed over 1 to S. Similarly, as UE is active after its located slot, the probability of scenario C is distributed over 2 to S. Based on this, we have

$$\begin{aligned} \overline{T}_{\text{A}} &= \sum_{i=1}^{S-1} p_{\text{A}}(i) t_{\text{A}}(i) \\ &= \sum_{i=1}^{S-1} \left(\frac{t'}{T} - \frac{t'}{TS} i \right) \left(\left(\frac{S}{2} t' + t \right) (1 - \epsilon^K) + \frac{t'}{2} \epsilon^K + \frac{\epsilon - \epsilon^{K+1}}{1-\epsilon} T - \frac{t'}{2} (1 + \epsilon^K) i \right), \end{aligned} \tag{2.56}$$

$$\begin{aligned}\overline{T}_{\mathrm{B}} &= \sum_{i=1}^{S} p_{\mathrm{B}}(i) t_{\mathrm{B}}(i) \\ &= \sum_{i=1}^{S} \frac{t'}{TS}\bigg((t-t')\big(1-\epsilon^{K-1}\big) + \frac{t'}{2} + \frac{1-\epsilon^{K}}{1-\epsilon}T - t'\epsilon^{K-1}i\bigg),\end{aligned} \tag{2.57}$$

$$\begin{aligned}\overline{T}_{\mathrm{C}} &= \sum_{i=2}^{S} p_{\mathrm{C}}(i) t_{\mathrm{C}}(i) \\ &= \sum_{i=2}^{S} \left(\frac{t'}{TS}i - \frac{t'}{TS}\right)\bigg((t-t')\big(1-\epsilon^{K-1}\big) + \frac{t'}{2} + \frac{1-\epsilon^{K}}{1-\epsilon}T - t'\epsilon^{K-1}i\bigg),\end{aligned} \tag{2.58}$$

and

$$\overline{T}_{\mathrm{D}} = p_{\mathrm{D}} t_{\mathrm{D}} = \frac{T - St'}{T}\bigg(\frac{\frac{1+\epsilon}{2} - \epsilon^{K}}{1-\epsilon}T - \frac{(S-1)\epsilon^{K-1}+1}{2}t' + \big(1-\epsilon^{K-1}\big)t\bigg). \tag{2.59}$$

When $K \to \infty$, we have

$$\begin{aligned}\overline{T}_{\text{standalone}} &= \sum_{i=1}^{S-1} p_{\mathrm{A}}(i) t_{\mathrm{A}}(i) + \sum_{i=1}^{S} p_{\mathrm{B}}(i) t_{\mathrm{B}}(i) + \sum_{i=2}^{S} p_{\mathrm{C}}(i) t_{\mathrm{C}}(i) + p_{\mathrm{D}} t_{\mathrm{D}} \\ &= \overline{T}_{\mathrm{A}} + \overline{T}_{\mathrm{B}} + \overline{T}_{\mathrm{C}} + \overline{T}_{\mathrm{D}} \\ &= t + \frac{1+\epsilon}{2(1-\epsilon)}T.\end{aligned} \tag{2.60}$$

Chapter 3

Joint Scheduling and Resource Allocation Optimization in 5G Millimeter Wave Heterogeneous Networks

3.1 Introduction

In the current sub-6 GHz cellular systems, spectral efficiency per link is approaching its theoretical limits [91]. A better exploitation of spectrum opportunities is a key enabler for cellular communications to meet the ever-increasing global mobile traffic demand. Besides aggregating carriers on both licensed and unlicensed spectrum, a

This chapter has been published in [11, 29]. I am the primary author of these works. Co-authors Dr. Emmanouil Pateromichelakis, Dr. Nikola Vučić, and Dr. Wen Xu have provided many valuable comments to derive the solution and improve this work, and Dr. Jian Luo and Prof. Giuseppe Caire are my supervisors. Except for the contributions in the previous publications, this chapter also extends the work in [29] with the design of a dynamic routing algorithm and the application to data delivery in vehicle platoon.

promising approach to provide higher transmission bandwidth is to extend cellular communications into mm-wave bands [92].

5G cellular communications are embracing mm-wave frequencies between 6 and 300 GHz, where the availability of large chunks of untapped bandwidth makes it possible to achieve the demand for gigabits per second data rates (Gbps) [10, 32]. Carrier frequencies up to 52.6 GHz with a bandwidth per single carrier up to 400 MHz has been standardized by the RAN of 3GPP 5G NR [7, 93]. The high potential of mm-wave communications to support the stringent data rate requirements for future cellular systems has been confirmed by tutorial articles, surveys, and pre-commercial trials [1, 94].

Such high frequencies, however, bring challenges in designing and deploying mm-wave communication systems. Specifically, the increase in isotropic path loss according to Friis' Law [82] and the propagation attenuation due to the atmospheric absorption of oxygen molecules and water vapor [33, 95] are anticipated at high frequencies. Moreover, mm-wave signals can be severely vulnerable to blockage caused by e.g. buildings and human body, which results in outages and intermittent channel quality [11, 96].

The combination of the high propagation loss and the blockage phenomenon advocates for a high-density deployment of infrastructure nodes [97]. However, the placement of additional macrocell BS involves significant cost and elaborate site-planning [98]. Small cells, as an available realization of network densification, offer a simpler cost-effective alternative to conventional cell splitting. In this regard, HetNets, where a core macrocell network seamlessly cooperates with mm-wave small cells, have been treated as one promising candidate of mm-wave cell deployment for universal coverage and augmenting capacity [99].

Network densification by HetNets is of little consequence unless it is complemented by connecting network nodes with backhauls. Nevertheless, equipping all small cells with high-performance fiber-based backhaul (with extremely high bandwidth and low latency) seems to be economically infeasible, as such ultra-dense de-

ployment leads to significantly high installation and operation expenditures. As an attractive cost-efficient substitute to the wired backhaul, self-backhauling provides technology- and topology-dependent coverage extension and capacity expansion to fully exploit the heterogeneity of the HetNets [100]. Leveraging the abundant available spectrum, the self-backhauling at mm-wave band is able to provide orders of Gbps capacity [101, 102]. 5G NR incorporates the concept of integrated access and backhaul (IAB) that supports multi-hop backhauling to provide range extension, and the flexibility in hop count is desirable depending on numerous factors such as frequency, cell density, propagation environment, and traffic load [103].

Another encouraging approach to cope with the high isotropic pathloss and the sensitivity to blockage effects is the exploitation of beamforming techniques to form narrow beams with high antenna gain for data transmissions [104]. This is possible as the small wavelength, which is one of the distinctive features of mm-wave bands, allows a large number of antenna arrays to be placed in a compact form factor. In general, directional antennas with beamforming reduce multi-user interference, where more simultaneous transmissions can be supported to exploit spatial multiplexing gain [37, 78].

Hereof, how to maximize the system performance of HetNets with self-backhauling and directional transmission becomes an interesting issue, particularly on the design of link scheduling, resource allocation, and path selection. A naive scheduling which lets macrocell BS serve all users in a round-robin (RR) fashion is neither practical nor efficient. By contrast, the limited interference at mm-wave bands makes it efficient to schedule simultaneous transmissions, where the same radio resource can be allocated to multiple links to improve spatial reuse [29]. At the same time, when the backhaul link, which connects the associated small cell AP of a user to macrocell BS, is weaker compared to the backhaul links to other nearby APs, a dynamic multihop routing is much more favorable as it alleviates the bottleneck at the BS [97]. An appropriated and well-designed joint link scheduling and resource allocation scheme, combined with multihop routing algorithm for path selection, yields an attractive,

practical, and often near-optimal solution to 5G mobile communication networks.

3.1.1 Related Works

The study of wireless backhaul has spanned the last two decades, aiming to replace the costly fixed links with more flexible wireless connections [105]. Recent research efforts, particularly at mm-wave bands, have addressed versatile aspects of wireless backhauling, including the design of high-throughput backhaul system [101], the qualitative overview of discovered challenges [106], the reveal of hidden advantages and shortcomings [107], as well as some preliminaries results [108]. Standard association such as 3GPP is currently focusing on tight integration between access and backhaul, which has been addressed as IAB in 5G NR [103], to overcome the limitation of traditional LTE by introducing the plug-and-play mannered self-backhauling AP. In particular, IAB enables both downlink and uplink transmission on backhaul and access links and supports time-, frequency-, and space-division multiplexing subject to half-duplex constraints.

As a natural step forward towards wireless backhauling, increasing cellular capacity through self-backhauled small cells has become the primary motivation of many previous works. Some studies emphasized the placement of relay nodes inside cells to improve signal quality at cell edges [109, 110], yet the capacity boost obtained through simple RF amplification and relaying is limited. More general approaches to enhance cellular capacity with multihop backhauling have been considered in [24, 111, 112], where computationally analytical models are developed to quantify interference, coverage, and data rate under varying parameters and network topologies. However, these works limit the number of links at each network node (BS, AP, UE, etc.) to a single steerable beam, which does not take advantage of potential spatial reuse provided by highly directional mm-wave arrays that allow simultaneous transmission or reception on multiple links at the same node.

When it comes to the exploitation of spatial reuse, the appropriate design of effi-

cient scheduling policy has been suggested as a key challenge in realizing the full benefits of the multiplexing gain brought by simultaneous links [113, 114, 115, 116, 117]. Authors in [113] and [114] proposed a multihop simultaneous transmission scheme to take advantage of spatial reuse and time-division multiplexing gain. Exploiting spatial multiplexing in device-to-device (D2D) transmissions has been studied in [115], where a centralized medium access control (MAC) scheduling scheme is proposed with a path selection criterion. Recently, [116] presented an energy efficient solution for the mm-wave backhauling scheme, in which the simultaneous transmissions are exploited for lower energy consumption and higher energy efficiency.

With the exception of some studies that cover limited models [115, 116], none of the above-mentioned works to date has considered the degree of spatial link isolation. In other words, interference in mm-wave communications has a much weaker effect than in sub-6 GHz small cell networks, but not negligible. However, some recent works even assume that inter-link interference levels can be ignored in mm-wave networks and thus links can be approximated as pseudo-wires [118]. This approach thus ignores inter-link interference between simultaneous transmissions (i.e., SNR instead of SINR), where the usual approach of guaranteeing a target SNR to each link turns out to yield an operating point that can be arbitrarily far from optimal for a Gaussian interference under the practical constraint of treating interference as Gaussian noise for the sake of complexity and robustness. By contrast, much better network throughput can be achieved by selecting a subset of active links in each slot and allocating positive power only to these selected links [119]. We think that the question of whether the fully isolated pseudo-wired hypothesis is realistic or not in certain scenarios remains open, as also addressed in [24], where the available capacity in simultaneously transmitted links should not be always considered as the same.

Furthermore, resource allocation and path selection for mm-wave HetNets have also been an active research topic [97, 105, 120, 121, 122, 123, 124, 125]. Frequency resource allocation has been addressed in [121] where a joint downlink cell asso-

ciation and wireless backhaul bandwidth allocation in a two-tier HetNet has been studied. Power allocation optimization problem has been formulated in [122] where two algorithms to solve the downlink sum rate maximization problem with respect to quality of service (QoS) requirements have been proposed. Some work [123] considered the problem of joint discrete power control and transmission duration allocation in self-backhauling mm-wave cellular networks and proposed both centralized and decentralized resource allocation algorithms, while others [124] concentrated on joint scheduling and power allocation optimization with an interference-aware hybrid scheduler to maximize capacity gain. Besides the consideration of resource allocation, path selection techniques are compared in [105] in terms of hop count and bottleneck SNR. Authors in [97] proposed a polynomial time algorithm for joint scheduling and routing, extending the work in [125], which is unlike the traditional NP-hard solutions [22].

Even though the aforementioned research studies cover a set of resource allocation and path selection schemes that enhance the system performance under different network configurations and constraint models, the authors focused on either one aspect of resource allocation (e.g., frequency allocation in [121] and power allocation in [122]), or limited combination (transmission duration and power allocation in [124], or scheduling and routing in [97]). The joint solution of scheduling, resource allocation, and routing, have not been considered in any of above. By contrast, we formulate the joint optimization model with an accurate schedule-dependent representation of data rate taking into account the resource allocation and path selection to derive our main results.

3.1.2 Contributions

In this chapter, we apply binary interference classification, i.e., an interference condition that prevents two links from being simultaneously active, which is widely used in the literature ([24, 97, 115, 116, 125, 126]) for designing scheduling algorithms, to the joint scheduling and resource allocation (JSRA) optimization. Nevertheless,

on top of the binary classification, we formulate the data rate of scheduled links as a function that depends on actual SINR, where links that experience inter-link interference are not completely blocked. Specifically, multiple links are allowed to be simultaneously transmitted provided that the mutual interferences among these links are below than a configurable threshold, which is usually considered in practical PHY implementation.

Based on this, we study the JSRA optimization for a typical HetNet with multihop backhauling structure. Each node in the network is subject to a fundamental scheduling constraint addressed for IAB, namely the half-duplex communication, so that nodes cannot transmit and receive at the same time. Point-to-multipoint (P2M) transmission is assumed where a node can support several links at a time to fully exploit spatial reuse. The constrained optimization problem is demonstrated as NP-hard and systematically decomposed, and a heuristic scheme including link scheduling, transmission duration allocation, and power allocation is proposed to maximize data rate. In conjunction with an efficient multihop routing algorithm for path selection, it is demonstrated that the data rate can be further enhanced. The proposed solutions are then validated against detailed system-level simulations, and the evaluation results demonstrate that the proposed algorithms achieve near-optimal data rate and with significantly lower latency. The main contributions of this chapter are summarized as follows:

- We formulate the joint optimization problem of scheduling and resource allocation into a MINLP problem, in which the data rate, determined by scheduling, transmission duration, and power allocation constraints, as well as by actual interference, is maximized by fully enabling simultaneous transmission to harvest the spatial multiplexing gain.
- The joint optimization problem is then demonstrated to be NP-hard. In order to obtain a feasible solution, we further decompose the problem into three subproblems: simultaneous transmission scheduling, transmission duration allocation, and power allocation. Based on this, we propose a heuristic scheduling

algorithm that is referred to as conflict graph maximum independent set (CG-MIS) algorithm, where transmission links together with their availability of being simultaneously scheduled, are abstracted in a directed graph (the conflict graph), and links with "no conflict" are greedily selected by the maximum independent set (MIS). CG-MIS algorithm allows links to be transmitted at the same time, wherein more transmission duration can be allocated on the simultaneous links which eventually enhances the data rate. In particular, the potential of CG-MIS algorithm in boosting the system performance of point-to-point (P2P) communications, where we are especially interested in packet delivery in a vehicle platoon, is studied as an extension to the P2M scenario of the considered HetNets.

- The CG-MIS scheduling algorithm with transmission duration and power allocation to solve the joint optimization problem is based on fixed route decisions, i.e., data delivered from a BS to a UE, or vice versa, is along predefined paths regardless of network load. This may degrade the system performance when the traffic from different users is congested at some nodes, e.g., the last hop to the BS. Hence, we propose a dynamic routing (DR) algorithm, where the selection of path depends on real-time network statistics and user traffic is routed along lightly loaded links, to investigate the ability of further improvement in data rate. For the comparison of system performance between the fixed and dynamic routing, latency is also included in addition to data rate as a key indicator to better characterize the system performance.
- Extensive simulations have been conducted under numerous system parameters to demonstrate the efficiency of the proposed solutions in achieving the considerably high gain of data rate compared to the classical multiplexing schemes and interference mitigation approaches. The system performance of the proposed scheme under different frame structures, duplex schemes, and case studies are also analyzed.

The remainder of this chapter is organized as follows: Section 3.2 presents the

system model and formulates the joint optimization problem of scheduling and resource allocation. The CG-MIS scheduling algorithm, the transmission duration allocation algorithm, and the power allocation algorithm are proposed in Section 3.3. In Section 3.4, we introduce the DR algorithm. The performance of the proposed algorithms are evaluated by extensive simulations in Section 3.5, followed by a summary concluding this chapter in Section 3.6.

3.2 System Overview and Problem Formulation

In this section, we first introduce the network model, the available connection between BS and UE, as well as the frame structure that incorporates the concept of space-division multiple access (SDMA) group in which multiple links are simultaneously scheduled. Then, we describe the channel model that is applied to the calculation of the achievable SINR and capacity of scheduled links. Finally, the optimization problem of JSRA is formulated, where the complexity of the optimization problem is demonstrated to be NP-hard, such that we are motivated to develop heuristic algorithms to solve the problem efficiently. The important notations and system parameters defined in this section are summarized in Table 3.1 and will be used in the rest of this chapter.

3.2.1 Network, Connection, and Frame Structure

We build our model on the assumption where we avail of an initial access stage [28] in which each node (BS, AP, or UE) in a network detects and identifies its set of neighbors, and a cell discovery procedure has been successfully performed by UEs through the schemes described in Chapter 2. We consider a typical HetNet that consists of a macrocell BS, a set of small cell APs, and UEs that are associated either directly with the BS, or with the geographically closest AP and connected to the BS via multihop [127, 128]. Note that the association information is obtained

Table 3.1: System Model Parameters

Notation	Description	Value
T	Frame length	10 ms
$\delta_i^{(k)}$	Schedule indicator of link i in SDMA group k	See (3.4)
g_i	Antenna gain of link i (Antenna array vertical × horizontal)	16×8 for BS/AP 4×4 for UE
l_i	Pathloss of a link at carrier frequency f and distance d_i	See (3.1)
f	Carrier frequency	28 GHz
c	Speed of light	3×10^8 m/s
n_{L}	Pathloss exponent	LOS, NLOS: 2.1, 3.17
SF	Shadowing factor	LOS, NLOS: 2.38, 6.44
$p(d_i)$	Probability of a link with distance d_i to be LOS	See (3.2)
d_1	Parameters in d_1/d_2 model	20
d_2	Parameters in d_1/d_2 model	39
SINR_i	SINR of link i	See (3.3)
p_i	Transmission power of link i	See (3.18)
η	Thermal noise power	2×10^{-11} W
$n^{(k)}$	Number of slots in SDMA group k	See (3.14)
r_i	Channel capacity of link i	See (3.5)
b_i	Allocated bandwidth of link i	See (3.17)
$\boldsymbol{\delta}$	Scheduling Policy	See Definition 3.2.1
$\boldsymbol{n}$	Slot allocation Policy	See Definition 3.2.2
$\boldsymbol{p}$	Power allocation Policy	See Definition 3.2.3
$\mathcal{M}_{s,k}$	Set of links simultaneously transmitted from node s in SDMA group k	See (3.11)
$P_{\max}$	Maximum transmission power of BS/AP Maximum transmission power of UE	1 W for BS/AP 0.1 W for UE
B	System bandwidth	1 GHz
σ	Inter-link interference threshold	−50 dB

by the cell discovery, and we also consider that all nodes remain at fixed locations for the period of interest of the JSRA algorithm. Fig. 3.1 shows a typical example of the considered HetNet with one BS, two APs and four UEs.

We represent the network as a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ indicates the set of nodes (BS, APs, and UEs) and $\mathcal{E}$ indicates the set of links. For the exampled network in Fig. 3.1, an abstracted graph model, which we refer to as *link graph*, is illustrated in Fig. 3.2. Admissible connections are BS⇌AP, BS⇌UE, and AP⇌UE, with both downlink and uplink traffic flows along arbitrary routes as considered in IAB of 5G NR. Furthermore, due to the half-duplex constraint, only non-sequential links (i.e. edges whose destination nodes are not the source nodes of others, e.g.

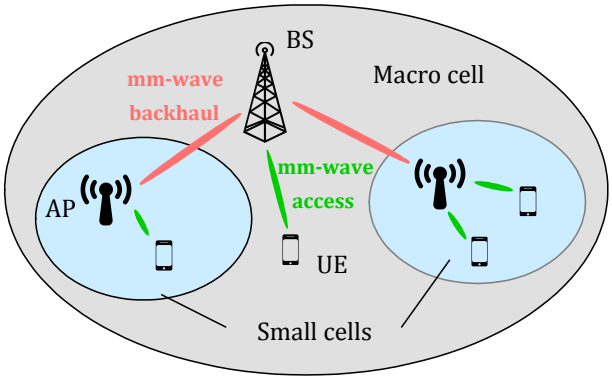


Figure 3.1: Illustration of an exampled HetNet consisting of one BS, two APs, and four UEs.

edge $A \rightarrow B$ and $A \rightarrow G$ in the link graph) can be active at the same time, as also addressed in IAB. Nevertheless, we also provide the performance analysis of HetNets will full-duplex constraints in Section 3.5.5 to investigate the efficiency of solutions considered in this chapter in achieving reasonable system performance under ubiquitous assumptions and configurations.

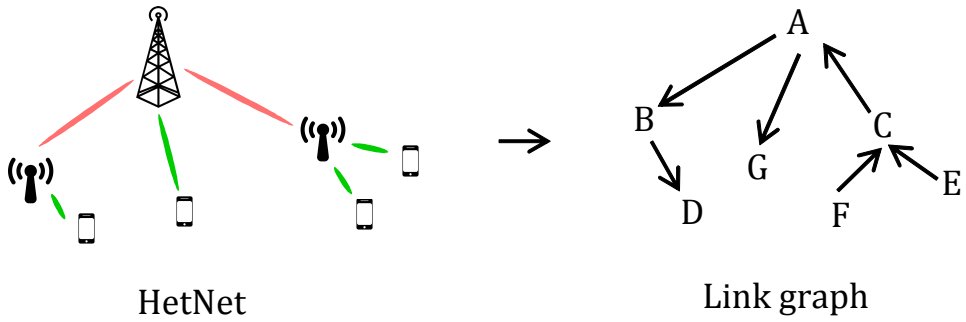


Figure 3.2: The HetNet is presented by a directed graph.

As addressed in Section 2.1 [103], BS and AP support mm-wave bands for backhaul and access transmission and reception (in-band backhauling). The transmission requests and corresponding time/frequency synchronization information are assumed to be collected by sub-6 GHz communications. For the data transmission of each UE associated with APs, a predefined route is selected, where we are able to focus on the JSRA optimization. Nevertheless, the design of DR is introduced in Section 3.4 as an extension to further improve the efficiency of the JSRA algorithm. More details will be elaborated in Section 3.4 and Section 3.5.

We consider a frame structure similar to Section 2.2.4 as shown in Fig. 3.3, where the system time is divided into consecutive frames with period T. Each cycle begins with a beacon phase, followed by a data transmission phase. Different from Chapter 2, in this chapter we concentrate on the data phase, which is modeled as a slotted-based time period in which transmissions between any valid pair of nodes can be allocated.

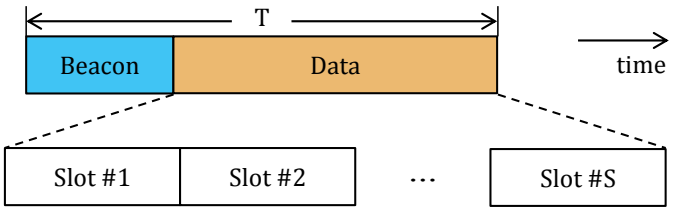


Figure 3.3: Frame structure including a slotted data phase.

TDMA is widely adopted for mm-wave channel access in 5G networks [113, 114, 116, 129]. Within the period of each frame, it is assumed that the network topology and channel condition remain unchanged [115, 130]. This assumption holds well for the case that all nodes remain at fixed locations, as mentioned above, with appropriately deployed BS and APs. In the TDMA scheme, each slot is exclusively occupied by a single link. An example of the slot allocation for six TDMA transmission links in a frame with ten-slot data phase is illustrated in Fig. 3.4, where the corresponding number of allocated slots are written below the links.

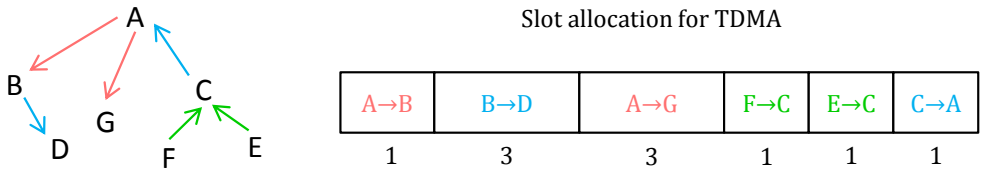


Figure 3.4: An example of slot allocation for TDMA.

By enabling the possibility of spatial multiplexing, multiple simultaneous transmissions can be scheduled in each slot. Hence, we can allocate more transmission duration to each link, such that the achievable data rate of each link is improved without other sophisticated scheme or algorithm. Fig. 3.5 provides an example of

the slot allocation for simultaneous transmission. Specifically, the number of slots allocated to link $A \to B$, $B \to D$, $A \to G$, $F \to C$, $E \to C$, and $C \to A$ are now 3, 3, 3, 4, 4, and 3, respectively, which are larger than or equal to the slots allocated in the TDMA case, which are 1, 3, 3, 1, 1, and 1, respectively.

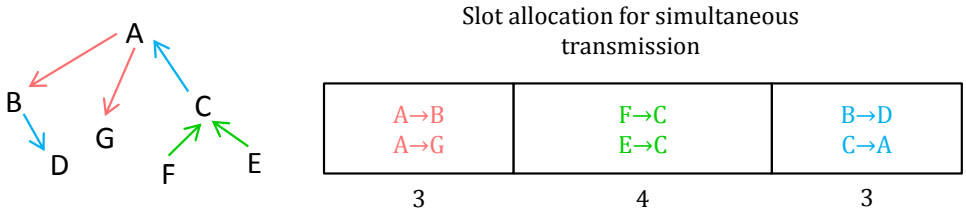


Figure 3.5: An example of slot allocation for simultaneous transmission.

3.2.2 Channel, Traffic, and Link SINR

For each pair of nodes that form a link $i = (m, n), m, n \in \mathcal{V}$, we define the logic indicator $\delta_i^{(k)} = 1$ if node m transmit towards node n (link i) at SDMA group k of a frame, and $\delta_i^{(k)} = 0$ otherwise. Here, a *SDMA group* is referred to as a transmission interval that consists of consecutive slots allocated to a link when enabling simultaneous transmission. For simplicity, in the rest of the chapter we use the term "group" to represent the SDMA group. In the example depicted in Fig. 3.5, the data phase consisting of ten slots is separated into three groups, in which different links are simultaneously scheduled. However, each link can be scheduled only once in a frame. Hence, the set of links allocated at group k is represented by the binary vector $\boldsymbol{\delta}^{(k)}$, also called a schedule[1].

A fundamental aspect of our models is that all nodes have adaptive beamforming capabilities as in [35, 131, 132]. This is a realistic assumption given the decreasing costs of circuitry for mm-wave frequencies, which will allow high dimensional antenna arrays to fit in a small form factor. This means that the antenna gain for a signal transmitted by a device will depend on the receiver and the intended destination,

[1]The terminology schedule $\boldsymbol{\delta}^{(k)}$ is simply the set of active links for group k, and a scheduling policy is the complete method for choosing schedules $\boldsymbol{\delta}^{(k)}$ across all time frames in order to solve the desired scheduling problem.

which can be realized by the installation of directional steerable antennas at all network nodes. In this chapter, we apply similar modeling of antenna gain as in Chapter 2 and denote the antenna gain of link i as g_i.

For an accurate mm-wave link capacity generation as a function of the distance between nodes, we compute the isotropic pathloss for link i transmitted from node m towards node n at distance d_i in meters, denoted as l_i, following the distribution in [78, 79] with the d_1/d_2 LOS probability model in [75, 76, 77], where link i transmitted from node m towards node n at distance d_i in meters is determined to be LOS or NLOS according to the LOS probability p_i, as follows:

$$l_i = \left(\frac{4\pi f}{c}\right)^2 \cdot {d_i}^{n_{\mathrm{L}}} \cdot \mathrm{SF}, \tag{3.1}$$

and

$$p(d_i) = \min\left(\frac{d_1}{d_i}, 1\right)\left(1 - e^{-d_i/d_2}\right) + e^{-d_i/d_2}. \tag{3.2}$$

Here, f indicates the carrier frequency in Hz, and n_{L} represents the pathloss exponent. c is the speed of light. Impact of objects such as trees, cars, etc., is modeled separately using the shadowing factor (SF). Further, an AWGN channel is assumed within all links, and the channel knowledge is assumed to be available at the BS. A similar expression can be found in Chapter 2.

Unlike the recent works [116, 118, 133] that have assumed a pseudo-wired behavior for mm-wave links with P2P transmission and the application of a power allocation mechanism in such a context is less relevant, we do not assume that the interference is negligible but rather compute real mutual interference between simultaneous links within a frame. This allows us to check the accuracy of non-pseudo-wired assumption and evaluate the efficiency of power allocation in enhancing the system performance. Then, the instantaneous SINR value of link i, denoted as SINR_i and computed based on the active links in the current network schedule $\boldsymbol{\delta}^{(k)}$, is modeled as

$$\mathrm{SINR}_i = \frac{p_i g_i l_i^{-1}}{\eta + \sum_j p_j g_j l_j^{-1}}, \tag{3.3}$$

where p_i, g_i, l_i, and η represent the transmission power, the antenna gain, the pathloss, the thermal noise power of link i, respectively. The experienced interference of link i is model by $\sum_j p_j g_j l_j^{-1}$ which summarizes the receive power of all interfered link j at link i. The main difference between mm-wave models and the models described in previous literature is the high spatial isolation, which causes the average sum of interfering power in (3.3) to shrink, but also results in stronger variations due to changes in scheduling.

3.2.3 Problem Formulation

We consider M transmission links to be scheduled in a given frame that consists of N slots. These slots are allocated into K groups, and the number of slots allocated in each group is denoted as $n^{(k)}$. Here, a group indicates a transmission interval that consists of consecutive slots and within each group, multiple links can be simultaneously scheduled, as defined in Section 3.2.2.

Remember that in Section 3.2.2, we defined a logic indicator $\delta_i^{(k)}$ that controls the scheduling policy of whether link i is scheduled in group k. Specifically, the policy is modeled as

$$\delta_i^{(k)} = \begin{cases} 1, & \text{Link } i \text{ is scheduled in group } k, \\ 0, & \text{Otherwise.} \end{cases} \tag{3.4}$$

Then, the actual capacity of link i, denoted as r_i, can be represented according to Shannon channel capacity equation as

$$r_i = b_i \cdot \log_2 \left(1 + \delta_i^{(k)} \mathrm{SINR}_i\right), \tag{3.5}$$

where b_i indicates the available bandwidth at link i in Hz.

Before providing the representation of the achievable rate of links in the considered frame, we first introduce the following definition:

Definition 3.2.1. *A scheduling policy $\boldsymbol{\delta}$ for all links in the considered frame is the*

set of schedule binary vector $\boldsymbol{\delta}^{(k)}$, $\forall k \in \{1, \ldots, K\}$.

Similarly, the slot allocation and the power allocation for all links in the considered frame are derived from:

Definition 3.2.2. *A slot allocation policy* $\boldsymbol{n}$ *for all links in the considered frame is the set of slot allocation for each group* $n^{(k)}$, $\forall k \in \{1, \ldots, K\}$.

Definition 3.2.3. *A power allocation policy* $\boldsymbol{p}$ *for all links in the considered frame is the set of power allocation for each link* p_i, $\forall i \in \{1, \ldots, M\}$.

Then, we can define the achievable data rate for all M links in the considered frame with K groups, given the scheduling policy $\boldsymbol{\delta}$, the slot allocation policy $\boldsymbol{n}$, and the power allocation policy $\boldsymbol{p}$, as

$$\sum_{i=1}^{M} \sum_{k=1}^{K} \frac{r_i n^{(k)}}{N}. \tag{3.6}$$

As the total number of slots N in each frame is fixed, the objective function, which is the achievable data rate defined in (3.6) and to be maximized, can be further simplified as

$$\sum_{i=1}^{M} \sum_{k=1}^{K} r_i n^{(k)}. \tag{3.7}$$

Now, we analyze the system constraints of the optimization problem. First of all, the scheduling policy $\boldsymbol{\delta}$ is ruled by the following two constraints:

$$\sum_{k=1}^{K} \delta_i^{(k)} = 1, \ \forall i \in \{1, \ldots, M\}, \tag{3.8}$$

and

$$\delta_i^{(k)} + \delta_j^{(k)} \leq 1, \ \forall \text{ sequential link } i \text{ and } j, \ \forall i, j \in \{1, \ldots, M\}, \forall k \in \{1, \ldots, K\}. \tag{3.9}$$

Here, (3.8) indicates that each link can be scheduled only once in a frame (in one of the K groups), as demonstrated in Section 3.2.2. However, each link scheduled in group k can be allocated with multiple slots, namely $n^{(k)}$. Further, due to half-duplex

constraint, the sequential links, i.e., edges whose destination nodes are the source nodes of others in the link graph (e.g. edge $A \rightarrow B$ and $B \rightarrow D$ in Fig. 3.2) cannot be scheduled in the same group, which is governed by the constraint demonstrated in (3.9).

Next, the constraint on the slot allocation policy $\boldsymbol{n}$, where the total number of allocated slots in all the K groups should not be larger than N (total number of slots in a frame), is represented as

$$\sum_{k=1}^{K} n^{(k)} \leq N. \tag{3.10}$$

Lastly, the summarized allocated power of the links transmitted from the same node, say node s, in group k, should not exceed the total available transmission power. Denoting the set of links simultaneously transmitted from node s in group k as $\mathcal{M}_{s,k}$, the constraint on the power allocation policy $\boldsymbol{p}$ is given by

$$\sum_{i \in \mathcal{M}_{s,k}} p_i \leq P_{\max}, \tag{3.11}$$

where $P_{\max}$ indicates the maximum transmission power of BS/AP/UE.

Finally, we can formulate our JSRA optimization problem to maximize the data rate r_i, under the constraints of scheduling policy $\boldsymbol{\delta}$, the slot allocation policy $\boldsymbol{n}$, and the power allocation policy $\boldsymbol{p}$, as follows:

Problem 3.1. *(JSRA optimization)*

$$\begin{aligned} \max_{\boldsymbol{\delta},\boldsymbol{n},\boldsymbol{p}} \quad & \sum_{i=1}^{M} \sum_{k=1}^{K} r_i n^{(k)}, \\ \text{s.t.} \quad & \text{constraints (3.8)–(3.11).} \end{aligned} \tag{3.12}$$

The maximization problem indicated in (3.12) with constraints (3.8)–(3.11) is a MINLP problem [134], where the variables are categorized as

- Integer variables: $n^{(k)}$,
- Continuous variables: p_i,

- Binary variables: $\delta_i^{(k)}$.

In general, one among the simplest MINLP problem is the 0–1 Knapsack problem, which is a class of problems that are typically NP-hard and concerned with selecting from a set of given items, each with a specified weight and value, a subset of items whose sum does not exceed the prescribed capacity and whose value sum is maximized [135, 136]. Nevertheless, the multiplication of the above variables in (3.12) further complicates the proposed optimization problem 3.1 and makes it even more complex than the 0–1 Knapsack problem. Specifically, there exists the third-order term $\delta_i^{(k)} p_i n^{(k)}$ in (3.12), in which the coupling exists among the constraints where slot and power allocation policies rely on the results of link scheduling. Therefore, the considered optimization problem 3.1 is also NP-hard and is not capable to be solved by traditional algorithms like dynamic programming or branch-and-bound. The proof of the unavailability of applying these algorithms to solve problem 3.1 is provided in Appendix 3.7.1. In the next section, we propose a heuristic JSRA algorithm to decouple the correlated variables and solve problem 3.1 efficiently with low complexity.

3.3 Scheduling and Resource Allocation Algorithm

As mentioned in Section 3.2, the scheduling policy $\boldsymbol{\delta}$, the slot allocation policy $\boldsymbol{n}$, and the power allocation policy $\boldsymbol{p}$ are three key and correlated terms for solving the optimization problem 3.1. However, the slot and power allocation policy cannot be determined until a scheduling decision is made. Specifically, the slot allocation depends on the demand of links that are scheduled in each group, and the number of scheduled links affects the power allocation policy of the corresponding group. Therefore, in this section we propose a heuristic JSRA algorithm to solve problem 3.1, where simultaneous transmission is fully exploited to increase the data rate. In Section 3.3.1, we first transfer the presented link graph by another directed graph, which we refer to as the *conflict graph*, to illustrate the "conflict" relationship

caused by scheduling constraints in (3.8) and (3.9) among difference links. Then, we propose a scheduling algorithm which we refer to as the CG-MIS algorithm introduced in Section 3.1.2, where MIS is employed to determine the scheduling policy in each group. Base on the scheduling result, the slot allocation algorithm and the power allocation algorithm are presented in Section 3.3.2 and Section 3.3.3, respectively, where the number of slots and transmission power are allocated to the links in different groups.

3.3.1 CG-MIS Scheduling Algorithm

As mentioned above, transmission link scheduling is constrained by (3.8) and (3.9). To better demonstrate the relationship among different links presented by these equations, we introduce an appropriate concept named conflict graph. Remember that in the link graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, as depicted in Fig. 3.2, nodes (set $\mathcal{V}$) represent BS, AP, and UE, and edges (set $\mathcal{E}$) represent transmission links among these nodes. In a conflict graph, as a comparison, each node represents one link in the network, and there is an edge connecting two nodes if "conflict" exists. Specifically, the conflict can be derived from the sequential links (e.g., backhaul and access links) as described in (3.9), due to half-duplex constraint, or from links that severely interfere with each other. For the latter case, we define an interference threshold σ as the criterion that judges whether an edge exists between two non-sequential links (nodes)[2]: An edge exists if the receive power of the interference signal of any of the two corresponding links is larger than σ. An example of conflict graph construction is illustrated in Fig. 3.6.

As shown in Figure 3.6, link $A \rightarrow B$ and $B \rightarrow D$ are sequential links and due to half-duplex constraint in (3.9), they cannot be simultaneously scheduled. For the interference constraint, we assume that severe interference is experienced by link

[2]The interference information is assumed to be acquired from e.g. interference sensing procedure. This issue, as well as other signaling designs, have been addressed by the other works of us and are beyond the scope of this chapter.

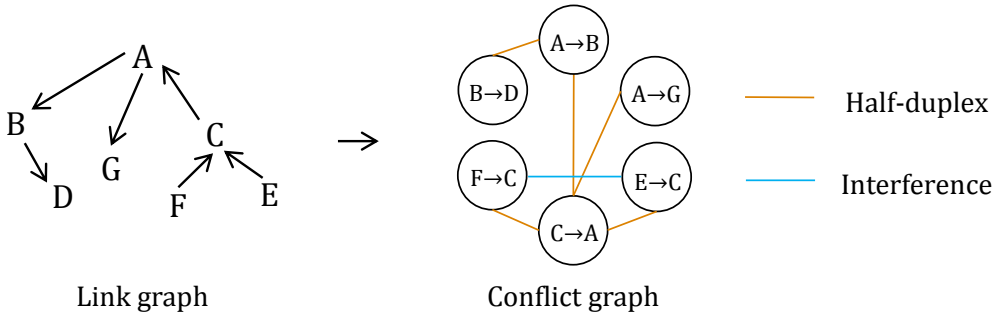


Figure 3.6: Conflict graph construction.

$F \rightarrow C$, as an example, when link $E \rightarrow C$ is scheduled to be transmitted at the same time.

Based on the conflict graph, we propose an MIS based scheduling algorithm, namely the CG-MIS algorithm, to distribute links into different groups. We first provide the definition of the independent set as follows [137]:

Definition 3.3.1. *In graph theory, an independent set is a subset of nodes in a graph, in which no pair of nodes (two nodes) are adjacent.*

Then, a MIS can be defined as:

Definition 3.3.2. *A maximum independent set is either an independent set such that adding any other node to the set forces the set to contain an edge or the set of all nodes of an empty graph.*

It is demonstrated in [138] that the computational complexity of finding the MIS of a general graph is NP-hard, and there does not exist efficient algorithms that are able to find the optimal solution in polynomial time [139]. Therefore, we utilize the minimum-degree greedy algorithm to solve the problem of finding the MIS of a conflict graph [140]. The algorithm iteratively picks the maximum number of nodes with the minimum degrees from the conflict graph, i.e., the links that are either most unlikely to be sequential or to be interfered by other links are simultaneously scheduled in each group, until all nodes have been traversed (every link is scheduled in one of the groups). The detail of the algorithm is elaborated as follows[3].

[3]The scheduling, the slot allocation, and the power allocation on all backhaul and access links

We denote the conflict graph as $\mathcal{G}_{\mathrm{C}}(\mathcal{V}_{\mathrm{C}}, \mathcal{E}_{\mathrm{C}})$, where $\mathcal{V}_{\mathrm{C}}$ and $\mathcal{E}_{\mathrm{C}}$ represent the set of nodes and the set of edges in the conflict graph, respectively. For any node $v \in \mathcal{V}_{\mathrm{C}}$, we define its neighbors as $\mathcal{N}_v$, which consists of all adjacent nodes of v. The degree of any node $v \in \mathcal{V}_{\mathrm{C}}$ is denoted by Δ_v.

We summarize the CG-MIS scheduling algorithm in Algorithm 3.1. In the algorithm, links are iteratively scheduled into each group until all links have been traversed, as indicated in line 1. At the beginning of each iteration, the set for recording the scheduled links in the corresponding group (say group k), denoted as $\mathcal{V}^k$, is initialized as empty. Line 5–9 describe the minimum-degree greedy scheduling algorithm of group k. In line 10, the traversed links are removed from the set $\mathcal{V}_{\mathrm{C}}$, which is evaluated as the non-traversed set $\mathcal{V}^{k,u}$ for next group in line 4.

The computational complexity of our algorithm is $\mathcal{O}(|\mathcal{V}_{\mathrm{C}}|^2)$, compared to $\mathcal{O}(|\mathcal{V}_{\mathrm{C}}|^2 \cdot 2^{|\mathcal{V}_{\mathrm{C}}|^2})$ of naive brute force scheme, where $|\mathcal{V}_{\mathrm{C}}|$ indicates the cardinality of the node set $\mathcal{V}_{\mathrm{C}}$, namely the number of nodes in the set. The performance analysis of the algorithm is presented as below.

Lemma 3.3.1. *Minimum-degree greedy algorithm outputs an independent set* $\{\mathcal{V}^k | k = 1, \ldots, K\}$ *such that* $|\mathcal{V}^k| \geq \frac{|\mathcal{V}_{\mathrm{C}}|}{\Delta+1}$ *where* Δ *is the maximum degree of any node in the graph.*

Proof. The proof is provided in Appendix 3.7.2. □

Corollary 3.3.1. *Minimum-degree greedy algorithm gives a* $\frac{1}{\Delta+1}$ *approximation for MIS in graphs of degree at most* Δ*.*

Proof. A straightforward result of Lemma 3.3.1. □

Having the performance approximation, we further define the performance ratio of the greedy algorithm as the worst-case ratio of the size of the optimal solution (naive brute force scheme) to the size of the greedy algorithm's solution. Then, the

are assumed to be centralized processed at the BS. The signaling for scheduling decision and resource allocation, which is similar to the interference sensing procedure, has also been addressed by the other works of us and are beyond the scope of this chapter.

Algorithm 3.1: CG-MIS Link Scheduling Algorithm

Input: Conflict graph $\mathcal{G}_{\mathrm{C}}(\mathcal{V}_{\mathrm{C}}, \mathcal{E}_{\mathrm{C}})$
Output: Scheduling policy $\boldsymbol{\delta}$

- $\mathcal{V}_{\mathrm{C}}$: Set of nodes
- $\mathcal{E}_{\mathrm{C}}$: Set of edges
- $\mathcal{V}^k$: Set of scheduled links in group k
- $\mathcal{V}^{k,u}$: Set of unscheduled links in group k
- Δ_v: Degree of node v
- $\mathcal{N}_v$: Neighbors of node v
- k: Iterator (index of group)

Initialization: $k = 0$
begin
1 **while** $\mathcal{V}_{\mathrm{C}} \neq \varnothing$ **do**
2 $k = k + 1$;
3 $\mathcal{V}^k = \varnothing$;
4 $\mathcal{V}^{k,u} = \mathcal{V}_{\mathrm{C}}$;
5 **while** $\mathcal{V}^{k,u} \neq \varnothing$ **do**
6 Get $v \in \mathcal{V}^{k,u}$ where $\Delta_v = \min_{v' \in \mathcal{V}^{k,u}} \Delta_{v'}$;
7 $\mathcal{V}^k = \mathcal{V}^k \cup v$;
8 $\mathcal{V}^{k,u} = \mathcal{V}^{k,u} - \{v \cup \mathcal{N}_v\}$;
9 **end**
10 $\mathcal{V}_{\mathrm{C}} = \mathcal{V}_{\mathrm{C}} - \mathcal{V}^k$;
11 **end**
12 **Return** $\mathcal{V}^k$ for each group k;
end

performance of the minimum-degree greedy algorithm is quantized by the following theorem.

Theorem 3.3.1. *Minimum-degree greedy algorithm achieves a* $\frac{\Delta+2}{3}$ *performance ratio for MIS in graphs of degree at most* Δ.

Proof. The proof is provided in Appendix 3.7.3. □

As the maximum degree of a graph varies with graph topology, we are interested in the performance comparison between greedy algorithm and brute force scheme in general case, where the performance ratio can be characterized by the average degree of the graph $\bar{d}$. This is also presented in the following theorem:

Theorem 3.3.2. *Minimum degree greedy algorithm achieves a* $\frac{2\bar{d}+3}{5}$ *performance ratio for MIS in graphs of average degree* $\bar{d}$.

Proof. The proof is provided in [140] in combination with a fractional relaxation technique in [141] and [142]. □

3.3.2 Slot Allocation Algorithm

With the simultaneous transmission scheduling result $\{\mathcal{V}^k\}$ obtained from Algorithm 3.1, we propose a proportional fair time resource allocation scheme to determine the transmission duration (slot) for each group. This algorithm calculates the number of slots for each link to satisfy its data rate requirement by TDMA (exampled in Fig. 3.4), assuming the maximum transmission power is allocated to each link, and proportionally allocates the slots to each group according to the maximum required number of slots of all links in the corresponding group. As more time slots are allocated to each link, the achievable data rate of these links are increased.

We denote the number of slots required by link i in TDMA scheme as n_i, then the maximum number of required slots among all links in group k, denoted as $n_{\max}^{(k)}$, can be obtained by

$$n_{\max}^{(k)} = \max_{i \in \mathcal{V}^k} n_i. \tag{3.13}$$

Based on this, the total N slots in a frame are allocated to each group proportionally to its maximum number of required slots $n_{\max}^{(k)}$. Hence, the number of slots allocated to all links in group k, denoted as $n^{(k)}$, can be calculated as

$$n^{(k)} = \left\lfloor \frac{n_{\max}^{(k)}}{\sum_k n_{\max}^{(k)}} \cdot N \right\rfloor, \tag{3.14}$$

where $\lfloor . \rfloor$ represents the floor function. Pseudo code of the time slot allocation algorithm is summarized in Algorithm 3.2. In the algorithm, firstly, the required number of slots of all links scheduled in each group are collected, as indicated in line 3–5. Then, for each group k, the maximum number of slots $n^{(k)}$ are calculated as in line 6. Based on this, the actual allocatable number of slots are determined for all groups, as indicated in line 9–12. As each link can only be scheduled into one group, the computational complexity of the algorithm is $\mathcal{O}(|\mathcal{V}_\mathrm{C}|)$.

Algorithm 3.2: Slot Allocation Algorithm

Input: Set of scheduled links in group $\{\mathcal{V}^k\}$, number of required slot n_i
Output: Slot allocation policy $\boldsymbol{n}$

- $n_{\max}^{(k)}$: Maximum number of slots of links in group k
- $n^{(k)}$: Allocated number of slots for links in group k
- k: Iterator (index of group)

Initialization: $k = 0$
begin
  while $k \neq K$ **do**
    $k = k + 1$;
    foreach link $i \in \mathcal{V}^k$ **do**
      Get the number of required slots n_i;
    end
    $n_{\max}^{(k)} = \max_{i \in \mathcal{V}^k} n_i$;
  end
  $k = 0$;
  while $k \neq K$ **do**
    $k = k + 1$;
    $n^{(k)} = \left\lfloor \frac{n_{\max}^{(k)}}{\sum_k n_{\max}^{(k)}} \cdot N \right\rfloor$;
  end
  Return $n^{(k)}$ for each k;
end

3.3.3 Power Allocation Algorithm

By enabling spatial multiplexing, at some nodes (BS, AP), there will be multiple links to be simultaneously transmitted. Hence, power control is required at these nodes to fulfill the power splitting in P2M transmission situation. Specifically, after link scheduling and slot allocation, constraints (3.8)–(3.10) are satisfied and consequently, the optimization problem 3.1 in (3.12) can be reformed as

Problem 3.2. *(Power allocation optimization)*

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{i \in \mathcal{V}^k} \sum_{k=1}^{K} r_i, \\ \text{s.t.} \quad & \sum_{i \in \mathcal{M}_{s,k}} p_i \leq P_{\max}. \end{aligned} \tag{3.15}$$

Here, in (3.15), the objective function first summarizes the achievable channel capacity instead of data rates, as the slot allocation policy of all links scheduled in group $\mathcal{V}^k$ (i.e., $\sum_{i\in\mathcal{V}^k}$) has been determined, and further summarizes all groups of a frame (i.e., $\sum_{k=1}^{K}$). As the power allocation of links in one group is independent of that in the others, and the total transmission power of BS/AP remains unchanged for each group, the above maximization problem 3.2 provided by (3.15) can be further relaxed as

Problem 3.3. *(Relaxed power allocation optimization)*

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{i\in\boldsymbol{M}_{s,k}} \log(1+\frac{g_i l_i^{-1}}{\eta}p_i), \\ \text{s.t.} \quad & \sum_{i\in\mathcal{M}_{s,k}} p_i \le P_{\max}. \end{aligned} \tag{3.16}$$

Here, we assume that the total system bandwidth B of node s is averaged to all links in $\mathcal{M}_{s,k}$, i.e.,

$$b_i = \frac{B}{|\mathcal{M}_{s,k}|}. \tag{3.17}$$

Hence, the term b_i is eliminated in (3.16) from (3.3). Moreover, the term $\sum_j p_j g_j l_j^{-1}$ that describes the experienced interference power of link i is also eliminated in (3.16), as all the other links scheduled in group V^k are isolated from link i, namely these links lead to interference power less than the threshold σ, and are allowed to be simultaneously transmitted with link i under tolerable interference level. Nevertheless, we still calculate the actual SINR of link i for performance evaluation in Section 3.5, as addressed in Section 3.1.2.

Denoting $\frac{g_i l_i^{-1}}{\eta}$ as γ_i, which we refer to as channel quality, the relaxed optimization problem 3.3 provided by (3.16) can be solved by the following theorem (known as water-filling solution).

Theorem 3.3.3. *The optimal solution of problem 3.3 is given by*

$$p_i^* = \max\Big\{\frac{1}{\phi^*} - \frac{1}{\gamma_i}, 0\Big\}, \tag{3.18}$$

where p_i^* is the optimal transmission power allocated to link i, and the optimal Lagrangian multiplier ϕ^* is given by

$$\sum_{i \in \mathcal{M}_{s,k}} \max\left\{\frac{1}{\phi^*} - \frac{1}{\gamma_i}, 0\right\} = P_{\max}. \tag{3.19}$$

Proof. The proof is provided in Appendix 3.7.4. □

The pseudo code of the power allocation algorithm is presented in Algorithm 3.3. In the algorithm, firstly, the channel quality γ_i is sorted in a descending order, as indicated in line 1. The reason behind this falls into the fact that according to (3.18), once the optimal Lagrangian multiplier ϕ^* is determined, all links that $\gamma_i \leq \phi^*$ will be allocated to zero power, and in turn, these links do not contribute in the calculation of the optimal Lagrangian multiplier ϕ^* in (3.19). With the descending γ_i, the corresponding Lagrangian multiplier ϕ_i of each link is derived as indicated in line 2–4. Then, the index of the optimal Lagrangian multiplier m is acquired by finding the largest Lagrangian multiplier greater than the reciprocal of the channel quality of the corresponding link, as indicated in line 5–7. Ultimately, the optimal Lagrangian multiplier ϕ^* is determined in line 8 and the transmission power is allocated to each link as indicated in line 9–11. Similar to the slot allocation algorithm, the computational complexity of the power allocation algorithm is also $\mathcal{O}(|\mathcal{V}_{\mathrm{C}}|)$, as each link can only be scheduled into one group.

3.4 Multihop Routing Algorithm Design for Path Selection

In the previous section, we assume that the multihop path connecting BS and UE is predefined, where the JSRA algorithm is designed based on a fixed network topology. This assumption does not fully exploit the freedom given by a reconfigurable mm-wave backhaul, which can be flexibly selected as an intermediate hop for the path connecting BS and UE, and the selection depends on real-time network load, i.e., the

Algorithm 3.3: Power Allocation Algorithm

Input: Channel quality γ_i, total transmission power $P_{\max}$, set of links $\mathcal{M}_{s,k}$ transmitted from sender s in group k

Output: Power allocation policy $\boldsymbol{p}$

- ϕ_i: Lagrangian multiplier for each link i
- m: Index of optimal Lagrangian multiplier
- ϕ^*: Optimal Lagrangian multiplier
- p_i: Allocated transmission power for link i

Initialization: $j = 0$

begin

Sort γ_i in descending order;

foreach link $i \in \mathcal{M}_{s,k}$ **do**

$\frac{1}{\phi_i} = \frac{P_{\max} + \sum_{j=1}^{i} \frac{1}{\gamma_j}}{i}$;

end

foreach ϕ_i **do**

Find m where $\frac{1}{\phi_m} > \frac{1}{\gamma_m}$ and $\frac{1}{\phi_m} \leq \frac{1}{\gamma_{m+1}}$;

end

$\frac{1}{\phi^*} = \frac{P_{\max} + \sum_{i=1}^{m} \frac{1}{\gamma_i}}{m}$;

foreach link $i \in \mathcal{M}_{s,k}$ **do**

$$p_i = \begin{cases} \frac{1}{\phi^*} - \frac{1}{\gamma_i}, & \text{for } i \in \{1, \ldots, m\} \\ 0, & \text{for } i \in \{m+1, \ldots, |\mathcal{M}_{s,k}|\} \end{cases};$$

end

Return p_i for each link i;

end

achievable data rate that varies for different paths consisting of different intermediate hops. In this section, we propose a DR algorithm for path selection, where the optimal path is greedily generated for each user in terms of achievable data rate taking into account the data rates achieved by applying the JSRA algorithm to the existing users in the network.

3.4.1 Preliminaries

As mentioned in Section 3.2.1, UEs are associated either directly with BS, or with the geographically closest AP and connected to the BS via multihop. Depending on different network layouts specified by various use cases, this assumption can be further modified as UEs are associated with the network node (BS/AP) from

which the strongest signal power is detected. Nevertheless, the first hop for the multihop path from UE to BS (or the last hop from BS to UE) is fixed, and the reconfigurability of intermediate hops is limited to the backhaul links. Clearly, there is no need to route the UEs that connect directly to the BS with a single hop.

In general, we expect our system could benefit from the DR algorithm to achieve a higher data rate compared to the fixed routing. As can be observed from (3.6), the data rate is proportional to the allocatable slots and the channel capacity. Moreover, in the proposed power allocation algorithm, the more connections a hop is shared by, the less transmission power each connection will get. Therefore, the situation that some hops have crowded with connections while others remain vacant seems not to be an attractive arrangement.

3.4.2 Comparison of Routing Algorithms

From the above description, one can see that the crux in our routing algorithm is finding the path between each UE and BS. Below we provide a discussion of classical routing algorithms and the insight of their availability in our system.

First and foremost, minimum weight spanning tree (MWST) algorithms appear to be not the qualified candidates, where an MWST is the spanning tree (a subgraph that is a tree and connects all nodes of a graph) with the minimum sum of weight [143]. Applying this to our system, each UE will find a path to communicate with the BS through some hops within the MWST. As each UE is willing to transmit on good links, setting the edge weight as 1/SINR of the corresponding hop seems reasonable, then finding the MWST turns to be finding the maximum SINR spanning tree.

However, this maximum SINR spanning tree maximizes only the summation of SINR that ensures the graph connected. It could happen that some of the hops are crowded with connections as they are the only ways for some UEs to reach the BS. Fig. 3.7 shows an example of maximum SINR spanning tree. As illustrated,

if MWST algorithm is implemented, the hop $D \rightarrow E$ with weight 4 is shared by the transmission flows from the BS (marked as node D) to node A, B, E, and F. This will lead to a risk of congestion in hop $D \rightarrow E$, where the resource allocated to all flows on this hop is reduced. By contrast, one can imagine that if the BS transmits data to node B in the blue path, the hop $D \rightarrow E$ will be shared by only two connections, and thus each connection will get more resource than the previous case.

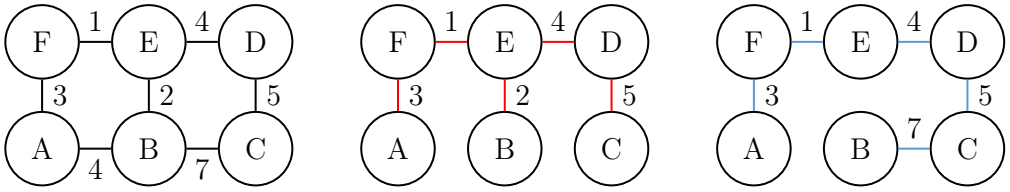


Figure 3.7: An example of maximum SINR spanning tree.

In other words, an MWST considers only the connectivity of a graph where not all edges are involved in the spanning tree. This leads to a waste of resource that inevitably causes the risk of congestion in some hops while others are totally unused. Hence, the allocatable resource for each UE is limited, which degrades the data rate correspondingly.

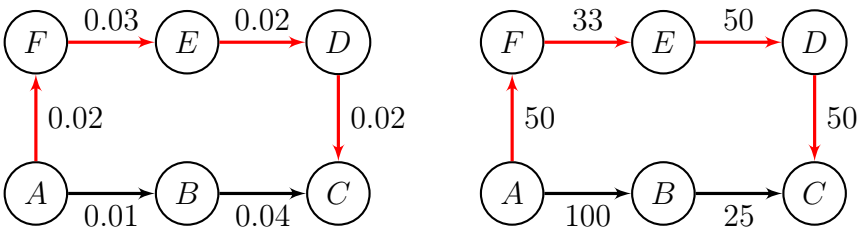
Furthermore, the minimum end-to-end outage routing algorithm and its derived versions are also not suitable. For a large-scale network, calculating all possible paths between BS and UEs is apparently a considerable work. In addition, when more BSs are involved in the network, the time needed for generating the path between all BS-UE pairs could double or even quadruple, as all paths for different BS to each single UE need to be found. Therefore, the minimum end-to-end outage routing algorithm seems to be not viable in our scenario. Besides, ad-hoc routing and N-hop routing reduce the computational complexity but also decrease system performance. The suboptimal solutions of these algorithms are not able to achieve the best data rate and should be improved. Hence, all the algorithms related to the minimum end-to-end outage are not feasible in our scenario.

At last, shortest path routing (SPR) algorithms seem more proper to be used in solving our problem compared to the above-mentioned algorithms, but still not fully fit. As one of the most famous shortest path algorithms, Dijkstra's algorithm [144] decreases the risk of congestion in the MWST algorithm by increasing the possibility of all edges in the network to become a hop of a path, where an individual path is selected for each UE to connect with BS. Meanwhile, Dijkstra's algorithm achieves less computational complexity than the other SPR algorithm, e.g. Bellman-Ford algorithm, which is ideal for large-scale networks.

However, there are still some aspects in Dijkstra's algorithm that cannot be performed as the path selection algorithm for our system. First of all, high SINR leads not necessarily to high data rate. As indicated in (3.6), the higher/more the SINR/slots achieved by/allocated on each link, the higher the achievable data rate will be. Dijkstra's algorithm finds the path with the smallest summation of edge weights, where most of the hops involved in the path experience relatively high SINRs. These high-SINR hops fulfill the requirement of good link quality but risk the potential congestions on these hops. Specifically, as the path for each UE is greedily selected taking into account all the existing UEs, these high-SINR hops are very likely to be selected for all UEs when applying Dijkstra's algorithm to find a path. It is obvious that in this case, the allocatable resources for transmission flows on those hops decrease, which eventually limits the achievable data rate.

More importantly, minimizing the summation of the edge weights does not necessarily maximize the achievable data rate even by setting the edge weight as 1/SINR, . Let us see an example in Fig. 3.8(a). When applying Dijkstra's algorithm, the path composed of black arrows will be selected as the path connecting node A and C, where the summarized edge weight is 0.05 compared to the other path that is composed of red arrows with 0.09 summarized edge weight. Assume that in the considered case of Fig. 3.8(a), the allocated slots of all hops are same. Then, the achievable data rates of these hops depend exclusively on their SINRs. By taking reciprocal of the edge weight, namely the SINR of each hop, the achievable data

rate for the black flow, given the fact that the achievable data rate is limited by the "weakest" hop on which the lowest data rate is achieved, is proportional to 25, which is less than the achievable data rate (proportional to 33) acquired by the red flow, as shown in Fig. 3.8(b). Obviously, the shortest path does not achieve the highest data rate, which is reversely achieved by the other path.



(a) Edge weight: 1/SINR (b) Edge weight: SINR

Figure 3.8: An example that shows the shortest path may not maximize data rate.

3.4.3 Routing Algorithm Design

Based on the above observations, we introduce our DR algorithm as follows. The network is first established as the link graph depicted in Fig. 3.2. With the network topology, the achievable data rate of each link is calculated as the initial value. The routing algorithm terminates when the path of all UEs are generated. A flow chart of the proposed DR algorithm is illustrated in Fig 3.9.

In the flow chart, the proposed routing algorithm first initializes the network where the set of UE, denoted as $\mathcal{U}$, is generated and the achievable data rate of each hop is calculated. Afterward, $\mathcal{U}$ is checked and if the set is empty, which means there is no UE that requires a multihop path to connect to the BS or all UEs have already found a path, the algorithm will terminate. Otherwise, we pick the first UE for which the path selection algorithm is executed to find the optimal path between the BS and the UE (UE_i). Then, the network is updated where the achievable data rate of each hop is recalculated. The algorithm terminates when all UEs in $\mathcal{U}$ have been distributed with a path. The two key steps of the proposed DR algorithm, namely path selection and network update, are elaborated in the following sections.

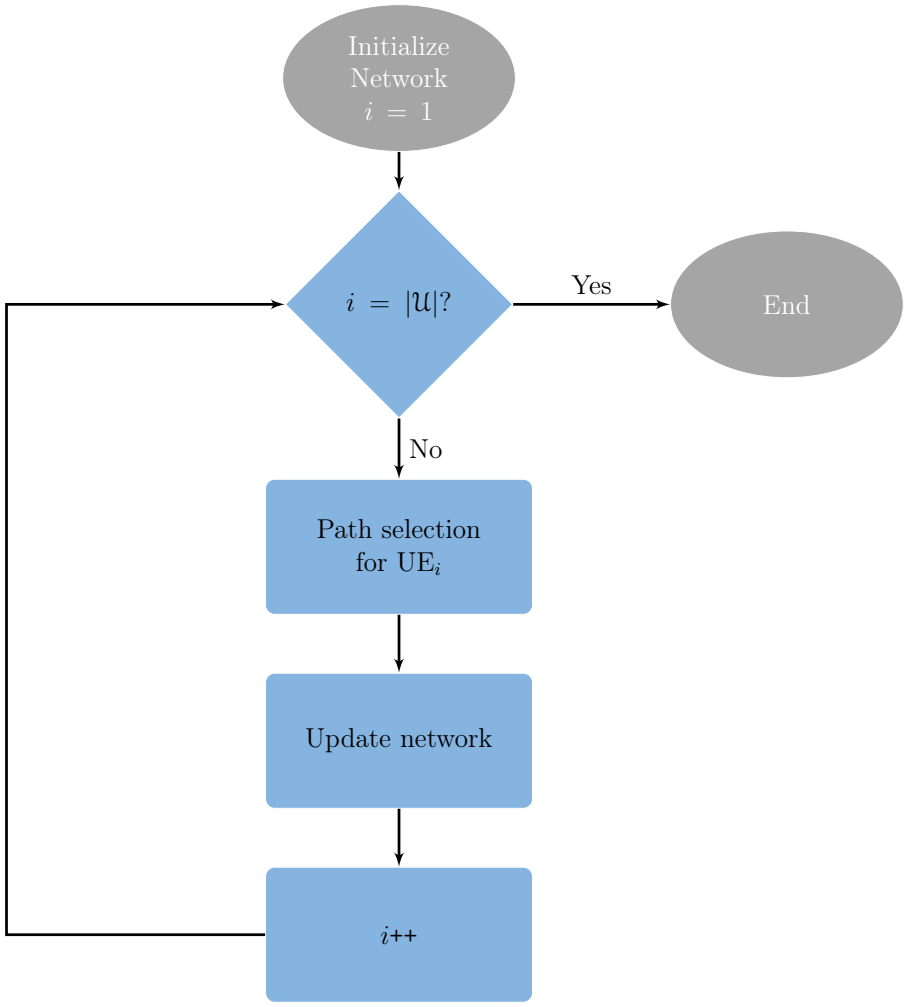


Figure 3.9: Flow chart of DR algorithm.

3.4.3.1 Path Selection Algorithm

For a multihop transmission between BS and UE, the source and destination node pair (s, d) is selected as (BS,UE) or (UE,BS) for either downlink or uplink situation. Further, the achievable data rate is used as edge weight. The pseudo code of the proposed path selection algorithm is provided in Algorithm 3.4.

In the beginning, both the set of traversed node $\mathcal{V}_{\mathrm{t}}$ and the set of next node $\mathcal{V}_{\mathrm{n}}$ are initialized as empty. The set of current node $\mathcal{V}_{\mathrm{c}}$ contains only the source node s, and for each node v except s, the weight of path starting from s and ending at it, referred to as $w(v)$ and defined as the minimum edge weights along the path (the

achievable rate of the path), is initialized as ∞, which means that in the beginning all the nodes are not connected. Then, the algorithm picks every v in the set of current node $\mathcal{V}_\mathrm{c}$ (starting from s), and finds all its neighboring node u in the set of non-traversed node $\mathcal{V}\backslash\mathcal{V}_\mathrm{t}$, as described in line 1–3. Afterwards, there are three criteria for judging whether u can be added to the path from s to v as the new destination node:

- The weight of u is infinity (the node has not been traversed) or adding u to the path will not change the weight of path $w(u)$, as indicated in line 4–7. In this case, the path from s to v then to u is one of the optimal path from s to u in terms of maximizing achievable data rate.
- Adding the node to the path will increase the weight of path $w(u)$. In this case, the optimal path from s to u in terms of maximizing the achievable data rate goes exclusively through v, as by adding the edge (v, u), the minimal weight of the path is increased. Therefore, the other paths from s to u through other parent nodes of u in the set $\mathcal{P}(u)$ should be eliminated and v is the only parent node of u, as described in line 8–11.
- Otherwise, adding u to the path may decrease the weight of path $w(u)$, which is not the desired result for maximizing the weight of the path and will not be considered.

When the sets of currents nodes $\mathcal{V}_\mathrm{c}$ has been traversed, it will be updated as the set of next nodes, which are generated by selecting the proper neighboring nodes of nodes in $\mathcal{V}_\mathrm{c}$ by the above criteria, as described in line 18–20. The algorithm terminates until all nodes in the network have been assigned a path to s.

Algorithm 3.4 is also illustrated in Fig. 3.10 for finding the best path from node S to node D. In the beginning, the set of current nodes $\mathcal{V}_\mathrm{c}$ contains only the source node S, as illustrated in step 1. The minimum weights of path ending to all nodes are set as ∞, as depicted in step 2. Then, the minimum weights of the two neighbors of S, namely node F and node A, are updated to 7 and 9, respectively, and S is added into the sets of parent nodes of A and F. Now the set of traversed nodes $\mathcal{V}_t$

Algorithm 3.4: Path Selection Algorithm

Input: Link graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$

Output: The optimal path ψ^*

- s: Source node
- d: Destination node
- $\mathcal{V}_\mathrm{t}$: Set of traversed nodes
- $\mathcal{V}_\mathrm{c}$: Set of current nodes
- $\mathcal{V}_\mathrm{n}$: Set of next nodes
- $\mathcal{P}(v)$: Set of parent nodes of node v
- $w(v)$: Minimum weight of path ending at node v
- $w_{v\text{-}u}$: Weight of edge (v, u)

Initialization: $\mathcal{V}_\mathrm{t} = \{\}$, $\mathcal{V}_\mathrm{n} = \{\}$, $\mathcal{V}_\mathrm{c} = \{s\}$, $w(v) = \infty$, $\forall v \in \mathcal{V} \backslash s$

begin

while $\mathcal{V}_\mathrm{t} \neq \mathcal{V}$ **do**
  foreach $v \in \mathcal{V}_\mathrm{c}$ **do**
    foreach $u \in \mathcal{V} \backslash \mathcal{V}_\boldsymbol{t}$ that is neighbor of v **do**
      if $w(u) = \infty$ or $w(u) = \min\{w_{v\text{-}u}, w(v)\}$ **then**
        $w(u) = \min\{w_{v\text{-}u},\, w(v)\}$;
        $\mathcal{P}(u) = \mathcal{P}(u) \cup \{v\}$;
        $\mathcal{V}_\mathrm{n} = \mathcal{V}_\mathrm{n} \cup \{u\}$;
      else if $w(u) < \min\{w_{v\text{-}u},\, w(v)\}$ **then**
        empty $\mathcal{P}(u)$;
        $\mathcal{P}(u) = \{v\}$;
        $\mathcal{V}_\mathrm{n} = \mathcal{V}_\mathrm{n} \cup \{u\}$;
      else
        continue;
      end
    end
    $\mathcal{V}_\mathrm{t} = \mathcal{V}_\mathrm{t} \cup \{v\}$;
  end
  empty $\mathcal{V}_\mathrm{c}$;
  $\mathcal{V}_\mathrm{c} = \mathcal{V}_\mathrm{n}$;
  empty $\mathcal{V}_\mathrm{n}$;
end
Return The optimal path $\psi^* : s \to \ldots \to \mathcal{P}(\mathcal{P}(d)) \to \mathcal{P}(d) \to d$;

end

contains the source S (shown in step 2 as red solid circle), and the set of current nodes contains A and F (shown in step 2 as red circle).

In the next step (step 3), similar procedures are carried out. Node G, E and B are marked with their own minimum weights 5, 3, and 6, respectively. Node A and F are added into the set of traversed nodes $\mathcal{V}_t$. Note that the weights of edge (A, E) as well as edge (F, E) are equal, thus the set of the parent nodes of E contains both A and F. The set of current nodes $\mathcal{V}_\mathrm{c}$ now consists of node G, E, and B.

Then, in step 4, the only non-traversed neighbor for node G is node H, thus H is marked with the minimum weight 4. After that, one would update the weight of the neighbors of E, namely node C and H. The minimum weight of node C is updated to 2, and its parent node is E. Moreover, as the weight of edge (E, H) is smaller than the minimum weight of H ($2<3$), thus H is no longer updated and its parent node is only the node G. Similarly, the minimum weight of C, as a neighbor of node B, remains unchanged, as walking through the edge (B, C) will decrease the minimum weight of C to 1. Naturally, B is not a parent node of C.

Finally, in step 5, the minimum weight of the destination node D is updated to 3 while ignoring the path from C. Then, all the nodes are traversed, and we can find the path from S to D by reversing the sequence of parent nodes, as shown in step 6.

3.4.3.2 Update Network

After the optimal path between BS and one UE is generated, the achievable data rate of each link in the network needs to be updated. This procedure utilizes the proposed JSRA algorithm, where every time the path for one UE has been determined, the achievable data rates of all links in the network are obtained by (3.6), according to the scheduling policy $\boldsymbol{\delta}$, the slot allocation policy $\boldsymbol{n}$, and the power allocation policy $\boldsymbol{p}$, which are acquired by running the JSRA algorithm for all the existing users in the network. In this way, the objective of DR is reached, where for each user the

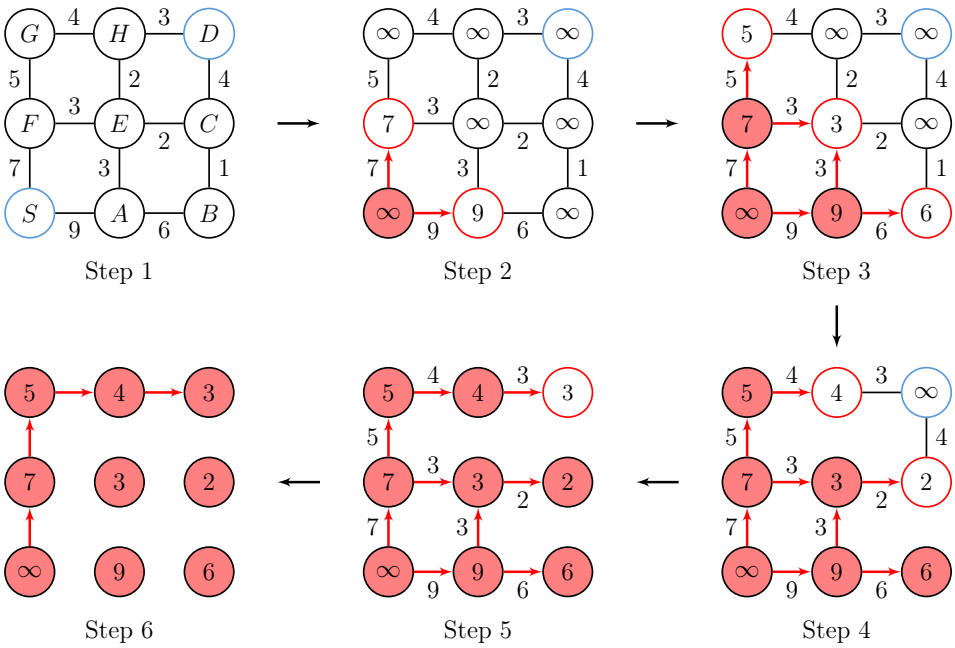


Figure 3.10: Illustration of an example of path selection algorithm.

optimal path is selected taking into account real-time network statistics (achievable data rate) and correspondingly the system performance is improved compared to the scenario of predefined routing.

3.5 Numerical Evaluation and Discussion

In this section, we evaluate the performance of the proposed JSRA algorithm, and the DR algorithm for path selection for the mm-wave HetNets with both downlink and uplink traffics. The system-level evaluation setup is described in Section 3.5.1. For the evaluation, we first compare the proposed algorithms with some benchmark schemes of multiplexing and interference mitigation and also with the approach that achieves the theoretical optimum, in terms of data rate and latency in Section 3.5.2 and Section 3.5.3, respectively. Then, the impacts of frame structure and duplex mode on data rate are addressed in Section 3.5.4 and Section 3.5.5, respectively. Finally, the capability of the proposed CG-MIS scheduling algorithm in improving

the system performance of P2P communications, for which we focus on the data delivery of a vehicular platoon, is demonstrated in Section 3.5.6.

3.5.1 Simulation Setup

The simulation results in this section adopt the system model, the JSRA algorithm, and the DR algorithm described in Section 3.2, Section 3.3 and Section 3.4, respectively. We consider a HetNet deployed under a single Manhattan Grid [145], where square blocks are surrounded by streets that are 200 meters long and 30 meters wide. As illustrated in Fig. 3.11, one BS and nine APs, which are marked as a black circle with black cross and black triangles, respectively, are located at the crossroads. 100 UEs are uniformly dropped in the streets marked as small blue crosses. In addition, green arrows illustrate mm-wave access links while red arrows depict mm-wave backhaul links, respectively.

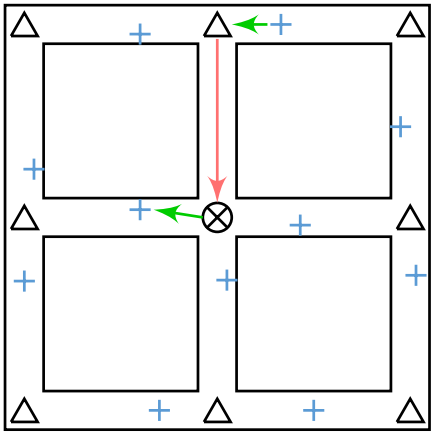


Figure 3.11: Simulation scenario with Manhattan Grid.

In particular, the system carrier frequency and bandwidth are $f = 28\,\text{GHz}$ and $B = 1\,\text{GHz}$, respectively. The adopted propagation and antenna model are explained in Section 3.2.2. It is worth noting that UEs are assumed to be almost stationary so the pathloss and shadowing values are fixed during the simulation. The type of user traffic is set as full buffer, and the default duplex mode is assumed to be half-duplex.

However, for different case studies addressed in Section 3.5.4 and Section 3.5.5, the default traffic type and duplex mode can be modified to investigate the efficiency of applying the proposed algorithms to ubiquitous system configurations. Simulation samples are averaged over 1000 independent snapshots. The default system parameter values are summarized in Table 3.1.

3.5.2 Case Studies: Data Rate

We begin with illustrating the data rate performance of the proposed algorithms. In Fig. 3.12, we plot the simulation results of the data rate for different schemes in the considered HetNet, versus the type of data rate. Here, the edge data rate is defined as the 5-th percentile point of the cumulative distribution function (CDF) of data rates. In particular, TDMA and eICIC schemes are selected as benchmark schemes for the performance comparison with the proposed algorithms. In order to address the dominant interference scenarios in HetNets, where small cell UEs experience strong interference from macrocell BS, eICIC techniques have been developed for LTE Release 10 and later release [146]. In time-domain eICIC method, the transmissions of victim users are scheduled in time-domain resources where the interference from other nodes is mitigated. One possible way to achieve this is to use the so-called almost blank subframe (ABS) at small cells, where in the ABSs only reference signals rather than control or data signals are transmitted.

As shown in Fig. 3.12, the proposed JSRA algorithm provides considerable improvement in both edge and average data rate compared to the benchmark schemes. Here, the JSRA algorithm is the algorithm described in Section 3.3 with a predefined path for each UE connecting to BS. Results are therefore demonstrated to be consistent with the theoretical inspection in Section 3.2 and the example illustrated in Fig. 3.5, where the exploitation of spatial multiplexing can enhance the data rate. Moreover, with the DR algorithm for path selection, the data rate is further boosted, which corresponds to our argument in Section 3.4 that the optimal path selected by the DR algorithm taking into account real-time network statistics improves the

data rate compared to the scheme of predefined routing.

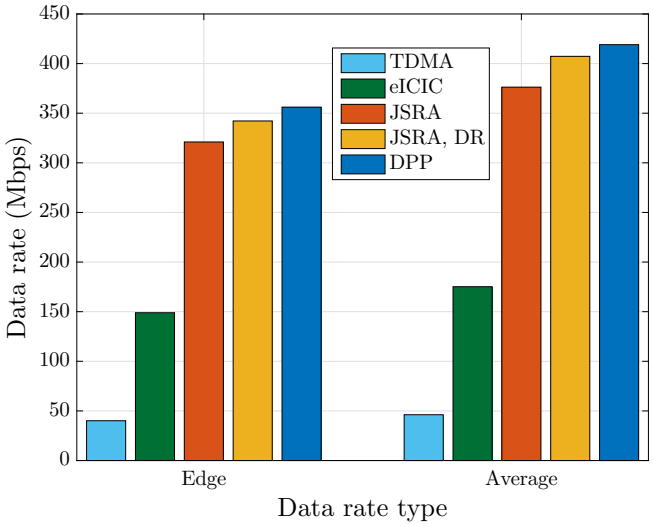


Figure 3.12: Performance of data rate for different schemes are compared in different types of data rates.

In particular, the proposed JSRA algorithm approximates closely to the optimal data rate, which is achieved by a general dynamic policy approach (known as DPP) with backpressure for the network utility maximization (NUM) problem [147]. Actually, the solution of the general NUM problem may be overly complex, as the long-term average data rate region is generally very difficult to be characterized. Hence, a method to solve the NUM problem with a dynamic policy, which is the DPP approach, is considered in [148], where at each time slot a max weighted sum (instantaneous) rate problem with respect to the single-hop instantaneous rates is solved, and the weights are recursively updated in terms of the flow queues at each node. While the DPP approach is (asymptotically) optimal for maximizing the data rate and outperforms the proposed JSRA algorithm (without and with the DR algorithm), due to the gain from slot-wise scheduling and resource allocation compared to our frame-wise approach, it is not optimal when we wish to guarantee other system performance, e.g. latency, which will be emphasized in Section 3.5.3.

In Fig. 3.13, we plot the simulation results of the data rate for different schemes in the considered HetNet, versus the number of users in the network. On the one

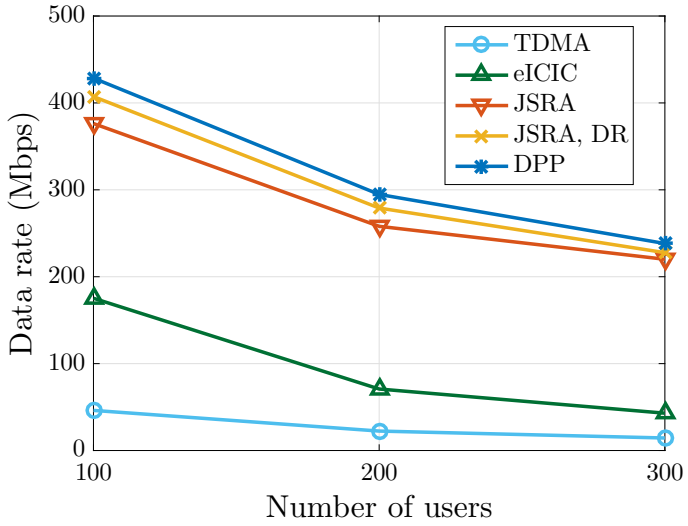


Figure 3.13: Performance of data rate for different schemes are compared in different number of users.

hand, as expected, increasing the number of users reduces the average data rate due to limited resources allocated to each user. However, the proposed JSRA algorithm (with and without the DR algorithm) still achieves high data rate in the case of 300 users and provides a significant improvement compared to the benchmark schemes. On the other hand, with the increased user density, the gap between the optimum (DPP) and the proposed algorithms shrinks. The reason behind this falls into the fact that when the number of users grows, the allocatable resource to each link in all schemes is limited and becomes the dominant factor in determining the data rate.

3.5.3 Case Studies: Latency

As described in Section 3.5.2, the DPP scheme is not optimal when we wish to minimize latency. The latency introduced by the scheme can be quantified as the queuing delay at each node, where by enlarging a control parameter that balances the greedy maximization of the utility function versus the "cost" of admitting more bits of each node, we can approach more closely the optimal NUM point, at the expenses of a longer queuing delay. In other words, DPP scheme achieves high data rate by sacrificing some flows, which might never be scheduled (always stay in the

queue), to maintain the stability of each transmission node in terms of the balance between influx and outflow bits.

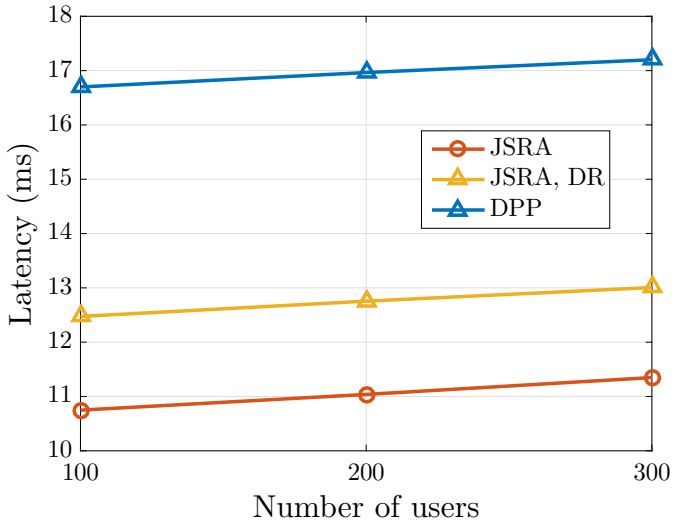


Figure 3.14: Performance of latency for different schemes are compared in different number of users.

In Fig. 3.14, we plot the simulation results of the latency for different schemes in the considered HetNet, versus the number of users in the network. Based on the results, we conclude that the JSRA algorithm achieves the lowest latency. By contrast, the JSRA algorithm with the DR algorithm, as well as DPP scheme, suffer from higher latency. The reason that DPP scheme experiences higher latency compared to JSRA (without and with DR algorithm) algorithm is stated above. The fact that JSRA algorithm with predefined routing outperforms that with the DR algorithm, in terms of latency, lies behind the criterion of path selection. Specifically, for the DR algorithm, a link that achieves higher data rate is more likely to be picked as an intermediate hop for a path, which may lead to the situation that the optimal path, in terms of maximizing achievable data rate, consists of a large number of hops and eventually causes higher latency compared to fixed routing with a limited number of hops.

3.5.4 Case Studies: Frame Structure

In the default configuration of frame structure, as described in Section 3.2.1 and depicted in Fig. 3.3, the entire system time is partitioned into frames and within the period of each frame, it is assumed that the network topology and channel conditions remain unchanged. However, the actual user demand, which is represented by the required transmission slot, varies across different frames. Therefore, the results of the slot allocation policy $\boldsymbol{n}$ for different frames may also be diverse. This brings an additional advantage for frame structure design, where the "switch point" for different nodes (BS/AP/UE), defined as the percentage of the number of downlink slots in each frame, can be flexibly adjusted. Specifically, as the number of slots allocated to the same link in different frames is changing, due to different scheduling and resource allocation results, the percentage of downlink slots in each frame can be also different. Hence, each node should be able to flexibly adjust the switch point according to the actually allocated slots for downlink and uplink transmissions.

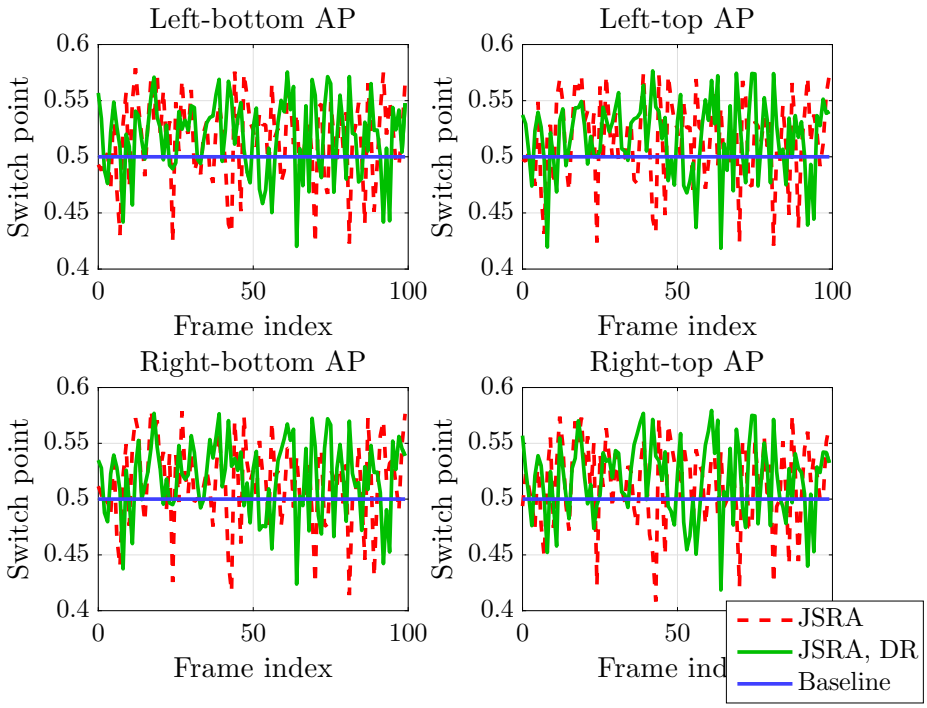


Figure 3.15: Switch point for different APs are illustrated across different frames.

In Fig. 3.15, we plot the simulation results of switch points for four APs in the considered HetNet, versus the index of successive frames. The total number of frames is selected as 100. The trend of switch point curves for the other BS/AP is expected to be similar to the depicted ones. As anticipated, the switch points of the selected APs fluctuate in the vicinity of the baseline (50%, which means that the number of downlink slots and the number of uplink slots are the same) for both algorithms, which shows the efficiency of the proposed algorithms in allocating slots to fulfill actual user demands.

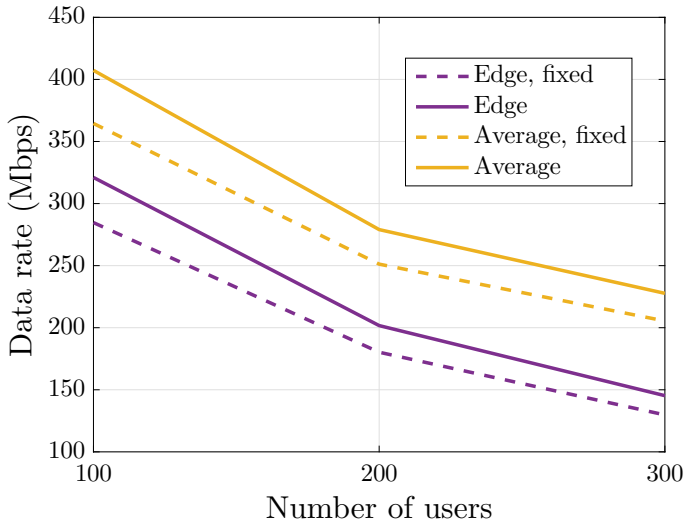


Figure 3.16: Performance of data rate for different frame structures are compared in different number of users.

In Fig. 3.16, we plot the simulation results of the data rate for JSRA algorithm with the normal frame structure and with frame structure with fixed switch point (50%) in the considered HetNet, versus the number of users in the network. Here, the scheduling policy $\boldsymbol{\delta}$ acquired by CG-MIS algorithm for the above two cases are the same, but the slot allocation for the fixed switch point case is restricted to allocating the total number of slots in each frame equally to all downlink and uplink links (namely 50% – 50%). In the figure, we notice that the proposed algorithm, owing to the capability of the flexible adjustment of downlink/uplink slots, yields both higher edge and average data rates compared to that of fixed switch point.

3.5.5 Case Studies: Duplex Mode

In addition to the default half-duplex mode, in Fig. 3.17 and Fig. 3.18 we plot the simulation results of the data rate for different duplex modes in the considered HetNet, versus the type of data rate and the number of users, respectively. Here, the perfect full-duplex refers to the case that a reception link is isolated from the simultaneously scheduled transmission link of the same node without interference, while the interfered (abbreviated as interf. in the figures) case incorporates a $-110\,\mathrm{dB}$ interference at the reception link caused by the simultaneously scheduled transmission link at the same node. The full-duplex modes are further classified into only enabling full-duplex mode at AP (green and red bars in Fig. 3.17 and Fig. 3.18) and at both AP and BS (yellow and blue bars). Results show that by relaxing the scheduling criterion for simultaneous transmission (from half-duplex to full-duplex, from full-duplex at only AP to at both AP and BS, etc.), the data rate is gradually improved.

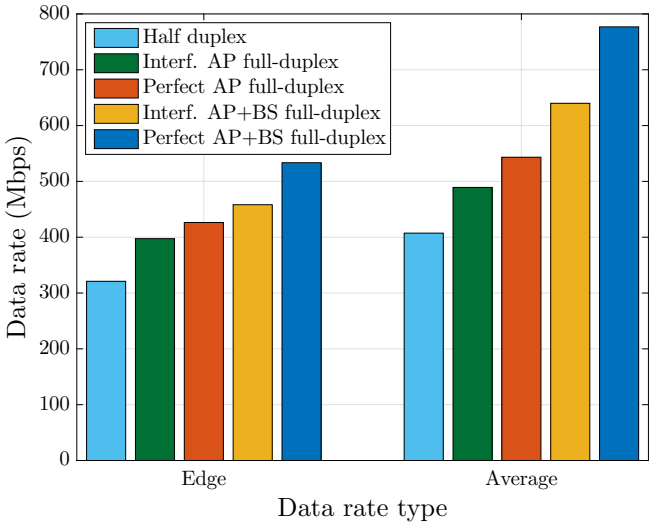


Figure 3.17: Performance of data rate for different duplex modes are compared in different type of data rates.

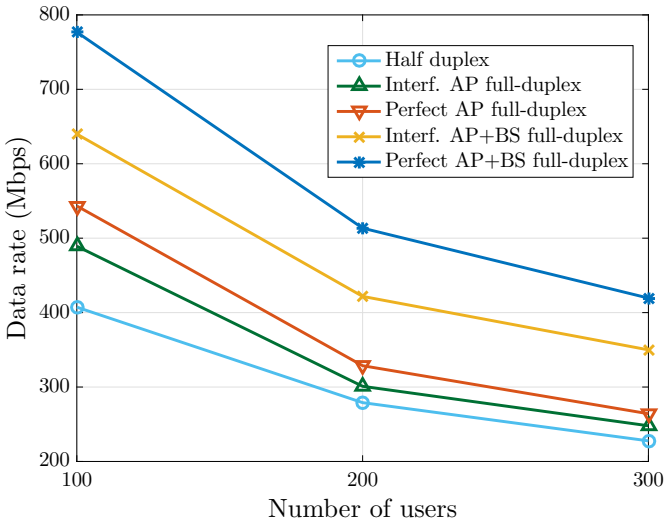


Figure 3.18: Performance of data rate for different duplex modes are compared in different number of users.

3.5.6 Case Studies: P2P Communications

In this section, we focus on the capability of the proposed JSRA algorithm in P2P communication scenario. Specifically, we consider the data delivery in a platoon of vehicles, which is currently intensively studied by standardization association and research activities [149, 150]. An example of a platoon with three vehicles is illustrated in Fig. 3.19, where pink ovals indicate BS beams and blue ovals depict vehicle beams.

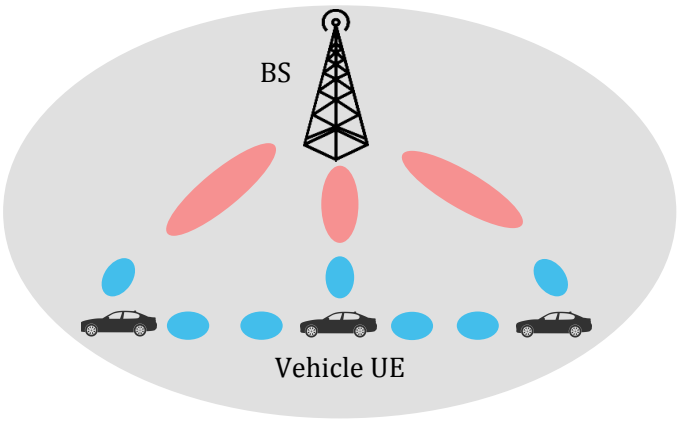


Figure 3.19: An example of a HetNet with a platoon of three vehicle UEs.

A typical 5G NR vehicle user may have multiple connections in different directions (forward/backward), where the antennas are assumed to be installed at the bumpers and ceiling of a vehicle. In this way, the nature of self-interference cancellation, where the interference from the forward transmission (from the antenna at the front bumper) at the backward reception (at the antenna at the rear bumper) is mitigated by the vehicle itself, enables the exploitation of full-duplex. Moreover, beamforming with directive transmission further reduces the self-interference level, which provides near-perfect interference cancellation. Nevertheless, as bumper antennas installed in vehicles are geographically separated and non-collocated, referring to the method of enabling simultaneous transmission and reception as full-duplex is no longer appropriate. Hence, the method in vehicular society is referred to as *bidirectional transmission* (BT), which is designed to be differentiated from the conventional full-duplex.

It is obvious that for vehicles in a platoon, P2P communications are more likely to be established, as vehicles move in a line and data is forwarded by the vehicles one after another. Therefore, the power allocation policy $\boldsymbol{p}$ is no longer one of the considered issues in verifying the capability of the proposed JSRA algorithm in P2P communication scenario. Further, the routing of data within a platoon is also fixed, where there is no degree of freedom in path selection as there is only one path for all data from one vehicle to another. Next, when the number of vehicles in a platoon increases, data delivery latency becomes a crucial target of performance optimization, for which we may concentrate on how to schedule the data streaming in a platoon.

Based on the above observations, in this section we study how the system performance of data delivery in a platoon benefits from the proposed CG-MIS scheduling algorithm. We consider an urban scenario, which is the same as the Manhattan Grid considered for pedestrian UEs in the previous sections, with a platoon of ten vehicles. Each vehicle has data to transmit to other vehicles in the platoon or to the BS, or vice versa. The data demand pops at random time of each frame (i.e., with

file transfer protocol (FTP) traffic type). Part of the system parameters for simulation setup in this subsection are same as Section 3.5.1, including carrier frequency, system bandwidth, and propagation and antenna model. The other vehicle-specific parameters are summarized in Table 3.2 referring to [149, 150]. Simulation samples are averaged over 1000 independent snapshots.

Table 3.2: Simulation Parameters for Data Delivery in Platoon of Vehicles

Description	Value
Vehicle speed	50 km/h
Distance between vehicles	Vehicle speed (in m/s) $\times 2$
Packet size	1520 bit, 2400 bit
Traffic type	FTP

In Fig. 3.20 and Fig. 3.21, we plot the simulation results of the latency and the throughput for different scheduling schemes in the platoon, versus the number of BT-enabled vehicles, respectively. On the one hand, the results suggest that the more BT-enabled vehicles, the lower the latency and the higher the throughput are. Here, for BT-enabled vehicles, transmission and reception links are simultaneously scheduled, which allows data to "flow" through these vehicles and eventually decreases the latency (half of the transmission time is saved in case the channel capacity of the transmission and reception links are the same). It is worth noting that even the self-interference of the BT-enabled vehicle is low, we still incorporate the interference when calculating the channel capacity of links among vehicles, which means that the actual latency for data passes through BT-enabled vehicle is the maximum of latency on the transmission and reception links.

On the other hand, the proposed CG-MIS scheduling algorithm outperforms other classical scheduling algorithms, namely RR and proportional fair algorithm, in both latency and throughput. The reason behind lies in the fact that CG-MIS scheduling algorithm groups links that are suitable to transmit simultaneously according to achievable channel capacity, which is better than the RR and the proportional fair algorithm that always transmits the first data flow in the queue and delays links with low achievable channel capacity due to interference, respectively.

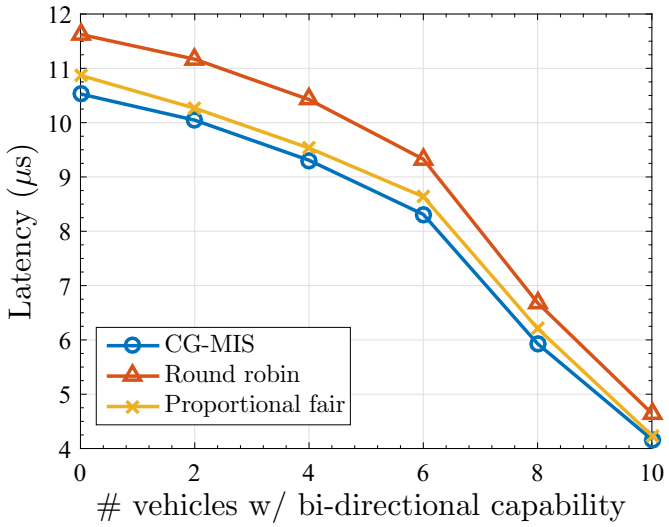


Figure 3.20: Performance of latency for different scheduling schemes are compared in different number of BT-enabled vehicles.

3.6 Summary

In this chapter, we have addressed the problem of maximizing the achievable data rate of mm-wave multihop HetNets considering both downlink and uplink transmissions on backhaul and access links. To solve the maximization problem in an efficient way, we proposed a joint scheduling and resource allocation algorithm, where the cross-layer optimization problem was decomposed into subproblems including link scheduling, time slot allocation, and transmission power allocation. The subproblems were then solved by the proposed maximum independent set based scheduling algorithm, the proportional fair slot allocation algorithm, and the water-filling power allocation algorithm, respectively. Based on this, a dynamic routing algorithm, which incorporates a path selection algorithm to greedily search the optimal path connecting BS and UE by considering real-time status of network load and traffic, was investigated to further improve the data rate achieved by the joint scheduling and resource allocation algorithm with static path selections. By evaluating the proposed algorithms via extensive simulations, we concluded that the proposed joint scheduling and resource allocation algorithm, with and without the dynamic routing algorithm, outperforms the benchmark schemes for multiplexing and interference

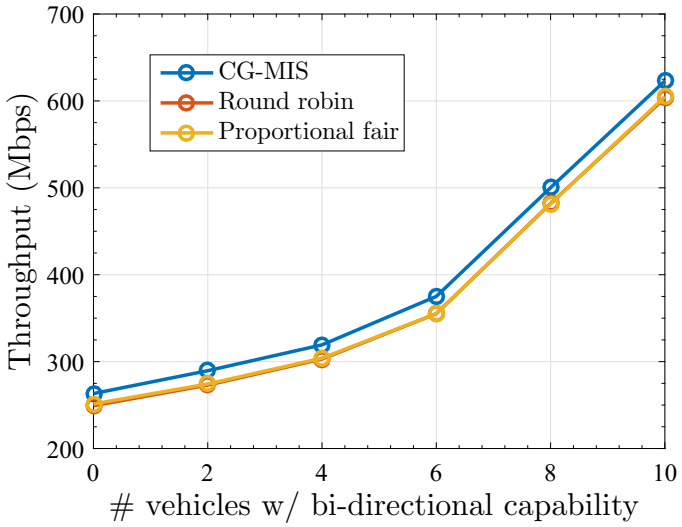


Figure 3.21: Performance of throughput for different scheduling schemes are compared in different number of BT-enabled vehicles.

mitigation, in terms of achievable data rate, and closely approach the theoretical optimum, yet with lower latency. Besides, the proposed algorithms enable the flexible adjustment of downlink and uplink slot allocation according to spontaneous user demand, which boosts the achievable data rate compared to fixed slot allocation arrangement that disregards actual network condition, and support both half- and full-duplex modes with considerable performance enhancements. In particular, the proposed algorithms can deliver significant flexibility in terms of fulfilling various performance requirements for both point-to-point and point-to-multipoint communications.

3.7 Appendix

3.7.1 Proof of Unavailability of Dynamic Programming and Branch-and-Bound Algorithms in Solving Problem 3.1

Depending on the characteristics of variables, the Knapsack problem can be classified as [135]:

- 0–1: Variables are selected from binary set.
- Bounded: Variables are selected from positive integer set with finite values.
- Unbounded: Variables are selected from positive integer set with infinite values.

Furthermore, the Knapsack problem can be also classified as

- Uncorrelated: Variables are not correlated.
- Correlated: At least one variable is linearly related to another variable.
- Subset sum: Values of variables are same.

To tackle with different classifications of the Knapsack problems, exact algorithms have been developed using the methods of dynamic programming and branch-and-bound. The method of dynamic programming is based on the idea that a global optimal solution can be constructed from the optimal solutions to subproblems. However, to achieve the integrity of optimal solutions, a bounding test is in general incorporated into the dynamic programming to discard unpromising subproblems, which leads to performance degrade. By contrast, the method of branch-and-bound subdivides (branches) the feasible solution set into successive subsets, instead of subproblems, and places bounds on the objective function over each subset.

For both approaches, the fundamental procedure is the enumeration of feasible solution sets, which is infeasible or extremely computationally complex for the unbounded Knapsack problems with correlated variables. Unfortunately, problem 3.1 is formulated as a mixed Knapsack-like problem with mixed bounded and unbounded variables, and the variables are correlated. Specifically, we have

- $n^{(k)}$ belongs to 0–1 variable.
- $\delta_i^{(k)}$ belongs to bounded variable, and is correlated with $n^{(k)}$.
- p_i belongs to unbounded variable, as it is a continuous variable and can be selected from the set of non-negative real numbers $[0, P_{\max}]$. It is also correlated with $n^{(k)}$.

In this case, it is not proper to apply either dynamic programming or branch-and-

bound to solve the problem 3.1. Reversely, the problem has to be decomposed into subproblems in which the scheduling policy $\boldsymbol{\delta}$, the slot allocation policy $\boldsymbol{n}$, and the power allocation policy $\boldsymbol{p}$ are separately optimized.

3.7.2 Proof of Lemma 3.3.1

Denote $\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k$ as the complement of set $\mathcal{V}^k$ for set $\mathcal{V}_{\mathrm{C}}$, namely $\mathcal{V}_{\mathrm{C}} = \mathcal{V}^k \cup \mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k$. Then, we upper bound the number of nodes in $\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k$, i.e. $|\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k|$, as follows.: A node u is in $\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k$ because it is removed as a neighbor of some node $v \in V^k$ when greedily added v to $\mathcal{V}^k$. Associate u to v. A node $v \in \mathcal{V}^k$ can be associated at most Δ times as it has at most Δ neighbors. Hence, we have $|\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k| \leq \Delta|\mathcal{V}^k|$. Then, as every node is either in $\mathcal{V}^k$ or $\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k$, we have $|\mathcal{V}_{\mathrm{C}}\backslash\mathcal{V}^k| + |\mathcal{V}^k| = |\mathcal{V}_{\mathrm{C}}|$ and therefore $(\Delta+1)|\mathcal{V}^k| \geq |\mathcal{V}_{\mathrm{C}}|$, which implies that $|\mathcal{V}^k| \geq \frac{|\mathcal{V}_{\mathrm{C}}|}{\Delta+1}$.

3.7.3 Proof of Theorem 3.3.1

The size of any MIS for graph $\mathcal{G}_{\mathrm{C}}(\mathcal{V}_{\mathrm{C}}, \mathcal{E}_{\mathrm{C}})$ is at least $\frac{|\mathcal{V}_{\mathrm{C}}|}{\Delta+1}$, as for each node added to the set, at most Δ others are removed as demonstrated in Corollary 3.3.1. This can be further improved by rudimentary arguments [151]: First, observe that greedy algorithm handles isolated and degree-zero node optimally, and that the independence number of a graph with minimum degree one is at most $\frac{\Delta|\mathcal{V}_{\mathrm{C}}|}{\Delta+2}$. Second, if a connected component contains a node of degree less than Δ, the greedy algorithm will never remove more than Δ nodes in any step. This then yields a ratio of $\frac{\Delta^2}{\Delta+2} \leq \Delta - 1$. If, however, each node is of degree Δ, the independence number can be seen to be at most $\frac{|\mathcal{V}_{\mathrm{C}}|}{2}$. This implies a ratio of $\frac{\Delta+1}{2}$, which is at most $\Delta - 1$ for all $\Delta \geq 3$. It is proved in [152] that the performance bound of greedy algorithm is proved to be greater than or equal to $\frac{|\mathcal{V}_{\mathrm{C}}|}{\bar{d}+1}$, where $\bar{d}$ is the average degree of nodes in $\mathcal{V}_{\mathrm{C}}$. The bound has been further tightened by [140] to $\frac{2}{\bar{d}+3}|\mathcal{V}_{\mathrm{C}}|$. Hence, the performance ratio of the greedy algorithm achieves $\frac{\Delta+2}{3}$, as described in Theorem 3.3.1.

3.7.4 Proof of Theorem 3.3.3

As the transmission power of a link is non-negative, the relaxed optimization problem 3.3 provided by (3.16) needs to be modified as

$$\begin{aligned} \max_{\boldsymbol{p}} \quad & \sum_{i \in \boldsymbol{M}_{s,k}} \log(1+\gamma_i p_i), \\ \text{s.t.} \quad & \sum_{i \in \mathcal{M}_{s,k}} p_i \leq P_{\max},\ p_i \geq 0,\ \forall i \in \mathcal{M}_{s,k}, \end{aligned} \tag{3.20}$$

where $\gamma_i = \frac{g_i l_i^{-1}}{\eta}$. For simplicity, we assume log = ln.

Observe that all inequality constraints functions in (3.20) are affine, and at least a set of $\{p_i\}$ exists such that

$$p_i = \frac{P_{\max}}{(|\mathcal{M}_{s,k}|+1)} > 0,\ \forall i \in \mathcal{M}_{s,k}, \tag{3.21}$$

and

$$\sum_{i \in \mathcal{M}_{s,k}} p_i \leq P_{\max}, \tag{3.22}$$

which implies the Karush-Kuhn-Tucker (KKT) conditions are necessary. Moreover, the objective function in (3.20) is the sum of convex functions and consequently convex as well. Therefore, the KKT conditions are also sufficient, in which we can use standard KKT form to solve the problem. They are concluded as follows:

- Primal Feasibility (PF):

$$\sum_{i \in \mathcal{M}_{s,k}} p_i \leq P_{\max}, \tag{3.23}$$

$$p_i \geq 0,\ \forall i \in \mathcal{M}_{s,k}. \tag{3.24}$$

- Dual Feasibility (DF):

$$-\frac{\gamma_i}{1+\gamma_i p_i} + \phi - \omega_i = 0,\ \forall i \in \mathcal{M}_{s,k}, \tag{3.25}$$

$$\phi \geq 0,\ \forall i \in \mathcal{M}_{s,k} \tag{3.26}$$

$$\omega_i \geq 0,\ \forall i \in \mathcal{M}_{s,k}. \tag{3.27}$$

- Complementary Slackness (CS):

$$\phi\Big(\sum_{i\in\mathcal{M}_{s,k}} p_i - P_{\max}\Big) = 0, \tag{3.28}$$

$$\omega_i p_i = 0,\ \forall i \in \mathcal{M}_{s,k}. \tag{3.29}$$

As the channel capacity r_i is increasing with p_i, the optimal power allocation for link i, denoted as p_i^*, should satisfy

$$\sum_{i\in\mathcal{M}_{s,k}} p_i^* = P_{\max}. \tag{3.30}$$

As a result, we can always increase some p_i without violating the first PF condition provided by (3.23) to increase the sum rate, in case

$$\sum_{i\in\mathcal{M}_{s,k}} p_i < P_{\max}, \tag{3.31}$$

which means the first CS condition provided by (3.28) is always satisfied and there is no further restriction on the Lagrangian multiplier ϕ besides the second DF condition provided by (3.26).

Actually, (3.26) also indicates that $\phi > 0$. Otherwise, assume $\phi = 0$, then the other two DF conditions, provided by (3.25) and (3.27), respectively, require

$$0 > -\frac{\gamma_i}{1+\gamma_i p_i} - \omega_i = 0,\ \forall i \in \mathcal{M}_{s,k}, \tag{3.32}$$

$$\omega_i \geq 0,\ \forall i \in \mathcal{M}_{s,k}, \tag{3.33}$$

which results in contradiction.

Now, let link j gets positive transmission power, i.e., $p_j > 0$. Then, according to (3.25) and (3.29), we have

$$\omega_j = 0, \tag{3.34}$$

and

$$\frac{\gamma_j}{1+\gamma_j p_j} = \phi, \tag{3.35}$$

which means

$$p_j = \frac{1}{\phi} - \frac{1}{\gamma_j}. \tag{3.36}$$

Note that this can hold only if

$$\frac{1}{\phi} - \frac{1}{\gamma_j} > 0, \tag{3.37}$$

or equivalently,

$$\gamma_j > \phi. \tag{3.38}$$

Otherwise, to satisfy the first DF condition provided by (3.25), we must have

$$p_j = 0, \tag{3.39}$$

and

$$\omega_j = \phi - \gamma_j. \tag{3.40}$$

In conclusion, the optimal solution of problem 3.3 provided by (3.16) is given by

$$p_i^* = \max\left\{\frac{1}{\phi^*} - \frac{1}{\gamma_i}, 0\right\}, \tag{3.41}$$

where the optimal Lagrangian multiplier ϕ^* can be derived from

$$\sum_{i \in \mathcal{M}_{s,k}} \max\left\{\frac{1}{\phi^*} - \frac{1}{\gamma_i}, 0\right\} = P_{\max}. \tag{3.42}$$

Chapter 4

Multihop Routing for Data Delivery in 5G Millimeter Wave V2X Networks

4.1 Introduction

Intelligent transportation system (ITS) revolutionizes the provisioning of diverse applications associated with driving safety, traffic management, and infotainment [153, 154]. These applications are beyond the far-fetched goals of academic, industry, and standardization groups, which aim to streamline the innovative operation of vehicle, facilitate safe and eco-friendly driving, and offer ubiquitous infotainment services for commuting passengers [155, 156, 157]. Vehicular ad-hoc networks (VANETs), where

This chapter has been published in [30]. I am the primary author of these works. Co-authors Zhongfeng Li, Dr. Richard A. Stirling-Gallacher and Dr. Wen Xu have provided insightful feedbacks to improve the presentation of the results, and Dr. Jian Luo and Prof. Giuseppe Caire are my supervisors. Except for the contributions in the previous publications, this chapter also extend the work in [30] with the development of distributed data delivery optimization, the design of distributed routing algorithm, and the support of wireless backhaul in data delivery.

vehicles behave as mobile sensors and/or data relays, paves a migration path to achieve these goals [158]. The ability of vehicles to assist data dissemination, maintain traffic connectivity, and promote the security of passengers, qualifies VANETs to be a promising candidate for inaugurating ITS [159]. Recent research efforts have placed a strong emphasis on novel VANETs architecture and implementation design, including specific areas such as broadcasting, routing, QoS, and security. If we take a close look at the advances of the aforementioned technologies, we find that the foundation for the realization of VANETs is provided by the features and characteristics offered by cutting-edge communication technologies along with the emergence of highly intelligent vehicles [160].

Typically, dedicated short-range communication (DSRC), which is supported by onboard units and allows vehicles to communicate with DSRC-equipped neighbors, has been widely applied for vehicle-to-vehicle (V2V) communications by the IEEE 802.11p standard [161]. DSRC exploits the flooding of information to achieve low latency, high reliability, and secure interoperability with very little interference [162]. In addition, an air interface named as PC5/sidelink has been defined in 3GPP cellular-V2X Release 14 [149], which supports both eNB-scheduled (referred to as mode 3) and autonomous (referred to as mode 4) resource allocation for the PC5/sidelink. Nevertheless, potential connectivity disruption as a result of high vehicle speed, time-varying vehicle density, and limited inter-vehicle contact time, confines V2V communications to applications and services with short communication range [163]. Fortunately, vehicle-to-infrastructure (V2I) communications represent a viable solution to bridge long-range vehicular connectivity by introducing stationary network entities, e. g., RSU, to exchange data with vehicles [149]. RSU collects road traffic information and interacts with vehicles and core network for various management and planning applications. Specifically, as a complement to V2V communications, RSU provides reliable interconnection to vehicles, receives their real-time access request, and allocates resource for the V2I communication links, as well as exchanges auxiliary information such as road traffic and routing

decision with vehicles [164].

By leveraging the complementary features of infrastructure-less V2V communications and infrastructure-assisted V2I communications, the use of hybrid vehicular communication, namely V2X, has been envisioned as a full-fledged solution to capture the connectivity, efficiency, and scalability of vehicular networks. The V2X communications, which are supposed to be capable of further incorporating vehicle-to-pedestrian (V2P) communications, vehicle-to-grid (V2G) communications, etc., support numerous attractive applications related to driving safety, advanced driving, and infotainment sharing by taking advantage of the unique features of all the above technologies. The potential of V2X communications not only helps to provide seamless connectivity for data delivery but also enables the carry-and-forward strategy to improve the delivery performance. Fig. 4.1 shows an example of V2X networks.

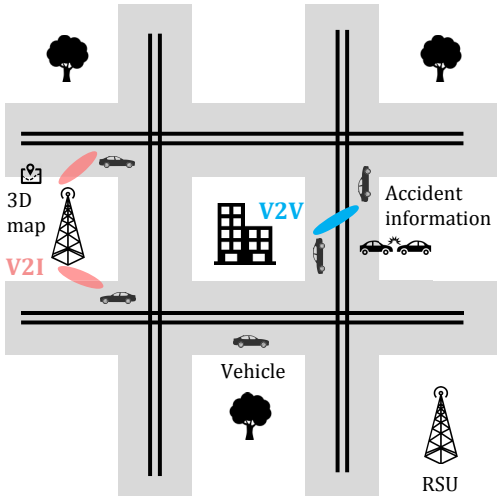


Figure 4.1: Illustration of an exampled V2X network incorporating both V2V and V2I communications.

5G mobile communications target to support an efficient data delivery with bulky data rate, high reliability, and low latency for diverse requirements of different V2X services and applications [11, 165]. Currently, 3GPP is pioneering advanced technologies and evolving the cellular-V2X communications towards 5G NR while maintaining backward capabilities [166]. Specifically, real-time services for

driving safety, such as collision avoidance and lane-change/merge notification, bring the requirement of low latency. Applications that focus on delivering large-size multimedia messages, e.g., danger notification for live traffic status in the vicinity of an accident scene, necessitates the provision of stable data delivery with ultra-low E2E latency [167]. Furthermore, non-emergency services including, e. g., video streaming, demand high data rate to fulfill capacity burst. Offering information like high-definition 3D maps of urban networks and in-car entertainment services greatly complicates the development of impeccable V2X data delivery systems in terms of satisfying the immense data rate [168]. Therefore, an efficient data delivery scheme to meet the diverse communication requirements in V2X networks is expected to be designed specifically taking into account the coexistence of low latency and high data rate.

Nevertheless, the data delivery in V2X networks is particularly challenging, as its performance depends highly on the efficiency of data routing [169]. On the one hand, pure inter-vehicle data delivery may introduce non-negligible latency because of frequent disconnection between vehicle pairs. On the other hand, the limited coverage of each RSU is a major concern of pure inter-RSU data delivery. Therefore, how to optimally design a multihop routing mechanism that minimizes latency as well as maximizes data becomes an interesting and challenging topic.

4.1.1 Related Works

Multihop routing for data delivery, particularly for vehicular networks, has been intensively investigated by recent research efforts [159, 167, 168, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184]. The knowledge of vehicular trajectories plays a key role in optimal data delivery, where the performance of routing algorithm relies heavily on the accuracy of vehicle mobility prediction [170, 171, 172, 173, 174]. Authors in [170] mined historical bus trajectory and exploited acquired patterns to build a probabilistic spatial-temporal graph model to provide available paths with the best QoS levels. To address the performance

limit of simple spacial distribution, authors in [171] and [172] developed accurate trajectory predictions by using multiple order Markov chains, and proposed routing algorithms taking full advantage of the predicted probabilistic vehicular trajectories.

It has been envisioned that the assistance of infrastructure is able to facilitate data delivery in terms of latency improvement [167, 175, 176, 177, 178]. Leveraging the proactive connectivity information from local infrastructures, authors in [175] developed a dynamic routing algorithm to reduce delivery latency. In addition, an infrastructure-assisted message dissemination framework has been proposed in [176] to guarantee timely and reliable delivery with a beaconing block schedule algorithm. Authors of [177] proposed both global and distributed routing algorithms for a co-existing packet forwarding and buffer allocation optimization problem. However, these works either assumed that the latency of V2V and V2I transmission can be ignored, or limited the latency to be considered in a single hop between vehicles, which is not proper for data delivery in large-scale vehicular networks where multihop transmissions are expected. More importantly, models applied to the above works depend on the prerequisite that the size of packets transmitted on V2V and V2I link is small enough, such that the data rate for delivering these packets is omitted. This assumption does not hold for the data delivery of services that rely on abundant data rate to guarantee the enormous requirement of data volume.

When it comes to achieving reasonable data rate for the data delivery in V2X networks, a rich body of earlier studies have tackled the problem of how "mobility improves data rate" in vehicular networks [159, 168, 179, 180]. A detailed analysis of achievable data rate considering various performance-impacting parameters under a cooperative communication strategy has been studied in [168]. A theoretical bound of achievable data rate was proposed in [159], where a stochastic model is formulated for deriving the existence probability of connectivity path between source and destination. Authors in [180] developed a mathematical framework to analyze asymptotic data rate scaling and exploited vehicle mobility to approach the optimal network data rate with data balancing. With the exception of some studies

that contributed to a limited investigation of latency performance [168], none of the above-mentioned works to date has considered the trade-off between low latency and high data rate of data delivery in V2X networks, which is supposed to be a key enabler in fully exploiting the mobility of vehicles complemented by the stability of infrastructures to improve data rate performance while keeping latency tolerable.

Store-carry-and-forward strategy, where data is stored at intermediate nodes along delivery and forwarded at a later time to another intermediate station or the final destination, has been recently recognized as a promising evolution path to improve data delivery efficiency, either in latency or data rate [181, 182, 183, 184]. By introducing the dropbox functionality of RSU, authors in [181] and [182] proposed scalable frameworks to minimize delivery latency and maximize network utility for large-scale vehicular networks. Similarly, a finite state decision conversion and a searching algorithm have been developed in [183] for data rate maximization. Data delivery for both message forward and backward phases has been addressed in [184], where a direction selection mechanism is modeled for latency reduction. Nevertheless, except the dropbox functionality that allows data to be stored at some cost, none of the aforementioned studies has considered a realistic model of RSU for providing V2I communications in terms of link establishment and resource allocation. Moreover, the ability of V2X networks to support ubiquitous connection to all devices in the network (especially for the coexistence of cellular and vehicular UEs addressed in 5G NR [166]), which is one of the key considerations in optimizing overall system performance in this chapter, has not been addressed in any of these works.

4.1.2 Contributions

In this chapter, we consider the fact that the infrequent encountering opportunities of sparsely distributed vehicles pose a bottleneck for the data delivery in V2X networks, and it is highly beneficial to exploit infrastructure such as RSU to collect vehicular mobility information and assist vehicle with data delivery. Different from

the previous studies [181, 182], where RSU is simply assumed to be a dropbox for data collection and temporary storage, in this chapter we treat RSU as a network entity that supports both V2I and cellular communications (e. g., integrated in gNB or transmission reception point (TRP) for 5G communications in 3GPP terminology [149]), and serves both vehicular and cellular users.

Besides, we develop a mathematical framework to analyze the data delivery performance in 5G mm-wave V2X networks, where data generated from a source node is assumed to be carried by vehicles moving along a multihop route directing to a destination node. Then, the expected latency and data rate of each hop of the route, distinguished by various vehicle mobilities and traffic situations, are derived taking into account the network coverage, vehicle arrival rate, and achievable data rate of different communication links. Based on this, the expected E2E latency and data rate are determined by adding the latency of each hop and by taking the minimum data rate of all hops, respectively, and are further transformed in closed-form expressions. Afterwards, both global and distributed optimization problems are formulated to minimize the delivery latency while maximizing the data rate, and the proposed problems are solved by convex optimization theory. Leveraging the solutions, we propose both global and distributed multihop routing algorithms to determine the optimal routes in terms of the maximum weighted sum of latency and data rate for the global and distributed optimization problem, respectively. The proposed algorithms are then validated against detailed system-level simulations, and the evaluation results demonstrate that our solutions achieve higher data rate as well as lower latency compared to other classical vehicular routing algorithms. Besides, the impact of broadcast scheme, vehicle arrival rate, and backhaul availability on the delivery performance are also analyzed. The main contributions of this chapter are summarized as follows:

- Development of an accurate analytical framework for the data delivery in V2X networks: The framework first considers *hop-wise* latency and data rate, which are divided into different scenarios distinguished by divergent vehicle mobility

patterns and data forwarding behaviors of each hop. Based on these, the expected *E2E* latency and data rate are also derived by adding the hop-wise latency and by minimizing the hop-wise data rate, respectively.

- Formulation of optimization problems that maximize the weighted sum of latency and data rate: Unlike some of the previous works that revolved around the feasibility study of data delivery without explicitly addressing delivery performance optimization ([171, 175, 176]), we obtain mathematical expressions of both hop-wise and E2E latency/data rate, based on rigorous derivations, and formulate optimization problems that maximize the weighted sum of latency and data rate considering both global and distributed scenarios, where the weighted sum is optimized in E2E manner and hop-wise manner, respectively.
- Leveraging the optimization problems and the corresponding solutions to propose multihop routing algorithms: The derived expressions of the latency and data rate are then transformed into closed-form for verifying the convexity of the proposed optimization problems that are then solved by convex optimization theory. Based on these, multihop routing algorithms to select the optimal route are addressed for both global and distributed data delivery in order to maximize the weighted sum of latency and data rate.
- A detailed system-level performance evaluation for data delivery performance: Extensive simulations have been conducted under numerous system parameters to demonstrate the efficiency of the proposed solutions and algorithms in achieving lower latency and higher data rate compared to the classical vehicular routing algorithms. The impact of broadcast scheme, vehicle arrival rate, and backhaul availability on the data delivery performance are also analyzed.

The remainder of this chapter is organized as follows: Section 4.2 presents the system model and Section 4.3 addresses the optimization problem formulation. In Section 4.4, we solve the formulated problems and propose the corresponding routing algorithms. The proposed algorithms are then evaluated by extensive simulations in Section 4.5, followed by a summary concluding this chapter in Section 4.6.

4.2 System Model

In this section, we introduce the network, traffic, data forwarding, and radio models considered for finding the best route and optimizing the data delivery. These models are the fundamentals of the analytical framework for the data delivery in V2X networks mentioned in Section 4.1.2, and are exploited as the preliminaries of the formulation of data delivery optimization problems in terms of latency and data rate and of the solution of the formulated problems and the corresponding multihop routing algorithms, which will be elaborated in Section 4.3 and in Section 4.4, respectively. Specifically, these models are applied to the derivation of delivery latency and data rate that will be detailed in Section 4.3.1 and Section 4.3.2, respectively. The important notations and system parameters defined in this section are summarized in Table 4.1 and will be used in the rest of this chapter.

Table 4.1: System Model Parameters

Notation	Description	Value
S	Source node	See Section 4.2.1
D	Destination node	See Section 4.2.1
γ_i	i-th route between S and D	See Section 4.2.1
n	Number of routes between S and D	12
T	Duration of vehicle staying in each hop	20 s
t	Global candidate discovery duration	$[0, T]$
ϵ	Decode error probability	10^{-3}

4.2.1 Network Model

We envision a scalable V2X network for data delivery incorporating both RSU-assisted and carry-and-forward strategies. The network is geographically and equally divided by the coverage of RSUs. Data is generated by a source node S and to be carried and forwarded by vehicles to a destination node D. In general, S and D can be either RSU or vehicle. Without loss of generality, we assume that both S and D are RSUs. The traffic information of all vehicles in the networks are available

at RSUs, however not all RSUs are necessarily interconnected[1]. We further assume n routes, denoted as a set $\mathbf{\Gamma} = \{\gamma_i | i = 1, \ldots, n\}$, between S and D. A route γ_i could be composed of multiple hops and each hop refers to a road segment within the coverage of corresponding RSU.

4.2.2 Traffic Model

The data generated from S will be carried by vehicles moving along a route directing to D. Similar to the previous works [168, 185, 186, 187], we assume that vehicles move at the same speed and stay within each hop for a constant duration T. This assumption holds well for highway or rural area, where RSUs are equidistantly deployed and vehicles on each lane move at the same speed with a slight deviation. For urban scenario where diverse coverages and speed limits are expected, a network can be partitioned into blocks and in each block, vehicles are likely to move at the similar speed due to speed limit along the isometric road segment between RSUs, hence the assumption is also valid. We further adopt the widely used traffic model where the number of vehicles arrives at a hop and heads to the next hop is Poisson distributed [168, 185, 186].

4.2.3 Data Forwarding Model

Consider one route out of the route set $\mathbf{\Gamma}$. As there might not exist a single vehicle that moves along all hops of the considered route, the data would need to be forwarded by the vehicle that carries the data, when it no longer heads to D, to another vehicle that does. We refer to the vehicle that carries data and moves along a hop of the route as the *courier* of the hop. If the courier moves towards the next hop of the route when leaving the current hop (the information about whether moving towards

[1]The data delivery problem with fully interconnected RSUs can be solved by routing algorithms for cellular networks of fixed network topology, which have been thoroughly studied and are trivial for V2X networks. Nevertheless, in this chapter we also consider partially and/or fully interconnected RSUs that enable data forwarding via wireless backhaul between RSU in Section 4.5.

the next hop or not can be acquired by e.g. navigation or pre-configured route of autonomous vehicle), the data is carried by the courier to the next hop and consequently the courier of the next hop is still this vehicle. Otherwise, when arriving at the current hop, the courier tries to discover another vehicle that moves towards the next hop, referred to as the *candidate* of the current hop, to forward the data. The discovery maintains at most a duration t (duration for the discovery process cannot exceed t), referred to as the *global candidate discovery duration*, and once succeeds, the courier forwards the data to the candidate via V2V communications within the remaining time it stays in the hop. Otherwise, it sends the data to the RSU that covers the hop via V2I communications within $T - t$, and then the RSU will find a suitable candidate heading to the next hop and forward the data. Fig. 4.2 shows an example of the candidate discovery.

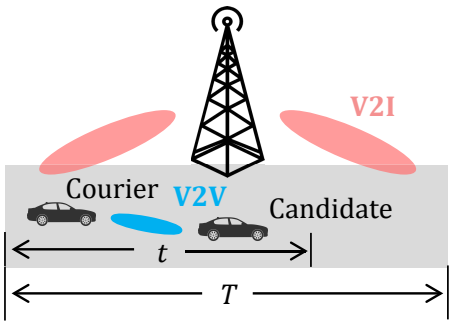


Figure 4.2: Illustration of candidate discovery within t by a courier to forward data when moving along the hop within T.

4.2.4 Radio Model

For candidate discovery, we further assume that when arriving at a hop, the courier starts to broadcast beacon for the candidate discovery assisted by RSU. Information encapsulated in the beacon could include, e.g., the direction of the next hop. Once the beacon has been received and successfully decoded, with an error probability ϵ, the candidate sends feedback to report the successful reception. The feedback is then decoded by the courier, also with the error probability ϵ, and in case of incorrect decoding or no feedback detected, the courier repeats the discovery trial until a

communication link between the courier and the candidate is established within t. The performance of broadcast and discovery has been studied in Chapter 1, and in this chapter we refer to the related analysis addressed there, including the trade-off between discovery latency and overhead, as well as the selection of the error probability ϵ.

4.3 Problem Formulation

The objective of this section is to characterize the data delivery problem by taking into account the models introduced in Section 4.2. Performance metrics, namely E2E latency and data rate, are analyzed in Section 4.3.1 and Section 4.3.2, respectively. Based on these, global and distributed data delivery problems that maximize the weighted sum of latency and data rate are formulated in Section 4.3.3 and Section 4.3.4, respectively. Without loss of generality, in the following sections we focus on a single route, which is referred to as the *typical* route and randomly picked from the routes set $\mathbf{\Gamma}$, between S and D. We further denote the number of hops of the typical route as k and the h-th hop of the typical route as hop h, respectively. Analysis of other routes can be derived similarly. The notations and parameters used for the analysis in this section are given in Table 4.2 and will be used in the rest of the chapter.

4.3.1 End-to-End Latency

As mentioned in Section 4.2.2, the arrival of vehicles at a hop follows a Poisson distribution. Here, we denote the arrival rate of vehicles that arrive at hop h and head to hop $h+1$ as $\lambda_{h,h+1}$. Based on the traffic model and data forwarding model described in Section 4.2.2 and Section 4.2.3, respectively, the E2E latency of the typical route is described in the following scenarios. Examples including courier moves to the next stop, successful discovery, and failed discovery are illustrated in

Table 4.2: Problem Formulation Parameters

Notation	Description	Value
k	Number of hops of the typical route	See Section 4.3
h	h-th hop of the typical route	See Section 4.3
$\lambda_{h,h+1}$	Vehicle arrival rate from hop h to $h+1$	$[0.05, 0.3]$
Deg_h	Number of way-out directions (except U-turn) for hop h	$\{1, 2, 3\}$
$\tau_{h,h+1}$	Inter-arrival time between courier and candidate arrival	See Section 4.3.1.2
Δt	Time of one discovery trial	See Section 4.3.1.2
m	Maximum number of discovery trials	$\lfloor \frac{t}{\Delta t} \rfloor$
$\tau_{h,h+1}^{(d,\mathrm{RSU})}$	Candidate discovery time for RSU in hop h	See (4.9)
r_{O}	Data rate of cellular communications	See Section 4.3.2.1
r_{V2V}	Data rate of V2V communications	See Section 4.3.2.2
$\tau_{h,h+1}^{(d,\mathrm{Veh})}$	Candidate discovery time for courier in hop h	See Section 4.3.2.2
r_{V2I}	Data rate of V2I communications	See Section 4.3.2.3
$\hat{t}_h$	Hop-wise candidate discovery duration of hop h	See Section 4.3.4

hop 1, 2, and 3 of Fig. 4.3, respectively.

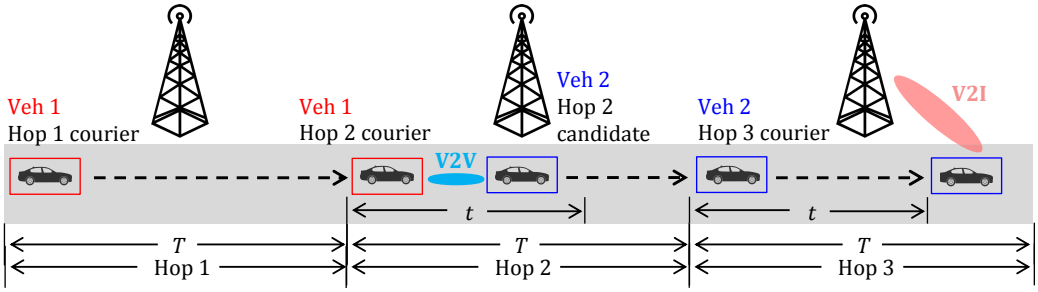


Figure 4.3: Illustration of a route with three hops that correspond to courier moves to the next hop, successful discovery, and failed discovery, respectively.

4.3.1.1 Courier Moves Towards the Next Hop

In this scenario, the courier carries the data to the next hop (hop $h+1$), as depicted in hop 1 of Fig. 4.3. Hence, there is no need for candidate discovery. Denoting the number of way-out directions except U-turn for hop h as Deg_h. Then, the probability of the event "Courier of hop h moves towards hop $h+1$", denoted as $P(\text{Courier: } h \to h+1)$, satisfies

$$P(\text{Courier: } h \to h+1) = \frac{1}{\mathrm{Deg}_h}, \tag{4.1}$$

which means when leaving hop h, the courier randomly selects a direction with equal probability. Clearly, the hop-wise latency of hop h for the event "Courier of hop h moves towards hop $h+1$", denoted as $L_{\text{Courier: } h \to h+1}$, is the duration of the courier staying in the hop, namely

$$L_{\text{Courier: } h \to h+1} = T. \tag{4.2}$$

4.3.1.2 Courier Succeeds in Candidate Discovery

In case the courier is not heading to the next hop, a candidate discovery for data forwarding is carried out by the courier within the discovery duration t, as illustrated in hop 2 of Fig. 4.3. When the number of arrivals in a given time interval follows Poisson distribution, inter-arrival times are known to have the exponential distribution [188]. Thus, the probability of the event "a candidate towards hop $h+1$ arrives at hop h within t on condition that courier of hop h does not move to hop $h+1$", denoted as $P(\tau_{h,h+1} \leq t)$, satisfies

$$\begin{aligned} P(\tau_{h,h+1} \leq t) &= P(\text{A candidate heading to } h+1 \text{ arrives at hop } h \text{ within } t) \\ &\quad \cdot \big(1 - P(\text{Courier: } h \to h+1)\big) \\ &= \left(\int_0^t \lambda_{h,h+1} e^{-\lambda_{h,h+1}\tau_{h,h+1}} \, \mathrm{d}\tau_{h,h+1} \right) \left(1 - \frac{1}{\text{Deg}_h} \right) \\ &= \big(1 - e^{-\lambda_{h,h+1} t}\big) \left(1 - \frac{1}{\text{Deg}_h} \right), \end{aligned} \tag{4.3}$$

where $\tau_{h,h+1}$ represents the time between the courier arrives at hop h and a candidate heading to $h+1$ arrives at hop h.

When arriving at hop h, the courier starts to broadcast beacon to discover a candidate for data forwarding, as described in Section 4.2.4. We denote the time of one discovery trial as Δt, which includes beacon broadcasting and feedback receiving time. Then, within the candidate discovery duration t, there would be maximally $m = \left\lfloor \frac{t}{\Delta t} \right\rfloor$ rounds of discovery trials, where $\lfloor . \rfloor$ represents the floor function. As one discovery trial consists of the beacon reception at candidate and feedback reception at courier, both with the error probability ϵ as discussed in Sec-

tion 4.2.4, the probability of a successful discovery trial is calculated as $(1-\epsilon)^2$. Then, the probability of the event "Discovery succeeds in m trials", denoted as $P(\text{Discovery succeeds in } m \text{ trials})$, is provided as

$$\begin{aligned} P(\text{Discovery succeeds in } m \text{ trials}) &= \sum_{i=0}^{m-1} \left(1-(1-\epsilon)^2\right)^i (1-\epsilon)^2 \\ &= \left(1-\left(1-(1-\epsilon)^2\right)^m\right). \end{aligned} \tag{4.4}$$

Correspondingly, the probability of the event "Courier of hop h successfully discovers a candidate", denoted as $P(\text{Success})$, can be derived as

$$\begin{aligned} P(\text{Success}) &= P(\tau_{h,h+1} \leq t) \cdot P(\text{Discovery succeeds in } m \text{ trials}) \\ &= \left(1-\frac{1}{\text{Deg}_h}\right)\left(1-e^{-\lambda_{h,h+1}t}\right)\left(1-\left(1-(1-\epsilon)^2\right)^m\right). \end{aligned} \tag{4.5}$$

Similarly to the previous scenario, the hop-wise latency of hop h for the event "Courier of hop h successfully discovers a candidate", denoted as L_{Success}, equals to the duration of the courier staying in the hop, i. e.,

$$L_{\text{Success}} = T. \tag{4.6}$$

4.3.1.3 Courier Fails in Candidate Discovery

In this scenario, the courier has to send the data to RSU as no proper candidate can be discovered within t, as drawn in hop 3 of Fig. 4.3. Accordingly, the probability of the event "no candidate towards hop $h+1$ arrives at hop h within t on condition that courier of hop h does not move to hop $h+1$", denoted as $P(\tau_{h,h+1} > t)$, satisfies

$$\begin{aligned} P(\tau_{h,h+1} > t) &= \big(1-P(\text{A candidate heading to } h+1 \text{ arrives at hop } h \text{ within } t)\big) \\ &\quad \cdot \big(1-P(\text{Courier: } h \to h+1)\big) \\ &= e^{-\lambda_{h,h+1}t}\left(1-\frac{1}{\text{Deg}_h}\right). \end{aligned} \tag{4.7}$$

Then, the probability of the event "Courier of hop h fails to discover a candidate", denoted as $P(\text{Failure})$, can be derived as

$$\begin{aligned} P(\text{Failure}) &= P(\tau_{h,h+1} \leq t)\big(1 - P(\text{Discovery succeeds in } m \text{ trials})\big) + P(\tau_{h,h+1} > t) \\ &= P(\tau_{h,h+1} \leq t)\big(1 - (1-\epsilon)^2\big)^m + P(\tau_{h,h+1} > t) \\ &= \left(1 - \frac{1}{\text{Deg}_h}\right)\big(1 - e^{-\lambda_{h,h+1}t}\big)\big(1 - (1-\epsilon)^2\big)^m + \left(1 - \frac{1}{\text{Deg}_h}\right)e^{-\lambda_{h,h+1}t}. \end{aligned} \tag{4.8}$$

As the data is not forwarded to any candidate within t, the courier will transmit the data to RSU in the remaining time $T - t$. Specifically, the courier has to stay in the hop for T anyway to move through the hop, which means that except the time for candidate discovery t, the remaining time for the courier to transmit the data to RSU is $T - t$. Afterwards, the RSU will find an appropriate candidate to forward the data and once found, the candidate receives the data from the RSU when moving along the hop within T. Therefore, the summarized hop-wise latency of hop h for the event "Courier of hop h fails to discover a candidate", denoted as L_{Failure}, is calculated as

$$L_{\text{Failure}} = T + \tau_{h,h+1}^{(d,\,\text{RSU})} + T = 2T + \tau_{h,h+1}^{(d,\,\text{RSU})}. \tag{4.9}$$

Here, $\tau_{h,h+1}^{(d,\,\text{RSU})}$ represents the time for RSU to find a candidate. Note that the first T in (4.9) does not overlap with $t_{h,h+1}^{(d,\,\text{RSU})}$ as in the remaining time after the courier has failed in candidate discovery, i.e. $T - t$, the RSU is receiving the data from the courier and not able to start to find a candidate.

Combining the three scenarios, the expected hop-wise latency for data delivery, denoted as $E(L_{h,h+1})$, can be derived as

$$\begin{aligned} E(L_{h,h+1}) = {} & P(\text{Courier: } h \to h+1)E(L_{\text{Courier: } h \to h+1}) + P(\text{Success})E(L_{\text{Success}}) \\ & + P(\text{Failure})E(L_{\text{Failure}}). \end{aligned} \tag{4.10}$$

In particular, $\tau_{h,h+1}^{(d,\,\text{RSU})}$ is also exponentially distributed due to Poisson distribution

of vehicle arrivals, and correspondingly we have

$$E(L_{\text{failure}}) = \int_0^t \left(2T + \tau_{h,h+1}^{(d,\,\text{RSU})}\right) \lambda_{h,h+1} e^{-\lambda_{h,h+1}\tau_{h,h+1}^{(d,\,\text{RSU})}} \cdot \mathrm{d}\tau_{h,h+1}^{(d,\,\text{RSU})}$$
$$= 2T + \frac{1}{\lambda_{h,h+1}}. \tag{4.11}$$

Finally, the expected E2E latency of the typical route, denoted as $\bar{L}$, is provided as

$$\bar{L} = \sum_{h=1}^{k} E(L_{h,h+1}). \tag{4.12}$$

4.3.2 End-to-End Data Rate

As mentioned in Section 4.1, we consider a realistic model of RSU in terms of providing V2I communications to vehicular users, which is different from the dropbox functionality considered in [181, 182, 183, 184] that only allows data to be stored at some cost. In this section, we study the data rate provided by the considered V2X networks for both vehicular and cellular users, where the overall system performance of the V2X networks is optimized, and promote our framework to be applied in realistic V2X networks that support ubiquitous connection to all devices. Similar to Section 4.3.1, three scenarios are addressed here for the analysis of the expected data rate of the typical route.

4.3.2.1 Courier Moves Towards the Next Hop

In this scenario, RSU is not requested by the courier for assisting the candidate discovery. Therefore, the RSU exclusively serves cellular users. Denoting the achievable data rate of the services provided for cellular users as r_{O}, the achievable hop-wise data rate of hop h for the event "Courier of hop h moves towards hop $h+1$", denoted as $C_{\text{Courier: }h\to h+1}$, is calculated as

$$C_{\text{Courier: }h\to h+1} = r_{\text{O}}. \tag{4.13}$$

Note that r_{O} indicates the overall achievable data rate of cellular users and can be obtained by e.g. taking average of data rates among all cellular users in the network.

4.3.2.2 Courier Succeeds in Candidate Discovery

When a communication link has been established between courier and candidate, the carried data is transmitted from the courier to the candidate via V2V communications with data rate r_{V2V}, which can be similarly obtained as r_{O} by e.g. taking average of data rates among all V2V communication links. Denoting the actual candidate discovery time for the courier of hop h as $\tau_{h,h+1}^{(d,\,\mathrm{Veh})}$, the achievable hop-wise data rate of hop h for the event "Courier of hop h successfully discovers a candidate", denoted as C_{Success}, can be written as

$$C_{\mathrm{Success}} = \frac{r_{\mathrm{V2V}}\left(T - \tau_{h,h+1}^{(d,\,\mathrm{Veh})}\right)}{T} + \frac{r_{\mathrm{O}}(T-t)}{T}. \tag{4.14}$$

Here, the data rate consists of two parts. The first term of the right hand side of (4.14) indicates that the V2V communication link maintains a duration of $T - \tau_{h,h+1}^{(d,\,\mathrm{Veh})}$, and the amount of data can be transmitted is calculated as $r_{\mathrm{V2V}}(T-\tau_{h,h+1}^{(d,\,\mathrm{Veh})})$. As the RSU of hop h is not aware of the actual candidate discovery time $\tau_{h,h+1}^{(d,\,\mathrm{Veh})}$, it preserves t for the candidate discovery and correspondingly, the data rate of applications/services provided by the RSU for cellular users is calculated as $r_{\mathrm{O}}(T-t)$.

4.3.2.3 Courier Fails in Candidate Discovery

In this scenario, the data is first transmitted to RSU after failed candidate discovery within t and then forwarded to a proper candidate. The RSU of hop h is correspondingly exclusively associated with the courier for data reception and candidate discovery, until the data is successfully forwarded to a candidate. Therefore, the amount of data transmitted from the courier to the RSU via V2I communications is $r_{\mathrm{V2I}}(T-t)$, where r_{V2I} indicates the data rate of V2I communications and can be obtained similarly as r_{V2V} and r_{O}. After finding a candidate within $\tau_{h,h+1}^{(d,\,\mathrm{RSU})}$

and forwarding the received data to the candidate within $T - t$ (The data rate between the candidate and the RSU is also assumed to be r_{V2I}), the RSU is able to serve cellular users with amount of data $r_{\mathrm{O}}t$. In summary, the achievable hop-wise data rate of hop h for the event "Courier of hop h fails to discover a candidate", denoted as C_{Failure}, can be written as

$$C_{\mathrm{Failure}} = \frac{r_{\mathrm{V2I}}(T-t)}{2T + \tau_{h,h+1}^{(d,\,\mathrm{RSU})}} + \frac{r_{\mathrm{O}}t}{2T + \tau_{h,h+1}^{(d,\,\mathrm{RSU})}}. \tag{4.15}$$

Combining the three scenarios, the expected hop-wise data rate for data delivery, denoted as $E(C_{h,h+1})$, can be derived as

$$\begin{aligned} E(C_{h,h+1}) = {} & P(\text{Courier: } h \to h+1) E(C_{\text{Courier: } h \to h+1}) + P(\text{Success}) E(C_{\text{Success}}) \\ & + P(\text{Failure}) E(C_{\text{Failure}}). \end{aligned} \tag{4.16}$$

For a multihop route between S and D, the achievable data rate is determined by the "weakest" hop in which the lowest rate is achieved. Therefore, the expected E2E latency of the typical route, denoted as $\bar{C}$, is calculated as

$$\bar{C} = \min_{\forall h} E(C_{h,h+1}). \tag{4.17}$$

4.3.3 Global Data Delivery Problem

Based on the analysis in Section 4.3.1 and Section 4.3.2, it is clear that the global candidate discovery duration t plays a vital role in determining the E2E latency and data rate. On the one hand, a larger t allows courier a better chance to find a candidate but leaves less time for data forwarding, which leads to data rate degrade. On the other hand, a decreased t brings more time for data forwarding and correspondingly enhances the achievable data rate, while an increased latency is expected due to less opportunity for successful candidate discovery. Therefore, a trade-off between latency and data rate to optimize the overall performance can be achieved by the adaptation of t. The global data delivery optimization problem is formulated as follows:

Problem 4.1. *(Global weighted sum maximization)*

$$\max_{t\in[0,T]} \alpha\bar{C} - (1-\alpha)\bar{L}. \tag{4.18}$$

Here, $\alpha \in [0,1]$ is the weight parameter. Distinguished by various use cases and application services, α can be flexibly adjusted, e. g., $\alpha = 1$ may refer to latency-tolerant but rate-sensitive use case, while $\alpha = 0$ may indicate real-time services which are keen to latency with relative low demand on data rate. Here, the term "global" indicates the universal configuration of the candidate discovery duration, where an identical candidate discovery duration t is applied to all hops in the typical route. Note that here $\bar{C}$ and $\bar{L}$ are normalized in $[0,1]$, where the motivation of the normalization lies in the fact that data ranges and units of latency and data rate are not directly comparable [189].

4.3.4 Distributed Data Delivery Problem

In addition to the global data delivery problem, we further propose the distributed data delivery problem, where the weighted sum of latency and data rate is hop-wisely maximized, compared to the E2E-wise maximization of weighted sum addressed in Section 4.3.3. The motivation of the distributed data delivery comes from the fact that the arrival rates $\lambda_{h,h+1}$, the actual candidate discovery time for RSU $\tau_{h,h+1}^{(d,\mathrm{RSU})}$, and the actual candidate discovery time for vehicle $\tau_{h,h+1}^{(d,\mathrm{Veh})}$ are diverse in each hop. Therefore, a hop-tailored maximization may benefit from a hop-specific configuration of discovery duration, as the hop-individual weighted sum could bring potential gain in increasing data rate while reducing latency.

By replacing t in (4.3), (4.4), (4.5), (4.7), and (4.8) with $\hat{t}_h$, which represents the *hop-wise candidate discovery duration* of hop h, we can get the hop-wise probability of successful discovery and the hop-wise probability of failed discovery, denoted as $\hat{P}(\mathrm{Success})$ and $\hat{P}(\mathrm{Failure})$, respectively. Similarly, by replacing t in (4.14) and (4.15) with $\hat{t}_h$, we can get the hop-wise achievable data rate of successful discovery and the

hop-wise achievable data rate of failed discovery, denoted as $\hat{C}_{\text{Success},h}$ and $\hat{C}_{\text{Failure},h}$, respectively. Then, the expected hop-wise latency and data rate, denoted as $\hat{L}_h$ and $\hat{C}_h$, respectively, satisfy

$$\hat{L}_h = P(\text{Courier: } h \to h+1)T + \hat{P}(\text{Success})T + \hat{P}(\text{Failure})\Big(2T + \frac{1}{\lambda_{h,h+1}}\Big), \quad (4.19)$$

and

$$\hat{C}_h = P(\text{Courier: } h \to h+1)r_{\text{O}} + \hat{P}(\text{Success})E(\hat{C}_{\text{Success},h}) + \hat{P}(\text{Failure})E(\hat{C}_{\text{Failure},h}). \quad (4.20)$$

Finally, the distributed data delivery optimization problem is formulated as follows:

Problem 4.2. *(Distributed weighted sum maximization)*

$$\max_{\hat{t}_h \in [0,T]} \alpha\hat{C}_h - (1-\alpha)\hat{L}_h. \quad (4.21)$$

Here, $\alpha \in [0,1]$ is the weight parameter. Similar to the global data delivery problem, a trade-off between latency and data rate to optimize the overall performance can be achieved by the adaptation of $\hat{t}_h$.

4.4 Data Delivery Optimization and Routing Algorithm Design

In this section, we propose the solutions of the data delivery optimization problems formulated in Section 4.3. The relevant expressions of latency and data rate described in Section 4.3 are reformed in closed-form in Section 4.4.1. Then, based on the reformed closed-form expressions of latency and data rate, in Section 4.4.2 we provide the convexity verification of the proposed optimization problems and solve both global and distributed weighted sum maximization by convex optimization theory. The design of the corresponding multihop routing algorithms based on the

optimal solutions proposed in Section 4.4.2 are addressed in Section 4.4.3.

4.4.1 Reformation of Problem Formulation

The optimization problem 4.1 and 4.2 are formulated in a sophisticated way in Section 4.3, and it is relatively hard to verify the convexity of the optimization problems determined in (4.18) and (4.21). In this section, the derived latency and data rate are reformed into the closed-form expressions.

4.4.1.1 Reformation of End-to-End Latency

For simplicity, let

$$\alpha_h = \frac{1}{\mathrm{Deg}_h}, \tag{4.22}$$

$$\beta_h(t) = e^{-\lambda_{h,h+1}t}, \tag{4.23}$$

$$\theta_h(t) = \big(1-(1-\epsilon)^2\big)^m, \tag{4.24}$$

$$\phi_h = T + \frac{1}{\lambda_{h,h+1}}. \tag{4.25}$$

Then, we transform the expected hop-wise latency derived in (4.10) as

$$\begin{aligned} E(L_{h,h+1}) &= P(\text{Courier: } h \to h+1)E(L_{\text{Courier: } h\to h+1}) + P(\text{Success})E(L_{\text{Success}}) \\ &\quad + P(\text{Failure})E(L_{\text{Failure}}) \\ &= \alpha_h T + (1-\alpha_h)\big(1-\beta_h(t)\big)\big(1-\theta_h(t)\big)\cdot T \\ &\quad + \Big((1-\alpha_h)\big(1-\beta_h(t)\big)\theta_h(t) + (1-\alpha_h)\beta_h(t)\Big)\cdot(T+\phi_h) \\ &= T + (1-\alpha_h)\phi_h\big(\beta_h(t)+\theta_h(t)-\beta_h(t)\theta_h(t)\big). \end{aligned} \tag{4.26}$$

Based on this, the expected E2E latency, which is depicted in (4.12), is now calculated as

$$\bar{L} = \sum_{h=1}^{k} E(L_{h,h+1}) = kT + \sum_{h=1}^{k}(1-\alpha_h)\phi_h\big(\beta_h(t)+\theta_h(t)-\beta_h(t)\theta_h(t)\big). \tag{4.27}$$

4.4.1.2 Reformation of End-to-End Data Rate

For simplicity, let

$$\zeta_h = (1-\alpha_h) r_{\mathrm{O}}, \tag{4.28}$$

$$\iota_h = \frac{r_{\mathrm{V2V}}\left(T - \frac{1}{\lambda_{h,h+1}}\right)}{T} + r_{\mathrm{O}}, \tag{4.29}$$

$$\kappa_h = \frac{r_{\mathrm{V2I}}}{2T + \frac{1}{\lambda_{h,h+1}}} T, \tag{4.30}$$

$$\nu_h(t) = -\frac{r_{\mathrm{O}}}{T} t, \tag{4.31}$$

$$\chi_h(t) = \frac{r_{\mathrm{O}} - r_{\mathrm{V2I}}}{2T + \frac{1}{\lambda_{h,h+1}}} t, \tag{4.32}$$

$$z(t) = \beta_h(t) + \theta_h(t) - \beta_h(t)\theta_h(t). \tag{4.33}$$

Then, similar to the reformation of the expected hop-wise latency, the expected hop-wise data rate, which is derived in (4.16), can be transformed as

$$\begin{aligned} E(C_{h,h+1}) &= P(\text{Courier: } h \to h+1) E(C_{\text{Courier: } h \to h+1}) + P(\text{Success}) E(C_{\text{Success}}) \\ &\quad + P(\text{Failure}) E(C_{\text{Failure}}) \\ &= \zeta_h + (1-\alpha_h)\big(1 - z_h(t)\big)\big(\iota_h + \nu_h(t)\big) + (1-\alpha_h) z_h(t)\big(\kappa_h + \chi_h(t)\big). \end{aligned} \tag{4.34}$$

As the expression of $E(C_{h,h+1})$ in (4.34) is a combination of multiplication and summation of hop-dependent terms (all terms with the subscript h), finding the closed-form of the minimum of $E(C_{h,h+1})$, namely $\bar{C}$, is still a very complicated problem. Therefore, we further reformed $\bar{C}$ as

$$\begin{aligned} \bar{C} &= P(\text{All success}) E(C_{\text{All success}}) + P(\text{All failure}) E(C_{\text{All failure}}) \\ &\quad + P(\text{Mixture}) E(C_{\text{Mixture}}), \end{aligned} \tag{4.35}$$

where $P(\text{All success})$, $P(\text{All failure})$, and $P(\text{Mixture})$ represent the probability of the event "Couriers of all hops succeed in candidate discovery", the probability of the event "Couriers of all hops fail in candidate discovery", and the probability of

the event "A mixture of all situations[2]", respectively, and are provided by

$$P(\text{All success}) = \prod_{h=1}^{k} P(\text{Success}), \tag{4.36}$$

$$P(\text{All failure}) = \prod_{h=1}^{k} P(\text{Failure}), \tag{4.37}$$

$$P(\text{Mixture}) = 1 - P(\text{All success}) - P(\text{All failure}). \tag{4.38}$$

Similarly, $C(\text{All success})$, $C(\text{All failure})$, and $C(\text{Mixture})$ represent the expected data rate of the corresponding events and are provided by

$$C_{\text{All success}} = \min_{\forall h} C_{\text{Success},h}, \tag{4.39}$$

$$C_{\text{All failure}} = \min_{\forall h} C_{\text{Failure},h}, \tag{4.40}$$

$$C_{\text{Mixture}} = \min_{\forall h,l,h\neq l} (r_{\text{O}}, C_{\text{Success},h}, C_{\text{Failure},l}). \tag{4.41}$$

Now, it is clear that the main task of finding the closed-form of the expected E2E data rate is to reform the minimization operators addressed in (4.39), (4.40), and (4.41).

a) Closed-form of $E(C_{\text{All success}})$.

Lemma 4.4.1. *Given a set of i.i.d geometrically distributed random variables* $\{X_i | i = 1, \ldots, n\}$ *with parameter* $p \in (0, 1)$*, the probability mass function (PMF) of* $Y = \max(X_1, \ldots, X_n)$*, denoted as* $f_Y(x)$*, satisfies*

$$f_Y(x) = \left(1 - (1-p)^x\right)^n - \left(1 - (1-p)^{x-1}\right)^n. \tag{4.42}$$

Proof. See Appendix 4.7.1. □

Considering the scenario of successful candidate discovery in all hops, the E2E data rate is limited by the "weakest" hop, at which the longest time for candidate

[2]Besides the above two particular scenarios, data delivery along the typical route is generally a mixture of successful and failed candidate discovery. Specifically, courier could find a candidate in some hops, not find any candidates in some other hops, and directly move to the next hop by itself as well.

discovery is consumed. Consequently, the goal of finding $\min_{\forall h} C_{\text{Success},h}$ in (4.39) turns to find $\max_{\forall h} \tau_{h,h+1}^{(d,\text{Veh})}$, which is solved by the following theorem.

Theorem 4.4.1. *Let $m_h \in \{1,\ldots,m\}, \forall h \in \{1,\ldots,k\}$ and $\xi = \max_{\forall h} m_h$ denote the actual number of discovery trials in hop h and the maximum of m_h, respectively. Let $p = (1-\epsilon)^2$. Then, the closed-form of $E(C_{All\ success})$ satisfies*

$$E(C_{All\ success}) = \frac{r_{\text{V2V}}\Big(T - E\Big(\max_{\forall h} \tau_{h,h+1}^{(d,\text{Veh})}\Big)\Big) + r_O(T-t)}{T}, \tag{4.43}$$

where

$$E\big(\max_{\forall h} \tau_{h,h+1}^{(d,\text{Veh})}\big) = \sum_{\xi=1}^{m} \xi\Big(\big(1-(1-p)^{\xi}\big)^k - \big(1-(1-p)^{\xi-1}\big)^k\Big)\cdot \Delta t. \tag{4.44}$$

Proof. See Appendix 4.7.2. □

b) Closed-form of $E(C_{\text{All failure}})$.

The E2E data rate of this scenario can be solved similarly. Specifically, finding $\min_{\forall h} C_{\text{Failure},h}$ in (4.40) turns to find $\max_{\forall h} \tau_{h,h+1}^{(d,\text{RSU})}$ where $\tau_{h,h+1}^{(d,\text{RSU})}$ follows a exponentially distribution with parameter $\lambda_{h,h+1}$.

Lemma 4.4.2. *Given a set of i.i.d exponentially distributed random variables $\{X_i | i = 1,\ldots,n\}$ with parameter λ_i, the probability density function (PDF) of $Z = \max(X_1,\ldots,X_n)$, denoted as $f_Z(x)$, satisfies*

$$f_Z(x) = \sum_{i=1}^{n}\Big(\lambda_i e^{-\lambda_i x}\prod_{j=1,j\neq i}^{n}\big(1-e^{-\lambda_j x}\big)\Big). \tag{4.45}$$

Proof. See Appendix 4.7.3. □

Given Lemma 4.4.2, the closed-form of the expected E2E data rate $E(C_{\text{All failure}})$ is solved by the following theorem.

Theorem 4.4.2. *Let $\eta = \max_{\forall h} \tau_{h,h+1}^{(d,\text{RSU})}, \tau_{h,h+1}^{(d,\text{RSU})} \in [0,\infty)$ and $\mu_h = \lambda_{h,h+1}, \forall h \in \{1,\ldots,k\}$, respectively. Then, the closed-form of $E(C_{All\ failure})$ satisfies*

$$E(C_{All\ failure}) = \frac{r_{\text{V2I}}(T-t) + r_O t}{2T + E\Big(\max_{\forall h} \tau_{h,h+1}^{(d,\text{RSU})}\Big)}, \tag{4.46}$$

where

$$E(\max_{\forall h} \tau_{h,h+1}^{(d,\,\mathrm{RSU})}) = \sum_{h=1}^{k} \mu_h \int_0^{\infty} \eta \Big(e^{-\mu_h \eta} \prod_{l=1, l\neq h}^{k} \big(1 - e^{-\mu_l \eta}\big) \Big) \mathrm{d}\eta. \tag{4.47}$$

Proof. See Appendix 4.7.4. □

c) Closed-form of $E(C_{\mathrm{Mixture}})$.

The derivation of the closed-form of the expected E2E data rate $E(C_{\mathrm{Mixture}})$ is a bit tricky as the hop latency for this scenario follows a combination of geometric distribution (successful discovery) and exponential distribution (failed discovery). To solve the problem, we first introduce the following lemma:

Lemma 4.4.3. *For a random variable X with non-negative values, the expectation of X, denoted as $E(X)$, satisfies*

$$E(X) = \int_0^{\infty} \big(1 - F_X(x)\big) \mathrm{d}x, \tag{4.48}$$

where $F_X(x)$ indicates the CDF of X.

Proof. See Appendix 4.7.5. □

Further, we notice that $E(C_{\mathrm{Mixture}})$ can be derived as

$$E(C_{\mathrm{Mixture}}) = E\Big(\min_{\forall h,l,h\neq l} (r_{\mathrm{O}}, C_{\mathrm{Success},h}, C_{\mathrm{Failure},l}) \Big). \tag{4.49}$$

Let $\rho = \min_{\forall h,l,h\neq l}(C_{\mathrm{Success},h}, C_{\mathrm{Failure},l})$, then we have

$$\begin{aligned} E(C_{\mathrm{Mixture}}) &= P(r_{\mathrm{O}} \leq \rho) r_{\mathrm{O}} + P(r_{\mathrm{O}} \geq \rho) E(\rho) \\ &= \big(1 - P(\rho \leq r_{\mathrm{O}})\big) r_{\mathrm{O}} + P(\rho \leq r_{\mathrm{O}}) E(\rho) \\ &= \big(1 - F_{\min_{\forall h} C_{\mathrm{Success},h}}(r_{\mathrm{O}}) F_{\min_{\forall l} C_{\mathrm{Failure},l}}(r_{\mathrm{O}})\big) r_{\mathrm{O}} \\ &\quad + F_{\min_{\forall h} C_{\mathrm{Success},h}}(r_{\mathrm{O}}) F_{\min_{\forall l} C_{\mathrm{Failure},l}}(r_{\mathrm{O}}) E(\rho). \end{aligned} \tag{4.50}$$

Here, the expectation of ρ can be derived by applying Lemma 4.4.3 as

$$E(\rho) = \int_0^{\infty} \big(1 - F_{\rho}(x)\big) \mathrm{d}x, \tag{4.51}$$

where

$$F_{\rho}(x) = 1 - \big(1 - F_{\min_{\forall h} C_{\text{Success},h}}(x)\big)\big(1 - F_{\min_{\forall l} C_{\text{Failure},l}}(x)\big)$$
$$= F_{\min_{\forall h} C_{\text{Success},h})}(x) + F_{\min_{\forall l} C_{\text{Failure},l}}(x) - F_{\min_{\forall h} C_{\text{Success},h}}(x) F_{\min_{\forall l} C_{\text{Failure},l}}(x). \tag{4.52}$$

In (4.50) and (4.52), $F_{\min_{\forall h} C_{\text{Success},h}}(x)$ and $F_{\min_{\forall l \neq h} C_{\text{Failure},l}}(x)$ can be calculated from $F_{\xi}(x)$ and $F_{\eta}(x)$, where $F_{\text{X}}(x)$ represents the CDF of random variable X. It is clear that $F_{\xi}(x)$ and $F_{\eta}(x)$ can be obtained by taking summation of (4.64) and integral of (4.69), respectively.

4.4.2 Optimal Data Delivery

4.4.2.1 Global Data Delivery Optimization

From (4.27), (4.43), (4.44), (4.46), (4.47), (4.50), (4.51), and (4.52), it is evident that the global candidate discovery duration t is the variable for adapting the E2E latency and data rate, given the typical route with k hops, the hop duration T, the error probability ϵ, the vehicle arrival rate $\lambda_{h,h+1}$, and the data rates r_{V2V}, r_{V2I}, and r_{O}. It is shown in Appendix 4.7.6 that the global optimization problem is convex and differentiable, which can be solved by convex optimization theory.

Theorem 4.4.3. *Let t_s be the stationary point of $\alpha\bar{C} - (1-\alpha)\bar{L}$, i. e., $\frac{\mathrm{d}(\alpha\bar{C} - (1-\alpha)\bar{L})}{\mathrm{d}t_s} = 0$. Then, t^* is the optimal solution for Problem 4.1, where*

$$t^* = \underset{t=\{0,t_s,T\}}{\arg\max}\, \alpha\bar{C} - (1-\alpha)\bar{L}, t_s \in [0,T]. \tag{4.53}$$

Proof. See Appendix 4.7.7. □

4.4.2.2 Distributed Data Delivery Optimization

Similar to the global data delivery problem, it is evident that the hop-wise candidate discovery duration $\hat{t}_h$ controls the hop-wise latency and data rate considering the

hop duration T, the error probability ϵ, the vehicle arrival rate $\lambda_{h,h+1}$, and the data rates r_{V2V}, r_{V2I}, and r_{O}. Specifically, the expected hop latency and data rate, which are $\hat{L}_h$ and $\hat{C}_h$ addressed in (4.19) and (4.20), respectively, can be derived as

$$\hat{L}_h = T + (1-\alpha_h)\phi_h\big(\hat{\beta}_h(t) + \hat{\theta}(t) - \hat{\beta}_h(t)\hat{\theta}(t)\big), \tag{4.54}$$

and

$$\hat{C}_h = \zeta_h + (1-\alpha_h)\big(1-\hat{z}(t)\big)\big(\iota_h + \hat{\nu_h}(t)\big) + (1-\alpha_h)\hat{z}(t)\big(\kappa_h + \hat{\chi_h}(t)\big), \tag{4.55}$$

where

$$\hat{\beta_h}(t) = e^{-\lambda_{h,h+1}\hat{t}_h}, \tag{4.56}$$

$$\theta_h(t) = \big(1-(1-\epsilon)^2\big)^{\left\lfloor \frac{\hat{t}_h}{\Delta t} \right\rfloor}, \tag{4.57}$$

$$\hat{\nu_h}(t) = -\frac{r_{\mathrm{O}}}{T}\hat{t}_h, \tag{4.58}$$

$$\hat{\chi_h}(t) = \frac{r_{\mathrm{O}} - r_{\mathrm{V2I}}}{2T + \frac{1}{\lambda_{h,h+1}}}\hat{t}_h,$$

$$\hat{z}(t) = \hat{\beta}_h(t) + \hat{\theta}(t) - \hat{\beta}_h(t)\hat{\theta}(t). \tag{4.59}$$

According to Chapter 2, Δt, which is the time of one discovery trial, depends on beam duration, frame length, and error probability, which are independent of the duration $\hat{t}_h$. Therefore, the distributed data delivery performance in terms of the weighted sum of the hop-wise latency and the hop-wise data rate, is maximized by determining the optimal $\hat{t}_h$. The convexity of the distributed optimization problem can be verified similarly to the global optimization problem 4.1 and is omitted here to avoid redundancy. Ultimately, the optimization problem 4.2 is solved by the following theorem:

Theorem 4.4.4. *Let $\hat{t}_{h,s}$ be the stationary point of $\alpha\hat{C}_h - (1-\alpha)\hat{L}_h$, i. e., $\frac{\mathrm{d}(\alpha\hat{C}_h-(1-\alpha)\hat{L}_h)}{\mathrm{d}\hat{t}_{h,s}} = 0$. Then $\hat{t}_h^*$ is the optimal solution for Problem 4.2, where*

$$\hat{t}_h^* = \underset{\hat{t}_h=\{0,\hat{t}_{h,s},T\}}{\arg\max}\ \alpha\hat{C}_h - (1-\alpha)\hat{L}_h, \hat{t}_{h,s} \in [0,T]. \tag{4.60}$$

4.4.3 Routing Algorithm Design

We summarize our routing algorithms to solve the global data delivery optimization problem 4.1 and the distributed data delivery optimization problem 4.2, based on Theorem 4.4.3 and Theorem 4.4.4, in Algorithm 4.1 and Algorithm 4.2, respectively. The algorithms can be e.g. executed at RSUs where vehicles have access to the routing information when entering in the corresponding coverage. In Algorithm 4.1, the set of routes $\mathbf{\Gamma}$ are planned and created considering all possible routes between S and D. Routes are iteratively selected from $\mathbf{\Gamma}$ for calculating the weighted sum of latency and data rate until all routes have been traversed, as indicated in line 1. For each route, the weighted sum $\alpha\bar{C}_i - (1-\alpha)\bar{L}_i$ and the corresponding optimal duration of the route t_i are obtained in line 3 and 4, respectively. The update of the overall maximum weighted sum *opt*, the optimal route γ^*, and the overall global optimal duration t^* are described in line 5–9. Similar procedures can be found in Algorithm 4.2.

4.5 Numerical Evaluation and Discussion

In this section, we evaluate the performance of the proposed routing algorithms for data delivery in V2X networks. The system-level evaluation setup is described in Section 4.5.1. For the evaluation, we first verify the performance of the global routing algorithm in achieving the maximum weighted sum of latency and data rate by the optimal candidate discovery duration t^* in Section 4.5.2. Then, we compare the proposed global routing algorithm with classical vehicular routing algorithms in terms of both weighted sum and latency in Section 4.5.3. In addition, we investigate the impact of broadcast schemes and vehicle arrival rates on data delivery performance in Section 4.5.4. Finally, in Section 4.5.5 and Section 4.5.6, we provide the performance comparison of data delivery with and without wireless backhaul for failed candidate discovery and the performance comparison of both global and distributed routing algorithms, respectively.

Algorithm 4.1: Global Optimization Algorithm

Input: The set of routes $\mathbf{\Gamma} = \{\gamma_i | i = 1, \ldots, n\}$

Output: The maximum weighted sum *opt*, the optimal route γ^*, and the global optimal candidate discovery duration t^*

- n: Number of routes between S and D
- $\lambda_{h,h+1}^{(i)}$: Arrival rate at h-th hop of route γ_i
- t_i: Global candidate discovery duration of route γ_i
- $\bar{L}_i$: Expected E2E latency of route γ_i
- $\bar{C}_i$: Expected E2E data rate of route γ_i
- i: Iterator

Initialization: $opt = 0$, $\gamma^* = \gamma_1$, $t^* = 0$, $i = 1$

begin

while $i \neq n$ **do**

Find $\lambda_{h,h+1}^{(i)}$, $\forall h \in \{1, \ldots, k_i\}$ in γ_i;

Obtain $\bar{L}_i$ and $\bar{C}_i$ by computing (4.27), (4.43), (4.44), (4.46), (4.47), (4.50), (4.51), and (4.52);

Obtain $\alpha\bar{C}_i - (1-\alpha)\bar{L}_i$ and t_i by solving $\frac{\mathrm{d}(\alpha\bar{C}_i - (1-\alpha)\bar{L}_i)}{\mathrm{d}t_i} = 0$;

if $\alpha\bar{C}_i - (1-\alpha)\bar{L}_i \geq opt$ **then**

$opt = \alpha\bar{C}_i - (1-\alpha)\bar{L}_i$;

$\gamma^* = \gamma_i$;

$t^* = t_i$;

end

$i = i + 1$;

end

Return *opt*, γ^* and t^*;

end

Algorithm 4.2: Distributed Optimization Algorithm

Input: The set of routes $\mathbf{\Gamma} = \{\gamma_i | i = 1, \ldots, n\}$

Output: The maximum weighted sum *opt*, the optimal route γ^*, and the set of hop-wise optimal candidate discovery durations $\hat{\boldsymbol{t}}^*$

- n: Number of routes between S and D
- $\lambda_{h,\,h+1}^{(i)}$: Arrival rate at h-th hop of route γ_i
- $\hat{t}_h$: Hop-wise candidate discovery duration in hop h
- $\hat{L}_h^{(i)}$: Expected hop-wise latency in hop h of route γ_i
- $\hat{C}_h^{(i)}$: Expected hop-wise data rate in hop h of route γ_i
- $\bar{L}_i$: Expected E2E latency of route γ_i
- $\bar{C}_i$: Expected E2E data rate of route γ_i
- $\hat{\boldsymbol{t}}_i$: Set of hop-wise candidate discovery durations of route γ_i
- i: Iterator

Initialization: $opt = 0$, $\gamma^* = \gamma_1$, $\hat{\boldsymbol{t}}^* = 0$, $i = 1$

begin

while $i \neq n$ **do**

Find $\lambda_{h,\,h+1}^{(i)}$, $\forall h \in \{1, \ldots, k_i\}$ in γ_i;

Obtain $\hat{L}_h^{(i)}$ and $\hat{C}_h^{(i)}$ by computing (4.54), (4.55), (4.56), (4.57), (4.58), and (4.59) $\forall h \in \{1, \ldots, k_i\}$;

Obtain $\alpha\hat{C}_h^{(i)} - (1-\alpha)\hat{L}_h^{(i)}$ and $\hat{t}_h$ by solving $\frac{\mathrm{d}\left(\hat{C}_h^{(i)} - (1-\alpha)\hat{L}_h^{(i)}\right)}{\mathrm{d}\hat{t}_h} = 0$, $\forall h \in \{1, \ldots, k_i\}$;

Obtain $\bar{L}_i$, $\bar{C}_i$, and $\hat{\boldsymbol{t}}_i$ by $\bar{L}_i = \sum_{h=1}^{k_i} \hat{L}_h^{(i)}$, $\bar{C}_i = \min_{\forall h \in \{1,\ldots,k_i\}} \hat{C}_h^{(i)}$, and $\hat{\boldsymbol{t}}_i = \{\hat{t}_h | h = 1, \ldots, k_i\}$;

if $\alpha\bar{C}_i - (1-\alpha)\bar{L}_i \geq opt$ **then**

$opt = \alpha\bar{C}_i - (1-\alpha)\bar{L}_i$;

$\gamma^* = \gamma_i$;

$\hat{\boldsymbol{t}}^* = \hat{\boldsymbol{t}}_i$;

end

$i = i + 1$;

end

Return *opt*, γ^*, and $\hat{\boldsymbol{t}}^*$;

end

4.5.1 Simulation Setup

The simulation scenario in this section is configured by adopting the system model described in Section 4.2. We consider an urban scenario deployment similar to the one proposed in [149], where square blocks are surrounded by streets that are 250 meters long and 20 meters wide. 9 RSUs marked as red triangles are located at the crossroads and 100 UEs marked as blue crosses are uniformly dropped in the street at the beginning of the simulation, as illustrated in Fig. 4.4. S and D are the upper-left RSU and the lower-right RSU depicted in the figure, respectively, and there are in total 12 loop-free routes between S and D. Channel model (LOS probability, pathloss, blockage model, etc.) is consistent with [149]. The default simulation parameters are summarized in Table 4.1, Table 4.2, and Table 4.3. Simulation samples are averaged over 1000 independent snapshots.

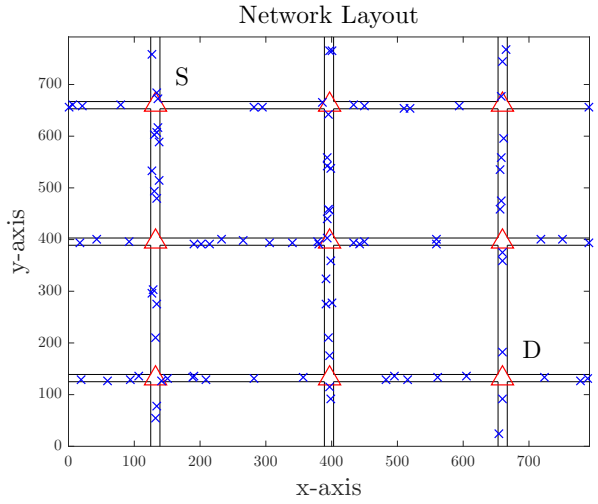


Figure 4.4: Illustration of simulation scenario.

Table 4.3: Simulation Parameters

Parameter	Value
Carrier frequency	28 GHz
System bandwidth	1 GHz
Vehicle speed	45 km/h
Antenna array (Vertical x Horizontal)	8×16 for RSU, 4×4 for UE
Maximum transmission power	30 dBm for RSU, 23 dBm for UE

4.5.2 Performance of Global Routing Algorithm

In Fig. 4.5 and Fig. 4.6, we plot the normalized weighted sum of E2E latency and E2E data rate, defined in (4.18), versus the choice of global candidate discovery duration t, with $\alpha = 0.5$ and $\alpha = \{0, 0.5, 1\}$, respectively. Here, $\alpha = 0.5$ indicates the weighted sum considering both latency and data rate, and $\alpha = 0$ and $\alpha = 1$ address the E2E latency and the E2E data rate, respectively. Remember that the motivation of the normalization lays in the fact that the data ranges of latency and data rate are not directly comparable.

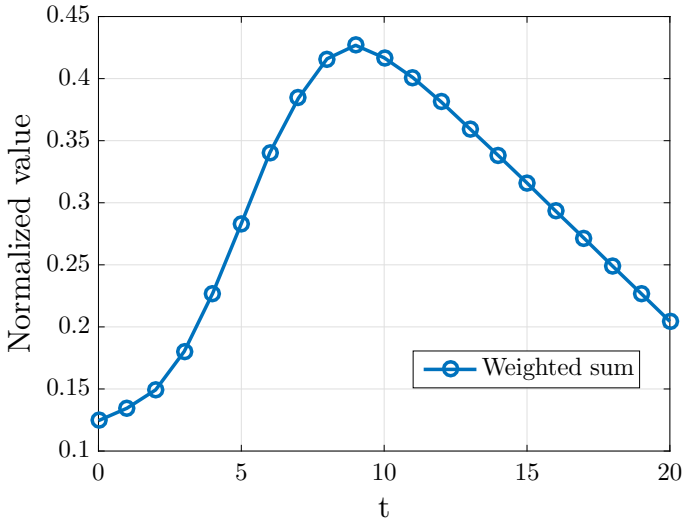


Figure 4.5: Performance of global routing algorithm is evaluated in normalized weighted sum with weight $\alpha = 0.5$.

Result in Fig. 4.5 shows that the solution of the global data delivery problem in Theorem 4.4.3 is accurate, where the optimal duration t^* exists and maximizes the weighted sum. With an increased t, a higher probability of successful candidate discovery can be expected where all hop latency equals to T, and thus the E2E latency is reduced and optimized when $t \to T$, as shown in Fig. 4.6. However, if both latency and data rate are to be considered, the optimal t^* is observed in the middle range of $[0, T]$ instead of $t^* \to T$ for minimizing latency, as a small value of t results in failed discovery in all hops and limits $r_{\mathrm{O}}t$ according to (4.15), while a

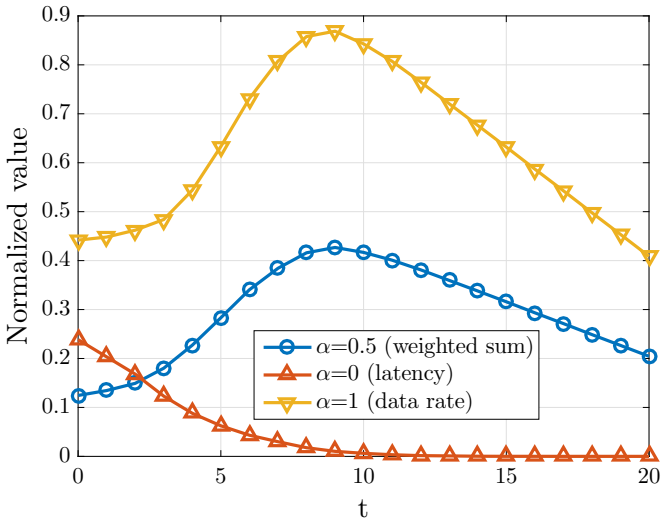


Figure 4.6: Performance of global routing algorithm is evaluated in normalized weighted sum with weight $\alpha = \{0, 0.5, 1\}$.

larger t degrades $r_{\mathrm{O}}(T-t)$ as indicated in (4.14).

In summary, these results indicate that the proposed routing algorithm is efficient for solving the global routing algorithm in terms of high data rate and low latency.

4.5.3 Comparison of Global Routing Algorithm and Other Vehicular Routing Algorithms

In this section, we compare the normalized weighted sum of the proposed global routing algorithm and other classical vehicular routing algorithms. In making the comparison, we consider two well-known routing algorithms [169]: the SPR that minimize the E2E latency, and the greedy perimeter stateless routing (GPSR) which is a classical geographical-based routing algorithm with high efficiency in data delivery.

To get some insights into the comparison, we plot the normalized weighted sum for different routing algorithms, versus the choice of candidate discovery duration t in Fig. 4.7 and Fig. 4.8, with $\alpha = \{0.5, 1\}$ and $\alpha = 0$ (latency), respectively. On the one hand, as depicted in Fig. 4.7, the proposed global algorithm outperforms the others in terms of the weighted sum with weight $\alpha = 0.5$ and $\alpha = 1$ (data rate), as

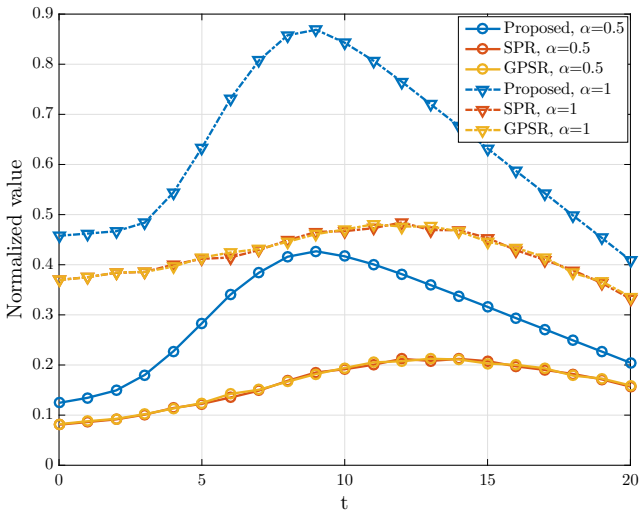


Figure 4.7: Performance of global routing algorithm and other routing algorithms are compared in normalized weighted sum with weight $\alpha = \{0.5, 1\}$.

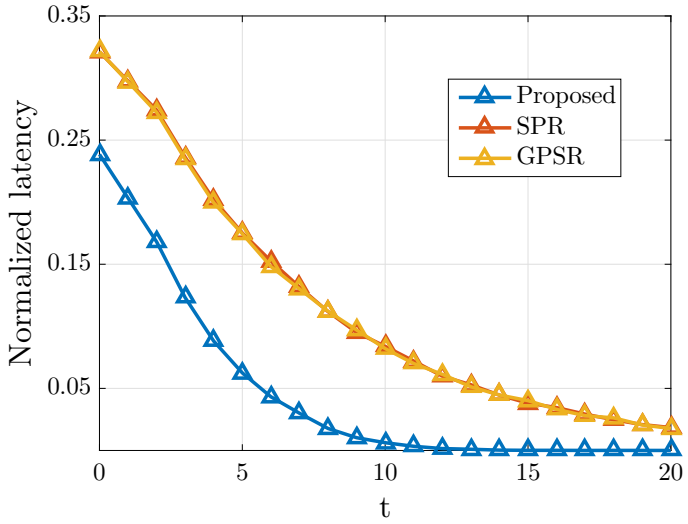


Figure 4.8: Performance of global routing algorithm and other routing algorithms are compared in normalized E2E latency.

the proposed algorithm selects the route that maximizes the weighted sum, which shows better performance than selecting the shortest route by SPR and GPSR. On the other hand, the proposed algorithm also achieves lower E2E latency compared to SPR and GPSR that are supposed to be able to minimize the latency, as shown in Fig. 4.8. The reason behind this lies in the fact that the geographically shortest route selected by SPR and GPSR is not necessarily the route that minimizes the latency.

4.5.4 Impact of Broadcast Schemes and Vehicle Arrival Rates on Data Delivery Performance

Fig. 4.9 and Fig. 4.10 plot the maximum normalized weighted sum calculated from (4.18) for different broadcast schemes described in Section 2.3, versus the number of simultaneous beams M for beamforming, with $\alpha = 0.5$ and $\alpha = 1$ (data rate), respectively. Here, TD, FD, CD, and SD represent the broadcast schemes addressed in Section 2.3 and multiplexed in time, time-frequency, time-code, and time-space, respectively. TD refers to the single beam exhaustive scan where $M = 1$, hence we plot a single circle instead of a curve to represent the maximum weighted sums achieved by TD in Fig. 4.9 and Fig. 4.10. For other schemes, multiple beams are formed simultaneously. Note that the E2E latency with $\alpha = 0$ is not compared as the minimum normalized latency is always 0.

According to the conclusion in Chapter 2 (addressed in Section 2.5), TD achieves the lowest average discovery latency which means that if the courier applies single beam exhaustive scan for candidate discovery, the time of one discovery trial Δt is minimized compared with other broadcast schemes. Moreover, FD and SD perform exactly same and worse than TD in terms of discovery latency, and the curve of SD locates in-between of TD and FD/CD.

In this way, TD decreases the number of maximum discovery trials m and the actual candidate discovery time $\tau_{h,h+1}^{(d,\mathrm{Veh})}$, and therefore leads to a higher probability

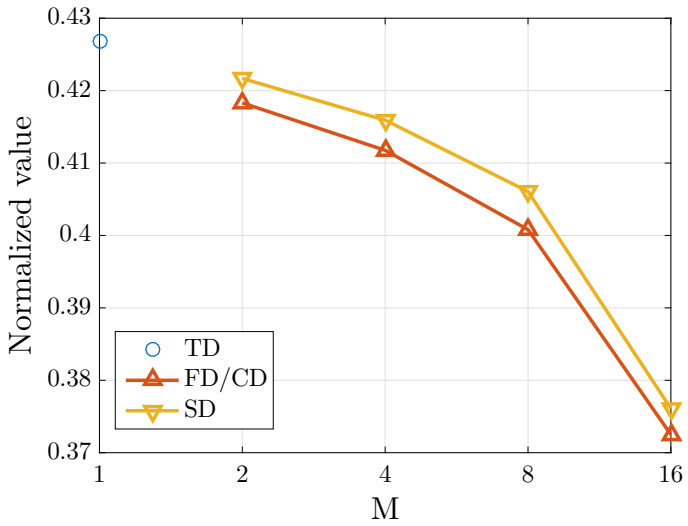


Figure 4.9: Performance of different broadcast schemes for candidate discovery are compared in the maximum normalized weighted sum with weight $\alpha = 0.5$.

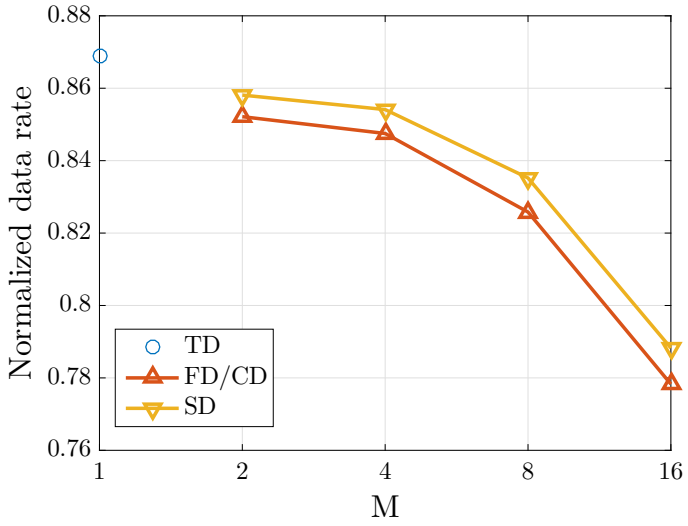


Figure 4.10: Performance of different broadcast schemes for candidate discovery are compared in the maximum normalized E2E data rate.

of successful candidate discovery, a lower E2E latency, and a higher hop-wise data rate of successful candidate discovery that are addressed in (4.5), (4.10), and (4.14), respectively, which eventually achieves the highest weighted sums for both $\alpha = 0.5$ and $\alpha = 1$, as illustrated in Fig. 4.9 and Fig. 4.10, respectively. By contrast, the weighted sums degrade when SD and FD/CD, which call for discovery latency penalty, are utilized for candidate discovery with more simultaneous beams. Therefore, the results in Fig. 4.9 and Fig. 4.10 recommend TD as the optimal beamforming and broadcast scheme for candidate discovery to explicitly achieve the best data delivery performance.

Fig. 4.11 plots the optimal candidate discovery duration t^* that optimizes te weighted sum with weight $\alpha = 0.5$ for different vehicle arrival rates. Here, the intervals of arrival rates for off-peak, in-between, and rush hour are defined as $[0.05, 0.15]$, $[0.1, 0.2]$, and $[0.2, 0.3]$ vehicles/second, respectively. According to the results in Fig. 4.11, we conclude that the optimal t^* is almost linearly decreased. The reason behind this falls in the fact that higher arrival rate improves the probability of successful candidate discovery in (4.5), and correspondingly relative small t is able to achieve the best data delivery performance.

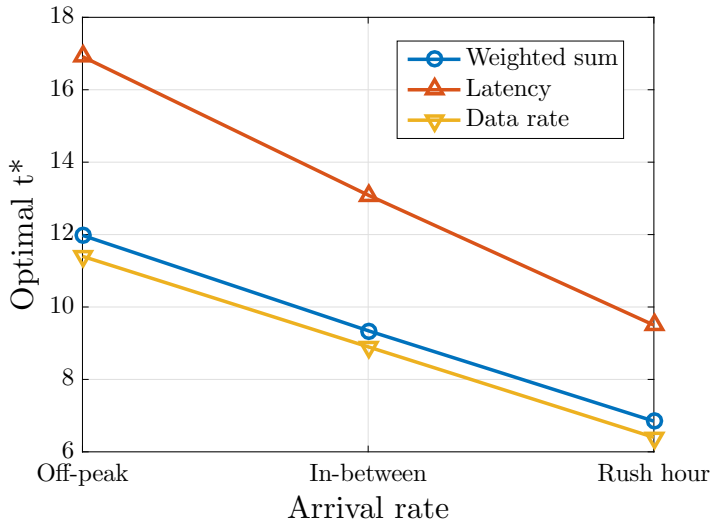


Figure 4.11: Comparison of the optimal t^* to achieve the maximum weighted sum for different vehicle arrival rates.

4.5.5 Performance of Wireless Backhauling for Data Delivery

Wireless backhauling, as an efficient alternative to expensive fiber connectivity among RSUs, provides coverage extension and capacity expansion with low latency, as addressed in Section 3.1. Therefore, instead of discovering a candidate to forward data received from a courier, RSU equipped with wireless backhaul is able to transmit the data directly to the RSU of the next hop in case they are interconnected via wireless backhaul. In Fig. 4.12, the normalized E2E latency ($\alpha = 0$) and the normalized E2E data rate ($\alpha = 1$) of global data delivery with (w/) backhaul and without (w/o) backhaul are compared, versus the choice of global candidate discovery duration t.

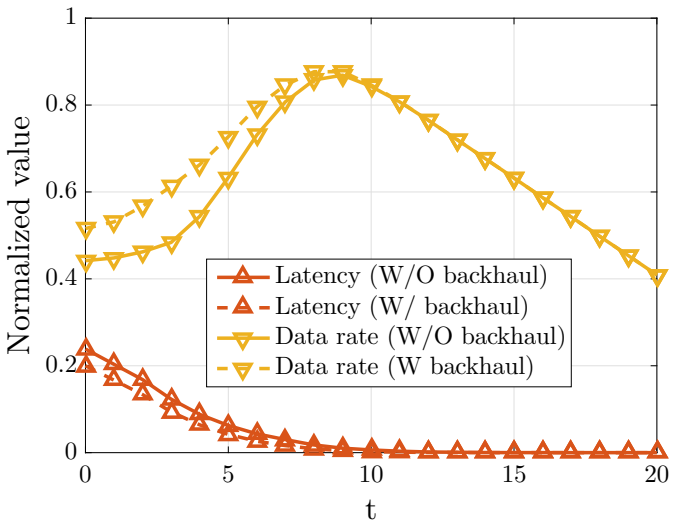


Figure 4.12: Performance of the data delivery with and without backhaul are compared in normalized latency and data rate.

On the one hand, the results in Fig. 4.12 suggest that incorporating backhaul data forwarding reduces the latency when t is small, i.e., the support of backhaul alleviates the burden of RSU in candidate discovery when courier fails. On the other hand, increased data rates are also observed when $t \to 0$, because a higher data rate is provided by the backhaul transmission compared to V2I communications. Note that for large values of t, candidate discovery tends to be successful in almost all

hops, therefore the data is carried and forwarded purely by vehicles without store-and-forward from RSU, and data delivery w/ and w/o backhaul perform exactly the same.

4.5.6 Comparison of Global and Distributed Routing Algorithm

In the end, the performance of global and distributed routing algorithms are compared in Table 4.4, Fig. 4.13, and Fig. 4.14. It is stated in Table 4.4 that around 4.7% gain of the maximum normalized weighted sum with weight $\alpha = 0.5$ and 2.9% gain with $\alpha = 1$ can be achieved by the distributed algorithm compared to the global one, respectively. The normalized E2E latency is minimized by both algorithms with equal value of 0 and thus not recorded in the table.

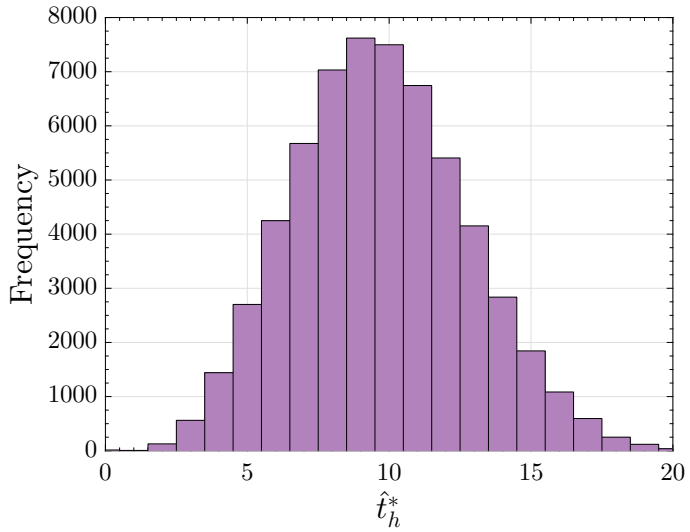


Figure 4.13: Illustration of the histogram of the optimal hop-wise candidate discovery duration $\hat{t}_h^*$ for minimizing hop-wise latency.

Table 4.4: Comparison of Global and Distributed Routing Algorithm

	Global	Distributed	Gain
$\alpha = 0.5$	0.4269	0.4469	4.7%
$\alpha = 1$	0.8688	0.8938	2.9%

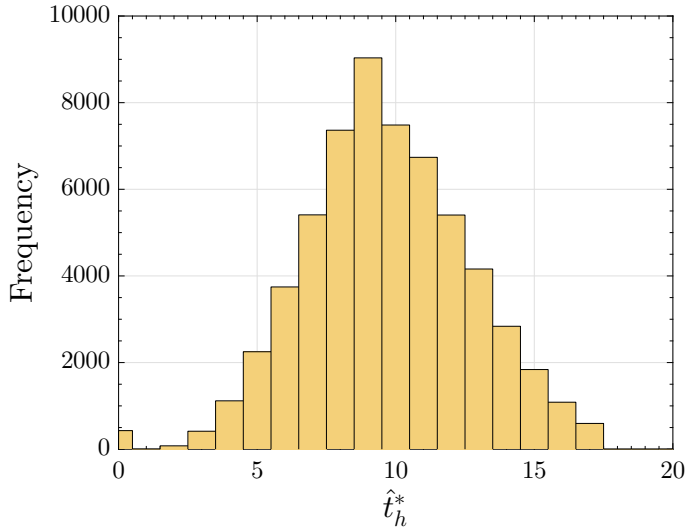


Figure 4.14: Illustration of the histogram of the optimal hop-wise candidate discovery duration $\hat{t}_h^*$ for maximizing hop-wise data rate.

Fig. 4.13 and Fig. 4.14 plot the histogram of the optimal hop-wise candidate discovery duration $\hat{t}_h^*$ for the distributed routing algorithm obtained by (4.60) across different snapshots for minimum E2E latency and maximum E2E data rate, respectively. We notice that both histogram figures yield distributions of optimal $\hat{t}_h^*$ similar to the curve in Fig. 4.6. Specifically, the most counts of the optimal $\hat{t}_h^*$ for both minimum latency and maximum data rate emerge at $\hat{t}_h^* = 9$, which is in consistency with the observation from Fig. 4.6 where $t^* = 9$ is the optimal value for minimizing latency and maximizing data rate with the global routing algorithm.

4.6 Summary

In this chapter, we investigated the data delivery problem in V2X networks. We proposed an analytical framework to derive mathematical expressions of expected delivery latency and data rate. Based on this, optimization problems for both global and distributed data delivery have been formulated to maximize the weighted sum of E2E latency and data rate and the weighted sum of hop-wise latency and data rare, respectively. Leveraging the reformation of closed-form expressions of the ex-

pected latency and data rate, the convexity of the proposed optimization problems are verified and correspondingly, the optimal solutions that maximize the weighted sum of both E2E and hop-wise latency and data rate are proposed. Afterwards, both global and distributed multihop routing algorithms are developed for solving the global and distributed optimization problem by obtaining the optimal global and hop-wise candidate discovery duration to select the optimal route, respectively. Finally, the extensive system-level simulations are conducted to evaluate the performance of the proposed algorithms by considering different classical vehicular routing algorithms, various beamforming configurations and vehicular arrival rates, and the support of wireless backhauling. It is demonstrated that the weighted sum is maximized by obtaining the optimal candidate discovery duration that varies with desired optimization objectives (latency and/or data rate). This implies that a larger candidate discovery duration is recommended for reducing the latency. If both latency and data rate are to be considered, a relatively smaller value of candidate discovery duration provides the flexibility to achieve a trade-off between latency and data rate. In other words, by configuring the candidate discovery duration, the optimization of different metrics can be fulfilled. Furthermore, the proposed routing algorithms provide considerable improvements over the classical vehicular routing algorithms in the sense of maximizing the data rate while minimizing the latency. Finally, the selection of different broadcast schemes and different vehicle arrival rates, and the availability of wireless backhauling, are demonstrated to influence the performance of latency and data rate, and the distributed multihop routing algorithm is shown to be able to provide lower latency and higher data rate compared to the global algorithm.

4.7 Appendix

4.7.1 Proof of Lemma 4.4.1

The CDF of the maximum Y can be calculated as:

$$\begin{aligned} P(Y \leq x) &= \prod_{i=1}^{n} P(X_i \leq x) = P(X \leq x)^n \\ &= \big(1 - (1-p)^x\big)^n. \end{aligned} \tag{4.61}$$

Hence, the PMF of Y can be derived as

$$\begin{aligned} f_{\mathrm{Y}}(x) &= P(Y \leq x) - P\big(Y \leq (x-1)\big) \\ &= \big(1 - (1-p)^x\big)^n - \big(1 - (1-p)^{x-1}\big)^n. \end{aligned} \tag{4.62}$$

4.7.2 Proof of Theorem 4.4.1

The expected E2E data rate $E(C_{\text{All success}})$ can be derived as

$$E(C_{\text{All success}}) = E(\min_{\forall h} C_{\text{Success},h}) = \frac{r_{\text{V2V}}\Big(T - E\Big(\max_{\forall h} \tau_{h,h+1}^{(d,\text{Veh})}\Big)\Big) + r_{\text{O}}(T-t)}{T}, \tag{4.63}$$

where $\tau_{h,h+1}^{(d,\text{Veh})} = m_h \Delta t$ for $m_h \in \{1, \ldots, m\}$. As Δt remains constant, finding $\max_{\forall h} \tau_{h,h+1}^{(d)}$ turns further to find $\xi = \max_{\forall h} m_h$.

Given Lemma 4.4.1, the PMF of ξ, denoted as $f(\xi)$ for $\xi \in \{1, \ldots, m\}$, is calculated as

$$f(\xi) = \big(1 - (1-p)^{\xi}\big)^k - \big(1 - (1-p)^{\xi-1}\big)^k. \tag{4.64}$$

Then, $\forall h \in \{1, \ldots, k\}$, we have

$$E\big(\max_{\forall h} \tau_{h,h+1}^{(d,\text{Veh})}\big) = \sum_{\xi=1}^{m} \xi \Delta t \cdot f(\xi) = \sum_{\xi=1}^{m} \xi \Big(\big(1 - (1-p)^{\xi}\big)^k - \big(1 - (1-p)^{\xi-1}\big)^k\Big) \cdot \Delta t. \tag{4.65}$$

4.7.3 Proof of Lemma 4.4.2

The CDF of the maximum Z can be calculated as:

$$p(Z \leq x) = \prod_{i=1}^{n} p(X_i \leq x) = \prod_{i=1}^{n} \left(1 - e^{-\lambda_i x}\right). \tag{4.66}$$

Hence, the PDF of Z can be derived as

$$f_{\mathrm{Z}}(x) = \frac{\mathrm{d}}{\mathrm{d}x} p(Z \leq x) = \frac{\mathrm{d}}{\mathrm{d}x} \prod_{i=1}^{n} \left(1 - e^{-\lambda_i x}\right) = \sum_{i=1}^{n} \left(\lambda_i e^{-\lambda_i x} \prod_{j=1, j\neq i}^{n} \left(1 - e^{-\lambda_j x}\right) \right). \tag{4.67}$$

4.7.4 Proof of Theorem 4.4.2

The expected E2E data rate $E(C_{\text{All faliure}})$ can be derived as

$$E(C_{\text{All failure}}) = E(\min_{\forall h} C_{\text{Failure},h}) = \frac{r_{\text{V2I}}(T-t) + r_{\text{O}} t}{2T + E\left(\max_{\forall h} \tau_{h,h+1}^{(d,\,\text{RSU})} \right)}. \tag{4.68}$$

Given Lemma 4.4.2, the PDF of η, denoted as $f(\eta)$ where $\eta \in [0, \infty)$ and $\forall h \in \{1, \ldots, k\}$, can be written as

$$f(\eta) = \sum_{h=1}^{k} \left(\mu_h e^{-\mu_h \eta} \prod_{l=1, l\neq h}^{k} \left(1 - e^{-\mu_l \eta}\right) \right). \tag{4.69}$$

Then, we have

$$\begin{aligned} E\big(\max_{\forall h} \tau_{h,h+1}^{(d,\,\text{RSU})} \big) &= E(\eta) \\ &= \int_0^\infty \eta f(\eta) \mathrm{d}\eta \\ &= \int_0^\infty \sum_{h=1}^{k} \left(\mu_h e^{-\mu_h \eta} \prod_{l=1, l\neq h}^{k} \left(1 - e^{-\mu_l \eta}\right) \right) \eta \mathrm{d}\eta \\ &= \sum_{h=1}^{k} \mu_h \int_0^\infty \eta \left(e^{-\mu_h \eta} \prod_{l=1, l\neq h}^{k} \left(1 - e^{-\mu_l \eta}\right) \right) \mathrm{d}\eta. \end{aligned} \tag{4.70}$$

For each μ_h in (4.70), the integral can be expanded, solved, and evaluated independently. Then, by taking the value of the integrals for different μ_h back to (4.70), we can get the result of $E(\max_{\forall h} \tau_{h,h+1}^{(d,\,\text{RSU})})$ and correspondingly the expected E2E data

rate $E\big(C_{\text{All faliure}}\big)$. As the mathematical expansion of (4.70) is very complicated, here we take $k = 2$ and $k = 3$ as two examples to show the form of integral summations. For a general value of k, similar forms of the expressions can be derived.

- $k = 2$:

$$E\big(\max_{\forall h} \tau_{h,h+1}^{(d,\,\text{RSU})}\big) = \frac{1}{\mu_1} + \frac{1}{\mu_2} - \frac{1}{\mu_1 + \mu_2}. \tag{4.71}$$

- $k = 3$:

$$\begin{aligned} E\big(\max_{\forall h} \tau_{h,h+1}^{(d,\,\text{RSU})}\big) = &\frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_3} \\ &- \frac{1}{\mu_1 + \mu_2} - \frac{1}{\mu_1 + \mu_3} - \frac{1}{\mu_2 + \mu_2} \\ &- \frac{1}{\mu_1 + \mu_2 + \mu_3}. \end{aligned} \tag{4.72}$$

4.7.5 Proof of Lemma 4.4.3

As

$$1 - F_{\text{X}}(x) = P(\text{X} \geq x) = \int_x^{\infty} f_{\text{X}}(y)\text{d}y \tag{4.73}$$

where $f_{\text{X}}(y)$ represents the PDF of X, then by taking integral on both sides of the equation, we have

$$\begin{aligned} \int_0^{\infty} \big(1 - F_{\text{X}}(x)\big)\text{d}x &= \int_0^{\infty} P(\text{X} \geq x)\text{d}x = \int_0^{\infty}\int_x^{\infty} f_{\text{X}}(y)\text{d}y\text{d}x \\ &= \int_0^{\infty}\int_0^{y} f_{\text{X}}(y)\text{d}x\text{d}y = \int_0^{\infty} y f_{\text{X}}(y)\text{d}y = \int_0^{\infty} x f_{\text{X}}(x)\text{d}x \\ &= E(\text{X}). \end{aligned} \tag{4.74}$$

4.7.6 Proof of Convexity of Data Delivery Optimization Problem

To prove the convexity of the global data delivery optimization problem, we incorporate the conclusions addressed in [190]:

1. Exponential function is convex: $e^{ax}, \forall a \in \mathbb{R}$.
2. Power function is convex: x^a, $\forall a \geq 1$.
3. Weighted sum of convex/concave functions is convex/concave: $\sum_{i=1}^{n} \omega_i f_i$.
4. Point-wise maximum/minimum of convex/concave functions is convex/concave: $\max_{\forall i} f_i$.

Based on these, we transform the expected E2E latency derived in (4.27) as follows:

$$\begin{aligned}\bar{L} &= kT + \sum_{h=1}^{k}(1-\alpha_h)\phi_h\big(\beta_h(t) + \theta_h(t) - \beta_h(t)\theta_h(t)big\big) \\ &= A + \sum_{h=1}^{k} B\big(\beta_h(t) + \theta_h(t) - \beta_h(t)\theta_h(t)\big). \end{aligned} \tag{4.75}$$

Here,

$$A = kT, \tag{4.76}$$

and

$$B = (1-\alpha_h)\phi_h, \tag{4.77}$$

can be treated as constant without the term t. $\beta_h(t)$ and $\theta_h(t)$ are the functions of t and convex according to 1). Then, the expected E2E latency $\bar{L}$, calculated as the weighted sum of convex function $\beta_h(t)$, $\theta_h(t)$ and product $\beta_h(t)\theta_h(t)$, is convex as stated in 3).

For the expected E2E data rate, similar demonstration of convexity can be deducted. Firstly, note that $\bar{C}$ in (4.17) can be reformed as

$$\begin{aligned}\bar{C} &= \min_{\forall h} E(C_{h,h+1}) \\ &= \min_{\forall h} \Big(\zeta_h + (1-\alpha_h)\big(1-\beta_h(t)\big)\big(1-\theta_h(t)\big) \cdot \big(\iota_h + \nu_h(t)\big) \\ &\quad + \Big((1-\alpha_h)\big(1-\beta_h(t)\big)\theta_h(t) + (1-\alpha_h)\beta_h(t)\Big) \cdot \big(\kappa_h + \chi_h(t)\big) \Big). \end{aligned} \tag{4.78}$$

Let

$$z(t) = \beta_h(t) + \theta_h(t) - \beta_h(t)\theta_h(t), \tag{4.79}$$

then $\bar{C}$ is further simplified as

$$\begin{aligned}\bar{C} &= \min_{\forall h} E(C_{h,h+1}) \\ &= \min_{\forall h} \Big(\zeta_h + (1-\alpha_h)\big(1-z(t)\big)\big(\iota_h + \nu_h(t)\big) + (1-\alpha_h) z(t)\big(\kappa_h + \chi_h(t)\big) \Big).\end{aligned} \tag{4.80}$$

By taking the second-order derivative, it can be shown that both $\big(1-z(t)\big)\big(\iota_h+\nu_h(t)\big)$ and $z(t)\big(\kappa_h + \chi_h(t)\big)$ are concave. Hence, the weighted sum of these two concave functions and the constant ζ_h is also concave due to 3), and the min operator $\min_{\forall h}(.)$ preserves the concavity of the weighted sum according to 4).

Finally, $\alpha\bar{C} - (1-\alpha)\bar{L}$, represented by the weighted sum of concave function $-\bar{L}$ and concave function $\bar{C}$, is a concave function in line with 3). In this way, the convexity of the optimization problem 4.1 is verified, and the weighted sum of the E2E latency and the E2E data rate is maximized by determining the optimal t^*.

4.7.7 Proof of Theorem 4.4.3

The optimization problem 4.1 can be refined as

$$\max_{t\in\mathbb{R}} \alpha\bar{C} - (1-\alpha)\bar{L} \quad \text{s.t.} \quad t \geq 0,\ t \leq T. \tag{4.81}$$

Observe that all inequality constraints functions are affine, and there exists a value of t such that $t = \frac{T}{2} \geq 0$ and $t = \frac{T}{2} \leq T$, which implies the KKT conditions are necessary. Moreover, the objective function is demonstrated as convex in Appendix 4.7.6. Therefore, the KKT conditions are also sufficient, in which we can use standard KKT form to solve the problem, similarly as in Section 3.7.4, as follows:

- PF

$$t \geq 0,\ t \leq T. \tag{4.82}$$

- DF

$$-\frac{\mathrm{d}(\alpha\bar{C}-(1-\alpha)\bar{L})}{\mathrm{d}t}-\psi+\omega=0,\ \psi\geq 0,\ \omega\geq 0. \tag{4.83}$$

- CS

$$\psi t=0,\ \omega t=0. \tag{4.84}$$

As the value of t is restricted in the set $[0,T]$, we can always take any t without violate PF conditions in case $t>0$ and $t<T$. Therefore, both CS conditions are always satisfied and there is no further restriction on the Lagrangian multiplier ψ and ω besides $\psi\geq 0$ and $\omega\geq 0$ from DF conditions. Actually, CS conditions also indicate that $\psi=0$ and $\omega=0$ for $t\in(0,T]$, where in order to satisfy DF conditions, $\frac{\mathrm{d}(\alpha\bar{C}-(1-\alpha)\bar{L})}{\mathrm{d}t}|_{t=t_s}$ must be zero, which leads to $t_s\in[0,T]$.

In conclusion, the optimal solution of the problem in (4.18) is given by

$$t^*=\underset{t=\{0,t_s,T\}}{\arg\max}\,\alpha\bar{C}-(1-\alpha)\bar{L}, t_s\in[0,T]. \tag{4.85}$$

Chapter 5

Conclusions

This dissertation has proposed several enabling technologies to address the challenges of the data delivery in 5G mm-wave mobile communication networks. The first contribution of this dissertation is the development of several new methodologies for the cell discovery analysis of mm-wave communication networks, including various broadcast schemes for delivering cell discovery information to users in the networks, and a new tractable analytical framework for investigating the performance of the mm-wave beamformed cell discovery and providing insights into realistic system design and 5G standardization activities. The second contribution, leveraging the user association achieved by the cell discovery, is the study of a new joint scheduling and resource allocation scheme for mm-wave HetNets, where heuristic scheduling, resource allocation, and routing algorithms are proposed as the solutions of a cross-layer optimization problem in terms of maximizing data rate. The third contribution, based on the optimal scheduling and resource allocation results, consists of a new tractable analytical framework that mathematically characterizes the data delivery performance in mobile communication networks, with a special focus on V2X networks, and two multihop routing algorithms that optimize the characterized data delivery performance by solving the formulated weighted sum maximization problems. In the rest of this chapter, we summarize the system design insights obtained from each contribution and discuss the prospective future works.

5.1 Summary

In Chapter 2, we have proposed an analytical framework to investigate the performance of mm-wave beamformed cell discovery. Specifically, several representative broadcast schemes are compared in terms of cell discovery latency and overhead. It is demonstrated that the discovery latency is optimized when the thinnest beam is formed. In addition, the beamforming architecture has no impact on the latency, which makes the performance of analog/hybrid beamforming the same as digital beamforming in terms of the latency. By contrast, a thinner beam results in a higher overhead. Moreover, multiple beam simultaneous scan leads to a latency penalty on the cell discovery but achieves a lower overhead compared to single beam exhaustive scan, and the best trade-off between the latency and overhead can be achieved by the spatial-division multiple beam scan. This trade-off can be further extended to frame structure design, where different frame structures result in a lower latency at the price of a higher overhead or vice versa. In particular, separating a beacon phase for broadcasting cell discovery signals into as many frames as possible is recommended for overhead-sensitive services, while keeping the beacon phase in each frame as long as possible under a certain overhead limit is suggested for low-latency applications. This can be potentially incorporated in future releases of 3GPP 5G NR for a flexible beacon adaptation to fulfill various performance requirements of different use cases. In the end, it has been demonstrated that the latency and overhead are relatively insensitive to extremely low block error rates.

In Chapter 3, we have addressed the problem of maximizing the achievable data rate of mm-wave HetNets considering both downlink and uplink transmissions on backhaul and access links. To solve the maximization problem in an efficient way, we have proposed a joint scheduling and resource allocation algorithm, which consists of the maximum independent set based scheduling algorithm, the proportional fair slot allocation algorithm, and the water-filling power allocation algorithm. In addition, a dynamic routing algorithm incorporating a path selection algorithm is investigated to further improve the data rate achieved by the joint scheduling and

resource allocation algorithm with predefined static routes. It is demonstrated that the proposed joint scheduling and resource allocation algorithm, with and without the dynamic routing algorithm, outperform the benchmark schemes, in terms of achievable data rate and closely approach the theoretical optimum, yet with lower latency. Besides, the proposed algorithms enable the flexible adjustment of downlink and uplink slot allocation and support both half- and full-duplex modes with considerable performance enhancements. In particular, the proposed algorithms are capable of fulfilling different performance requirements for not only point-to-multipoint communications, but also point-to-point communications with the special focus on data delivery in a vehicle platoon that is currently studied by 5G standardization associations and research activities.

In Chapter 3, we have extended the study of data delivery into V2X networks, where an analytical framework is proposed to mathematically characterize the delivery latency and data rate, which are two of the key performance indicators considered in ongoing 5G V2X standardization activities. Based on this, optimization problems for both global and distributed data delivery are formulated to minimize the latency while maximizing the data rate. The convexity of the proposed optimization problems is then verified with the help of closed-form expressions of the latency and data rate, and the optimization problems are solved by convex optimization theory, followed by corresponding multihop routing algorithms inspired by the proposed solutions. It is demonstrated that the latency and data rate are able to be optimized by selecting different values of the optimal duration, which is a key enabler utilized in the framework to optimize the data delivery performance, and the proposed routing algorithms provide considerable improvements over classical vehicular routing algorithms in the sense of minimizing the latency while maximizing the data rate. Furthermore, the selection of different broadcast schemes and vehicle arrival rates, and the availability of wireless backhauling, are also demonstrated to influence the data delivery performance.

5.2 Future Directions

The cell discovery analysis proposed in Chapter 2 only allows users to receive cell discovery signals by quasi-omnidirectional antennas, or assumes a certain correlation of transmission and reception beam scan where either transmission beam scan along all angular directions can be finished within one reception beam duration or the other way around. Therefore, future work can leverage the proposed analytical framework to investigate the cell discovery performance considering general beam scan approaches without specific correlations of angular speed at both AP and UE side. Also, the framework proposed in Chapter 2 is limited to the analysis of cell discovery, where the subsequent phases, such as carrier sensing and random access, have not been addressed in the dissertation. Motivated by this observation, we may further extend the proposed framework to the complete initial access procedure to investigate the performance of mm-wave initial access for the considered networks. It would also be interesting to include different system models, e.g., to address other types of spatial locations, blockage models, and/or fading channels. In addition, studies on pre-coding for the inter-beam interference cancellation/coordination of the multiple simultaneous beam scan could be also included.

In the HetNet considered in Chapter 3, the system performance in terms of data rate is addressed by the joint scheduling and resource allocation optimization. A multihop dynamic routing algorithm has been proposed after determining the scheduling and resource allocation policies to further enhance the achievable data rate of the considered network. This may limit the system performance, as the scheduling, resource allocation, and routing policies are individually investigated, which does not fully exploit the global optimality that may be brought from a joint optimization. Therefore, future work can jointly optimize the link scheduling, resource allocation, and routing strategy to achieve an even better overall system performance compared to the current solutions, while a potential increase in computational complexity should be also considered. In addition, the solutions proposed in Chapter 3 rely on a fixed user association that only allows users to connect

with one mm-wave BS. By enabling users to associate with multiple mm-wave BSs (i.e., multi-point connectivity), better robustness to interference and more flexible re-association/handover can be achieved to further boost the system performance.

For the work on the data delivery in multihop V2X networks in Chapter 4, there are several open issues left behind. First, currently only one type of vehicle is considered. Future work can incorporate more types of vehicles with different mobility patterns to investigate the efficiency of the proposed delivery strategy and routing algorithms in achieving competitive latency and data rate performances. Second, we build the analytical framework based on a probability distribution approach with the derivation of expectation values, while more realistic statistical analysis driven by ground-truth data, such as machine learning methods for the classification of vehicle mobility patterns, should be taken into consideration to check the feasibility of the proposed strategies and algorithms in more accurate radio and network models with relaxed assumptions. It would also be interesting to include other system models, such as highway, or extend the framework by investigating other data delivery performance metrics, e.g., delivery ratio and throughput.

References

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, Feb. 2016.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[3] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2016–2021," https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf, Sept. 2017.

[4] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5G era," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 90–96, Feb. 2014.

[5] METIS, "Deliverable D8.4: METIS final project report," https://www.metis2020.com/wp-content/uploads/deliverables/METIS_D8.4_v1.pdf, Apr. 2015.

[6] 5GPPP mmMAGIC, "Deliverable D6.6: Final mmMAGIC system concept," https://bscw.5g-mmmagic.eu/pub/bscw.cgi/d215278/mmMAGIC_D6.6.pdf, July 2017.

[7] 3GPP, "NR; Overall description; Stage-2," TS 38.300, Oct. 2018.

[8] Y. Li, "Over-the-air backhaul for dense mobile communication networks operating at 70 GHz," Master's thesis, Technische Universität München, Oct. 2014.

[9] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[10] T. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. Wong, J. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[11] Y. Li, E. Pateromichelakis, N. Vucic, J. Luo, W. Xu, and G. Caire, "Radio resource management considerations for 5G millimeter wave backhaul and access networks," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 86–92, June 2017.

[12] A. Alejos, M. Sanchez, and I. Cuinas, "Measurement and analysis of propagation mechanisms at 40 GHz: Viability of site shielding forced by obstacles," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3369–3380, Nov. 2008.

[13] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.

[14] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[15] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam

management for 3GPP NR at mmWave frequencies," *IEEE Communications Surveys Tutorials*, pp. 1–1, Sept. 2018.

[16] Y. Li, J. Luo, M. H. Castaneda, N. Vucic, W. Xu, and G. Caire, "Analysis of broadcast signaling for millimeter wave cell discovery," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Sept. 2017, pp. 1–5.

[17] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "Initial access frameworks for 3GPP NR at mmWave frequencies," in *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, June 2018, pp. 1–8.

[18] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, June 2011.

[19] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Network*, vol. 29, no. 2, pp. 6–14, Mar. 2015.

[20] Z. Gao, L. Dai, D. Mi, Z. Wang, M. A. Imran, and M. Z. Shakir, "MmWave massive-MIMO-based wireless backhaul for the 5G ultra-dense network," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 13–21, Oct. 2015.

[21] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfrojd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 44–51, May 2014.

[22] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "How many hops can self-backhauled millimeter wave cellular networks support?" *arXiv preprint arXiv:1805.01040*, May 2018.

[23] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2196–2211, Oct. 2015.

[24] J. García-Rois, F. Gómez-Cuba, M. R. Akdeniz, F. J. González-Castaño, J. C. Burguillo, S. Rangan, and B. Lorenzo, "On the analysis of scheduling in dynamic duplex multihop mmWave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6028–6042, Nov. 2015.

[25] J. Du, E. Onaran, D. Chizhik, S. Venkatesan, and R. A. Valenzuela, "Gbps user rates using mmWave relayed backhaul with high-gain antennas," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1363–1372, June 2017.

[26] M. N. Kulkarni, J. G. Andrews, and A. Ghosh, "Performance of dynamic and static TDD in self-backhauled millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6460–6478, Oct. 2017.

[27] E. Guttman, "5G New Radio and system standardization in 3GPP," https://www.itu.int/en/ITU-T/Workshops-and-Seminars/201707/Documents/Eric_Guttman_5G%20New%20Radio%20and%20System%20Standardization%20in%203GPP.pdf, Dec. 2017.

[28] Y. Li, J. Luo, M. H. C. Garcia, R. Böhnke, R. A. Stirling-Gallacher, W. Xu, and G. Caire, "On the beamformed broadcasting for millimeter wave cell discovery: Performance analysis and design insight," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7620–7634, Nov. 2018.

[29] Y. Li, J. Luo, W. Xu, N. Vucic, E. Pateromichelakis, and G. Caire, "A joint scheduling and resource allocation scheme for millimeter wave heterogeneous networks," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2017, pp. 1–6.

[30] Y. Li, J. Luo, Z. Li, R. A. Stirling-Gallacher, W. Xu, and G. Caire, "Multihop routing in hybrid vehicle-to-vehicle and vehicle-to-infrastructure networks," in *2018 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.

[31] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *2007 IEEE 65th Vehicular Technology Conference (VTC-Spring)*, Apr. 2007, pp. 1234–1238.

[32] S. Rangan, T. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[33] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, June 2011.

[34] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[35] W. Roh, J. Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, Feb. 2014.

[36] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152–1163, June 2014.

[37] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[38] H. Jung, Q. Li, and P. Zong, "Cell detection in high frequency band small cell networks," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Nov. 2015, pp. 1762–1766.

[39] M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5G mmWave cellular networks," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 40–47, Nov. 2016.

[40] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, "Initial access in millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7926–7940, Dec. 2016.

[41] 3GPP, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); overall description," TS 36.300, Oct. 2018.

[42] ——, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); medium access control (MAC) protocol specification," TS 36.321, Oct. 2018.

[43] ——, "Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (E-UTRAN); radio resource control (RRC) protocol specification," TS 36.331, Sept. 2018.

[44] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, S. S. Panwar, and T. S. Rappaport, "Directional cell discovery in millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 6664–6678, Dec. 2015.

[45] Q. C. Li, H. Niu, G. Wu, and R. Q. Hu, "Anchor-booster based heterogeneous networks with mmWave capable booster cells," in *2013 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 93–98.

[46] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 136–143, Dec. 2014.

[47] 3GPP, "NR; Physical layer; General description," TS 38.201, Jan. 2018.

[48] ——, "Study on New Radio access technology," TS 38.912, July 2018.

[49] IEEE Standard 802.11ad, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," Dec. 2012.

[50] T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer, "IEEE 802.11ad: Directional 60 GHz communication for multi-Gigabit-per-second Wi-Fi [invited paper]," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 132–141, Dec. 2014.

[51] IEEE Standard 802.15.3c, "Part 15.3: Millimeter-wave-based alternative physical layer extension," Oct. 2009.

[52] H. Shokri-Ghadikolaei, L. Gkatzikis, and C. Fischione, "Beam-searching and transmission scheduling in millimeter wave communications," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 1292–1297.

[53] B. Li, Z. Zhou, W. Zou, X. Sun, and G. Du, "On the efficient beam-forming training for 60 GHz wireless personal area networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 504–515, Feb. 2013.

[54] C. Jeong, J. Park, and H. Yu, "Random access in millimeter-wave beamforming cellular networks: Issues and approaches," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 180–185, Jan. 2015.

[55] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, Oct. 2015.

[56] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmWave communications," in *2014 48th Asilomar Conference on Signals, Systems and Computers*, Nov. 2014, pp. 1926–1930.

[57] V. Raghavan, J. Cezanne, S. Subramanian, A. Sampath, and O. Koymen, "Beamforming tradeoffs for initial UE discovery in millimeter-wave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 543–559, Apr. 2016.

[58] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, "Comparative analysis of initial access techniques in 5G mmWave cellular networks," in *2016 Annual Conference on Information Science and Systems (CISS)*, Mar. 2016, pp. 268–273.

[59] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.

[60] X. Song, S. Haghighatshoar, and G. Caire, "A scalable and statistically robust beam alignment technique for millimeter-wave systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4792–4805, July 2018.

[61] Y. Li, J. G. Andrews, F. Baccelli, T. D. Novlan, and C. J. Zhang, "Design and analysis of initial access in millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6409–6425, Oct. 2017.

[62] M. Hussain and N. Michelusi, "Throughput optimal beam alignment in millimeter wave networks," in *2017 Information Theory and Applications Workshop (ITA)*, Feb. 2017, pp. 1–6.

[63] D. D. Donno, J. P. Beltrán, D. Giustiniano, and J. Widmer, "Hybrid analog-digital beam training for mmWave systems with low-resolution RF phase

shifters," in *2016 IEEE International Conference on Communications Workshops (ICC)*, May 2016, pp. 700–705.

[64] I. Filippini, V. Sciancalepore, F. Devoti, and A. Capone, "Fast cell discovery in mm-wave 5G networks with context information," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1538–1552, July 2018.

[65] A. Capone, I. Filippini, and V. Sciancalepore, "Context information for fast cell discovery in mm-wave 5G networks," in *2015 European Wireless Conference*, May 2015, pp. 1–6.

[66] A. S. Marcano and H. L. Christiansen, "Macro cell assisted cell discovery method for 5G mobile networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC-Spring)*, May 2016, pp. 1–5.

[67] W. B. Abbas and M. Zorzi, "Context information based initial cell search for millimeter wave 5G cellular networks," in *2016 European Conference on Networks and Communications (EuCNC)*, June 2016, pp. 111–116.

[68] A. Alkhateeb, Y. H. Nam, M. S. Rahman, J. Zhang, and R. W. Heath, "Initial beam association in millimeter wave cellular systems: Analysis and design insights," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2807–2821, May 2017.

[69] S. Akoum, O. E. Ayach, and R. W. Heath, "Coverage and capacity in mmWave cellular systems," in *2012 46th Asilomar Conference on Signals, Systems and Computers*, Nov. 2012, pp. 688–692.

[70] T. Bai, A. Alkhateeb, and R. W. Heath, "Coverage and capacity of millimeter-wave cellular networks," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 70–77, Sept. 2014.

[71] M. D. Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5038–5057, Sept. 2015.

[72] A. Guo and M. Haenggi, "Asymptotic deployment gain: A simple approach to characterize the SINR distribution in general cellular networks," *IEEE Transactions on Communications*, vol. 63, no. 3, pp. 962–976, Mar. 2015.

[73] B. Błaszczyszyn, M. K. Karray, and H. P. Keeler, "Wireless networks appear Poissonian due to strong shadowing," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4379–4390, Aug. 2015.

[74] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: Mm-Wave beam steering without in-band measurement," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2015, pp. 2416–2424.

[75] 3GPP, "Study on 3D channel model for LTE," TS 36.873, Jan. 2018.

[76] ——, "Study on channel model for frequencies from 0.5 to 100 GHz," TS 38.901, June 2018.

[77] M. K. Samimi, T. S. Rappaport, and G. R. MacCartney, "Probabilistic omnidirectional path loss models for millimeter-wave outdoor communications," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 357–360, Aug. 2015.

[78] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Transactions on Communications*, vol. 63, no. 9, pp. 3029–3056, Sept. 2015.

[79] S. Sun, T. A. Thomas, T. S. Rappaport, H. Nguyen, I. Z. Kovacs, and I. Rodriguez, "Path loss, shadow fading, and line-of-sight probability models for 5G urban macro-cellular scenarios," in *2015 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–7.

[80] J. Park, S. L. Kim, and J. Zander, "Tractable resource management with uplink decoupled millimeter-wave overlay in ultra-dense cellular networks,"

IEEE Transactions on Wireless Communications, vol. 15, no. 6, pp. 4362–4379, June 2016.

[81] A. K. Gupta, J. G. Andrews, and R. W. Heath, "On the feasibility of sharing spectrum licenses in mmWave cellular systems," *IEEE Transactions on Communications*, vol. 64, no. 9, pp. 3981–3995, Sept. 2016.

[82] J. D. Kraus, *Antennas*. McGraw-Hill Education, 1988.

[83] J. Luo, N. Vucic, M. Castaneda, Y. Li, and W. Xu, "Initial access assisted by an auxiliary transceiver for millimeter-wave networks," in *2017 21th International ITG Workshop on Smart Antennas (WSA)*, Mar. 2017, pp. 1–6.

[84] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-advanced for mobile broadband*. Academic press, 2013.

[85] S. C. Yang, *3G CDMA2000: wireless system engineering*. Artech House, 2004.

[86] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[87] 5GPPP mmMAGIC, "Deliverable D4.2: Final radio interface concepts and evaluations for mm-wave mobile communications," https://bscw.5g-mmmagic.eu/pub/bscw.cgi/d214055/mmMAGIC_D4.2.pdf, June 2017.

[88] J. Vihriala, N. Cassiau, J. Luo, Y. Li, Y. Qi, T. Svensson, A. Zaidi, K. Pajukoski, and H. Miao, "Frame structure design for future millimetre wave mobile radio access," in *2016 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[89] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE*

Transactions on Wireless Communications, vol. 16, no. 3, pp. 1508–1522, Mar. 2017.

[90] A. J. Viterbi, *CDMA: principles of spread spectrum communication.* Addison Wesley Longman Publishing Co., Inc., 1995.

[91] A. Khandekar, N. Bhushan, J. Tingfang, and V. Vanghi, "LTE-Advanced: Heterogeneous networks," in *2010 European Wireless Conference*, Apr. 2010, pp. 978–982.

[92] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, July 2015.

[93] 3GPP, "NR; User Equipment (UE) radio transmission and reception," TS 38.101, Oct. 2018.

[94] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[95] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Bjoernson, K. Yang, C. I, and A. Ghosh, "Millimeter wave communications for future mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909–1935, Sept. 2017.

[96] S. Sun, T. S. Rappaport, R. W. Heath, A. Nix, and S. Rangan, "MIMO for millimeter-wave wireless communications: Beamforming, spatial multiplexing, or both?" *IEEE Communications Magazine*, vol. 52, no. 12, pp. 110–121, Dec. 2014.

[97] D. Yuan, H. Lin, J. Widmer, and M. Hollick, "Optimal joint routing and scheduling in millimeter-wave cellular networks," in *2018 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2018, pp. 1205–1213.

[98] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[99] E. Hossain and M. Hasan, "5G cellular: Key enabling technologies and research challenges," *IEEE Instrumentation Measurement Magazine*, vol. 18, no. 3, pp. 11–21, June 2015.

[100] E. Pateromichelakis, A. Maeder, A. D. Domenico, R. Fritzsche, P. de Kerret, and J. Bartelt, "Joint RAN/Backhaul optimization in centralized 5G RAN," in *2015 European Conference on Networks and Communications (EuCNC),*, June 2015, pp. 386–390.

[101] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Network*, vol. 28, no. 6, pp. 6–11, Nov. 2014.

[102] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[103] 3GPP, "NR; Study on integrated access and backhaul," TS 38.874, Nov. 2018.

[104] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[105] M. Polese, M. Giordani, A. Roy, D. Castor, and M. Zorzi, "Distributed path selection strategies for integrated access and backhaul at mmWaves," *arXiv preprint arXiv:1805.04351*, May 2018.

[106] U. Siddique, H. Tabassum, E. Hossain, and D. I. Kim, "Wireless backhauling of 5G small cells: Challenges and solution approaches," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 22–31, Oct. 2015.

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

[108] C. Dehos, J. L. González, A. D. Domenico, D. Kténas, and L. Dussopt, "Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5G mobile communications systems?" *IEEE Communications Magazine*, vol. 52, no. 9, pp. 88–95, Sept. 2014.

[109] B. Li, D. Zhu, and P. Liang, "Small cell in-band wireless backhaul in massive MIMO systems: A cooperation of next-generation techniques," *IEEE Transactions on Wireless Communications*, vol. 14, no. 12, pp. 7057–7069, Dec. 2015.

[110] H. Tabassum, A. H. Sakr, and E. Hossain, "Analysis of massive MIMO-enabled downlink wireless backhauling for full-duplex small cells," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2354–2369, June 2016.

[111] M. E. Rasekh, D. Guo, and U. Madhow, "Joint routing and resource allocation for millimeter wave picocellular backhaul," *arXiv preprint arXiv:1809.07470*, Sept. 2018.

[112] A. Sharma, R. K. Ganti, and J. K. Milleth, "Joint backhaul-access analysis of full duplex self-backhauling heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1727–1740, Mar. 2017.

[113] J. Qiao, L. Cai, X. Shen, and J. W. Mark, "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011.

[114] J. Qiao, L. Cai, and X. Shen, "Multi-hop concurrent transmission in millimeter wave WPANs with directional antenna," in *2010 IEEE International Conference on Communications (ICC)*, May 2010, pp. 1–5.

[115] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2052–2069, Oct. 2015.

[116] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, Y. Zhu, and D. O. Wu, "Energy-efficient scheduling for mmWave backhauling of small cells in heterogeneous cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2674–2687, Mar. 2017.

[117] R. Taori and A. Sridharan, "Point-to-multipoint in-band mmWave backhaul for 5G networks," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 195–201, Jan. 2015.

[118] J. Kim and A. F. Molisch, "Quality-aware millimeter-wave device-to-device multi-hop routing for 5G cellular networks," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 5251–5256.

[119] X. Yi and G. Caire, "Optimality of treating interference as noise: A combinatorial perspective," *IEEE Transactions on Information Theory*, vol. 62, no. 8, pp. 4654–4673, Aug. 2016.

[120] W. Xia, J. Zhang, S. Jin, C. Wen, F. Gao, and H. Zhu, "Large system analysis of resource allocation in heterogeneous networks with wireless backhaul," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 5040–5053, Nov. 2017.

[121] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3251–3268, May 2016.

[122] W. Hao and S. Yang, "Small cell cluster-based resource allocation for wireless backhaul in two-tier heterogeneous networks with massive MIMO," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 509–523, Jan. 2018.

[123] Y. Liu, X. Fang, and M. Xiao, "Discrete power control and transmission duration allocation for self-backhauling dense mmWave cellular networks," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 432–447, Jan. 2018.

[124] S. Goyal, P. Liu, and S. Panwar, "Scheduling and power allocation in self-backhauled full duplex small cells," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–7.

[125] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 910–917, Sept. 1988.

[126] A. Zhou, M. Liu, Z. Li, and E. Dutkiewicz, "Cross-layer design for proportional delay differentiation and network utility maximization in multi-hop wireless networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1446–1455, Apr. 2012.

[127] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, Sept. 2008.

[128] M. Cudak, A. Ghosh, T. Kovarik, R. Ratasuk, T. Thomas, F. Vook, and P. Moorut, "Moving towards mmWave-based Beyond-4G (B-4G) technology," in *2013 IEEE 77th Vehicular Technology Conference (VTC-Spring)*, June 2013, pp. 1–5.

[129] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Wireless Networks*, vol. 21, no. 8, pp. 2657–2676, Nov. 2015.

[130] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 5221–5225.

[131] S. Han, C. I, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[132] A. L. Swindlehurst, E. Ayanoglu, P. Heydari, and F. Capolino, "Millimeter-wave massive MIMO: The next wireless revolution?" *IEEE Communications Magazine*, vol. 52, no. 9, pp. 56–62, Sept. 2014.

[133] H. Ju, B. Liang, J. Li, and X. Yang, "Dynamic power allocation for throughput utility maximization in interference-limited networks," *IEEE Wireless Communications Letters*, vol. 2, no. 1, pp. 22–25, Feb. 2013.

[134] J. Lee and S. Leyffer, *Mixed integer nonlinear programming.* Springer Science & Business Media, 2011, vol. 154.

[135] L. Caccetta and A. Kulanoot, "Computational aspects of hard knapsack problems," *Nonlinear Analysis: Theory, Methods, and Applications*, vol. 47, no. 8, pp. 5547–5558, Nov. 2001.

[136] D. Pisinger, "Where are the hard knapsack problems?" *Computers and Operations Research*, vol. 32, no. 9, pp. 2271–2284, Sept. 2005.

[137] D. B. West *et al.*, *Introduction to graph theory.* Prentice hall Upper Saddle River, 2001, vol. 2.

[138] R. M. Karp, "Reducibility among combinatorial problems," in *Complexity of computer computations.* Springer, 1972, pp. 85–103.

[139] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of combinatorial optimization.* Springer, 1999, pp. 1–74.

[140] M. M. Halldórsson and J. Radhakrishnan, "Greed is good: Approximating independent sets in sparse and bounded-degree graphs," *Algorithmica*, vol. 18, no. 1, pp. 145–163, 1997.

[141] D. S. Hochbaum, "Efficient bounds for the stable set, vertex cover and set packing problems," *Discrete Applied Mathematics*, vol. 6, no. 3, pp. 243–254, 1983.

[142] G. L. Nemhauser and L. E. Trotter Jr, "Vertex packings: Structural properties and algorithms," *Mathematical Programming*, vol. 8, no. 1, pp. 232–248, 1975.

[143] R. G. Gallager, P. A. Humblet, and P. M. Spira, "A distributed algorithm for minimum-weight spanning trees," *ACM Transactions on Programming Languages and systems (TOPLAS)*, vol. 5, no. 1, pp. 66–77, 1983.

[144] D. B. Johnson, "A note on dijkstra's shortest path algorithm," *Journal of the ACM (JACM)*, vol. 20, no. 3, pp. 385–388, 1973.

[145] Z. Pi and F. Khan, "System design and network architecture for a millimeter-wave mobile broadband (MMB) system," in *34th IEEE Sarnoff Symposium*, May 2011, pp. 1–6.

[146] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Q. S. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 22–30, June 2011.

[147] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[148] L. Georgiadis, M. J. Neely, L. Tassiulas *et al.*, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[149] 3GPP, "Study on LTE-based V2X services," TS 36.885, July 2016.

[150] ——, "Study on evaluation methodology of new V2X use cases for LTE and NR," TS 37.885, Oct. 2018.

[151] H. U. Simon, "On approximate solutions for combinatorial optimization problems," *SIAM Journal on Discrete Mathematics*, vol. 3, no. 2, pp. 294–310, 1990.

[152] P. Turán, "On an external problem in graph theory," *Mat. Fiz. Lapok*, vol. 48, pp. 436–452, 1941.

[153] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, and T. Weil, "Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions," *IEEE Communications Surveys Tutorials*, vol. 13, no. 4, pp. 584–616, July 2011.

[154] W. Liang, Z. Li, H. Zhang, S. Wang, and R. Bie, "Vehicular ad hoc networks: Architectures, research issues, methodologies, challenges, and trends," *International Journal of Distributed Sensor Networks*, vol. 11, no. 8, p. 745303, Aug. 2015.

[155] S. Zeadally, R. Hunt, Y.-S. Chen, A. Irwin, and A. Hassan, "Vehicular ad hoc networks (VANETS): Status, results, and challenges," *Telecommunication Systems*, vol. 50, no. 4, pp. 217–241, Aug. 2012.

[156] S. K. Bhoi and P. M. Khilar, "Vehicular communication: A survey," *IET Networks*, vol. 3, no. 3, pp. 204–217, Aug. 2013.

[157] B. T. Sharef, R. A. Alsaqour, and M. Ismail, "Vehicular communication ad hoc routing protocols: A survey," *Journal of network and computer applications*, vol. 40, pp. 363–396, Apr. 2014.

[158] S. Al-Sultan, M. M. Al-Doori, A. H. Al-Bayatti, and H. Zedan, "A comprehensive survey on vehicular ad hoc network," *Journal of network and computer applications*, vol. 37, pp. 380–392, Jan. 2014.

[159] R. Atallah, M. Khabbaz, and C. Assi, "Multihop V2I communications: A feasibility study, modeling, and performance analysis," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2801–2810, Mar. 2017.

[160] R. Hussain and S. Zeadally, "Autonomous cars: Research results, issues and future challenges," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2018.

[161] IEEE Standard 802.11p, "Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications amendment 6: Wireless access in vehicular environments," pp. 1–51, July 2010.

[162] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1162–1182, July 2011.

[163] A. M. Vegni and T. D. Little, "Hybrid vehicular communications based on V2V-V2I protocol switching," *International Journal of Vehicle Information and Communication Systems*, vol. 2, no. 3-4, pp. 213–231, Dec. 2011.

[164] A. Abdrabou and W. Zhuang, "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 129–139, Jan. 2011.

[165] Q. Zhao, Y. Zhu, C. Chen, H. Zhu, and B. Li, "When 3G meets VANET: 3G-assisted data delivery in VANETs," *IEEE Sensors Journal*, vol. 13, no. 10, pp. 3575–3584, Oct. 2013.

[166] 3GPP, "Study on NR vehicle-to-everything (V2X)," TS 38.885, Oct. 2018.

[167] Y. Ni, J. He, L. Cai, and Y. Bo, "Delay analysis and message delivery strategy in hybrid V2I/V2V networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–6.

[168] J. Chen, G. Mao, C. Li, A. Zafar, and A. Y. Zomaya, "Throughput of infrastructure-based cooperative vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 2964–2979, Nov. 2017.

[169] J. Liu, J. Wan, Q. Wang, P. Deng, K. Zhou, and Y. Qiao, "A survey on position-based routing for vehicular ad hoc networks," *Telecommunication Systems*, vol. 62, no. 1, pp. 15–30, May 2016.

[170] F. Zhang, B. Jin, Z. Wang, H. Liu, J. Hu, and L. Zhang, "On geocasting over urban bus-based networks by mining trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1734–1747, June 2016.

[171] Y. Zhu, Y. Wu, and B. Li, "Trajectory improves data delivery in urban vehicular networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 4, pp. 1089–1100, Apr. 2014.

[172] L. Yao, J. Wang, X. Wang, A. Chen, and Y. Wang, "V2X routing in a VANET based on the hidden Markov model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 889–899, Mar. 2018.

[173] O. Choi, S. Kim, J. Jeong, H. Lee, and S. Chong, "Delay-optimal data forwarding in vehicular sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6389–6402, Aug. 2016.

[174] G. Sun, Y. Zhang, D. Liao, H. Yu, X. Du, and M. Guizani, "Bus-trajectory-based street-centric routing for message delivery in urban vehicular ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 8, pp. 7550–7563, Aug. 2018.

[175] N. Alsharif and X. Shen, "*i*CAR-II: Infrastructure-based connectivity aware routing in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4231–4244, May 2017.

[176] B. Liu, D. Jia, K. Lu, H. Chen, R. Yang, J. Wang, Y. Barnard, and L. Wu, "Infrastructure-assisted message dissemination for supporting heterogeneous driving patterns," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 10, pp. 2865–2876, Oct. 2017.

[177] Y. Wu, Y. Zhu, and B. Li, "Infrastructure-assisted routing in vehicular networks," in *2012 IEEE Conference on Computer Communications (INFOCOM)*, Mar. 2012, pp. 1485–1493.

[178] M. Wang, H. Shan, R. Lu, R. Zhang, X. Shen, and F. Bai, "Real-time path planning based on hybrid-VANET-enhanced transportation system," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 5, pp. 1664–1678, May 2015.

[179] M. Garetto and E. Leonardi, "Restricted mobility improves delay-throughput tradeoffs in mobile ad hoc networks," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5016–5029, Oct. 2010.

[180] M. Wang, H. Shan, T. H. Luan, N. Lu, R. Zhang, X. Shen, and F. Bai, "Asymptotic throughput capacity analysis of VANETs exploiting mobility diversity," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 9, pp. 4187–4202, Sept. 2015.

[181] J. He, L. Cai, P. Cheng, and J. Pan, "Delay minimization for data dissemination in large-scale VANETs with buses and taxis," *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 1939–1950, Aug. 2016.

[182] M. Xing, J. He, and L. Cai, "Utility maximization for multimedia data dissemination in large-scale VANETs," *IEEE Transactions on Mobile Computing*, vol. 16, no. 4, pp. 1188–1198, Apr. 2017.

[183] ——, "Maximum-utility scheduling for multimedia transmission in drive-thru Internet," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2649–2658, Apr. 2016.

[184] J. He, L. Cai, J. Pan, and P. Cheng, "Delay analysis and routing for two-dimensional VANETs using carry-and-forward mechanism," *IEEE Transactions on Mobile Computing*, vol. 16, no. 7, pp. 1830–1841, July 2017.

[185] A. B. Reis, S. Sargento, F. Neves, and O. K. Tonguz, "Deploying roadside units in sparse vehicular networks: What really works and what does not," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 6, pp. 2794–2806, July 2014.

[186] Z. Zhang, G. Mao, and B. D. O. Anderson, "Stochastic characterization of information propagation process in vehicular ad hoc networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 122–135, Feb. 2014.

[187] W. Wang, S. S. Liao, X. Li, and J. S. Ren, "The process of information propagation along a traffic stream through intervehicle communication," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 1, pp. 345–354, Feb. 2014.

[188] L. Kleinrock, *Queueing systems, volume 2: Computer applications*. Wiley New York, 1976, vol. 66.

[189] R. T. Marler and J. S. Arora, "The weighted sum method for multi-objective optimization: New insights," *Structural and multidisciplinary optimization*, vol. 41, no. 6, pp. 853–862, 2010.

[190] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.