Enabling Millimeter Wave Communication for 5G Cellular Networks: MAC-layer Perspective

by

Jian Qiao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering
Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Data traffic among mobile devices increases dramatically with emerging high-speed multimedia applications such as uncompressed video streaming. Many new applications beyond personal communications involve tens or even hundreds of billions wireless devices, such as wireless watch, e-health sensors, and wireless glass. The number of wireless devices and the data rates will continue to grow exponentially. Quantitative evidences forecast that total data rate by 2020 will be 1000 times of current 4G data rate. Next generation wireless networks need fundamental changes to satisfy the overwhelming capacity demands.

Millimeter wave (mmWave) communication with huge available bandwidth is a very promising solution for next generation wireless networks to overcome the global bandwidth shortage at saturated microwave spectrum. The large available bandwidth can be directly translated into high capacity. mmWave communication has several propagation characteristics including strong pathloss, atmospheric and rain absorption, low diffraction around obstacles and penetration through objects. These propagation characteristics create challenges for next generation wireless networks to support various kinds of emerging applications with different QoS requirements. Our research focuses on how to effectively and efficiently exploit the large available mmWave bandwidth to achieve high capacity demand while overcoming these challenges on QoS provisioning for various kinds of applications.

This thesis focuses on MAC protocol design and analysis for mmWave communication to provide required capacity and QoS to support various kinds of applications in next generation wireless networks. Specifically, from the transmitter/receiver perspective, multi-user beamforming based on codebook is conducted to determine best transmission/reception beams to increase network capacity considering the mutual interferences among concur-

rent links. From the channel perspective, both interfering and non-interfering concurrent links are scheduled to operate simultaneously to exploit spatial reuse and improve network capacity. Link outage problem resulting from the limited diffraction capability and low penetration capability of mmWave band is addressed for quality provisioning by enabling multi-hop transmission to replace the link in outage (for low-mobility scenarios) and buffer design with dynamic bandwidth allocation among all the users in the whole coverage area (for high-mobility scenarios). From the system perspective, system structure, network architecture, and candidate MAC are investigated and novel backoff mechanism for CSMA/CA is proposed to give more transmission opportunity to faraway nodes than nearby nodes in order to achieve better fairness and higher network capacity. In this thesis, we formulate each problem mentioned above as an optimization problem with the proposed algorithms to solve it. Extensive analytical and simulation results are provided to demonstrate the performance of the proposed algorithms in several aspects, such as network capacity, energy efficiency, link connectivity and so on.

Acknowledgements

Since starting at the University of Waterloo, many of my lifelong lessons were taught by individuals who are probably unaware of their impressions on me over the years.

I express the deepest appreciation and gratitude to Prof. Sherman (Xuemin) Shen and Prof. Jon W. Mark, who have provided their exceptional insight and consultation in the development of my research interests. Both of them positively influenced me in many ways. Their novel ideas, infectious enthusiasm and intellectually stimulating discussions kept me motivated and encouraged through the entire life of my PhD studies.

I would like to acknowledge Prof. Lin X. Cai, Prof. Lei Lei, Dr. Bin Cao, Prof. Kan Zheng, Prof. Hao Liang, and Prof. Zhiguo Shi, who have been invaluable as a research partner and friend. This thesis is based upon my research with them, and only with their dedication, wisdom, and inspiration, could I have completed everything on time.

Thank you Prof. Abdallah Shami, Prof. Zhou Wang, Prof. Sagar Naik, and Prof. Raouf Boutaba for reading my thesis, attending my PhD seminar and PhD defense, and of course, for your insights.

I would like to thank every member in the Broadband Communications Research Group (BBCR) at the University of Waterloo who gave me great supports for my research and personal life.

Thank you everyone for your belief in me.

Dedication

For the most loving Father and Mother, who sacrificed all for our destinies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Future fifth generation (5G) cellular networks are being developed to achieve very high network capacity in order to satisfy dramatically increasing mobile traffic demands brought on by the continuing advances and discoveries in computing and communications [1–3]. Future 5G cellular networks come with fundamental changes and are expected to provide multi-Gbps data rate for each mobile user to support high-speed multimedia applications with stringent Quality of Service (QoS) requirements. For example, uncompressed video streaming requires mandatory data rate of 1.78/3.56 Gbps. The newly emerged bandwidth-intensive applications (such as mobile could, video streaming, ubiquitous health care, and real-time interactive game) create unprecedented challenges for wireless service providers on global bandwidth shortage [2]. Therefore, more spectrum availability is essential in 5G cellular networks.

Millimeter wave (mmWave) communication has very large available bandwidth (multi-GHz), which can be directly translated to overwhelming capacity. Thus, mmWave communication is a very promising solution for future 5G cellular networks to overcome the global bandwidth shortage at saturated microwave spectrum for most commercial access

technologies [1]. Multi-Gbps transmission at mmWave band has been realized in both indoor (e.g., wireless personal area networks) [4] and outdoor (e.g., wireless mesh networks) systems [5]. The availability of mmWave spectrum and the recent advances of RFIC design motivate intensive interest of both industry and academia in leveraging mmWave communication for future 5G cellular networks. 5G cellular networks with mmWave communication are expected to have the main characteristics of highly directional antennas at both wireless devices and base stations, lower link outage probability, extremely high data rate in most coverage area, and higher aggregate capacity for many simultaneous users. mmWave mesh networks can be used as wireless backbone to provide rapid deployment and mesh-like connectivity.

In order to effectively and efficiently utilize the wireless medium among multiple users with various QoS requirements, a simple but robust, efficient and fair sharing medium access control(MAC) protocol for the future 5G cellular networks will be in high demand taking into account the unique features of mmWave communication.

1.1 5G Cellular Networks

In this section, we review the evolution of cellular networks, describe what will 5G cellular networks be, and summarize the standardization activities for 5G, followed by the key enabling technologies to achieve the expected performances of 5G cellular networks.

1.1.1 The Evolution of Cellular Networks

To date, four generations of cellular networks have been deployed in the world. Each generation emerges about every 10 years to significantly improve the transmission rate

Table 1.1: Evolution of 1G through 4G Cellular Networks

Generation	Features	Applications
1G	Deployed in the 1980s. Analogy technology.	Voice communication.
2G	Deployed in the 1990s. Digital modulations. Primary Technologies are IS-95 CDMA and GSM.	Voice SMS and low-rate data.
3G	144 kbps for mobile, 384 kbps pedestrian, and 2 Mbps for indoor. CDMA2000, WIMAX, and UMTS-HSPA.	New applications, such as video conference, location-based service.
4G	Require ability of 40 MHz channel with high spectral efficiency. LTE, LTE-A and IEEE 802.16.m.	Higher rate data hundreds of Mbps.

and support more applications. For example, in USA, first generation analog FM cellular systems appeared in 1981; second generation digital technology appeared in 1992; 3G appeared in 2001, and 4G LTE-A was implemented in 2011. Table 1.1 summarizes the evolution of cellular networks from 1G to 4G from the aspects of implemented key technologies and mainly supported applications.

The first generation cellular networks were designed for the application of voice call based on analog communication. The second generation cellular networks can support both voice call and low rate data services (such as text) through digital modulations and time division multiple access (TDMA) or code division multiple access (CDMA) with

improved spectral efficiency. 3G introduced high-speed Internet access, improved video and audio streaming capabilities by using technologies such as Wideband CDMA (WCDMA) and High Speed Packet Access (HSPA) which is the combination of High Speed Downlink Packet Access (HSDPA) protocol and High Speed Uplink Packet Access (HSUPA) protocol. An improved 3GPP (3rd Generation Partnership Project) standard, Evolved HSPA, was released in late 2008 with subsequent worldwide utilization beginning in 2010. HSPA has been deployed in over 150 countries by more than 350 communications service providers (CSP) on multiple frequency bands and is now the most extensively sold radio technology worldwide [6], although LTE is closing the gap rapidly.

The International Mobile Telecommunications-Advanced (IMT-Advanced) standard is defined by the ITU and is considered as a 4G mobile communications technology. IMT-Advanced standard includes capabilities outstripping those of IMT-2000 (3G) mobile communication. It should be noted that there is no universal definition for 4G cellular networks. 3GPP also provides a 4G-capable mobile broadband system, named Long Term Evolution (LTE). LTE is an orthogonal frequency-division multiplexing (OFDM)-based radio access technology supporting a scalable transmission bandwidth up to 20 MHz and advanced multi-antenna transmission. Multiple-Input Multiple-Output (MIMO) is involved as a key technology to achieve high data rates in 4G cellular networks by enabling multi-stream transmission. It can obtain high spectrum efficiency, improved link quality, and the adaptation of radiation patterns for signal gain and interference mitigation by adaptive beamforming using antenna arrays [7]. The combination of HSPA and LTE will increase the peak mobile data rates of the two systems, with data rates exceeding 100 Mbps, and will also allow for optimal dynamic load balancing between the two technologies [6]. To improve the existing LTE network, LTE-Advanced (LTE-A) is defined to IMT-Advanced requirements, which will be theoretically capable of peak throughput rates exceeding 1

Gbps. LTE-A supports heterogeneous networks with co-existing large macro, micro, and pico cells, and WiFi access points (AP). Low cost deployment will be realized by self-organizing features and relays.

1.1.2 What Will 5G be?

The annual visual network index (VNI) reports of Cisco show the quantitative evidence that the wireless data explosion is real and will continue [8]. As the mobile traffic demand increases dramatically every year, wireless carriers must be prepared to support up to a thousand-fold increase in total mobile traffic by 2020. To achieve one thousand-fold capacity by 2010, 5G cellular network researchers need to find new wireless spectrum beyond the 4G standard. Fundamental changes are expected to implement in future 5G cellular networks since incremental improvements and small amount of new spectrum would not satisfy the capacity demands.

In order to understand the engineering challenges of 5G, it is necessary to identify the performance requirements. In the following, we discuss the performance requirements in terms of data rate, latency, and energy cost. It should be stressed that not all of these performances aspects need to be satisfied simultaneously since different applications have different performance requirements. For example, very high-rate applications such as high definition video streaming may have relaxed latency and reliability requirements compared to driverless cars or public safety applications, where latency and reliability are paramount but lower data rates can be tolerated.

1)**Data Rate:** It is unquestionable that the data rate plays the most significant role in the design of 5G cellular networks since mobile traffic explosion is the main driver behind 5G. Data rate can be measured in several different ways, and 5G cellular networks should

have a specific target for each such data rate metric:

- **Aggregate data rate or area capacity** is defined as the total amount of data the network can support, characterized in bits/s per unit area. Generally, the aggregate data rate will need to increase by 1000 times from 4G to 5G.

- **Edge rate** is the worst data rate a user experiencing within the range of the network, and so is an important metric and has a concrete engineering meaning. Targets for the 5G edge rate range from 100 Mbps to 1 Gbps. Considering the fact that current 4G systems have a typical edge rate of about 1 Mbps, this requirement needs at least 100 times advance. Please note that the exact number of edge rate greatly depends on network load, cell size, and other factors.

- **Peak rate** is the best data rate of a user under any conceivable network configuration. The peak rate is a marketing number, devoid of much meaning to engineers and likely to be in the range of tens of Gbps.

2)**Latency:** Current 4G round-trip latencies are on the order of about 15 ms, and are based on the 1 ms subframe time with necessary overheads for resource allocation and medium access control. The latency of 4G can satisfy the requirements of most current applications. However, driven by the emerging new applications including round-way gaming and wearable wireless devices, the anticipated latency for 5G cellular networks is around 1 ms, an order of magnitude faster than 4G. In addition to shrinking down the subframe structure, such severe latency constraints may have important implications on design choices at several layers of the protocol stack and the core network?

3)**Energy and Cost:** For 5G cellular networks, it is ideally expected that the per-link energy consumption should decrease. Since the per-link data rates being offered will be

increasing by about 100 times, this means that the energy consumption per bit and cost per bit need to decrease by at least 99 percentages. Some promising technologies such as mmWave communication and small cell would provide reasonable cost and power scaling. Specifically, mmWave spectrum should be 10100 times cheaper per Hz than the spectrum below 3 GHz for 3G/4G cellular network because mmWave communication has very large available bandwidth to achieve high data rate rather than achieving high data rate by increasing the transmission power. In addition, small cells should be 10100 times cheaper and more power efficient than macrocells. A major cost consideration for 5G, even more so than in 4G due to the new BS densities and increased bandwidth, is the backhaul from the network edges into the core.

Among these performance requirements, it is essential to improve the aggregate data rate and per-flow data rate which are the motivation of 5G and the main differences of 5G compared with 4G cellular networks. Therefore, in this thesis, we focus on the data rate aspect of 5G cellular networks.

1.1.3 5G Standardization Activities

Several forums and large projects have been organized to build the overall picture of 5G [1,3] and investigate the key enabling technologies [9,10]. Specifically, 5G scenarios and requirements have been released by the aforementioned EU project METIS (Mobile and wireless communications Enablers for the Twenty-twenty Information Society) [9]. Meanwhile, China IMT-2020 promotion group has been established based on IMT-Advanced promotion group. 5G has been referred to as "IMT-2020" in many industry forums and international telecommunications union (ITU) working groups [10] aiming to start 5G deployments around 2020. To identify 5G user requirements and to elaborate a standards agenda driven

by the requirements, the ETSI (European Telecommunications Standards Institute) held a future mobile summit in Nov. 2013. The summit concluded that an evolution of LTE may not be sufficient to meet the anticipated 5G requirements. "5G forum" has been established in June 2013 in Korea and "2020 and Beyond Ad-Hoc" has been established in October 2013 in Japan, both aiming to next generation wireless systems.

1.1.4 Enabling Technologies for 5G

Among the outlined requirements in Sec. 1.1.2, the one certainly getting most attention is the need to radically higher data rates. Given a area A, the total data rate R_{total} can be given as

$$R_{total} \approx \frac{\eta}{A} N_{BS} W_{sys} \tag{1.1}$$

where η is the spectrum efficiency, N_{BS} is the number of base stations in the area, and W_{sys} is the system bandwidth, respectively. Significant improvement on aggregate data rate can be achieved through combined gains in three categories:

- Dense deployment with small cell to increase the number of base stations in the area;
- Moving toward and into mmWave spectrum to obtain the huge and clean system bandwidth;
- Taking advantages of advances in MIMO (such as massive MIMO) to increase spectral efficiency (support more bits/s/Hz per node).

In the following, we discuss the potential enabling technologies related to these categories.

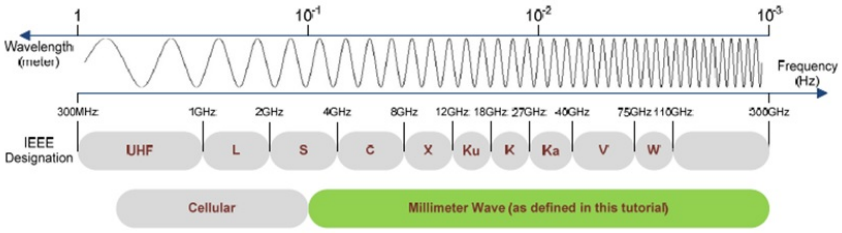

Figure 1.1: Micro-wave and mmWave Frequency Spectrum

mmWave Communication

Fig. 1.1 shows the frequency spectrum including both micro-wave frequency band and mmWave frequency band. The micro-wave spectrum which has been used for most commercial networks has become nearly fully occupied, in particular at peak times and in peak markets. Regardless of the efficiency of densification and offloading, much more bandwidth is highly required. Fortunately, vast amounts of available spectrum do exist in the mmWave range. Strictly, mmWave spectrum includes the frequency from 30 GHz to 300 GHz with the wavelength in millimeter order (from 1mm to 1 cm). Due to similar behavior, the spectrum between 3GHz and 300GHz is referred to as mmWave frequency band with the effort to explore new spectrum for mobile broadband communication. To realize mmWave communication in 5G cellular networks fully satisfying the requirements of 5G cellular networks, the main challenges need to be overcome are mmWave propagation features (such as high propagation loss and blockage), novel transceiver design, and co-existence with 4G. More details about mmWave communications would be discussed in Sec. 1.2.

Another reason that mmWave communication attracts more and more attention in both academia and industry is the recent success in building inexpensive and low power mmWave transceiver components with CMOS technology [11, 12]. This advance makes it feasible to use the mmWave radio for mass deployment in 5G cellular networks.

Massive MIMO

MIMO (multiple-input and multiple-output) exploits the spatial multiplexing gain when multiple antennas are available at both base station and mobile devices. In single-user MIMO (SU-MIMO), the dimensions are limited by the number of antennas that can be integrated on a mobile device. Multi-user MIMO (MU-MIMO) can be enabled by having each base station communicate with several users simultaneously, thus MU-MIMO can effectively pull together the antennas at the users and overcome this limitation. Then, the signaling dimensions are defined as the smaller number of aggregate antennas between at the users and at the base station. Furthermore, coordinated multipoint (CoMP) transmission/reception in LTE supports multiple base stations cooperating and acting as a single effective MIMO transceiver, thus some of the interferences in the system are turned into useful signals.

In [13], the number of antennas is increased by more than an order of magnitude. Each base station is equipped with a number of antennas much larger than the number of active users per timefrequency resource. The condition that accurate channel estimation can be obtained for tens of users per resource, drives the number of antennas for each base station into the hundreds. This idea was initially called "large-scale antenna systems" and known as "massive MIMO" nowadays. Massive MIMO can offer the following advantages:

1. Massive MIMO can increase the system capacity by 10 times or more and simultaneously improve the energy-efficiency in the order of 100 times;
2. Inexpensive, low-power components can be realized in massive MIMO system;
3. Massive MIMO enables a significant reduction of latency on the air interface;
4. Massive MIMO simplifies the multiple-access layer.

There are some significant problems that urgently need to be solve in massive MIMO system: making many low-cost low-precision components work effectively together, acquisition and synchronization for newly-joined terminals, the exploitation of extra degrees of freedom provided by the excess of service-antennas, reducing internal power consumption to achieve total energy efficiency reductions, and finding new deployment scenarios.

Small Cell

According to (2.3), another effective way to increase the network capacity is to make the cells smaller. The effectiveness of this approach has been demonstrated over several generations of cellular networks [14]. For example, in Japan, the spacing between base stations can be as small as two hundred meters, giving a coverage area well under a tenth of a square km [14]. Networks are now rapidly evolving to include nested small cells such as picocells (range under 100 meters) and femtocells (WiFi-like range), as well as distributed antenna systems that are functionally similar to picocells from a capacity and coverage standpoint but have all their baseband processing at a central site and share cell IDs [15].

It can achieve numerous benefits by making the cell range shorter. The most important benefit is the spectrum reuse in the whole coverage area. In addition, by dense base station deployment, fewer number of users are competing for the resources of each base station. This allows each base station to devote its resources and its backhaul connection to relatively smaller number of users. As the densification becomes extreme, some challenges arise:

- Ideally, the effective increase in data rate is expected to be proportional to the increase of the number of base stations (e.g., base station deployment density). However, at microwave frequency band, the gain in SINR by increasing base station deployment

density is not enough to keep up with the decrease in small-cell utilization [14], especially for the case that each base station is associated with fewer number of users and becomes more lightly loaded. Therefore, how to preserve the expected cell-splitting gains particularly for low-power nodes is a very important issue;

- Networks will continue to become increasingly heterogeneous as moving toward 5G [2]. Therefore, a key feature will be increased integration between different networks, with a typical 5G device capable of supporting multiple networks, such as new 5G standard (e.g., at mmWave frequencies), 3G, 4G LTE, several types of WiFi, and perhaps direct device-to-device (D2D) communication, all across a great many spectral bands. Hence, determining which standard and spectrum to utilize and which base station or users to associate with will be a truly complex task for the network;

- Clearly, the continued network densification and increased heterogeneity make challenges on supporting mobility. Although a hefty share of data is served to stationary indoor users, the support of mobility and keeping connectivity is always the most important challenging issue for 5G cellular networks, especially for mmWave communication with highly directional antenna.

- Affording the rising costs by deploying more base stations associated with the installation, maintenance and backhaul connection.

In summary, the potential technologies, small cell, massive MIMO, and mmWave communication can achieve much higher system capacity which is the key feature of 5G cellular networks, although there are several challenges to be overcome for each potential technology. mmWave communication is the fundamental technology to achieve more gains on system capacity with huge available bandwidth, compared with the other two. mmWave communication would be more promising if small cell and/or massive MIMO are applied

on it. Therefore, this thesis focuses on the fundamental research issues on mmWave communication for 5G cellular networks.

1.2 mmWave Communication

The recent advances on mmWave Radio Frequency Integrated Circuits (RFIC) design [11] and the high available bandwidth (around several GHz) motivate researchers and engineers around the world to put intensive attention and imagination to mmWave communications [5, 16–22]. Multi-Gbps transmission at mmWave band has been realized in both indoor (e.g., wireless personal area networks) [22] and outdoor (e.g., wireless mesh networks) systems [5]. As discussed in Sec. 1.1.4, mmWave communication is a fundamental, promising, and practical candidate for future 5G cellular networks.

Generally, the unlicensed mmWave spectrum provides the following advantages:

- Huge frequency allocation. Typically serval GHz in most regions of the world.
- Much more power available. Unlike the strict transmit power restrictions on UWB unlicensed operation, the mmWave band allows an effective isotropic radiated power that is significantly greater.
- Clean spectrum, no incumbents. There are not any widely deployed mmWave radiators in both indoor and outdoor environments, so there is less chance for interference.
- High frequency with small wavelength allows small, high-gain antennas. A 25-dB gain antenna has an effective aperture approximately one square inch. High-gain antennas allow high EIRP with low-power RF amplifiers

- High propagation loss allows overlapping networks that do not interfere each other much. Because the antennas are highly directional at mmWave frequencies, spatial reuse is enabled.

Apart from this, mmWave communication shares the common features with other wireless systems. These include: (1) high error rate and bursty errors; (2) location-dependent and time-varying wireless link capacity; (3) half-duplex communication; (4) user mobility; and (5) power constraints of mobile users.

1.2.1 Early Activities Using mmWave Spectrum

Due to the high achievable data rates, mmWave communication has been attracting more and more efforts in research, development, regulation, and standardization for wireless personal area networks (WPANs) and wireless local area networks (WLANs) in the past decade. Several standardization groups, such as ECMA international TC48 [23], IEEE 802.15.3c [24], and IEEE 802.11ad VHT [25] have been built to realize multi-Gbps WiFi in mmWave band. Each of them is influenced heavily by its technical heritage in the respective standard body. ECMA TC48 is a 60 GHz PHY and MAC standard to provide high rate wireless personal area network (WPAN) transport. The MAC in ECMA is rooted in WiMedia Ultra-Wideband (UWB) MAC with necessary changes to accommodate the characteristics of 60 GHz. The IEEE 802.15.3c Task Group is developing a mmWave-based alternative PHY to work with the existing 802.15.3 WPAN MAC Standard. IEEE 802.11ad is another installment of the successful 802.11 (i.e., Wi-Fi) family and so it will maintain its affinity with 802.11 in many aspects. One key advantage of IEEE 802.11ad over the other standardization activities in the 60 GHz networks is that it builds on the already existing strong market presence of Wi-Fi in the 2.4/5 GHz bands. LMDS (local

multipoint distribution service) operating on frequencies from 28 to 30 GHz was standardized by the IEEE 802 LAN/MAN Standards Committee through the efforts of the IEEE 802.16.1 Task Group ("Air Interface for Fixed Broadband Wireless Access Systems" for 1066 GHz). LMDS uses a cellular infrastructure, with multiple base stations supporting point-to-multipoint communication to small customer transceivers. The Federal Communications Commission (FCC) auctioned two LMDS licenses per market (basic trading areas). The A license includes a total of 1.15 GHz bandwidth, and consists of the 27.528.35 GHz, 29.129.25 GHz, and 31.07531.225 GHz bands. The B license is 150 MHz wide, covering the 31.031.075 GHz and 31.22531.3 GHz bands.

The fundamental researches in mmWave WPANs/WLANs, such as transceiver design, beamforming, neighbor discovery, transmission scheduling, resource allocation and so on, provide the preliminary results to motivate mmWave communication for 5G cellular networks.

1.2.2 mmWave Spectrum Standardization for 5G

Spectrum standardization and harmonization efforts for 5G have begun within the ITU. Studies are under way on the feasibility of bands above 6 GHz, including technical aspects such as channel modeling, semiconductor readiness, coverage, mobility support, potential deployment scenarios and coexistence with existing networks. To be available for 5G, mmWave spectrum has to be repurposed by national regulators for mobile applications and agreement must be reached in ITU world radiocommunication conferences (WRC) on the global bands for mmWave communications. These processes tend to be tedious and lengthy, and there are many hurdles to clear before the spectrum can indeed be available. On the ITU side, WRC-18 is shaping up as the time and venue to agree on mmWave

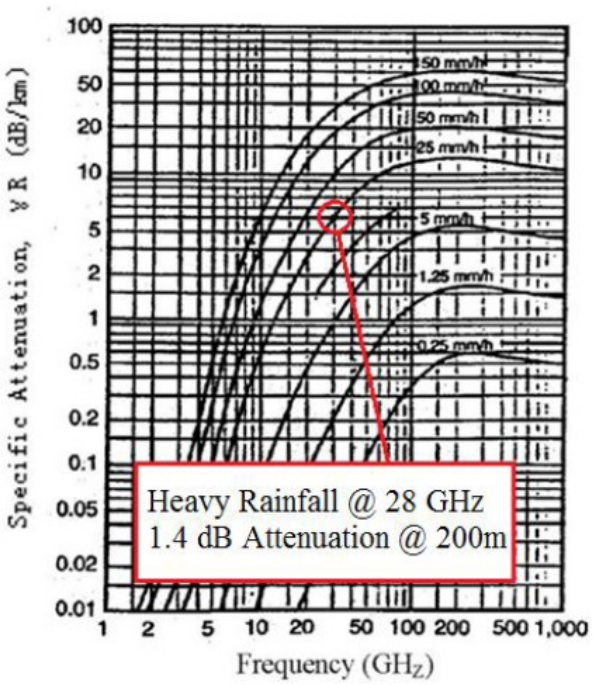

Figure 1.2: Rain attenuation in dB/km across frequency at various rainfall rates

spectrum allocations for 5G.

In addition to the ITU, many national regulators have also started their own studies on mmWave spectrum for mobile communications. In the USA, the technological advisory council of the federal communications committee (FCC) has carried out extensive investigations on mmWave technology in the last few years and it is possible that FCC will issue a notice of inquiry in 2014 [1], which is always the first step in FCC's rule making process for allocation of any new frequency bands. As discussed above, it is also unclear how such bands will be allocated or even how they should be allocated, and the technical community should actively engage the FCC to make sure they are allocated in a manner conducive to meeting 5G requirements.

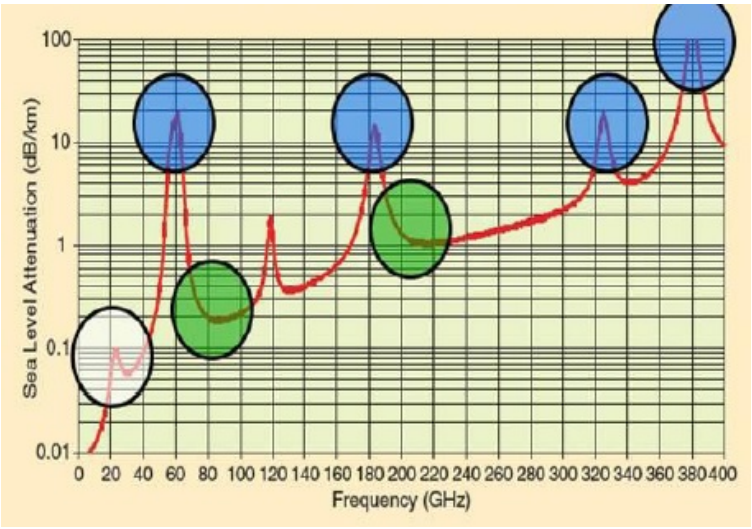

Figure 1.3: Atmospheric absorption across mmWave frequencies in dB/km

1.2.3 Propagation Characteristics of mmWave Spectrum

Typically, mmWave spectrum includes the huge spectrum from around 30 GHz to 300 GHz. The propagation characteristics of mmWave communication are different over such huge frequency band. Fig. 1.2 and Fig. 1.3 show the rain attenuation and atmospheric absorption characteristics over different frequency bands [26, 27]. A doubt on using mmWave communication for 5G cellular networks is that rain and atmosphere make mmWave communication not suitable for mobile communications. However, considering the fact that today's cell sizes in urban environments are on the order of 200 m, it becomes clear that mmWave communication can overcome these issues.

In Fig. 1.3, it can be seen that for cell sizes on the order of 200 m, atmospheric absorption does not create significant additional path loss for mmWave communication, particularly at 28 GHz and 38 GHz. Only 7 dB/km of attenuation is expected due to heavy rainfall rates of 1 inch/hr for cellular propagation at 28 GHz, which translates to only 1.4

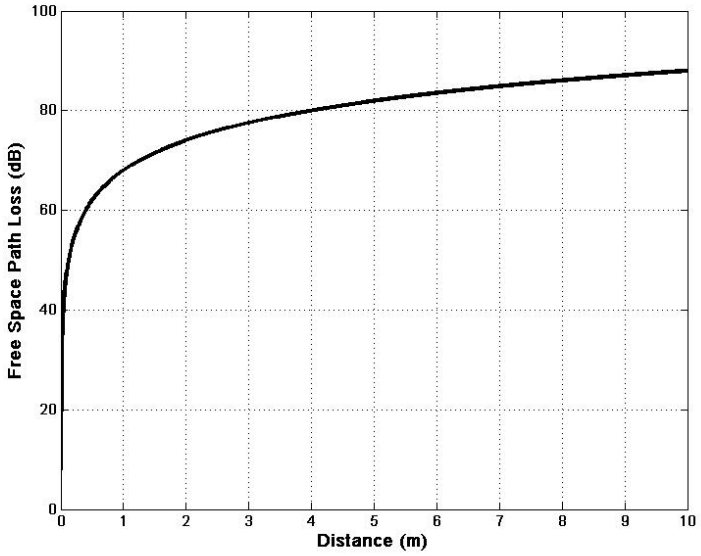

Figure 1.4: Free Space Path loss at 60 GHz

dB of attenuation over 200 m distance. Work by many researchers has confirmed that for short distances (less than 1 km), rain attenuation will present a minimal effect on the propagation of mmWaves at 28 GHz to 38 GHz for small cells [26].

One fundamental distinguishing feature of mmWave communications is the high propagation loss. As the free space propagation loss increases proportionally as the square of the carrier frequency, the propagation loss at mmWave band is much higher than that at lower frequency bands, e.g., 28 dB higher at 60 GHz than that at 2.4 GHz as shown in Fig. 1.4. The relatively short wavelengths in mmWave bands impose challenges such as greater signal diffusion and difficulty in diffracting around obstacles. Non-line-of-sight (N-LOS) transmissions in mmWave channels suffer from significant attenuation and a shortage of multipaths [28]. Therefore, mmWave systems rely on line-of-sight (LOS) transmissions to achieve the high data rate. The obstacles and moving people can easily block the

LOS transmission, and greatly reduce the transmission data rate. mmWave signals have difficulties on penetrating through solid building materials, e.g., at 40 GHz, 178 dB attenuation for brick wall [2, 29]. Such high attenuation makes mmWave connection unavailable sometimes in the environment with dense high-rise buildings.

1.2.4 Impact of mmWave on 5G

- **Directional networking:**due to the high propagation loss (including free space path loss, rain attenuation and atmospheric absorption) in mmWave band, directional antenna with high directivity gain can be utilized to combat the severe propagation loss in order to achieve high data rate. Thus the whole networks should be directional, such as directional MAC, directional routing, directional handoff, directional transmission/receiption.

- **Spatial reuse:** the high propagation loss and the utilization of directional antenna result in relatively low multi-user interference (MUI), so that more concurrent transmissions can be supported to exploit the spatial reuse and further improve the aggregate data rate.

- **Small cell:** due to the high propagation loss, mmWave base stations need to be deployed densely with small cell to provide the high aggregate capacity and required edge rate for high-rate multimedia applications. The densification could put heavy stress on the cost to build, operate and upgrade the network.

- **Link outage:** mmWave communication has limited diffraction capability resulting from the short wavelength while mmWave signals have difficulties on penetrating through solid building materials. The limited diffraction capability and penetration

capability make mmWave signal suffer from frequent link outages in metropolis areas with lots of high-rise buildings and moving people.

- **HetNet:** current 4G cellular networks can provide seamless coverage and reliable communications because of the lower frequency band. For smooth and cost-efficient transition from 4G to 5G, the hybrid architecture of 4G+mmWave is promising to achieve seamless coverage and high rate in most coverage area. The management information and low-rate applications (such as voice, text, and web browser) are transmitted in the 4G networks while the mmWave bands are available for high-rate multimedia applications.

1.2.5 What will be mmWave 5G?

As 5G cellular networks based on mmWave communication are developed and implemented, we believe the main differences compared to 4G will be the use of much greater spectrum allocations at untapped mmWave frequency bands, highly directional beamforming antennas at both the mobile device and base station, longer battery life, lower outage probability, much higher bit rates in large portions of the coverage area, lower infrastructure costs, and higher aggregate rate for many simultaneous users in both licensed and unlicensed spectrum (e.g. the convergence of WiFi and cellular). The backbone networks of 5G in some areas will move from copper and fiber to mmWave wireless connections, allowing rapid deployment and mesh-like connectivity.

1.3 MAC-layer Challenges on mmWave 5G

The main design objective for mmWave 5G cellular networks is to provide multiple Gbps data rate for mobile users to support multimedia applications with QoS requirement. There are several challenges on mmWave communication to achieve the design objective.

First, although mmWave communication can provide high aggregate capacity because of the huge available bandwidth, network capacity needs to be further improved to satisfy the increasing mobile traffic demands. Thanks to high propagation loss and directional antenna, spatial reuse can be exploited by allowing concurrent transmissions to improve aggregate capacity. Meanwhile, allowing multiple communication links to transmit simultaneously results in a higher MUI, which can decrease the system throughput. How to improve the system capacity of mmWave 5G cellular networks by properly enabling concurrent transmissions is an important and challenging issue.

Second, directional antenna radiates much greater power in certain directions for increased performance on transmission/reception while reducing interference from unwanted sources. Since the transmitter and the receiver do not know the locations of each other, how to let both the transmitter and the receiver exactly direct their beams towards each other to achieve higher network capacity is an important and challenging issue. The procedure to determine an antenna beam toward a certain direction is beamforming. Furthermore, beamforming for concurrent communication links considering the mutual interferences are more complex.

Third, quality provisioning is another challenging issue which need to be addressed at MAC layer due to the vulnerable mmWave channel at PHY layer. The limited penetration ability and diffraction ability in mmWave band result in frequent link outage, especially for users with high mobility in metropolitan areas. How to overcome the link outage problem

is very important to provide the required QoS for various kinds of applications.

Fourth, in the contention-based MAC, the channel is usually captured by the packet with the strongest power level in the presence of contending transmissions. Since mmWave communication suffers severe propagation loss over distance, the channel is usually captured by the nearby users in mmWave 5G cellular networks. It is necessary to design an efficient CSMA/CA protocol with fairness among the users.

Therefore, to realize the reliable multiple Gbps transmission for mmWave 5G cellular networks, we need to keep network connectivity, conduct accurate beamforming, improve MAC efficiency and resource utilization efficiency, and so on.

Due to the unique characteristics of mmWave system, legacy MAC protocols originally designed for narrowband mobile systems supporting applications with lower data rate, need to be adjusted to satisfy the required performances for bandwidth-intensive applications. The following figure (Fig. 1.5) shows the issues related to MAC we would like to investigate, the performance gains we want to achieve, and the corresponding challenges we would like to overcome. The issues to be addressed are, but not limited to, concurrent beamforming protocol, concurrent transmission scheduling, quality provisioning, and MAC protocol design with fairness. With beamforming, the transmitter and receiver can attain maximum SINR, thus achieve higher transmission data rate. Additionally, the total setup time can be shorten considerably if the beamforming procedure of all the communication pairs are conducted jointly rather than separately, which can improve MAC efficiency.

Link outage problem can be dealt with by replacing the unavailable link with multiple short hops. Replacing single long hop with properly selected multiple short hops can increase the flow throughput even if the source and the destination are within one-hop range. Enabling multi-hop transmission is helpful to keep network connectivity for vul-

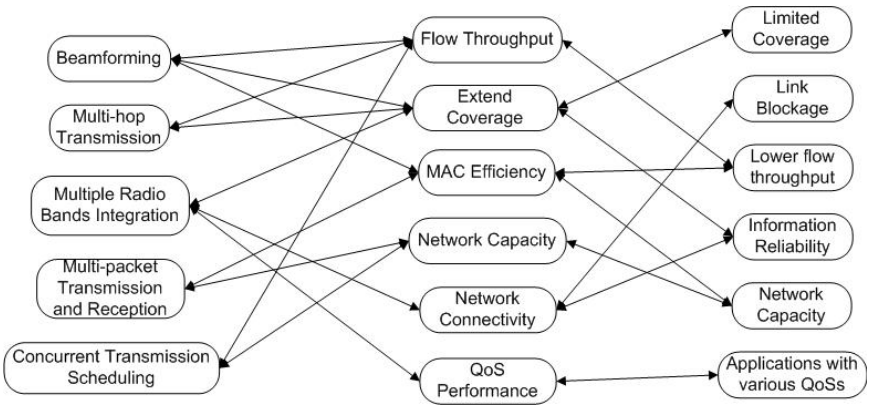

Figure 1.5: Research Motivations on 5G Cellular Networks

nerable mmWave channel. Another effective method to mitigate the link outage effect is to install a buffer in mobile device to store the data transmitted with extra allocated mmWave bandwidth when mobile user is in the areas with mmWave connection available. Larger allocated mmWave bandwidth can result in more data stored in the buffer, and thus the mobile users can maintain video playout quality for a specific time duration when they move into areas where mmWave connection is unavailable. The playout buffer is charged/discharged by dynamic bandwidth allocation to each moving user to satisfy the video playout quality requirement.

Concurrent transmission scheduling is an effective way to improve network capacity in mmWave 5G cellular networks. The mutual interferences among concurrent links can be mitigated by either selecting the non-interfering links to operate concurrently or using optimization model to selecting both interfering links and non-interfering links to operate simultaneously to achieve maximum aggregate capacity.

For contention based MAC protocol in mmWave 5G cellular networks, we propose a new backoff mechanism giving higher transmission priority to the distant nodes to achieve

better system throughput and fairness, considering the high propagation loss of mmWave communications. Specifically, different from traditional backoff mechanism in CSMA/CA, the contention window would be increased after a transmission success while the contention window would be decreased after a transmission success.

1.4 Thesis Focus and Structure

The thesis mainly focuses on MAC-layer modeling, design and analysis. Specifically, from transmitter/receiver perspective, multi-user beamforming based on codebook is conducted to determine best transmission/reception beams to increase network capacity; From channel perspective, both interfering and non-interfering concurrent links are enabled to exploit spatial reuse and improve network capacity. Link outage problem is addressed by enabling multi-hop transmissions (for low-mobility scenarios) and buffer design with bandwidth allocation (for high-mobility scenarios); From system perspective, system structure, network architecture, and candidate MAC are investigated and novel backoff mechanism for CSMA/CA is proposed to achieve better fairness and higher throughput.

1.5 Organization of Thesis

The remainder of the thesis is organized as follows. In Chapter 2, the system model, which is the platform of our research, is described. In Chapter 3, quality provisioning in 5G cellular networks is investigated to support various kinds of applications. Network capacity improvement by spatial reuse is presented in Chapter 4 with proposed concurrent transmission scheduling algorithms. In Chapter 5, multi-user beamforming is addressed to select best tranmsission/receiption beam patterns for multiple concurrent links to achieve

maximum system throughput. In addition, backoff mechanism for CSMA/CA in mmWave communication is re-designed to improve fairness and throughput in Chapter 6, followed by conclusions and further research topics in Chapter 7.

Chapter 2

Problem Definition and System Model

In this chapter, the research problems addressed in the thesis are defined and the system model is described on the heterogeneous network architecture, MAC structure, signal propagation model, transmission data rate, and directional antenna model.

2.1 Problem Definition

Sec. 1.3 and Sec. 1.4 mention the general research issues in the thesis. This section describes the definitions of the problems addressed in the thesis.

2.1.1 Quality Provisioning in 5G Cellular Networks

mmWave communication can achieve very high data rate with sensitivity to link blockage and attenuation loss. The mobile users would suffer intermittent mmWave connections in

the environment with obstacles and buildings. Most of high-rate applications are expected to operate in mmWave networks. The most challenging problem is how to provide quality provisioning at MAC layer for multimedia applications with high data rate requirement (such as uncompressed video streaming) in mmWave networks.

For low-mobility scenarios, in order to keep network connectivity, the link in outage is replaced by a multi-hop path. A hop selection metric is essential to determine the hops considering the flow throughput and load-balancing. Based on the hop selection metric, a multi-hop structure is necessary to detect the link in outage, select the relays, and build the multi-hop path.

For high-mobility scenarios, in order to mitigate effect of the high dynamics of mmWave channel, a buffer system is embedded in each mobile device to receive the data packets from mmWave channel. The basic idea is that extra bandwidth is allocated to mobile users to charge the buffer when mmWave connections are available. The saved data in the buffer is used to overcome frequent link outages due to high user mobility. How to properly and dynamically allocate bandwidth among moving users to charge/discharge the buffer to optimize video quality of all the users is an challenging problem.

2.1.2 Network Capacity Improvement by Spatial Reuse

For mmWave communication, the high propagation loss and directional antenna result in relatively low multi-user interference (MUI). Multiple communication links can operate simultaneously to improve system capacity. Allowing concurrent transmissions leads to higher MUI, which can decrease the system capacity. How to schedule appropriate concurrent transmissions to improve system performances is an important and challenging issue.

Due to antenna's directivity, any two communication links can operate simultaneously without mutual interferences under the condition that any transmitter is outside the beamwidth of the other receiver or does not direct its beam to the other receiver if it is within the beamwidth of the other receiver. A concurrent transmission scheduling scheme is necessary to determine a set of active links without MUI in each time slot in a spatial time-division multiple-access (STDMA) system, in order to achieve optimal system capacity while satisfying the data rate requirement of each user.

Non-interfering concurrent transmission scheduling is easy to implement without fully utilizing the resources. To further improve system capacity, both interfering and non-interfering links can be scheduled concurrently if the MUI is properly managed. How to schedule multiple concurrent links to operate simultaneously in order to achieve optimal resource utilization is an important issue.

2.1.3 Codebook-based Concurrent Beamforming

Since concurrent transmissions are desirable in mmWave networks to improve network capacity, beamforming for concurrent links should consider the MUI to decide the best transmission/reception beam for each link. The codebook-based concurrent beamforming problem is formulated as an optimization problem by maximizing the sum rates of the concurrent links. An algorithm is necessary to solve the optimization problem, and a comprehensive beamforming protocol is desired to setup directional concurrent links on MAC layer in a distributive manner.

2.1.4 CSMA/CA-based MAC Protocol Design

mmWave communication has high propagation loss over distance. If multiple users send packets to the same user simultaneously, the channel would be captured by the users with stronger power, which are likely the users nearby the destination. Traditional backoff mechanisms in CSMA/CA give high transmission priority after transmission success and lower transmission priority after transmission failure, which can result in serious unfairness and significant throughput reduction (if considering multi-packet reception) due to high propagation loss of mmWave band. A novel backoff mechanism in CSMA/CA is necessary to adjust the transmission probability aiming to improve fairness and system throughput.

2.2 System Model

2.2.1 Heterogeneous Network Architecture

Current 4G cellular network operating in microwave frequency band can provide seamless coverage and reliable communication. To achieve smooth and cost-efficient transition from 4G to 5G, 5G cellular networks use the hybrid 4G+mmWave system structure as shown in Fig. 2.1 to provide ubiquitous coverage and high rate in most coverage areas. 5G cellular networks are expected to support multiple kinds of applications. Low rate applications (such as voice call and Web browsing) can be supported with perfect QoS in 4G network. Most of high-rate applications, such as high-speed file downloading and high definition TV (HDTV), are expected to operate in mmWave networks because of the high achievable data rate.

The hybrid system is composed of 4G base stations, mmWave base stations, and mobile stations. 4G based stations are deployed in the center of each cell while mmWave base

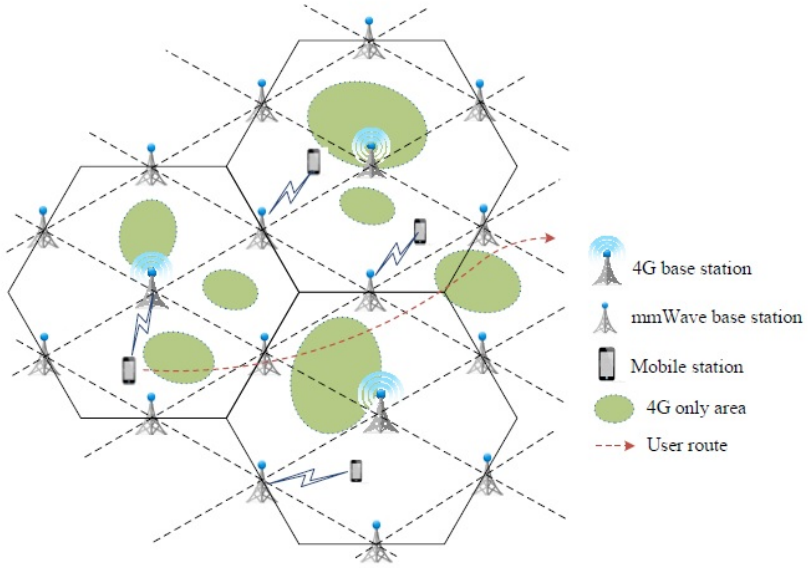

Figure 2.1: 5G Heterogeneous System Architecture with mmWave Coverage Holes

stations are deployed densely in small cell with grid topology to provide high data rates and aggregate capacity. Each mobile station has both mmWave access mode and 4G access mode, with fast mode transition between them. All mobile stations and 4G base stations have omni-directional antenna for 4G communications. All mobile stations and mmWave base stations are equipped with electronically steerable directional antennas with beamforming technologies for mmWave communication.

The basic topology of each mmWave cell as shown in Fig. 2.2 is composed of several wireless nodes and a single mmWave base station. The base station and wireless nodes are equipped with an electronically steerable directional antenna. Both transmitters and receivers direct their beams towards each other for data transmission. It supports the communication between BS and wireless node, and the communication between wireless nodes (D2D mode). Moving obstacles greatly degrade the transmission data rate when

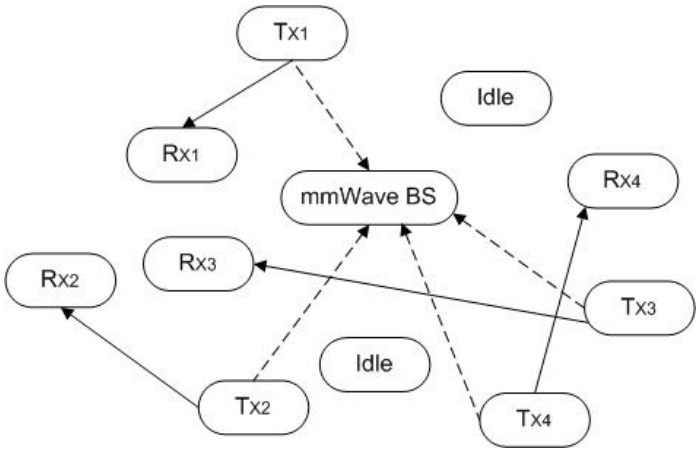

Figure 2.2: mmWave Cell Topology

they are located in the line-of-sight (LOS) link between the transmitter and the receiver, thus the data transmission for bandwidth-intensive applications relies on LOS link.

2.2.2 Timing Structure for mmWave Cell

Fig. 2.3 shows the timing structure of each mmWave cell. Time is partitioned into superframes composed of three phases: the Beacon period (BP) for network synchronization, control messages broadcasting, and scheduling decision distribution; the contention access period (CAP) for applications without performance guarantee; and the channel time allocation period (CTAP) composed of z channel time slots for bandwidth-intensive and delay-sensitive applications, respectively. Each time slot in CTAP is a TDMA slot granted by mmWave BS for multiple communication links. The beamforming operation is processed for all the active links at the beginning of each TDMA time slot.

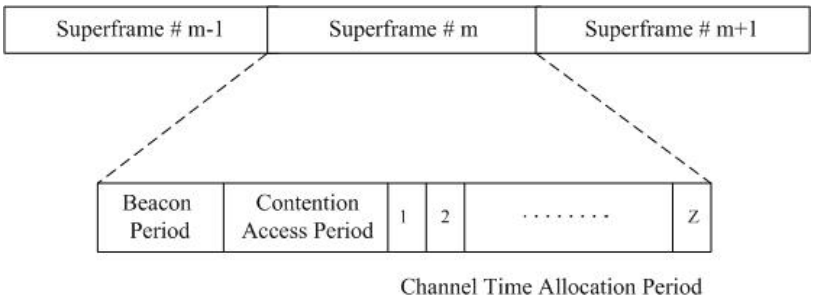

Figure 2.3: Superframe-based Timing Structure for Each mmWave Cell

2.2.3 mmWave Propagation and Coverage

Slow fading is considered in mmWave propagation. Instantaneous pathloss values in dB of mmWave signal are generated using statistical model [30]

$$PL(d)[dB] = \Theta + 20log_{10}(f) + 10\phi log_{10}(d) + \mathbb{E} \tag{2.1}$$

where ϕ is the pathloss exponent; Θ is the parameter specified for scenario and antenna configuration; and $\mathbb{E}$ is a random variable which describes shadow fading effect with $\mathbb{E} = 0$ for LOS scenario. For outdoor scenario, because of the directional antenna and high frequency band, mmWave propagation is short of reflections and multi-paths. Thus, only LOS transmissions are considered in mmWave outdoor communication [5]. For indoor environments [30], $\mathbb{E}$ follows Gaussian distribution in dB (log-normal distribution in absolute scale) with zero mean and shadow fading standard deviation σ for NLOS scenario. The pathloss parameters for LOS scenario and indoor NLOS scenario are shown in Table 2.1 [30].

With above pathloss model, the whole area in Fig. 2.1 can be classified into two categories (namely single-coverage areas and double-coverage areas) based on the instantaneous

Table 2.1: PROPAGATION PARAMETERS

LOS	Θ_{LOS}=32.5	ϕ_{LOS}=2.0	σ_{LOS}=0
NLOS	Θ_{NLOS}=44.7	ϕ_{NLOS}=1.5	σ_{NLOS}=3.3

pathloss. The areas satisfying

$$PL(d)[dB] \geq PL_{thr} \tag{2.2}$$

are considered as single-coverage areas, where PL_{thr} is the pathloss threshold to determine if the mmWave connection is available. Other areas with $PL(d)[dB] < PL_{thr}$ are defined as double-coverage areas. As shown in Fig. 2.1, there are some single-coverage areas without mmWave connection disjointedly distributed in whole area. Mobile users pass through the single-coverage and double-coverage areas alternatively.

The capacity of an additive white Gaussian noise (AWGN) channel with broadband interference assumed as Gaussian distribution is given by:

$$C = W \log_2[1 + \frac{P_R}{(N_0 + I)W}] \tag{2.3}$$

where P_R is the received signal power corresponding to $10^{PL(d)}$, W is the system bandwidth, N_0 and I are the one-side power spectral densities of white Gaussian noise and broadband interference, respectively. According to (2.1), the flow throughput reduction over distance is more serious in mmWave communication due to its large bandwidth and high frequency.

2.2.4 Antenna Model

We apply an ideal "flat-top" model for directional antenna. Every node employs an antenna with N beams, each of which spans an angle of Δ with $\Delta\alpha > \frac{2\pi}{N}$. A total of N beams can collectively maintain the seamless coverage of the entire plane. Directional antennas are

characterized by their pattern functions that measure the power gain $G(\phi)$ over the angle ϕ. The normalized pattern function is defined as

$$g(\phi) = \frac{G(\phi)}{G_{max}} \tag{2.4}$$

where

$$G_{max} = max_{\phi} G(\phi) \tag{2.5}$$

In an ideal case, the antenna gain is constant, i.e., unit gain within the beamwidth and zero outside the beamwidth [31],

$$g(\phi) = \begin{cases} 1, & \mid \phi \mid \leq \frac{\Delta\phi}{2} \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

where Δ is the antenna beamwidth. In our system, the transmitter and receiver antenna gains are $G_T = G_R = 1$ within the antenna beamwidth, while $G_T = G_R = 0$ outside the beamwidth.

Chapter 3

Quality Provisioning in 5G Cellular Networks

mmWave communication has huge available bandwidth to satisfy capacity demands of 5G. However, a few challenges on network connectivity could be created by the unique propagation characteristics of mmWave band. First, the mmWave band has limited diffraction capability because of the short wavelength. The line-of-sight (LOS) transmissions can be easily blocked by obstacles and moving people. Since non-line-of-sight (NLOS) transmissions suffer from significant attenuation and a shortage of multipaths [28], link outage can happen if LOS link is blocked. Second, mmWave signals have difficulties on penetrating through solid building materials, e.g., at 40 GHz, 178 dB attenuation for brick wall [2, 29]. Such high attenuation makes mmWave connection unavailable sometimes in the environment with dense high-rise buildings. The limited diffraction capability and penetration capability can result in mmWave link outage.

Multiple wireless networks coexist in 5G cellular networks [2, 3], such as 4G cellular

networks and mmWave networks. 4G cellular networks have seamless coverage and provide reliable connection with relatively low data rate. mmWave networks can achieve very high data rate with sensitivity to link blockage and attenuation loss. The integration of multiple radio access technologies is an appealing approach to support various emerging applications with different QoS requirements. 4G cellular networks operating can support low-rate applications with perfect QoS. Most of high-rate applications are expected to operate in mmWave networks. The limited diffraction capability and penetration capability of mmWave frequency band lead to frequency link outage for mobile users, which can be tolerated by high-rate applications without constant data rate requirement, such as file downloading. The most challenging problem is how to provide quality provisioning for multimedia applications with high constant data rate requirement (such as video streaming). The fundamental question is how to maintain the video playout quality for 5G users suffering from mmWave link outage.

3.1 Video Quality Provisioning with Low Mobility

For low-mobility scenarios, in order to keep network connectivity, the link in outage is replaced by a multi-hop path determined by the proposed novel hop selection metric aiming to improve the flow throughput and achieve load-balancing. A multi-hop structure is designed based on the proposed hop selection metric to detect the link in outage, select the relays, and build the multi-hop path. Theoretical analysis based on geographic interference demonstrates that by properly replacing single long hop with multiple short hops, the opportunities for spatial reuse and time division multiplexing gain can be improved.

3.1.1 Hop Selection Metric

To achieve high transmission data rate for each traffic flow, short links are usually preferred for hop selection. Therefore, more hops may be involved in the transmission of each flow, which results in heavy traffic loads in the network. In addition, if traffic aggregates at a node, congestion may occur and this node becomes a bottleneck for the network. Therefore, we need to select appropriate relay hops to improve the network throughput, considering both the link length and the traffic loads at the node.

When the mmWave base station receives a transmission request, it determines the appropriate relaying hops based on the network information, including the distance from one node to all its neighbors, antenna directions steering towards their neighbors, and the traffic load of each node. The traffic load of each node is defined as the number of transmissions running or to run at each node. More concurrent transmissions can be supported by well balancing the traffic loads in the network. To determine the relaying hops for a pair of transmitter and receiver, a weighted graph is generated by the base station. The weight associated with link $(A \rightarrow B)$ between nodes A and B is given by

$$w(A,B) = \frac{d^n(A,B)}{E[d^n]} + \frac{F(B)}{E[F]} \tag{3.1}$$

where $d(A,B)$ is the length of link $(A \rightarrow B)$, $F(B)$ is the traffic load of node B, $E[d^n]$ is the average link length to the power n among all the links in the network currently, and $E[F]$ is the average node traffic load, which is defined as the summation of the traffic loads of all the nodes in the network divided by the number of nodes. We use normalized value for hop selection to smooth the large difference between the node's loads and link lengths. To improve flow throughput and achieve load balancing, the transmission rate and the node's load should jointly contribute to the hop selection. The transmission data

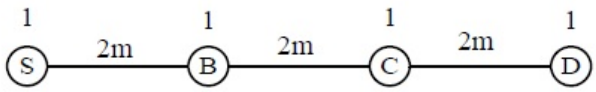

Figure 3.1: Illustration for hop selection

rate is a function of d^n, which is included in the metric (3.1) to favor multi-hop high data rate transmissions. For example, as shown in Fig. 3.1, there are four options from the source to the destination, $S \to B \to C \to D$, $S \to B \to D$, $S \to C \to D$, and $S \to D$. According to the hop selection metric in (3.1), given $n = 2$ for example, the weights of the options are in the following sequence $W_{S\to D} > W_{S\to B\to D} = W_{S\to C\to D} > W_{S\to B\to C\to D}$ while $W_{S\to B\to C\to D} > W_{S\to C\to D} = W_{S\to B\to D} > W_{S\to D}$ if the item of d^n ($n \geq 2$) in (3.1) is replaced by link length d. Due to the severe propagation loss at mmWave band, the option $S \to B \to C \to D$ is more likely to achieve higher flow throughput in comparison with the others. The hops with minimum summation weights from the source to the destination are chosen for the traffic flow to transmit data. The hops selected for each traffic flow heavily depend on the network topology, but the proposed metric provides important insights on hop selection in densely deployed networks: 1) the selected relay nodes should be close to or on the line between the source and the destination; 2) the accumulated weights achieve the minimum value if the lengths of selected hops for each flow are almost the same, based on the inequality of arithmetic and geometric means (AM-GM inequality); and 3) with a greater number of short hops for each traffic flow, the summation of link length to the power n for each flow decreases while the summation of node loads increases. This tradeoff bounds the number of hops for each flow.

3.1.2 Impact of Multi-hop on Concurrent Transmissions

We use multi-hop transmissions to address the link outage problem and combat severe propagation loss at mmWave band to improve flow throughput. Concurrent transmissions are more favorable than serial TDMA transmissions in terms of network throughput [19,21]. In this section, we analyze the performance of the proposed multi-hop transmission scheme on spatial reuse and time division multiplexing gain.

Spatial Multiplexing

To exploit the spatial multiplexing gain, communication links that do not interfere with each other can operate concurrently over the same mmWave channel. With short link length, the overlap area between the beamwidth of the transceivers and the area of WPAN is likely to be smaller in comparison to that with long link length. Therefore, there are probably more concurrent links in the whole area, according to the concurrent transmission condition. Hence, the channel can be utilized more efficiently by replacing a single long hop with multiple short hops using the relaying mechanism. In the following, we derive the average concurrent transmission probability as a function of link lengths to show the spatial multiplexing gain.

Consider an indoor circle area S with radius R. As shown in Fig. 3.2, there are two pairs of transmitters and receivers randomly distributed in the circle area S. The transmitter and the receiver direct their beams to each other for data transmission. S_{R1} is the overlap area between S and the beamwidth of receiver R_1, while S_{T2} is the overlap area between S and the beamwidth of transmitter T_2. The probability that transmitter T_2 is outside the beamwidth of receiver R_1 is $1-\frac{S_{R1}}{S}$. The probability that T_2 is located in the beamwidth of R_1 but does not direct its beam to R_1 is $\frac{S_{R1}}{S}(1-\frac{S_{T2}}{S})$. Then the probability

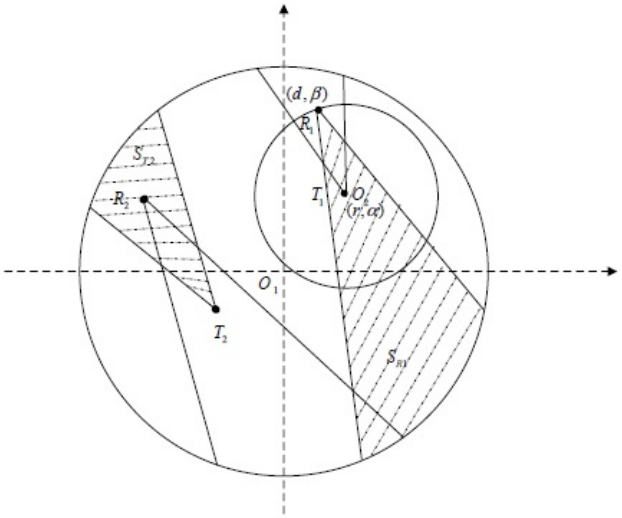

Figure 3.2: Concurrent transmissions Scheduling

that transmitter T_2 does not interfere with receiver R_1 is

$$\begin{aligned} P &= (1 - \frac{S_{R1}}{S}) + \frac{S_{R1}}{S}(1 - \frac{S_{T2}}{S}) \\ &= 1 - \frac{S_{R1} \times S_{T2}}{S^2} \end{aligned} \tag{3.2}$$

Similarly, the probability that transmitter T_1 does not interfere with receiver R_2 is $P' = 1 - \frac{S_{R2} \times S_{T1}}{S^2}$. Therefore, $Q = PP'$ is the probability that two links do not conflict with each other, thus they can operate concurrently.

A polar coordinate system O_1 is established with the center of circle area S as the origin. Then the position of transmitter T_1 is given by (r, α) in system O_1. Similarly, a polar coordinate system O_2 is built with transmitter T_1 as the origin and the position of the corresponding receiver R_1 is given by (d, β) in the polar coordinate system O_2. d is the distance between transmitter T_1 and receiver R_1, i.e., link length. Due to symmetry, we only consider the case $0 \leq \alpha \leq \frac{\pi}{2}$. To make the analysis tractable, the overlapping area

of S_{R1} is approximated as a triangle as follows,

$$\begin{cases} \frac{1}{2}g(a,b,\theta)g(a,b,-\theta), & -\frac{\pi}{2}+\frac{\theta}{2}<\beta<\frac{\pi}{2}-\frac{\theta}{2} \\ -\frac{1}{2}g(a,b,\theta)g(-a,-b,-\theta), & \frac{\pi}{2}-\frac{\theta}{2}<\beta<\frac{\pi}{2}+\frac{\theta}{2} \\ \frac{1}{2}g(-a,-b,\theta)g(-a,-b,-\theta), & \frac{\pi}{2}+\frac{\theta}{2}<\beta<\frac{3\pi}{2}-\frac{\theta}{2} \\ -\frac{1}{2}g(-a,-b,\theta)g(a,b,-\theta), & \frac{3\pi}{2}-\frac{\theta}{2}<\beta<\frac{3\pi}{2}+\frac{\theta}{2} \end{cases} \tag{3.3}$$

where $a = r\cos\alpha + d\cos\beta$, $b = r\sin\alpha + d\sin\beta$, θ is the antenna beamwidth, and the function $g(a,b,\theta)$ is defined as

$$\begin{aligned} g(a,b,\theta) = \cos(\beta-\frac{\theta}{2}) \mid \sqrt{\frac{R^2}{\cos^2(\beta-\frac{\theta}{2})} - [a\tan(\beta-\frac{\theta}{2})-b]^2} \\ +a+b\tan(\beta-\frac{\theta}{2}) \mid \end{aligned} \tag{3.4}$$

Substituting (3.4) into (3.3), S_{R1} becomes a function of r, α, d, β and θ, i.e., $S_{R1} = G(r,\alpha,d,\beta,\theta)$. To obtain the average probability $\overline{P}$, we have

$$\overline{P} = 1 - \frac{\overline{S_{R1}} \times \overline{S_{T2}}}{S^2}. \tag{3.5}$$

In the following, we derive $\overline{S_{R1}}$ and $\overline{S_{T2}}$ to obtain $\overline{P}$. Integrating $G(r,\alpha,d,\beta,\theta)$ with respect to β yields

$$G(r,\alpha,d,\theta) = \int_{\beta} f(\beta)G(r,\alpha,d,\beta,\theta)d\beta \tag{3.6}$$

where $f(\beta)$ is the probability density function of β. The receiver is uniformly distributed on the circle with d as the radius and the transmitter as the origin, or on part of the circle

within the WPAN area if $R - r < d < R + r$, i.e., the two circles in Fig. 3.2 intersect each other. Therefore, the pdf of β is

$$f(\beta) = \begin{cases} \dfrac{1}{2\pi}, & (0 \le \beta < 2\pi,\ 0 < d \le R - r) \\ \dfrac{1}{\beta_2 - \beta_1}, & (\beta_1 \le \beta \le \beta_2,\ R - r < d < R + r) \\ 0, & (R + r \le d) \end{cases} \tag{3.7}$$

where β_1 is determined by $\cos\beta_1 = \frac{(R^2-r^2-d^2)\cos\alpha - \sin\alpha\sqrt{4R^2r^2-(R^2+r^2-d^2)^2}}{2rd}$ and $\sin\beta_1 = \frac{(R^2-r^2-d^2)\sin\alpha + \cos\alpha\sqrt{4R^2r^2-(R^2+r^2-d^2)^2}}{2rd}$ while $\cos\beta_2 = \frac{(R^2-r^2-d^2)\cos\alpha + \sin\alpha\sqrt{4R^2r^2-(R^2+r^2-d^2)^2}}{2rd}$ and $\sin\beta_2 = \frac{(R^2-r^2-d^2)\sin\alpha - \cos\alpha\sqrt{4R^2r^2-(R^2+r^2-d^2)^2}}{2rd}$ give the value of β_2. Then, $\overline{S_{R1}}$ is given by

$$\begin{aligned} \overline{S_{R1}} &= G(d,\theta) \\ &= \int_0^R \int_0^{2\pi} f(r,\alpha) G(r,\alpha,d,\theta) d\alpha dr \\ &= 4\int_0^R \int_0^{\frac{\pi}{2}} f(r,\alpha) G(r,\alpha,d,\theta) d\alpha dr \end{aligned} \tag{3.8}$$

where $f(r,\alpha)$ is the joint pdf of r and α. The transmitter is uniformly distributed in the WPAN area, thus the joint probability density function of r and α is

$$f(r,\alpha) = \frac{r}{\pi R^2} \quad (0 \le \alpha < 2\pi,\ 0 < r \le R) \tag{3.9}$$

Similarly, we can also obtain $\overline{S_{T2}}$ as a function of link length d' between R_2 and T_2. Therefore, the average probability $\overline{P}$ is a function of the link length d and d'. Following the same procedure, we can also obtain $\overline{P'}$ as a function of link length d' and d. The average probability that two links can operate concurrently ($\overline{Q} = \overline{PP'}$) is a function of the link lengths of the two links. This function indicates how the spatial multiplexing gain varies with link lengths on average. Fig. 3.3 shows the numerical results of the average probability $\overline{Q}$ that concurrent transmission in two communication links can take place as a function of

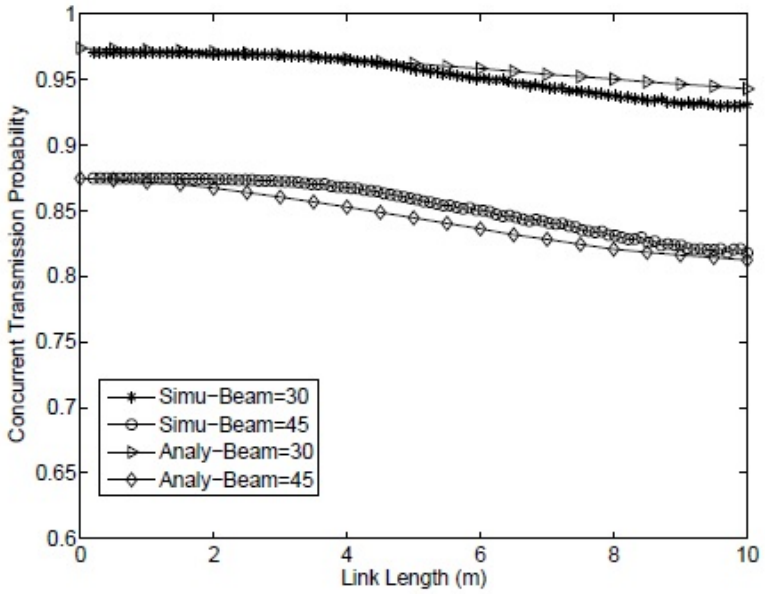

Figure 3.3: Probability Q versus link length

the link lengths with different antenna beamwidths (to simply show the numerical results, let $d = d'$). The probability $\overline{Q}$ decreases as the link length becomes larger. Therefore, multi-hop transmission can exploit more spatial reuse, i.e., it can support more concurrent links in the network.

Time Division Multiplexing

The data transmission rate is a function of the link length. Thus, data transmissions over each link may require a different number of time slots. In Fig.3.4 (a), a schedule is shown for single-hop concurrent transmission. The required number of time slots of links 2, 4, 5 and 8 are far fewer than the number of time slots reserved for the groups they are scheduled in. It can be expected to utilize the resource more efficiently by properly breaking the single

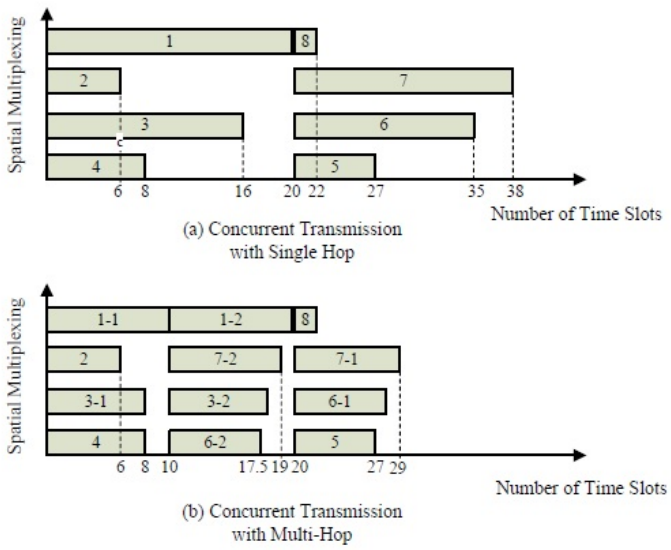

Figure 3.4: Illustration for time division multiplexing

long hop into multiple short hops. For example, as shown in Fig. 3.4 (b), by breaking a long hop (i.e., hops 1, 3, 6 and 7) into two short hops, the total number of time slots for eight flows is reduced from 38 to 29 although the number of groups increases from 2 to 3. Replacing single long hops with multiple short hops can save transmission time, thus the flow throughput and network throughput are enhanced greatly.

To make the performance of single hop concurrent transmission and multi-hop concurrent transmission schemes comparable with time division multiplexing, we consider that the average number of concurrent links for each transmission group in multi-hop transmission is the same as that in single-hop transmission. N flows are divided into N' ($N' > N$) links for multi-hop transmissions. To make the analysis tractable, we study a conservative case that there are no concurrent transmissions for intra-flow links. In this case, all the links in the same group are independent. In 5G cellular networks, mmWave base stations are densely deployed in small cell. Thus, only a few links are involved in each flow. The

limited number of hops for each flow and the half-duplex transmission make the intra-flow links difficult to operate concurrently.

Let X denote the number of reserved slots for a group, which is defined as the maximum number of required time slots among all the links in the group. The cumulative distribution function (cdf) of X is given by

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(X_1 \leq x, X_2 \leq x \dots X_M \leq x) \end{aligned} \tag{3.10}$$

where X_m $(m = 1, 2 \dots M)$ is the number of slots requested by the m^{th} link in the group. The X_m's are i.i.d. random variables with pdf $\widetilde{f}(x)$ for single-hop concurrent transmission and pdf $\overline{f}(x)$ for multi-hop concurrent transmission. Therefore, we have

$$F_X(x) = \prod_{m=1}^{M} P(X_m \leq x) = [\widetilde{F}(x)]^M \tag{3.11}$$

where $\widetilde{F}(x) = \int_0^x \widetilde{f}(x)dx$. The pdf of X is thus given by

$$f_X(x) = \frac{dF_X(x)}{dx} = M\widetilde{f}(x)[\widetilde{F}(x)]^{M-1} \tag{3.12}$$

The expected number of time slots reserved for each group is

$$\begin{aligned} E[X] &= \int_{x_a}^{x_b} x f_X(x)dx \\ &= \int_{x_a}^{x_b} xM\widetilde{f}(x)[\widetilde{F}(x)]^{M-1}dx \end{aligned} \tag{3.13}$$

where x_a and x_b are respectively the minimum and maximum required numbers of slots among all the hops in a group. Then we have $\widetilde{F}(x_a) = 0$ and $\widetilde{F}(x_b) = 1$. The total required number of slots for N flows using single-hop concurrent transmission is

$$T_{SHCT} = \frac{N}{M}E[X] = N\int_{x_a}^{x_b} x[\widetilde{F}(x)]^{M-1}\widetilde{f}(x)dx \tag{3.14}$$

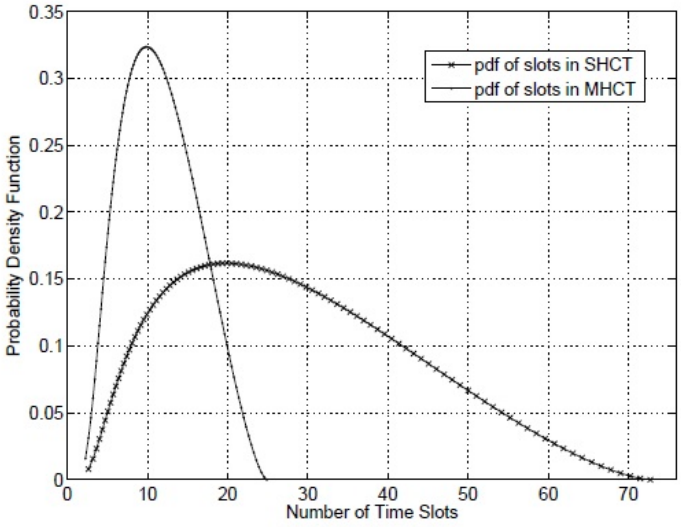

Figure 3.5: Probability density function of required number of time slots

Similarly, the total number of time slots for N flows using multi-hop concurrent transmission is given by

$$T_{MHCT} = N' \int_{x'_a}^{x'_b} x[\overline{F}(x)]^{M-1}\overline{f}(x)dx \tag{3.15}$$

where $\overline{F}(x) = \int_0^x \overline{f}(x)dx$.

In (3.15), the total number of time slots for N flows with multi-hop concurrent transmission depends on the number of hops, the average number of concurrent transmissions, and the required number of time slots for each hop. Therefore, the hop selection metric plays a crucial role in achieving the time division multiplexing gain in multi-hop concurrent transmission. In [32], the pdf of the distance between two nodes randomly picked in a circle area with radius R is $f(d) = \frac{4d}{\pi R^2}cos^{-1}(\frac{d}{2R}) - \frac{2d^2}{\pi R^3}\sqrt{1-\frac{d^2}{4R^2}}$. The pdf of link length in the multi-hop concurrent transmission (MHCT) is obtained by shrinking the pdf of link length in single-hop concurrent transmission (SHCT) by a half, $f(d) = \frac{16d}{\pi R^2}cos^{-1}(\frac{d}{R}) - \frac{16d^2}{\pi R^3}\sqrt{1-\frac{d^2}{R^2}}$.

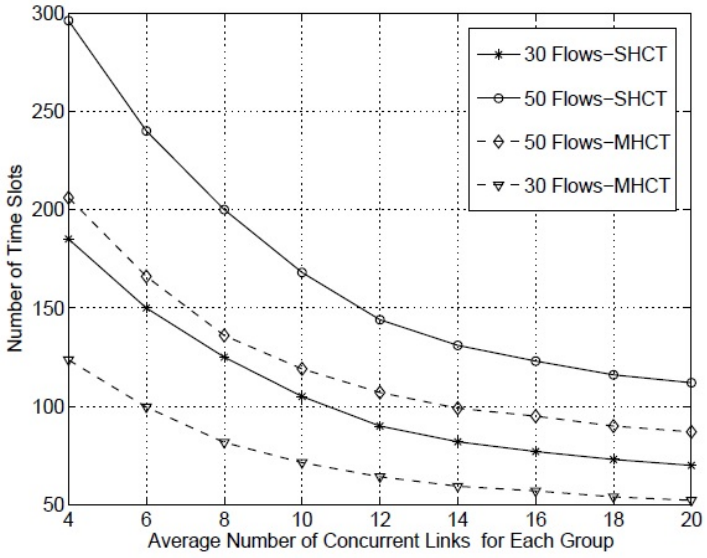

Figure 3.6: Time division multiplexing

Without loss of generality, we choose $N' = 2N$ to show the numerical results. The transmission rate is calculated according to (2.3). With the same traffic demands for each flow, we can get $\overline{f}(x)$ and $\widetilde{f}(x)$ numerically as shown in Fig. 3.5. The required numbers of time slots for various number of traffic flows with the MHCT scheme and the SHCT scheme are shown in Fig. 3.6. We can see that the proposed multi-hop concurrent transmission scheme greatly improves the time division multiplexing gain. With the MHCT scheme, it takes less time to transmit the same number of traffic flows, thus the network throughout and flow throughput are enhanced.

3.2 Video Quality Provisioning with High Mobility

For high-mobility scenarios, We embed a buffer in mobile device to store the data transmitted with extra allocated mmWave bandwidth when mobile user is in the areas with

mmWave connection available. Larger allocated mmWave bandwidth can result in more data stored in the buffer, and thus the mobile users can maintain video playout quality for a specific time duration even they move into areas where mmWave connection is unavailable. On the other hand, with larger allocated mmWave bandwidth to each user, fewer users can be supported by the network, causing inefficient resource utilization. Additionally, due to finite buffer size, larger allocated bandwidth can result in packet overflow and quality loss for the multimedia applications. Therefore, how to properly allocate the mmWave bandwidth to dynamically charge or discharge the buffer to maintain the playout quality of multimedia applications is an important and challenging issue.

3.2.1 Related Work

Quality provisioning in 3G/4G cellular networks has been intensively studied in [33–38]. Quality provisioning for various kinds of applications is achieved in 3G/4G cellular networks by load-balancing between cellular networks and wireless local area networks (WLANs). One category of research assumes the mobile users can obtain the bandwidth from a single access network for a specific application [33–35]. In [33], the economic model is used to allocate radio resource between CDMA-based 3G cellular network and WLAN network, based on pricing and system utilities. The proposed admission control scheme in [33] aims to optimize the total welfare of the two networks considering the characteristics of each network. In [34], the performance of WLAN-first admission scheme is analyzed for cellular and WLAN integrated networks for voice and low-rate data services (such as Web browsing). In [35], a generalized admission control scheme is considered to analyze the dependence of resource utilization and the impact of user mobility and traffic characteristics on admission parameters. The resource allocation solutions belonging to this category consider the available networks individually to satisfy the bandwidth requirements of each

application, without fully exploiting the available resources of multiple networks.

Another category of resource allocation for quality provisioning in 3G/4G cellular networks assumes that each mobile station can jointly utilize the available resources from multiple networks (referred to as multihoming capability) [36, 37, 39]. In [39], multimedia applications attain the bandwidth from 4G cellular network and WLAN to maximize the system utility based on the load characteristics of each network. Literature [36] provides a practical resource allocation scheme in a distributed manner for 3G mobile users with multihoming capability, considering both constant bit rate (CBR) and variable bit rate (VBR) traffic. Most of the existing works on quality provisioning for 3G/4G cellular network focus on how to allocate the resources of the heterogeneous networks without addressing link outage problem which does not occur frequently in microwave frequency band.

The link outage problem has been investigated in mmWave WLANs/WPANs [22, 40]. A hop selection metric is proposed to select appropriate relay hops to replace the blocked link and improve flow throughput [22]. A multi-hop concurrent transmission scheduling scheme is then proposed to allow multiple hops to operate simultaneously in order to improve system capacity. In [40], cross-layer modeling and design approaches are used to address the problem of directionality and link blockage due to the limited ability to diffract around obstacles. It uses minimum number of hops for data transmission. Specifically, if the LOS link is available between the source node and the destination node, single hop transmission is adopted; otherwise, an intermediate node is selected as the relay to build a two-hop path to replace the blocked link. The relaying method is not suitable to solve mmWave link outage problem in 5G cellular networks since it involves long setup time. The relaying method needs to discover the blocked link, select the relays, and re-build the path. These procedures need much longer time to setup the alternative path especially with the directional communications. Users with high mobility suffer from very frequent link

outage and the relaying method should be repeated very frequently, resulting in significant video quality degradation. To the best of our knowledge, an effective solution for mmWave link outage problem in 5G cellular networks is highly in demand to maintain the playout quality for high-rate multimedia applications.

3.2.2 Buffer System

Frequent link outage could occur to mmWave signal because of the limited diffraction and penetration capability. mmWave transmission rate could descend dramatically with link outage, which results in video frozenness. To improve user's video playout quality, we propose an integrated system with playout buffer to deal with mmWave link outage problem resulting from the variations of mmWave channel. As shown in Fig. 3.7, the data is transmitted through mmWave channel to the buffer with finite buffer size. Since mmWave connection is considered to be available if the path loss is under a specific threshold, it is assumed that the data packets transmitted to the buffer through mmWave channel can be decoded correctly and video packet loss comes from the buffer overflow. The data packets in the buffer are combined into frames and injected into the video player. Video playout quality can be maintained for a specific period after mmWave link outage occurs, with the data packets stored in the buffer. Mobile users move alternatively between double-coverage areas and single-coverage areas, and the playout buffer can be charged or discharged dynamically. The video can suffer frozenness if the buffer is empty, and video has quality loss if buffer is overflowed.

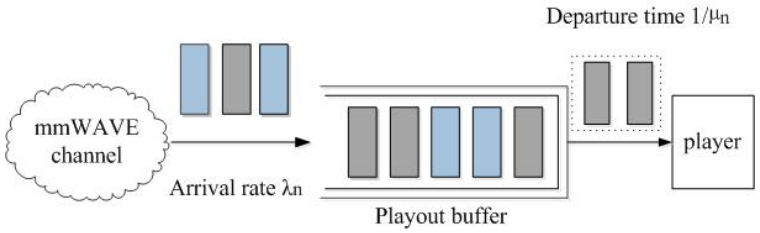

Figure 3.7: Buffer System to Maintain QoS at End User

3.2.3 Problem Formulation

In this section, the problem of bandwidth allocation and video playout quality provisioning for $\mathcal{N}$ mobile users dynamically moving in the whole coverage area is formulated as a MDP problem composed of *the state space*, *the control space*, *the state transition probabilities* and *the cost function*.

A two-dimensional 5G system with a set of M mmWave base stations ($B = \{B_1, B_2, ...B_M\}$) is considered in this paper. Each mmWave base station B_m has a specific capacity C_m with $C_m \in \mathfrak{C} = \{C_1, C_2, ...C_M\}$. The capacity C_m can be affected by several factors, such as transmission range and building distributions in the cell. It is assumed that the capacity set $\mathfrak{C}$ is available for a given 5G system. The m^{th} cell covered by base station B_m is partitioned into S_m regions. The total numfber of regions is $S = \sum_{m=1}^{M} S_m$ and we denote each region as $\boldsymbol{s} \in \mathcal{S} = \{1, 2, 3, ..., S\}$. Each region $\boldsymbol{s}$ is small enough so that the propagation effects on all signals in region $\boldsymbol{s}$ can be considered to be equivalent. Therefore, the users in each region $\boldsymbol{s}$ can be assumed to have the same instantaneous transmission rate. With AWGN channel, the instantaneous transmission data rate (denoted as $R_{\boldsymbol{s}}$) for all users in each region $\boldsymbol{s}$ can be estimated by Shannon capacity formula. Since mmWave network adopts TDMA-based [18] (or STDMA-based [4,41]) MAC, the allocated bandwidth to the n^{th} mobile user U_n in region $\boldsymbol{s}$ can be a portion of the instantaneous rate $R_{\boldsymbol{s}}$. Given a set of

available rates in each region $\mathbb{R}_{\boldsymbol{s}} = \{\frac{1}{\mathcal{K}}R_{\boldsymbol{s}}, \frac{2}{\mathcal{K}}R_{\boldsymbol{s}}, ..., R_{\boldsymbol{s}}\}$, the allocated bandwidth to the n^{th} mobile user U_n in region $\boldsymbol{s}$ denoted by $B_{n,\boldsymbol{s}}$ could be $B_{n,\boldsymbol{s}} = \frac{k}{\mathcal{K}}R_{\boldsymbol{s}}$ with $k = 1, 2, ..., \mathcal{K}$, where k is the number of time slots allocated to mobile user U_n and $\mathcal{K}$ is the total number of time slots in each frame. To make the analysis trackable, $R_{\boldsymbol{s}} = 0$ if region $\boldsymbol{s}$ is a single-coverage area.

The State Space

Given a set of regions in the whole area $\mathcal{S} = \{1, 2, 3, ..., \boldsymbol{s}, ..., S\}$ and a set of active users $\mathcal{U} = \{U_1, U_2, ..., U_n, ...U_{\mathcal{N}}\}$, bandwidth allocation in 5G system at any particular time is then characterized by a $\mathcal{N} \times S$ matrix:

$$\mathbb{B} = \| B_{n,\boldsymbol{s}} \|_{\mathcal{N}\times S} = \begin{Vmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,S} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,S} \\ \vdots & \vdots & \ddots & \vdots \\ B_{\mathcal{N},1} & B_{\mathcal{N},2} & \cdots & B_{\mathcal{N},S} \end{Vmatrix} \tag{3.16}$$

where $B_{n,\boldsymbol{s}}$ is the allocated bandwidth to user U_n in region $\boldsymbol{s}$ and is defined as

$$B_{n,\boldsymbol{s}} = \begin{cases} \dfrac{k}{\mathcal{K}}R_s, & (U_n \text{ in } \boldsymbol{s}) \\ 0, & (\text{otherwise}) \end{cases} \tag{3.17}$$

$B_{n,\boldsymbol{s}} = 0$ occurs either user U_n is not in region $\boldsymbol{s}$ or mmWave connection is unavailable in region $\boldsymbol{s}$ even if user U_n is located in region $\boldsymbol{s}$. Since each user U_n can only locate in one specific region at any particular time, at most one element in each row of matrix $\mathbb{B}$ is non-zero.

To describe the random movement of the users $\mathcal{U} = \{U_1, U_2, ..., U_n, ...U_{\mathcal{N}}\}$ among the

regions $\mathcal{S} = \{1, 2, 3, ..., \boldsymbol{s}, ..., S\}$, the following $\mathcal{N} \times S$ matrix is defined,

$$\mathbb{W} = \| w_{n,\boldsymbol{s}} \|_{\mathcal{N}\times S} = \begin{Vmatrix} w_{1,1} & \cdots & w_{1,S} \\ \vdots & \ddots & \vdots \\ w_{\mathcal{N},1} & \cdots & w_{\mathcal{N},S} \end{Vmatrix} \tag{3.18}$$

where

$$w_{n,\boldsymbol{s}} = \begin{cases} -1, & (U_n \text{ moves out of } \boldsymbol{s}) \\ 1, & (U_n \text{ moves into } \boldsymbol{s}) \\ 0, & \text{(otherwise)}. \end{cases} \tag{3.19}$$

For the n^{th} user U_n moving from region $\boldsymbol{s}$ to $\boldsymbol{s}^{'}$, we have $w_{n,\boldsymbol{s}} = -1$ and $w_{n,\boldsymbol{s}'} = 1$. Since the matrix $\mathbb{W}$ describes the movement for a particular user U_n, there are only two elements in the matrix $\mathbb{W}$ with non-zero value and all the other elements are zero.

The state of mmWave bandwidth allocation for all the mobile users in 5G cellular networks is composed of the following two components:

- The primary state component $\mathbb{B}$ describes the allocated bandwidth to all the users $\mathcal{U}$ in the whole area $\mathcal{S}$;
- The random state component $\mathbb{W}$ describes the movement of a particular user among different regions.

Therefore, the state space $\mathcal{D}$ composed of all possible combinations of the primary state component and random state component can be given by

$$\mathcal{D} = \{i = (\mathbb{B}, \mathbb{W}) \mid \mathbb{B} \in \mathcal{B}, \mathbb{W} \in \mathcal{W}\} \tag{3.20}$$

where $\mathcal{B}$ is the set of all bandwidth allocation solutions based on the available rates $\mathbb{R}_{\boldsymbol{s}}$ in each region, and $\mathcal{W}$ is the set of all events consisting of a single user moving from one region to another.

The time duration for the user U_n staying in region $\boldsymbol{s}$ until it moves to region $\boldsymbol{s}^{'}$ is independent of the previous visit regions and the previous length of connection, and is exponentially distributed with rate $\xi_{n\boldsymbol{s}\boldsymbol{s}'}$. The probability that a mobile user U_n currently in region $\boldsymbol{s}$ moves to region $\boldsymbol{s}^{'}$ is independent of the regions U_n has already visited and is given to be $\boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'}$ with $\sum_{\boldsymbol{s}'} \boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'} = 1$. Based on these assumptions, the next state $j = (\mathbb{B}, \mathbb{W})$ only depends on the current state i and the allocated bandwidth for the user moving into the new region, i.e., Markov property.

This paper considers the general cases that the knowledge of where a particular user U_n is heading or the history of U_n's locations are unavailable. However, our methodology extends to cases where such information is available. The statistical parameters regarding these information, such as $\boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'}$ and $\xi_{n\boldsymbol{s}\boldsymbol{s}'}$, can be obtained either through measurements or other means. Since we consider the bandwidth allocation problem for a number of mobile users in 5G cellular networks, the admission control for new transmission arrivals and connection endings for existing users are beyond the scope of this paper.

The Control Space

For each state $i = (\mathbb{B}, \mathbb{W}) \in \mathcal{D}$, the available control set $A(i)$ is the available rate set in the region $\boldsymbol{s}^{'}$ if region $\boldsymbol{s}^{'}$ is the moving-into region determined by $\mathbb{W}$. Specifically, $A(i) = \mathbb{R}_{\boldsymbol{s}'} = \{\frac{1}{\mathcal{K}} R_{\boldsymbol{s}'}, \frac{2}{\mathcal{K}} R_{\boldsymbol{s}'}, ..., R_{\boldsymbol{s}'}\}$ if $w_{n,\boldsymbol{s}'} = 1$ with $w_{n,\boldsymbol{s}'} \in \mathbb{W} \in i$. Therefore, the control space is

$$\mathcal{A} = \bigcup_{i \in \mathcal{D}} A(i) = \bigcup_{\boldsymbol{s}' \in \mathcal{S}} \mathbb{R}_{\boldsymbol{s}'} \tag{3.21}$$

Given the state i with the random component $\mathbb{W}$ describing user's movement from region $\boldsymbol{s}$ to region $\boldsymbol{s}^{'}$, the control of state i determines a selected rate from the set of available rates $\mathbb{R}_{\boldsymbol{s}'}$ in region $\boldsymbol{s}^{'}$. The selected rate is the newly allocated bandwidth to the user U_n

moving into region $\boldsymbol{s}'$.

The selected rate for each user passing different regions has significant impact on the playout quality of multimedia applications. Specifically, if the packet arrival rate into the buffer (which can be translated from the selected rate) is less than the packet departure rate, the playout buffer would be discharged and it is possible that the video is frozen when the buffer is empty. The larger allocated bandwidth can smooth the playout quality and charge the buffer, while it would result in less bandwidth allocated to other users in the same region, which causes frozenness for other users. Therefore, how to select the bandwidth from available rate set for the users going through different regions to optimize the playout quality of all the users is a challenging and important issue. Note that all the possible controls would be zero if mmWave signal is unavailable in region $\boldsymbol{s}'$.

The State Transition Probabilities

For any bandwidth allocation configuration $\mathbb{B}$, the time duration until the next random event $\mathbb{W}$ occurs depends on the movements of all the users in the network. For each user, the arrival of random event is a Poisson process with rate of $\boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'}\xi_{n\boldsymbol{s}\boldsymbol{s}'}$. The overall rate $\Lambda_{\mathbb{B}}$ at which random events occur starting from a bandwidth allocation configuration $\mathbb{B}$ is the sum of the rates of all possible events and is given by

$$\Lambda_{\mathbb{B}} = \sum_{n=1}^{\mathcal{N}} \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \mathrm{I}_{n,\boldsymbol{s}}\, \boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'}\xi_{n\boldsymbol{s}\boldsymbol{s}'} \tag{3.22}$$

where $\mathrm{I}_{n,\boldsymbol{s}}$ indicates if user U_n is in the $\boldsymbol{s}^{th}$ region and defined as

$$\mathrm{I}_{n,\boldsymbol{s}} = \begin{cases} 1, & (U_n \text{ in } \boldsymbol{s}) \\ 0, & (\text{otherwise}) \end{cases} \tag{3.23}$$

Assuming the control takes effect immediately, the total state transition rate from state i under control $a \in A(i)$ to next state is then $\Lambda_{\mathbb{B}}$. We denote the expected time duration of the transition from state i under control a to the next state as $\overline{\tau}_i(a)$

$$\overline{\tau}_i(a) = \frac{1}{\Lambda_{\mathbb{B}}} \tag{3.24}$$

For state i, given the random state component $\mathbb{W}^{(i)}$ with $w_{n,\boldsymbol{s}}^{(i)} = -1$ and $w_{n,\boldsymbol{s}'}^{(i)} = 1$, the primary sate component $\mathbb{B}^{(i)}$, and the control $a \in A(i) = \mathbb{R}_{\boldsymbol{s}'} = \{\frac{1}{\mathcal{K}}R_{\boldsymbol{s}'}, \frac{2}{\mathcal{K}}R_{\boldsymbol{s}'}, ..., R_{\boldsymbol{s}'}\}$, the next primary state component $\mathbb{B}^{(j)} = \| B_{\widehat{n},\widehat{\boldsymbol{s}}} \|_{\mathcal{N}\times S}$ can be given as

$$B_{\widehat{n},\widehat{\boldsymbol{s}}}^{(j)} = \begin{cases} B_{n,\boldsymbol{s}'}^{(i)} + a * w_{n,\boldsymbol{s}'}^{(i)}, & (\widehat{n} = n, \widehat{\boldsymbol{s}} = \boldsymbol{s}') \\ B_{n,\boldsymbol{s}}^{(i)}(1 + w_{n,\boldsymbol{s}}^{(i)}), & (\widehat{n} = n, \widehat{\boldsymbol{s}} = \boldsymbol{s}) \\ B_{\widehat{n},\widehat{\boldsymbol{s}}}^{(i)}, & \text{(otherwise)}. \end{cases} \tag{3.25}$$

The probability distribution of the next random event $\mathbb{W}^{(j)}$ depends on the bandwidth allocation configuration resulting from the current state i and control $a \in A(i)$. Given $\mathbb{B}^{(j)}$ by (3.25), the state transition probability from state $i = (\mathbb{B}^{(i)}, \mathbb{W}^{(i)}) \in \mathcal{D}$ to state $j = (\mathbb{B}^{(j)}, \mathbb{W}^{(j)}) \in \mathcal{D}$ is the rate at which the event occurs divided by the total transition rate:

$$p(j \mid i, a) = p_{i,j}(a) = \frac{1}{\Psi} \frac{\mathrm{I}_{n,\boldsymbol{s}} \boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'} \xi_{n\boldsymbol{s}\boldsymbol{s}'}}{\sum_{n=1}^{\mathcal{N}} \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \mathrm{I}_{n,\boldsymbol{s}} \boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'} \xi_{n\boldsymbol{s}\boldsymbol{s}'}} \tag{3.26}$$

where $1/\Psi$ is the probability of selecting one rate from Ψ available rates in region $\boldsymbol{s}'$ if region $\boldsymbol{s}'$ is double-coverage area with $R_{\boldsymbol{s}'} > 0$. Ψ is maximum integer within the range of $0 \leq \Psi \leq \mathcal{K}$ satisfying the capacity constraint of mmWave base station. Specifically, assuming the moving in region $\boldsymbol{s}'$ is within the coverage of mmWave base station B_m, Ψ is the maximum k satisfying the constraint of

$$(\sum_{\boldsymbol{s}\in B_m} \sum_{n=1}^{\mathcal{N}} B_{n,\boldsymbol{s}}) - B_{n,\boldsymbol{s}} + \frac{k}{\mathcal{K}} R_{\boldsymbol{s}'} \leq C_m \tag{3.27}$$

if the moving out region $\boldsymbol{s}$ and the moving in region $\boldsymbol{s}^{'}$ are within the same mmWave base station B_m. Otherwise, if only the moving in region $\boldsymbol{s}^{'}$ is within the coverage of mmWave base station B_m, the constraint should be

$$(\sum_{\boldsymbol{s}\in B_m}\sum_{n=1}^{\mathcal{N}} B_{n,\boldsymbol{s}}) + \frac{k}{\mathcal{K}} R_{\boldsymbol{s}'} \leq C_m. \tag{3.28}$$

For the scenario that the moving in region $\boldsymbol{s}^{'}$ is single-coverage area, the only control we can select is $a = 0$. Thus, the state transition probability is

$$p(j \mid i, a) = p_{i,j}(a) = \frac{\mathrm{I}_{n,\boldsymbol{s}} \boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'} \xi_{n\boldsymbol{s}\boldsymbol{s}'}}{\sum_{n=1}^{\mathcal{N}} \sum_{\boldsymbol{s}\in\mathcal{S}} \sum_{\boldsymbol{s}'\in\mathcal{S}} \mathrm{I}_{n,\boldsymbol{s}} \boldsymbol{q}_{n\boldsymbol{s}\boldsymbol{s}'} \xi_{n\boldsymbol{s}\boldsymbol{s}'}} \tag{3.29}$$

with $R_{\boldsymbol{s}'} = 0$.

The Cost Function

The objective is to properly select controls at each state to maximize the playout quality, which is the smooth playout duration over the whole transmission period for all the users in the network under the capacity constraint of each mmWave base station B_m. To evaluate the smoothness of multimedia applications at the user, we define ***quality consistency*** as the percentage of time when there is no frozenness or/and no packet dropping at the user out of the total operation period of video traffic.

With the proposed buffer system, there are $\mathcal{N}$ queues in the system with each queue corresponding to one user. Fig. 3.8 shows the timing structure of the system and recursive queuing model for the user U_n with packet arrivals, departures, and queue state. The timing of the whole system is divided into stages. Each stage is characterized by a specific state. The system is in the same stage until state transition occurs. A new stage starts (previous stage ends at the same time) when the system enters a new state with any user moving into a new region. We need to allocate a new bandwidth for the user moving from

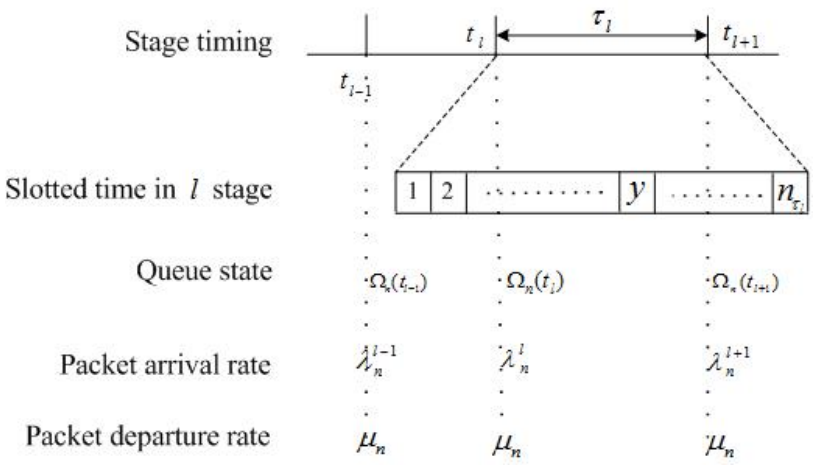

Figure 3.8: The Recursive Queue Model for User n

previous region to new region. In other words, in each stage, all the users utilize the same bandwidths as those in previous stage except for the user triggering the new stage. Let t_l denote the starting time of the l^{th} stage of the MDP, thus $\tau_l = (t_{l+1} - t_l)$ is the duration for the l^{th} stage. Time set $\mathcal{T} = \{t_1, t_2, ..., t_l, ..., t_L\}$ indicates the starting time for each stage where L is the total number of stages. For the l^{th} stage corresponding to the state with user U_n moving from region $\boldsymbol{s}$ to region $\boldsymbol{s}'$, the stage duration can be given as

$$\tau_l = \frac{1}{\xi_{nss'}} \tag{3.30}$$

The queue state at the beginning of each stage is used in the cost function as shown in the following of the paper. Therefore, we recursively derive the queue state based on the slotted time structure resulting from TDMA-based MAC in mmWave networks. As shown in Fig. 3.8, there are n_{τ_l} time slots during the l^{th} stage and each time slot has a fixed duration T with $\tau_l = T n_{\tau_l}$. It is assumed that the packets move out the queue at the beginning of each time slot and the arriving packets enter the queue throughout the time slot. For user U_n in the y^{th} time slot of the l^{th} stage, let $\Upsilon_n^{(l)}(y)$ denote the number

of packets left in the queue after $V_n^{(l)}(y)$ packets move out of the queue, then we have

$$\Upsilon_n^{(l)}(y) = \max\{0, \Omega_n^{(l)}(y-1) - V_n^{(l)}(y)\} \tag{3.31}$$

where $\Omega_n^{(l)}(y-1)$ is the queue length for user U_n at the end of the $(y-1)^{th}$ time slot. The number of free slots in the queue at the beginning of y^{th} slot is thus:

$$F_n^{(l)}(y) = G - \Upsilon_n^{(l)}(y) = G - \max\{0, \Omega_n^{(l)}(y-1) - V_n^{(l)}(y)\}. \tag{3.32}$$

where G is the buffer size. If the number of arriving packets in the y^{th} slot $A_n^{(l)}(y)$ is larger than $F_n^{(l)}(y)$, only $F_n^{(l)}(y)$ packets enter the queue and the other $[A_n^{(l)}(y) - F_n^{(l)}(y)]$ packets are dropped. Therefore, the recursion of the queue state can be given as follows,

$$\Omega_n^{(l)}(y) = \min\{G, \max\{0, \Omega_n^{(l)}(y-1) - V_n^{(l)}(y)\} + A_n^{(l)}(y)\}. \tag{3.33}$$

Based on the recursion equation (3.33), we can obtain the queue state $\Omega_n(t)$ for the user U_n at anytime t corresponding to the y^{th} time slot in l^{th} stage, which can be defined as

$$\Omega_n(t) \triangleq \Omega_n^{(l)}(y), \text{with } t \geq 0. \tag{3.34}$$

$\Omega_n(t)$ can be affected by the initial queue state and the packet arrival/departure rates. It is known that the packet delay mainly results from the randomness of packet arrival and departure. Since the emerging high-rate video applications have stringent requirements on delay and data rate, massive packets come intensively to the buffer with much lower randomness. Taking into account the fact that the scale of stage duration is much larger than the scale of packet randomness, the randomness of packet arrival and departure is not considered in the cost of each stage, in order to make the analysis tractable. Let $\mathbb{X}(t)$ represent the state for the system at time t with $\mathbb{X}(t) = (\mathbb{B}^{(t)}, \mathbb{W}^{(t)}) \in \mathcal{D}$, and $a(t)$ represents the control at time t for state $\mathbb{X}(t)$ with $a(t) \in \mathcal{A}$. The allocated bandwidth for

user U_n in the l^{th} stage can be obtained from $\mathbb{X}(t)$ and $a(t)$. Specifically, $\mathbb{B}^{(t)}$ describes the allocated bandwidth for each user in different regions; $\mathbb{W}^{(t)}$ determines the user to re-select the allocated bandwidth; and the control $a(t)$ determines the new allocated bandwidth to the moving user selected by $\mathbb{W}^{(t)}$. Let $\lambda_n^{(l)}$ denote the packet arrival rate of user U_n in the l^{th} stage. $\lambda_n^{(l)}$ can be determined by the bandwidth allocation as follows:

$$\lambda_n^{(l)} = \begin{cases} a(t_l)/Z, \ (\text{if } w_{n,\boldsymbol{s}}^{(t_l)} = 1 \text{ and } w_{n,\boldsymbol{s}'}^{(t_l)} = -1), \\ \sum\limits_{\boldsymbol{s}\in\mathcal{S}} B_{n,\boldsymbol{s}}^{(t_l)}/Z, \qquad\qquad (\text{otherwise}). \end{cases} \tag{3.35}$$

where Z is the packet size.

The video application on each mobile user U_n has a specific data rate requirement r_n with $n \in \{1, 2, ...\mathcal{N}\}$. The packet departure rate for the n^{th} queue does not depends the stage and can be given as $\mu_n = r_n/Z$. With packet arrival rate $\lambda_n^{(l)}$, packet departure rate μ_n, and initial queue state $\Omega_n(t_l)$, the l^{th} stage cost of user U_n (denoted by G_n^l) comes from the following two aspects:

- Quality loss duration is the total time duration when video quality loss results from the arrival packet dropping if the buffer is full.
- Frozenness duration is the total time duration when the video suffers from frozenness if there is no packets in the buffer for playout.

If $\lambda_n^{(l)} \geq \mu_n$, there is no frozenness duration for U_n in the l^{th} stage and the stage cost is the quality loss duration. The time to fully charge the buffer in the l^{th} stage (denoted by $t_c^{(l)}$) with initial buffer state $\Omega_n(t_l)$ is given as

$$t_c^{(l)} = \frac{G - \Omega_n(t_l)}{\lambda_n^{(l)} - \mu_n} \tag{3.36}$$

If the stage duration $\tau_l \leq \frac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}$, there is no packet dropping and the stage cost $G_n^l = 0$. Otherwise, with the stage duration $\tau_l > \frac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}$, the total number of dropping packets is $(\lambda_n^{(l)} - \mu_n)[\tau_l - \frac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}]$. Thus, the stage cost in terms of quality loss duration is given as

$$G_n^l = \frac{(\lambda_n^{(l)} - \mu_n)[\tau_l - \frac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}]}{\mu_n}. \tag{3.37}$$

If $\lambda_n^{(l)} < \mu_n$, there isn't quality loss duration for U_n in the l^{th} stage and the stage cost is the frozenness duration. The time to fully discharge the buffer in the l^{th} stage (denoted by $t_d^{(l)}$) with initial buffer state $\Omega_n(t_l)$ is given as

$$t_d^{(l)} = \frac{\Omega_n(t_l)}{\mu_n - \lambda_n^{(l)}} \tag{3.38}$$

If the stage duration $\tau_l \leq \frac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}$, there are always packets in the buffer to transmit and there isn't frozenness duration in the l^{th} stage with $G_n^l = 0$. Otherwise, with the stage duration $\tau_l > \frac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}$, the total number of extra packets to be provided for smooth quality at user is $(\mu_n - \lambda_n^{(l)})[\tau_l - \frac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}]$. Thus, the stage cost in terms of frozenness duration is given as

$$G_n^l = \frac{(\mu_n - \lambda_n^{(l)})[\tau_l - \frac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}]}{\mu_n}. \tag{3.39}$$

Therefore, the stage cost of user U_n in the l^{th} stage can be summarized as

$$G_n^l = \begin{cases} \dfrac{(\lambda_n^{(l)} - \mu_n)[\tau_l - \frac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}]}{\mu_n}, (\lambda_n^{(l)} \geq \mu_n,\ \tau_l > \dfrac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}), \\ 0, \qquad\qquad (\lambda_n^{(l)} \geq \mu_n,\ 0 < \tau_l \leq \dfrac{G-\Omega_n(t_l)}{\lambda_n^{(l)}-\mu_n}), \\ 0, \qquad\qquad (\lambda_n^{(l)} < \mu_n,\ 0 < \tau_l \leq \dfrac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}), \\ \dfrac{(\mu_n - \lambda_n^{(l)})[\tau_l - \frac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}]}{\mu_n}, \quad (\lambda_n^{(l)} < \mu_n,\ \tau_l > \dfrac{\Omega_n(t_l)}{\mu_n-\lambda_n^{(l)}}). \end{cases} \tag{3.40}$$

Our goal is to select a proper control at each possible state that minimizes the expected average cost per unit time. Given the set of initial queue state for all the users $\{\Omega_1(0), \Omega_2(0), ..., \Omega_{\mathcal{N}}(0)\}$ and the set of required data rates $\{r_1, r_2, r_3, ..., r_{\mathcal{N}}\}$, we can obtain the stage costs for all the users in the whole process including L stages. Therefore, the objective cost function can be the sum of the stage costs of all the users over total time as follows (P1):

$$\lim_{L\to\infty} \frac{1}{E(t_L)} \sum_{l=1}^{L} \sum_{n=1}^{\mathcal{N}} G_n^l. \tag{3.41}$$

Note that the objective function describes the average cost per unit time, resulting from the packet dropping and frozenness. The quality consistency can be obtained by $(1 - \text{P1})$. A policy ψ determines the controls for all the stages of P1. By minimizing P1, we can maximize the playout quality of all the users.

For mmWave 5G cellular networks with dense population, the bandwidth resources are not sufficient for all the users to achieve smooth playout quality. Thus, the dominant cost in each stage is the frozenness duration. It is obvious that the optimal policy ψ^* for P1 will cause a biased solution, i.e., the bandwidth allocation would incline towards those users with lower required data rates while some unlucky users will suffer from lots of frozenness periods. To ensure long-term fairness among the competing users, a parameter ρ based on the weighted fair queuing [42] is adopted. The objective cost function can be given as (P2):

$$\lim_{L\to\infty} \frac{1}{E(t_L)} \sum_{l=1}^{L} \sum_{n=1}^{\mathcal{N}} \rho_n^{(l-1)} G_n^l. \tag{3.42}$$

where $\rho_n^{(l-1)}$ is the coefficient for G_n^l, obtained according to the bandwidth allocation records of previous $(l-1)$ stages for user U_n. Specifically, we have

$$\rho_n^{(l-1)} = \frac{\beta_n}{[(1 - g_n^{(l-1)}) + \epsilon]^\alpha} \tag{3.43}$$

where ϵ is a small positive scalar to prevent zero denominator, β_n is a weight for user U_n to provide differentiated services, and α is a parameter to make a tradeoff between fairness and total cost. $g_n^{(l-1)}$ is the average cost per unit time during the previous $(l-1)$ stages and is given as

$$g_n^{(l-1)} = \frac{\Sigma_{h=1}^{l-1} G_n^h}{\Sigma_{h=1}^{l-1} t_h} \tag{3.44}$$

With a large value of α, users with less accumulated smooth duration (calculated by $1 - g_n^{(l-1)}$) in the previous stages have larger values of ρ, so that they have less chance to obtain cost in the following stages. Thus, better fairness can be achieved. When $\alpha = 0$, the bandwidth resources are allocated to users without taking into account the history of the obtained cost.

3.2.4 Solving the MDP Model

In Sec. 3.2.3, we formulated dynamic bandwidth allocation problem as an average cost Markov decision process, including the state space $\mathcal{D}$, the set of available controls $\mathcal{A}(i)$ for each state $i \in \mathcal{D}$, the state transition probabilities $p_{i,j}(a)$, and the cost function. In this section, we discuss how to solve the Markov decision progress with dynamic programming method.

Bellman's Equation

Bellman's Equation is an effective dynamic programming method to solve the MDP problem achieving optimality. Given any stationary policy ψ, $\psi(i) \in A(i)$ is the control of state i. The average cost obtained by minimizing (3.41) or (3.42) under policy ψ is denoted by v^{ψ} while v^* is the average cost under an optimal policy ψ^*. The average cost v^{ψ} is independent of the initial state for a stationary system. Consequently, we use "differential

cost" to compare different states. The differential cost starting in state i under a policy ψ is denoted by $h^{\psi}(i)$ while the differential cost under an optimal policy ψ^* is $h^*(i)$. The optimal differential cost set $\vec{h}^* = \{h^*(i)|i \in \mathcal{D}\}$ and scalar v^* satisfy the Bellman's equation. Specifically, we have

$$h^*(i) = \min_{a \in A(i)} \{G(i,a) - v^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{D}} p_{i,j}(a) h^*(j)\}, \quad \text{for all } i \in \mathcal{D} \tag{3.45}$$

where $G(i,a)$ is the stage cost corresponding to state i and is given as

$$G(i,a) = \sum_{n=1}^{\mathcal{N}} G_n^{(i)}. \tag{3.46}$$

Given the optimal average cost v^* and the optimal differential cost set $\vec{h}^* = \{h^*(i)|i \in \mathcal{D}\}$, the optimal control $\psi(i)$ can be obtained by minimizing the current stage cost minus the expected average cost of the stage plus the remaining expected differential cost based on the possible resulting states, i.e.,

$$\psi^*(i) = \arg \min_{a \in A(i)} \{G(i,a) - v^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{D}} p_{i,j}(a) h^*(j)\}, \quad \text{for all } i \in \mathcal{D}. \tag{3.47}$$

Specifically, if the instantaneous rate in the destination region $\boldsymbol{s}'$ is $R_{\boldsymbol{s}'} \neq 0$,

$$\psi^*(i) = \arg \min_{a \in \mathbb{R}_{\boldsymbol{s}'}} \{G(i,a) - v^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{D}} p_{i,j}(a) h^*(j)\}, \quad \text{for all } i \in \mathcal{D}; \tag{3.48}$$

Otherwise, the destination region is single-coverage area with $R_{\boldsymbol{s}'} = 0$ and there is no control to exercise, i.e.,

$$\psi^*(i) = 0 \tag{3.49}$$

Several methods can be used to obtain the optimal average cost v^* and the optimal differential cost set $\vec{h}^* = \{h^*(i)|i \in \mathcal{D}\}$. In this thesis, we adopt the iteration method to achieve optimal v^* and $\vec{h}^*$. Initially, we set $v_0^* = 0$ and $\vec{h}_0^* = \mathbf{0}$. For the β^{th} iteration

step, each $h^*_\beta(i)$ for all $i \in \mathcal{D}$ and v^*_β are calculated according to (3.45) based on the newly obtained values by traversing all the possible controls of state i. If the obtained control set in the β^{th} iteration step is the same as that of the $(\beta-1)^{th}$ iteration step, the iteration is considered to be converged and the achieved v^*_β and $\vec{h}^*_\beta$ are the optimal values. The following pseudo codes shown in Algorithm 1 describe the procedure to obtain the optimal v^* and $\vec{h}^*$.

The computational complexity of the iteration algorithm is $O(\|A\|\|\mathcal{D}\|^2)$ [42], where $\|.\|$ is the cardinality of a set. The number of controls for each state i is the number of available rates in the moving in region $\boldsymbol{s}^{'}$. Thus the total number of controls in the control space $\mathcal{A}$ is $\|\mathcal{A}\| \approx \mathcal{K}S$. Each state $i \in \mathcal{D}$ is determined by both $\mathbb{B}^{(i)}$ and $\mathbb{W}^{(i)}$. $\mathbb{B}^{(i)}$ describes the location of $\mathcal{N}$ users among S regions and the allocated bandwidth for each user with $\|\mathbb{B}^{(i)}\| \approx S^{\mathcal{N}}\mathcal{K}^{\mathcal{N}}$. $\mathbb{W}^{(i)}$ determines the moving user from $\mathcal{N}$ users and the moving in/out regions with $\|\mathbb{W}^{(i)}\| \approx \mathcal{N}S(S-1)$. The cardinality of $\mathcal{D}$ is $\|\mathcal{D}\| \approx S^{\mathcal{N}}\mathcal{K}^{\mathcal{N}}\mathcal{N}S(S-1) = S^{\mathcal{N}+1}\mathcal{K}^{\mathcal{N}}\mathcal{N}(S-1)$. Therefore, the computational complexity of the iteration algorithm to obtain the optimal v^* and $\vec{h}^*$ is

$$\begin{aligned} O(\|A\|\|\mathcal{D}\|^2) &\approx O(\mathcal{K}S\{S^{\mathcal{N}+1}\mathcal{K}^{\mathcal{N}}\mathcal{N}(S-1)\}^2) \\ &\approx O(\mathcal{K}^{2\mathcal{N}+1}S^{2\mathcal{N}+5}\mathcal{N}^2) \end{aligned} \tag{3.50}$$

From (3.50), it is shown that the iteration algorithm to obtain optimal v^* and $\vec{h}^*$ introduces extremely high computational complexity. The computation can be done off-line with high-speed computer systems (which are separate and independent of the real system), before the real system starts to operate. Once the iterations are completed, the optimal decision at each stage can be determined quickly according to Bellman's equation. However, even off-line computation, the iteration complexity required to determine the

Algorithm 1 Policy Iteration Procedure

BEGIN:

Initialize $\vec{h}_0^* = \mathbf{0}$ and $v_0^* = 0$;

repeat

  In the β^{th} iteration step;

  Vector $\vec{a}_\beta = \{a_\beta(1), ..., a_\beta(\Upsilon + 1)\}$ indicates the selected controls for $h_\beta^*(i)$ of all the Υ states and v_β;

  for index number $i = 1$ to $i = \Upsilon + 1$ **do**

    if $1 \leq i \leq \Upsilon$ **then**

      Calculate $h_\beta^*(i)$ according to

$$h_\beta^*(i) = \min_{a \in A(i)} \{G(i,a) - v_{\beta-1}^* \overline{\tau}_i(a) + \sum_{j \in \mathcal{D}} p_{i,j}(a) h_{\beta-1}^*(j)\};$$

      Record $a_\beta(i)$;

    else

      Calculate v_β^* according to

$$v_\beta^* = \min_{a \in A(i)} \frac{[G(i,a) - h_{\beta-1}^*(i) + \sum_{j \in \mathcal{D}} p_{i,j}(a) h_{\beta-1}^*(j)]}{\overline{\tau}_i(a)};$$

      Record $a_\beta(\Upsilon + 1)$;

    end if

  end for

until $\vec{a}_{\beta-1} = \vec{a}_\beta$ indicates the convergency of the iteration.

END;

optimal average cost v^* and the optimal differential cost set $\vec{h}^*$ is overwhelming due to the exponential complexity as shown in (3.50). As a result, lower-complexity method is necessary. In next section, approximation method is used to significantly reduce the complexity by reducing the number of states taking into account the unique characteristics of mmWave network.

An Approximate Solution

Equation (3.50) shows that the iteration complexity greatly depends on the number of mobile users $\mathcal{N}$ in the system, the number of available rates $\mathcal{K}$ in each region, and the the number of regions S of the whole area. In this section, we give an approximate solution for the MDP problem by reducing the number of regions S and the number of available rates $\mathcal{K}$ in each region, given the number of mobile users $\mathcal{N}$.

In 5G cellular networks, mmWave base stations are likely to be densely deployed in small cell to provide high capacity to satisfy the demand of ultra-high traffic volume density. In addition, mmWave communication range is confined due to the severe propagation loss in mmWave frequency band. Therefore, the transmission range of mmWave communication in 5G cellular networks is relatively shorter than the range of previous generations of cellular networks. The shorter transmission distance results in less differences on the pathloss over distance and so does the transmission data rate. Thus, the whole double-coverage region is considered to have uniform instantaneous data rate of mmWave transmission denoted by $R_{\boldsymbol{d}}$, where the subscript $\boldsymbol{d}$ indicates the double-coverage region. Therefore, the whole single-coverage area (or the whole double-coverage area) is considered to be a single region since each region has the same instantaneous transmission rate. In the whole 5G coverage area, there are only two regions, namely, single-coverage region and double-coverage region. With the simplifications on instantaneous transmission rate, the total number of regions

is reduced from S to 2 by considering mmWave propagation characteristics. With above approximations, the set of regions in the whole area is $\mathcal{S} = \{\boldsymbol{s}, \boldsymbol{d}\}$ and the available rate set in double-coverage area is $\mathbb{R}_{\boldsymbol{d}}$. For each state $i = (\mathbb{B}, \mathbb{W}) \in \mathcal{D}$, $\mathbb{B}$ is simplified as

$$\mathbb{B} = \begin{Vmatrix} B_{1,\boldsymbol{s}} & B_{1,\boldsymbol{d}} \\ B_{2,\boldsymbol{s}} & B_{2,\boldsymbol{d}} \\ \vdots & \vdots \\ B_{\mathcal{N},\boldsymbol{s}} & B_{\mathcal{N},\boldsymbol{d}} \end{Vmatrix} = \begin{Vmatrix} 0 & B_{1,\boldsymbol{d}} \\ 0 & B_{2,\boldsymbol{d}} \\ \vdots & \vdots \\ 0 & B_{\mathcal{N},\boldsymbol{d}} \end{Vmatrix} \tag{3.51}$$

where $B_{n,\boldsymbol{d}} \in \mathbb{R}_{\boldsymbol{d}}$ if user U_n is in double-coverage region while $B_{n,\boldsymbol{d}} = 0$ if user U_n is in single-coverage region; $\mathbb{W}$ can be simplified as

$$\mathbb{W} = \begin{Vmatrix} w_{1,\boldsymbol{s}} & w_{1,\boldsymbol{d}} \\ \vdots & \vdots \\ w_{n,\boldsymbol{s}} & w_{n,\boldsymbol{d}} \\ \vdots & \vdots \\ w_{\mathcal{N},\boldsymbol{s}} & w_{\mathcal{N},\boldsymbol{d}} \end{Vmatrix} = \begin{Vmatrix} 0 & 0 \\ \vdots & \vdots \\ w_{n,\boldsymbol{s}} & w_{n,\boldsymbol{d}} \\ \vdots & \vdots \\ 0 & 0 \end{Vmatrix} \tag{3.52}$$

where $w_{n,\boldsymbol{s}} = 1$ and $w_{n,\boldsymbol{d}} = -1$ if user U_n moves from double-coverage region to single-coverage region while $w_{n,\boldsymbol{s}} = -1$ and $w_{n,\boldsymbol{d}} = 1$ if user U_n moves from single-coverage region to double-coverage region.

The iteration complexity shown in (3.50) can be further reduced if we choose the control for each state from fewer number of available rates in the double-coverage region. The number of available rates in the double-coverage region determines the fineness of bandwidth allocation. In the MDP model in Sec. 3.2.3, the available rates depend on the number of time slots in each superframe, i.e., the available rate set in double-coverage area is $\{\frac{1}{\mathcal{K}}R_{\boldsymbol{d}}, \frac{2}{\mathcal{K}}R_{\boldsymbol{d}}, ..., R_{\boldsymbol{d}}\}$. In the approximate solution, we adopt larger fineness of rate selection, i.e., the available rate set is $\mathbb{R}_{\boldsymbol{d}} = \{\frac{\eta}{\mathcal{K}}R_{\boldsymbol{d}}, \frac{2\eta}{\mathcal{K}}R_{\boldsymbol{d}}, ..., \frac{Q\eta}{\mathcal{K}}R_{\boldsymbol{d}}\}$ with $Q = \lfloor \frac{\mathcal{K}}{\eta} \rfloor$ where

$\eta \gg 1$. Thus the number of available rates in the double-coverage region can be reduced from $\mathcal{K}$ to Q. The simplified control space is $\mathcal{A} = A(s) \cup A(d) = \{0, \frac{\eta}{\mathcal{K}} R_{\boldsymbol{d}}, \frac{2\eta}{\mathcal{K}} R_{\boldsymbol{d}}, ..., \frac{Q\eta}{\mathcal{K}} R_{\boldsymbol{d}}\}$.

For the case that 20 users move in the area with 100 regions, the iteration complexity after simplification is $1/(2^{86}5^{136})$ of the original iteration complexity if 10 rates are available for the users to select in double-coverage region. In the approximation solution, we consider the case that all the users are uniformly distributed in the whole area, thus the capacity constraint is for all the users in the whole coverage area. The optimal policy can be obtained by Bellman's equation based on the MDP model described in Sec. 3.2.3 and Sec. 3.2.3 with the simplified states.

Implementation Considerations

To implement the proposed method on video quality provisioning, there is a controller connected with all the mmWave base stations in the coverage area. In each mmWave base station, there is a index table saving the information of each state and the corresponding control obtained by off-line calculation with Bellman's equation. When a user U_n enters the new region, the associated mmWave base station determines the new state based on previous state and the moving in region of the user U_n. Then, the associated mmWave base station allocates the bandwidth to the user U_n immediately according to the index table. The movement of the user U_n is reported to the controller by the associated mmWave base station. The controller broadcasts the new state to all the mmWave base stations and the whole system waits for another user entering the new region to trigger the incoming new state. Please note that there is only one user with mmWave bandwidth re-allocation when system enters a new state and other users have the same bandwidth allocations as previous state. The user entering a new region can obtain the newly allocated bandwidth immediately and does not need to wait for the updated information from the controller to

start bandwidth allocation.

3.2.5 Numerical Results

In this section, numerical results are provided to demonstrate the performance of the bandwidth allocation solution obtained from the MDP model with state aggregation, compared with another two solutions, namely random policy solution and normal transmission solution. Random policy solution is based on the proposed buffer system and randomly selects the allocated bandwidth of the user among available rates in each MDP stage. In normal transmission solution, the buffer system is not installed in the user to mitigate the effect of link outage and the allocated bandwidth for each user is based on the required data rate of video applications. The performance of the obtained optimal policy is mainly demonstrated by quality consistency. Additionally, we show the distribution of the frozenness duration, the impact of buffer size on quality consistency, and the number of supported traffic in mmWave network.

The numerical computations are conducted in a square area with side length of 10 Km. Initially, a number of users are randomly located in the whole area with no packets in each buffer. mmWave base stations are distributed in the whole area in grid topology with the distance of 500 m. The instantaneous data rate in double-coverage region is estimated by

$$R \approx W \log_2[1 + \frac{P_T G_R G_T \lambda^2}{16\pi^2 (N_0 + I) W d^n}] \tag{3.53}$$

where P_T is the transmitted power, G_T and G_R are respectively the antenna gains of the transmitter and receiver. λ is the wavelength, W is the system bandwidth, N_0 and I are the one-side power spectral densities of white Gaussian noise and broadband interference, respectively. d is the transmission distance and the maximum link operation range is

Table 3.1: PROPAGATION RELATED PARAMETERS

Parameters	**Value**
Central frequency (f)	30 GHz
Spectrum bandwidth (W)	2 GHz
Transmission power (P_T)	0.5mW
Background noise (N_0)	-134dBm/MHz
Maximum distance (d)	250 m
Link outage threshold (PL_{thr})	-40 dBm
Transmitter antenna gain (G_T)	1
Receiver antenna gain (G_R)	1

adopted for link budget to determine the instantaneous rate in the double-coverage area. The propagation-related parameters are shown in Table 5.2.

The time durations of each user staying in single-coverage region and double-coverage region are exponentially distributed with rate ξ_{nsd} and ξ_{nds}, respectively. Each ξ_{nsd} for user U_n is randomly selected within the range of $[\frac{1}{15}, \frac{1}{30}]$ while each ξ_{nds} for user U_n is also randomly selected within the range of $[\frac{1}{80}, \frac{1}{120}]$. The required data rate of the multimedia application in each user is randomly selected from the range of [1.5 Gbps, 2 Gbps]. The system-related parameters are shown in Table 3.2.

Table 3.2: SYSTEM RELATED PARAMETERS

Parameters	**Value**
Packet size (Z)	10^3 bits
Buffer size (G)	16×10^6 packets
Capacity of double-coverage region	200 Gbps
Number of available rates (Q)	10
Time slot duration	$18\mu s$

Quality Consistency

Fig. 3.9 shows the numerical results on quality consistency of the four solutions with various numbers of video traffics. The optimal policy and approximate solution achieve much better quality consistency than random policy and normal transmission. As more traffic is involved, the quality consistency goes down when the total allocated bandwidth reaching the system capacity. By allocating more bandwidth than the required bandwidth to the users in the double-coverage region, the stored data in the buffer can be used to maintain video playout quality in single-coverage region. The proposed buffer system can mitigate the impact of link outage on video playout quality at the cost of the storage space and extra bandwidth.

Fig. 3.10 shows the impact of buffer size on maintaining quality consistency for both optimal policy solution and random policy solution. It is shown that as the buffer size increases, the quality consistency can be improved. Large buffer size has the robustness on allocated bandwidth of each user to improve quality consistency. Specifically, on one hand, large buffer has lower probability to be overflowed, which results in lower packet dropping

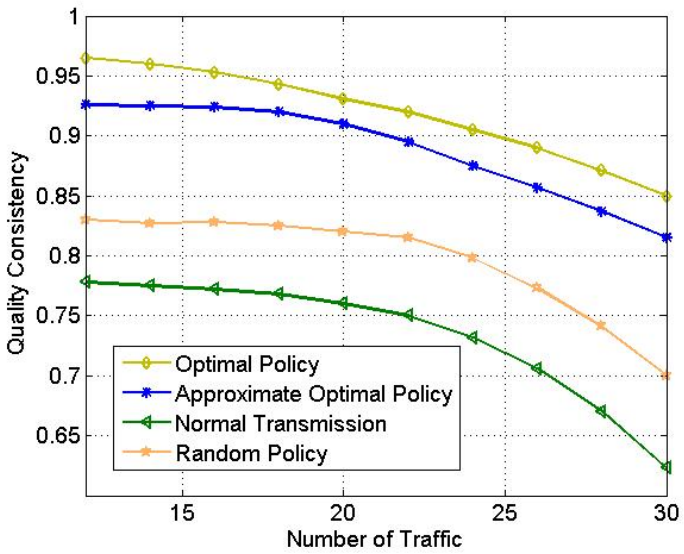

Figure 3.9: Quality Consistency via the Number of Traffic

probability; On the other hand, large buffer can store more packets for users spending longer time in single-coverage area without suffering from frozenness. The extreme case is that we can achieve 100 percent quality consistency for a stationary system if we use infinite buffer at the user and allocate very high bandwidth to each user.

Frozen Duration Distribution

Fig. 3.11 shows the probability distribution of the normalized frozenness period suffered by all the users in the network. The probability distribution of frozenness duration shows that the obtained optimal policy with buffer system can greatly reduce the frozenness duration compared with the other two solutions, since most of the frozenness durations in optimal policy solution are in the short duration range. The proposed method using dynamic bandwidth allocation to maintain video quality is very effective when mmWave

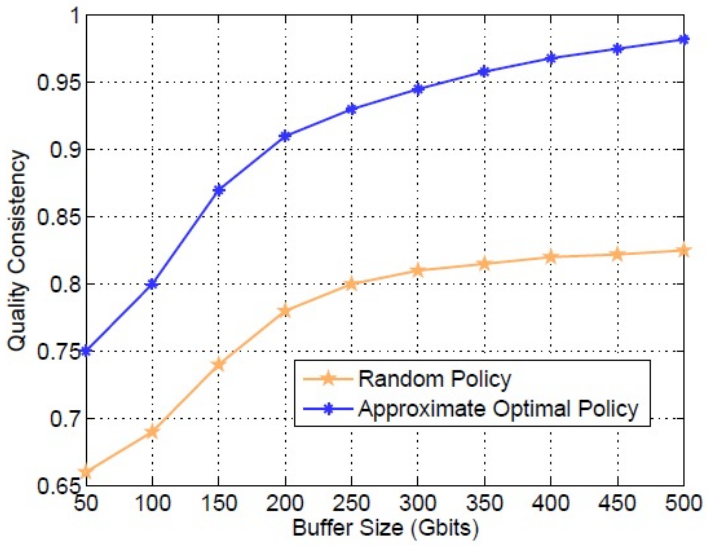

Figure 3.10: Quality Consistency of Different Buffer Sizes

links suffering from frequent link outage. The random policy solution can reduces the frozenness duration compared with normal transmission solution because of the buffer system.

Number of Supported Traffic

The number of traffic supported successfully in the network based on the required bandwidth is shown in Fig. 3.12. It can be seen that the network with optimal policy solution can support less number of traffic than that in normal transmission solution since larger bandwidth need to be allocated to users for video quality provisioning. Both the video frozenness duration and the probability of video frozenness can be significantly reduced at the cost of extra allocated bandwidth.

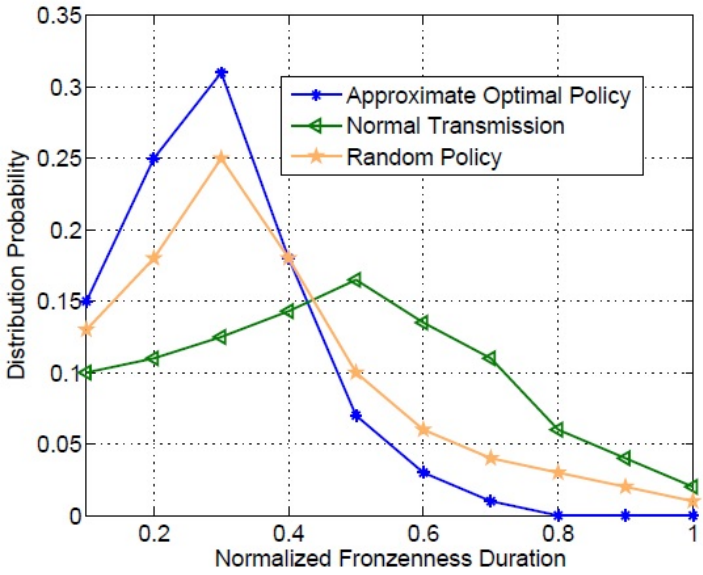

Figure 3.11: Probability Distribution of Frozenness Duration

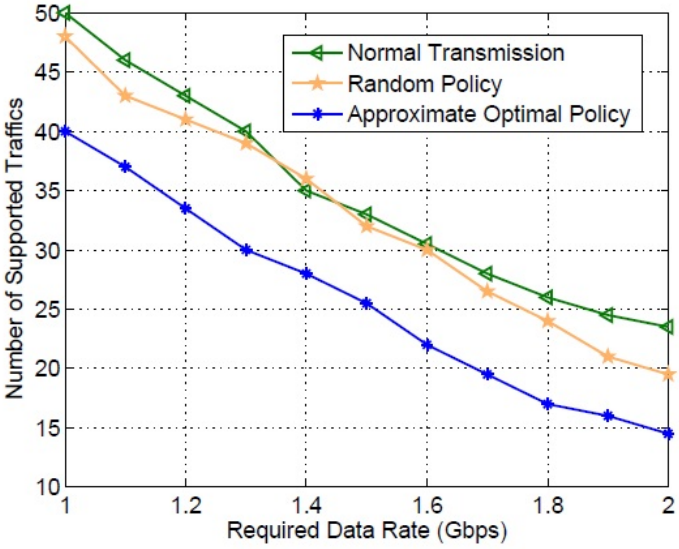

Figure 3.12: Number of Supported Traffic

3.3 Conclusions of the Chapter

In this chapter, we dealt with quality provisioning problem for mobile users in mmWave 5G cellular networks. For high-mobility scenarios, the problem of dynamic bandwidth allocation in mmWave band for users alternatively entering single-coverage region or double-coverage region is formulated as a MDP model. By solving the MDP model, we can obtain the optimal policy indicating the allocated bandwidth to each user in different regions. The obtained optimal policy can dynamically charge or discharge the buffer to maintain the video playout quality at the user. The numerical results demonstrate that the proposed methodology on video playout quality provisioning can effectively mitigate the impact of frequent mmWave link outage and significantly improve the smoothness of video quality.

The proposed method on playout quality provisioning is effective for non-real time video applications of the users with high mobility. For low-mobility users, it is possible that the users stay in the areas with mmWave link outage for much longer time. In this case, multi-hop relaying is an effective method to maintain network connectivity by replacing the outage link with multiple available links. Longer setup time of multi-hop relaying has less impact on network performance since it does not require frequent path re-selections for low-mobility scenarios. Real-time multimedia applications of high-mobility users can be supported by switching the underlaying network from mmWave to 4G to overcome frequent mmWave link outage problem. However, in this case, the playout quality of multimedia applications would descend because of the lower transmission rate of 4G network.

Chapter 4

Network Capacity Improvement by Spatial Reuse

mmWave communication has high propagation loss resulting from strong pathloss, atmospheric and rain absorption. High gain directional antenna is used to combat the severe propagation loss and achieve high data rate. The high propagation loss and directional antenna result in relatively low multi-user interference (MUI). Therefore, multiple communication links can operate simultaneously to improve system capacity by exploiting spatial reuse. On the other hand, allowing multiple communication links to transmit data at the same time leads to higher MUI, which can decrease the system capacity. How to schedule appropriate concurrent transmissions to improve system performances, e.g., network capacity and the data rate of each link, is an important and challenging issue.

4.1 Non-interfering Concurrent Transmission Scheduling

Due to antenna's directivity, any two communication links can operate simultaneously without mutual interferences under the condition that any transmitter is outside the beamwidth of the other receiver or does not direct its beam to the other receiver if it is within the beamwidth of the other receiver. We propose a concurrent transmission scheduling algorithm to allow non-interfering communication links to operate simultaneously in each time slot in a spatial time-division multiple-access (STDMA) system to significantly increase the system capacity. The protocol is designed based on the proposed concurrent transmission scheduling algorithm for the hybrid MAC structure of CSMA/CA+STDMA. Non-interfering concurrent transmission scheduling is easy to implement without fully utilizing the resources.

4.1.1 Concurrent Transmission Scheduling Scheme

Based on the hop selection metric in the Chapter 3, we propose a multi-hop concurrent transmission (MHCT) scheme within each cell as follows: initially, the base station activates all the beams to collect network topology information and node's load information. The base station maintains these information and updates it when there is any change. Multimedia applications (i.e., uncompressed video streaming and kiosk high speed downloading) usually are long-lived flows and their statistical traffic demands do not change frequently. Without loss of generality, we consider the scenarios in which each node sends its traffic load information to the base station every m superframes. When a transmission request is received, the base station calculates $E[d^n]$ and $E[F]$ based on the traffic load and

topology information received in the random access period, and then generates a weighted graph where the weight of each link is calculated according to (3.1). The base station chooses the path with the lowest accumulated weights for each flow, using the Dijkstra Algorithm. Then the flow uses this route for data transmission thereafter until the network topology information and node's load information change. When a new flow arrives, the base station will re-schedule the transmission route for each flow as the network topology and traffic loads have changed.

The base station then checks the concurrent transmission condition and schedules non-conflicting links to operate concurrently. During the random access period, base station collects the network information and transmission requests, based on which base station allocates network resources and distributes the scheduling information to nodes during the beacon period in one of the future superframes. It is possible that a node sends a transmission request to the base station but does not receive the scheduling information within a certain time interval (e.g., a random access period plus a beacon period) because moving obstacles in the room may block the LOS link between the node and the base station. In this case, the node needs to send its request in the next random access period. During the data transmission period, the base station switches to the directional mode for data transmission. Nodes transmit data in the allocated time slots according to the scheduling information. If the data transmission is not successful due to blocked LOS by moving obstacles, the failure should be reported to the base station during the next random access period so that the base station can re-schedule the transmission in the next transmission period.

Next, we develop a scheduling algorithm for concurrent transmissions. For a given network topology, some hops in dense areas may have a lower probability to operate concurrently with other hops. Therefore, we should give higher priority of scheduling to the

hops with less transmission opportunity but with heavy traffic loads. To achieve this, we sort hops in a descending order of transmission loads to guarantee that the hops with the highest loads will be scheduled first. The detailed algorithm is described as follows. Initially, there are L slots in a transmission period. A transmission request $r_{i,j}$ (i.e., the j-th hop of flow i) needs $n(i,j)$ slots. The base station sequentially checks the hops to be scheduled in the descending order of their traffic loads and schedules non-conflicting links to operate concurrently according to the concurrent transmission condition: concurrent transmission is feasible if and only if any transmitter is outside the beamwidth of the other receiver or does not direct its beam to the other receiver if it is within the beamwidth of the other receiver. Two adjacent links which share one node cannot operate concurrently due to half-duplex operations of wireless transmission. If a link does not conflict with all existing hops in the group, this link is added in the group for concurrent transmissions and the base station updates the reserved slots for this group based on the maximum number of required time slots of all the hops in the group. If a hop is not added in the current group, the base station will sequentially check the following groups for concurrent transmissions. If a hop can not be added in any existing groups, the base station will create a new group for this hop if the number of available slots in the superframe is sufficient. Otherwise, the base station rejects the request due to limited network resources. In this case, the base station will remove all hops involved in the flow where the rejected hop belongs to. The pseudo code for the concurrent transmission scheduling algorithm is shown in Algorithm 3.

4.1.2 Simulation Results

We compare the proposed MHCT scheme with two other transmission schemes, namely, SHCT [21] and single hop transmission (SHT). In the SHCT scheme, if LOS link is un-

Algorithm 2 Concurrent Transmission Scheduling Scheme

BEGIN:

1: Base station receives $r_{i,j}$ requesting $n(i,j)$ time slots
2: **for** All non-empty group (T_b! = Null) **do**
3: **if** $r_{i,j}$'s link does not conflict with those of all existing links in T_b **then**
4: **if** $r_{i,j}$ does not have shared nodes with other links in T_b **then**
5: **if** $r_{i,j}$ requires extra slots, $n(i,j) - n(b) > 0$ **then**
6: **if** Available slots $L \geq n(i,j) - n(b)$ **then**
7: Schedule $r_{i,j}$ in group T_b;
8: Update $T_b = T_b \bigcup \{r_{i,j}\}$;
9: Update the available slots $L = L - [n(i,j) - n(b)]$;
10: Update $n(b) = n(i,j)$;
11: Update the allocated slots for $r_{i,j}$;
12: Sort all hops in the decreasing order of loads.
13: Go to END;
14: **else**
15: Go to line 26;
16: **end if**
17: **else**
18: Schedule $r_{i,j}$ in T_b;
19: Update $T_b = T_b \bigcup \{r_{i,j}\}$;
20: Update the allocated slots for $r_{i,j}$;
21: Sort all hops in the decreasing order of loads.
22: Go to END;
23: **end if**
24: **end if**
25: **end if**
26: Next Group;
27: **end for**
28: **if** The number of available slots is larger than $n(i,j)$ **then**
29: Start a new group $T(k) = \{r_{i,j}\}$;
30: **else**
31: Reject request $r_{i,j}$ and release the reserved resources for all the hops in flow i;
32: **end if**

END;

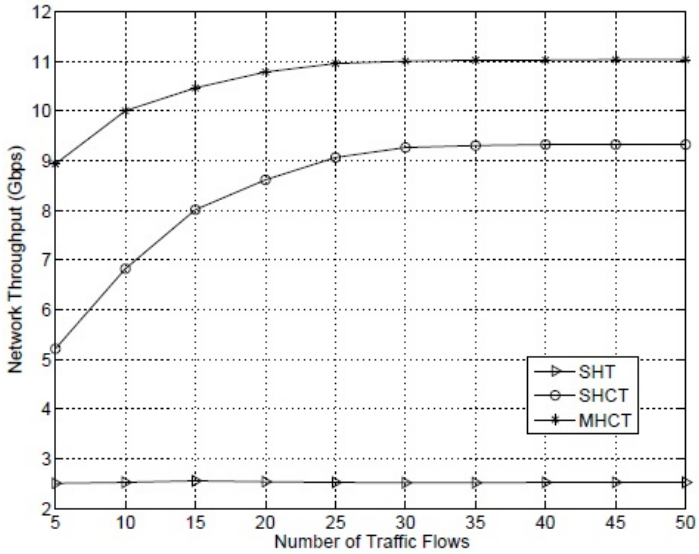

Figure 4.1: Network throughput versus the number of traffic flows

available, a neighboring node is randomly selected to relay the traffic [40]. We use the SHT scheme operating in a TDMA mode as the baseline for comparison.

The network throughput of the three schemes is shown in Fig. 4.1. In the SHT scheme, there is at most one transmission at any time in the network. Concurrent transmissions can improve the spatial reuse and significantly increase the network throughput. It is also shown that the proposed MHCT scheme achieves higher network throughput compared with the SHCT scheme because, with the MHCT scheme, the number of concurrent links is larger and the transmission rate of each link is higher due to the short link length.

4.2 Interfering Concurrent Transmission Scheduling

To further improve system capacity, both interfering and non-interfering links can be scheduled concurrently if the MUI is properly managed. We formulate general concurrent transmission scheduling into an optimization problem, by maximizing the number of flows successfully scheduled in the network while satisfying flow throughput requirements. The optimization problem considering the MUI depending on different active link set is solved by dynamic programming approach with flip-based scheduling algorithm to achieve sub-optimal solution with significantly reduced complexity.

4.2.1 Related Work

The problem of concurrent transmission scheduling has been investigated extensively in the literature [18, 21, 43–45]. Multiple links are scheduled to transmit concurrently to improve the system throughput, based on the derived exclusive region or a link selection metric in [21, 43]. A concurrent transmission scheduling algorithm is proposed in [44] in wireless networks with rate adaptation, but it does not consider the unique features of mmWave systems, e.g., high propagation loss and the use of directional antenna. In [18], multiple communication links are scheduled in the same time slot if the accumulated interference in this slot is below a specific threshold. The optimal concurrent transmission scheduling problem can be converted to a Knapsack problem which is known to be NP-complete. With the knowledge of the hardness of concurrent transmission scheduling and the fact that the real time scheduling decision should be made within a few milliseconds, we propose an efficient heuristic scheduling algorithm with reasonable computational complexity.

4.2.2 Optimal Scheduling Problem Formulation

There are N transmission requests submitted to the base station and each of them specifies a required minimal throughput R^i_{min}. The CTAP of each superframe contains M time slots. For the k^{th} time slot of the CTAP in a superframe, the scheduling decision can be represented by a control vector $U_k = [u_{k,1}, u_{k,2}, ...u_{k,N}]$, where $u_{k,i} = 1$ if flow i is scheduled in the k^{th} time slot, otherwise $u_{k,i} = 0$. To maximize the total throughput, we first formulate the concurrent scheduling problem as an optimization problem (P1):

$$\max_{u_{k,i}\in\{0,1\}} \sum_{k=1}^{M}\sum_{i=1}^{N} R_{k,i} \tag{4.1}$$

where

$$R_{k,i} = \eta W \cdot log_2(1 + \frac{k u_{k,i} P_t d_{i,i}^{-\gamma}}{W N_0 + b\sum_{l\neq i} f_{l,i} u_{k,l} k P_t d_{l,i}^{-\gamma}}) \tag{4.2}$$

The optimization problem is a non-linear integer programming problem. The transmission data rate of each flow in a time slot can not been determined until scheduling decision for this slot is made because of the MUI. The concurrent transmission scheduling problem is similar to $0-1$ Knapsack problem [46] in the case that items (flows) can be added into the knapsack (allocated to the slot). The objective is to maximize the total profit (total throughput) with the weight (interference) constraints. It is well-known that the Knapsack problem is NP-complete [46]. The concurrent transmission scheduling problem here is even more difficult than the Knapsack problem, since the profits of items (flow throughput) would change with the selected subsets of flows. Therefore, the existing approximation algorithms for Knapsack problems cannot be applied directly. In addition, the optimal solution may be unfair because the time slots are more likely to be allocated to those flows with a higher throughput while some unlucky flows will be starved. In other words, the throughput requirements of the starved flows can not be satisfied.

The bandwidth-intensive applications supported by mmWave networks require multi-Gbps throughput, e.g., the mandatory data rate for uncompressed video streaming is 1.78 or 3.56 Gbps. The transmission demands need to be satisfied to provide the required quality of service (QoS) for the applications. With TDMA scheme, to achieve the high flow throughput, each flow will be allocated a greater number of time slots. Thus, we formulate an optimal scheduling problem to find maximum number of flows scheduled in the network, subject to the minimum throughput requirement R^i_{min} of each flow. The optimal scheduling problem with the constraints is formulated as (P2):

$$\max_{u_{k,i} \in \{0,1\}} \sum_{i=1}^{N} I_i \tag{4.3}$$

where

$$I_i = \begin{cases} 1, & F_i \geq R^i_{min}; \\ 0, & \text{otherwise}; \end{cases} \tag{4.4}$$

I_i indicates whether the transmission demand of flow i is satisfied. A flow is scheduled successfully if and only if its minimum throughput requirement is ensured. By allowing a number of flows to transmit concurrently in each time slot, we try to maximize the number of flows in the network aiming to utilize the resource efficiently. F_i denotes the throughput of flow i according to the scheduling results and is defined as

$$F_i = \frac{\sum_{k=1}^{M} R_{k,i} \Delta T}{T_{BP} + T_{CAP} + M\Delta T} \tag{4.5}$$

where T_{BP} and T_{CAP} are the time duration for beacon period and contention access period, respectively, and ΔT is the time duration of each time slot in CTAP period. $R_{k,i}$ is given in (4.2). Maximizing the number of flows successfully scheduled in the network (P2) has advantages on fairness and resource utilization efficiency, compared with maximizing the

network throughput with constraints that $F_i \geq R^i_{min}$ $(i = 1, 2, 3...N)$ (P1 with constraints). First, P1 with constraints is for the case that the network resources are sufficient to meet the transmission demands of all flows in the network, which holds well only for the scenario that a small number of flows need to be scheduled in the networks. In this case, the transmission demands can be satisfied even without efficient scheduling. Thus, it is more interesting in considering a more challenging case where the network resources are limited compared with the intensive traffic demands of users. Second, to maximize the network throughput, network resources (time slots) would be allocated to the flows with a higher transmission data rate while satisfying the throughput requirement of each flow. The extra throughput of a flow makes little contributions on improving the network performance. To utilize the resource efficiently, we allocate the resources which are assigned to flows to obtain extra flow throughput, to other flows. Therefore, the network can support more users with limited amount of resources.

The problem of P2 is a non-convex integer programming problem and is NP-hard. The searching space is $2^{N \cdot M}$ if exhaustive searching approach is implemented. Given the hardness of solving the optimization problem of P2 in polynomial time, we propose a practical heuristic scheduling algorithm to obtain the scheduling decision with full consideration of the unique features of mmWave networks.

4.2.3 Scheduling Algorithm Design

In this section, we present a heuristic scheduling algorithm for optimization problem (P2), considering the unique characteristics of mmWave networks.

Since the scheduling problem has a slotted structure in the time domain, we make the scheduling decision slot by slot instead of optimizing all the N flows in M time slots si-

multaneously. The decomposed method can solve the problem in an iterative manner by reducing the searching space of the states from $2^{N \cdot M}$ to 2^N. The problem of maximizing network throughput based on the slotted structure has been studied in [44]. In this thesis, we propose a slot-based concurrent transmission scheduling algorithm to allow as many flows scheduled in the network as possible, considering the minimum throughput requirement for each flow. For the k^{th} time slot, the scheduling algorithm generates the decision vector U_k that indicates which flows are scheduled in the k^{th} slot. The schedule for the data transmission period is $\widetilde{U}_{N \times M} = [U_1, U_2 ... U_k ... U_M]$.

To schedule as many flows as possible within the data transmission period while satisfying the minimum throughput requirement of each flow, we have the following requirements on algorithm design: 1) transmitting as much data as possible in each slot; and 2) reducing the number of *fractional flows*, where *fractional flow* is defined as a flow which is scheduled in the network, but its transmission throughput requirement is not satisfied. The fractional flow results in inefficient resource utilization since the resources (time slots) are allocated to it but the fractional flow can not support the specific applications.

During the CAP period of the m^{th} superframe, the base station receives a number of transmission requests, each of which needs a specific required minimum flow throughput. The base station makes the scheduling decision for the $(m+1)^{th}$ CTAP before the $(m+1)^{th}$ beacon period, during which the distributes the scheduling information to all the nodes in the network. The nodes start data transmission according to the scheduling information in the $(m+1)^{th}$ CTAP period. To transmit as much data as possible in each time slot, let flow i be active in slot k if the profit of adding flow i is greater than the degradation of throughput it causes to other pre-selected flows. Based on this approach, we propose a flip-based algorithm to obtain the set of flows active in each slot. Initially, we set $U_k = \overrightarrow{0}$. For each flow i $(i = 1, 2 ... N)$, $u_{k,i}$ is set to 1 if adding it to the active set can increase the

Algorithm 3 Concurrent Transmission Scheduling Algorithm

BEGIN:

1: Base station receives transmission request $r_i(i = 1, 2...N)$ requiring minimum throughput R^i_{min}
2: Initialize the scheduling basis vector $\mathbf{B}_1 = \overrightarrow{0}$
3: **for** non-scheduled slot $k(1 \leq k \leq M)$ **do**
4: **if** $B_k = B_{k-1}$ **then**
5: $U_k = U_{k-1}$
6: go to line 20
7: **else**
8: **repeat**
9: **for** flow i=1 to N **do**
10: **if** $b_{k,i} = 1$ or $b_{k,i}$ is marked as unchanged **then**
11: Keep the corresponding $b_{k,i}$ the same
12: **else**
13: **if** flow i does not have shared nodes with other links scheduled in slot k **then**
14: set $b_{k,i} = u_{k,i}$ according to

$$\arg\max_{u_{k,i}\in\{0,1\}} \{R_{k,i} + \sum_{l=1,l\neq i}^{N} R_{k,l}\}$$

15: **end if**
16: **end if**
17: **end for**
18: **until** $B_k = [b_{k,1}, b_{k,2}, ...b_{k,N}]$ converges
19: update $U_k = B_k$
20: update $F_i = \frac{\sum_{n=1}^{k} R_{n,i}\Delta T}{T_{BP}+T_{CAP}+M\Delta T}$
21: **if** any $F_i \geq R^i_{min}$ **then**
22: change $b_{k,i}$ from 1 to 0, and mark it as unchanged
23: **end if**
24: update $B_{k+1} = B_k$
25: update k=k+1
26: **if** $k > M$ **then**
27: go to END
28: **end if**
29: **end if**
30: **end for**

END;

throughput in slot k. Otherwise, we set $u_{k,i} = 0$. The above process is repeated until U_k converges. Therefore, the searching complexity for each slot changes from 2^N to $O(T_1 N)$, where T_1 is the number of iterations for U_k to converge. For the first slot of the CTAP, we use the flip-based algorithm to obtain the set of active flows. We keep this set of flows active for a number of following time slots until one flow's minimum throughput requirement is satisfied. These time slots have the same active flow set since there are the same flows available to be scheduled. In the next time slot, the set of active flows is re-determined. We do not flip $u_{k,i}$ if flow i's minimum throughput requirement is satisfied and set $u_{k,i} = 0$ for all the following time slots. Also, to minimize the number of fractional flows, $u_{k,i}$ is set to 1 if flow i is scheduled in the previous time slot but its transmission requirement is not satisfied. We only flip the variables of $u_{k,i}$ if the corresponding flow i is not scheduled in the previous slots, to obtain better throughput performance in slot k. Then, the same schedule is used for the following time slots until the throughput requirement of one flow is satisfied. The above procedure is repeated until all the time slots have been scheduled. The pseudo code for the concurrent transmission scheduling algorithm is shown in Algorithm 3.

4.2.4 Performance Evaluation

In this section, we describe the performance evaluation procedure and present the simulation results for the proposed scheme compared with other concurrent transmission scheduling schemes. We evaluate the performance of the proposed scheme in terms of the number of scheduled flows and the network throughput in a 10×10 m^2 area. The base station is placed in the center and wireless nodes are randomly distributed in the area. Each node is equipped with a steerable directional antenna with a beamwidth of 60°. There are sufficient number of flows with constant bit rates, and we schedule them in the network. The source and destination of each flow are randomly selected. The required throughput

Table 4.1: SIMULATION PARAMETERS

Parameters	**Symbol**	**Value**
System bandwidth	W	1200 MHz
Transmission power	P_T	0.1mW
Background noise	N_0	-134dBm/MHz
Pathloss exponent	n	2
Reference distance	d_{ref}	1.5m
Pathloss at d_{ref}	PL_0	71.5 dB
Slot time	ΔT	$18\mu s$
Beacon period	T_{bea}	50 μs
Random access period	T_{ran}	800 μs
Number of slots in transmission period	$\overline{N}$	1000

of each flow is uniformly distributed between 1.5 Gbps and 3.5 Gbps. All senders use the same transmission power level for transmission. A typical physical layer parameter setting of mmWave systems is shown in Table 4.1.

We compare the proposed concurrent transmission scheduling scheme with three other transmission schemes, namely, the exhaustive search, exclusive region based scheduling scheme [43], and TDMA scheme. Exhaustive search takes long time to achieve the optimal solution, exclusive region based scheduling only allows non-interfering links to transmit concurrently, and TDMA scheme supports only one active link in each time slot.

Fig. 4.2 shows the number of flows scheduled successfully in the network. The total

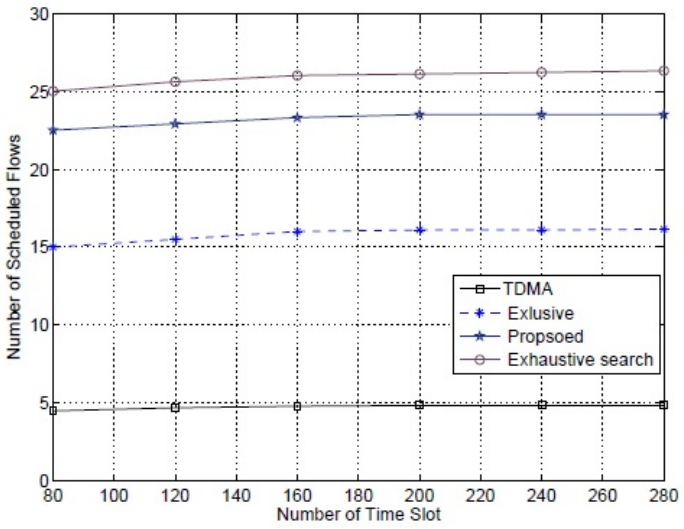

Figure 4.2: Number of Scheduled Flows

number of flows is $N = 50$. It is shown that the proposed scheme can utilize the resource more efficiently than exclusive region based scheduling and TDMA scheme since it can schedule more flows in the network, each of which needs a minimum throughput requirement.

The network throughput is shown in Fig. 4.3. For TDMA scheme, there is at most one transmission at any time slot in the network. Concurrent transmissions can enhance the spatial reuse and significantly increase the network throughput. Our proposed scheme can further improve the network throughput by allowing both interfering links and non-interfering links to transmit simultaneously, compared with exclusive region based scheduling [43].

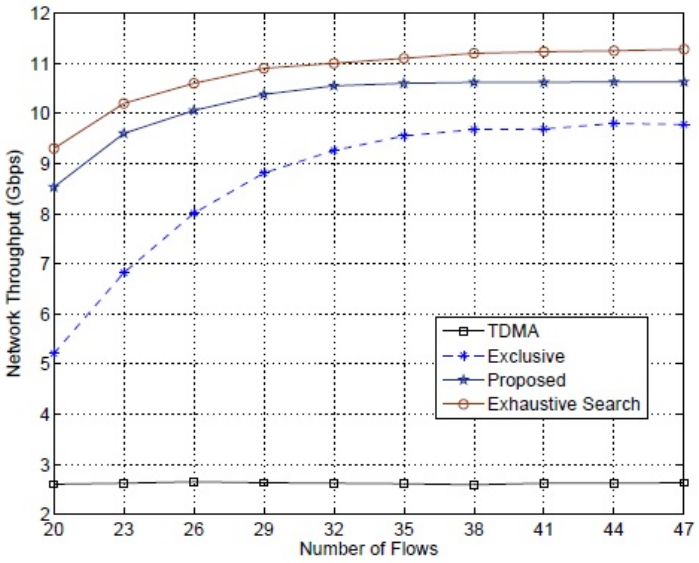

Figure 4.3: Network Throughput in mmWave Network

4.3 Conclusions of the Chapter

In this chapter, we exploit the spatial reuse to improve system capacity by enabling both non-interfering concurrent links and interfering concurrent links. The proposed concurrent transmission scheduling schemes considering the unique characteristics of mmWave networks (i.e., high pathloss, directional antenna, and large required flow throughput) can significantly improve system capacity while satisfying the flow throughput requirements.

Chapter 5

Codebook-based Concurrent Beamforming

Beamforming determines a beam toward a certain direction formed by multiple antennas to maximize the transmission rate. The antenna gains of the transmitter (TX) and the receiver (RX) have significant impact on the transmission data rates. Thus beamforming protocol is required to select the best transmission and reception beams, according to the selection metric [47, 48], e.g., signal-to-noise ratio (SNR). Because of directional antennas and the high propagation loss, multiple communication links can operate simultaneously to exploit the spatial reuse potential [49]. Appropriate concurrent transmissions are desirable in directional mmWave networks to improve the network capacity [18]. The high propagation loss accompanying the limited penetration and diffraction capability confines the network coverage in a shorter range. The mutual interferences generating from the concurrent links in short coverage area can greatly affect the concurrent throughput. Beamforming for multiple concurrent links (named as concurrent beamforming) should consider the mutual interferences to decide the best transmission/reception beam for each

link, instead of enabling the beamforming for each link in isolation. The best transmission beam can improve the link rate while it generates the interferences to other receivers to reduce the rates of other links. How to select appropriate beams for concurrent links to optimize sum rates is an important and challenging issue.

Concurrent beamforming in 60 GHz band has several challenges: (1) The antenna array forms a specific beam pattern according to the weight vector calculated based on the measuring signals' angle of departure (AOD)/angle of arrival (AOA), or acquisition of the entire channel state information (CSI) matrices. It introduces high calculation load and large overhead. (2) the concurrent links make beamforming even more difficult, because the mutual interferences are unknown until the best transmission and reception beam patterns for all the concurrent links are finalized. and (3) no complete MAC protocol to setup multiple directional communication links is available.

To reduce the realization complexity on beamforming, we adopts the beam-switching operations based on the pre-defined beam steering vectors (i.e., codebook) without the necessity of AOD/AOA or CSI estimation. This chapter has the following main contributions. First, the codebook-based concurrent beamforming problem is formulated as an optimization problem by maximizing the sum rates. Second, to reduce the complexity and the total setup time (meaningful on reinforcing transmission efficiency and reducing energy consumption), we decompose concurrent beamforming into multiple single-link beamforming and propose an iterative searching algorithm to quickly achieve the sub-optimal beam sets. Third, with the proposed algorithm, a comprehensive beamforming protocol is proposed to determine the transmission/reception beam sets operating on MAC layer in a distributive manner. Finally, the analytical evaluation demonstrates that the proposed protocol can significantly reduce the searching space.

5.1 Related Work

Beamforming techniques at 60 GHz band have been actively studied in [47, 48, 50–54] and specified in standards, including IEEE802.15.3c (TG3c) for WPANs, IEEE802.11ad (T-Gad) and Wireless Gigabit Alliance (WiGig) for WLANs. One type of beamforming for mmWave networks is in accordance with the measurement of signals' AOD/AOA and the entire CSI to decide the beam direction [50–54]. A general beamforming problem was formulated in [55] and an automatic alignment mechanism was provided to adaptively track the transmitter location and steers the beam to maximize the received power. In [50], experimental results on AOD estimation and optimal beamforming are presented for 60 GHz smart antennas. Beamforming is accomplished with the control of amplitudes and phases of millimeter-wave signals by means of a maximum directivity beamformer algorithm. The main challenge on this type of beamforming methods is that the channel sounding procedure based on transmission and reception between all pairs of individual elements is not always possible due to the high propagation loss in mmWave band. In addition, the high calculation loads increase the setup time of beamforming and hinder the applications sensitive to delay (e.g., wireless display).

Recently, another type of beamforming protocols realized on MAC layer is proposed in [47, 48, 52, 53] based on switched antenna array with a structured codebook. In [47], by enabling sector-level and beam-level searching, the proposed beamforming protocol targets to minimize the set-up time and to mitigate the high pathloss of 60GHz WPAN systems. In [48], multi-level heuristic searching algorithm are proposed to determine the best transmission/reception beam pattern while reducing the setup time considerably in comparison with exhaustive search. A simple and efficient antenna weight vector training algorithm is proposed in [53] using only one antenna weight vector feedback to obtain the

best transmission and reception antenna weight vector pair. In [47], the beam switching process with pre-specified beam codebooks is described to identify the best beam-pair for data transmissions. The Rosenbrock numerical algorithm is used to implement beam searching, which can reduce the searching spaces and improve the probability of success.

To the best of our knowledge, the previous works on beamforming protocols for indoor mmWave networks are designed for one pair of transmitter and receiver (i.e., single link). Concurrent beamforming is highly required since concurrent transmissions are desired in indoor mmWave networks to increase network capacity, provide higher transmission rates, and support more users, especially for the densely populated networks. It is necessary to jointly consider the beamforming for all the transmission links operating simultaneously, in order to mitigate the mutual interferences and achieve better system performance.

5.2 Beamforming Model

We consider a system with L multiple concurrent communication links. Asymmetric channel is considered since the channels of both communication directions may not be reciprocal. Fig. 5.1 shows the beamforming model for link l between transmitter WDEV_l^T and receiver WDEV_l^R, i.e., aiming to obtain the best transmission beam pattern for WDEV_l^T and the best reception beam pattern for WDEV_l^R. Multiple digital-to-analog converters (DACs) or analog-to-digital converters (ADCs) consume high power because the multiple DACs and ADCs operate at several Giga-samples per second at baseband. To reduce the overhead and energy consumption, the beamforming is operated in radio frequency (RF) band, including only a single RF chain. At the transmitter, the signal after baseband processing is up-converted to the RF band. The RF signal is phase-shifted by applying the transmit weight vector and is transmitted via the MIMO channel. At the receiver, the received RF

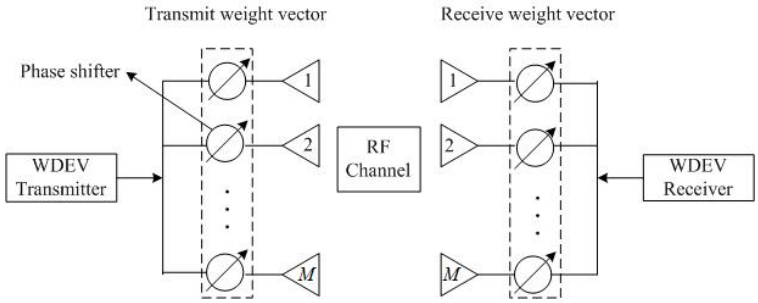

Figure 5.1: Beamforming Structure Model

signal is phase-shifted by the receive weight vector and combined in the RF domain. Then the combined signal is down-converted for baseband processing.

5.3 Codebook Design

The beamforming training with exact phase shift and amplitude adjustment results in large training data and high overheads. Consequently, in both the standards (IEEE 802.15.3c and IEEE 802.11ad) and recent papers [47, 48, 52], 60 GHz band communication prefers the codebook-based beamforming by giving each antenna a phase shift without amplitude adjustment. A codebook is a matrix, in which each column (a weight vector) indicates a phase shift for each antenna element and forms a specific beam pattern. For M antennas with N weight vectors ($N \geq M$), the codebook used in IEEE 802.15.3c WPANs is given by the following matrix:

$$\mathbf{W} = |w_{(m,u)}| = |j^{\lfloor \frac{m \times mod((u+N/2),N)}{N/4} \rfloor}| \tag{5.1}$$

where N is also the total number of beam patterns, $m \in \{0, 1, 2...M-1\}$, $u \in \{0, 1, 2...N-1\}$, and $j = \sqrt{-1}$. The above codebook can result in losing beam gain in some direction-

s [47] if N or M is larger than 4. The reason is that the optimal beam pattern may not be achieved with a limited number of phase shifts for each antenna elements.

To achieve uniform antenna gain in different directions, we use DFT-based codebook design [56]. With the same notations in (5.1), the DFT-based codebook can be given by the following matrix:

$$\mathbf{W} = |w_{(m,u)}| = |e^{-j2\pi(m-1)(u-1)/N}| \tag{5.2}$$

where $m = 1, 2...M$ and $u = 1, 2...N$. The weight vectors in DFT-based codebook are composed of complex numbers which assign the unit amplitude and specific phase shift for each antenna element without losing beam gain in the designed directions.

5.4 Concurrent Beamforming Formulation

Conventionally, beamforming is to search the best transmission/reception beam pattern for each communication link to optimize a cost function, such as signal-to-interference plus noise ratio (SINR). With concurrent links, the SINR of each link depends on both the beam selection of this link and the beam selection of other concurrent links because of the possible mutual interferences. Directional antenna radiates much greater power in certain directions for increased performance on transmission/reception while reducing interference from unwanted sources. Each transmitter WDEV_k^T in link k can generate interference to the receiver WDEV_l^R in link l if they are active simultaneously. The amount of interference depends on the antenna gains of the transmission and reception directions. Since the ultimate goal of beamforming is to transmit more data in the network, we use the sum data rate as the cost function to select the best transmission and reception beam patterns for the concurrent links, instead of the data rate for each link. Given a set of L concurrent links in each time slot, the following sets are defined:

- $B_T = [n_t^1, n_t^2, ...n_t^l, ...n_t^L]$ is the set of beam patterns selected for the L links for data transmission;

- $B_R = [n_r^1, n_r^2, ...n_r^l, ...n_r^L]$ is the set of corresponding reception beam patterns for the L links.

By applying the Shannon capacity formula, the transmission rate of link l can be estimated as

$$\begin{aligned} R_l &= \eta W \cdot log_2(1 + SINR_l) \\ &= \eta W \cdot log_2(1 + \frac{P(n_t^l, n_r^l)}{WN_0 + \sum_{k \neq l} P(n_t^k, n_r^l)}) \end{aligned} \tag{5.3}$$

where W is the system bandwidth, $\eta \in (0, 1)$ is the coefficient describing the efficiency of the transceiver design, and N_0 the one-side power spectral density of white Gaussian noise. $P(n_t^k, n_r^l)$ denotes the received power at the WDEV_l^R from WDEV_k^T. n_t^k and n_r^l are the selected beam patterns for WDEV_k^T and WDEV_l^R, respectively. In (5.3), $n_t^k, n_r^l \in \{1, 2...N\}$ with $k, l = 1, 2...L$.

To maximize the sum rates, we formulate the concurrent beamforming problem as an optimization problem (P):

$$\max \sum_{l=1}^{L} \eta W \cdot log_2(1 + \frac{P(n_t^l, n_r^l)}{WN_0 + \sum_{k \neq l} P(n_t^k, n_r^l)}) \tag{5.4}$$

$$\text{s.t.} \qquad 1 \leq n_t^l, n_r^l, n_t^k \leq N \tag{5.5}$$

$$n_t^l, n_r^l, n_t^k \in \mathcal{Z}^+ \tag{5.6}$$

where $\mathcal{Z}^+$ is the set for all positive integers. This optimization problem is a non-linear integer programming problem. The concurrent beamforming problem is similar to Knapsack

problem [46] in the case that items (antenna patterns) can be put into the knapsack (active in the channel). The objective is to maximize the total profit (sum rates) with the total weight (interference) constraints. It is proved that the Knapsack problem is NP-complete [46]. Concurrent beamforming is even more difficult than the Knapsack problem because the profits of items (rate) is not determined until the active beam patterns of other links are determined. As a result, the concurrent beamforming problem can not be solved by applying existing approximation algorithms for Knapsack problems. There are two main challenges to solve the optimization problem on concurrent beamforming: 1) it is NP-hard without polynomial time solution; 2) the received power $P(n_t^k, n_r^l)$ is dependent on both the transmission/reception beam patterns and the 60 GHz indoor channel status. To obtain the best transmission and reception beam pattern, the channel status has to be available. However, it is almost impractical to obtain the updated channel status for indoor environment due to the moving people and obstacles [28]. Therefore, in this thesis, we propose a MAC-layer beamforming protocol independent of the channel status.

5.5 Iterative Searching Algorithm

In conventional codebook-based beamforming protocol with exhaustive search for single link, the transmitter sends the training sequence with one of its beam patterns, while the receiver attempts to listen to it with different beam patterns. The above process is repeated until all the beam patterns have been tried by the transmitter. Then the best transmission/reception beam pattern can be found by detecting the best SINR at the receiver. The total number of transmission attempts is $N_r \times N_t$ for each link if N_t and N_r denote the number of beam patterns at the transmitter and receiver, respectively. For L concurrent transmission links with mutual interference, there should be $(N_r \times N_t)^L$ total

Algorithm 4 Iterative Searching Algorithm for Beamforming

BEGIN:

L links are scheduled to be active in a time slot of CTAP.

Initialize the beam sets $B_T = \overrightarrow{0}$ and $B_R = \overrightarrow{0}$

repeat

  for link $l = 1$ to L (l is link number index) **do**

    if $n_t^l = 0$ and $n_r^l = 0$ **then**

      set n_t^l and n_r^l according to

$$\arg\max_{n_t^l, n_r^l \in \{1 \ldots N\}} \sum_{\beta=1}^{l} \eta W \cdot log_2(1 + \frac{P(n_t^\beta, n_r^\beta)}{WN_0 + \sum_{k<\beta} P(n_t^k, n_r^\beta)})$$

    else

      set n_t^l and n_r^l according to

$$\arg\max_{n_t^l, n_r^l \in \{1 \ldots N\}} \sum_{\beta=1}^{L} \eta W \cdot log_2(1 + \frac{P(n_t^\beta, n_r^\beta)}{WN_0 + \sum_{k\neq\beta} P(n_t^k, n_r^\beta)})$$

    end if

  end for

until $B_T = \{n_t^l\}$ and $B_R = \{n_r^l\}$ converges

END;

transmission attempts to select the best set of transmission/reception beam patterns. The exhaustive search method is not feasible since the computation and communication loads grow exponentially w.r.t. the number of concurrent links.

To obtain the best transmission beam set B_T^* and reception beam set B_R^*, concurrent beamforming problem is decomposed and conducted link by link to reduce the complexity and setup time, instead of optimizing beam selection for all the L links simultaneously. The decomposed method can solve the problem in an iterative manner and achieve sub-optimal solutions. There are many efficient beamforming protocols (e.g., multi-stage scheme [47] and multi-level scheme [48]) to attain the best transmission/reception beam patterns for single link beamforming. Hence, it is assumed that the efficient beam pattern selection protocol for single communication pair is available and the proposed iterative searching algorithm is to find the beam sets B_T^* and B_R^* for L concurrent links.

The beam sets B_R and B_T are initialized as $B_T = \overrightarrow{0}$ and $B_R = \overrightarrow{0}$. In the initial searching round, the beamforming procedure of the first link is conducted with the beamforming protocol for single communication pair, then $B_T[1]$ and $B_R[1]$ are updated with the selected beam patterns. For the following l^{th} $(1 < l \leq L)$ link, the beamforming is conducted with the single-link beamforming protocol considering the interferences generated from the concurrent links whose beamforming has been determined. The single-link beamforming of link l conducted in concurrent beamforming uses the sum rates as the the selection metric rather than the SINR of link l. By trying different pair of transmission/reception beam in link l, the beam pair resulting maximum sum rates is selected. Then, $B_T[l]$ and $B_R[l]$ are updated. After the transmission and reception beam patterns for all the L links are determined based on the above procedure, another iterative searching round is re-conducted considering the interferences from the other $L-1$ links since there is a predetermined beam pattern for the other $L-1$ links. In each iterative searching round, the obtained beam

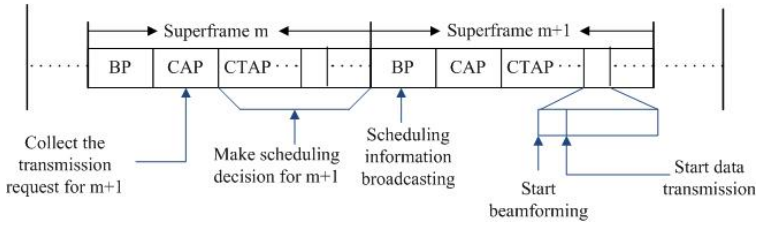

Figure 5.2: Procedure for Scheduling and Beamforming

sets $\widetilde{B}_T$ and $\widetilde{B}_R$ is better than the beam sets B_T and B_R of the previous round in terms of the total throughput.

The above process is repeated until the beam set B_T and B_R converge (i.e., no change of B_T and B_R will result in higher sum rates of all the L links.). In other words, for a specific searching round, if the beam patterns of all the links are not updated, the beam sets B_T and B_R can be determined to converge. The step by step description of the proposed iterative searching algorithm is shown in Algorithm 4.

Remarks: The iterative searching algorithm ensures that we can not find a better transmission/reception beam set by conducting beamforming link by link. However, it is possible to attain a better solution by conducting beamforming for two or more links simultaneously. Therefore, the proposed iterative searching algorithm only obtains a locally optimal solution for concurrent beamforming and can not ensure the global optimality of the solution.

5.6 Beamforming Protocol

In this section, we present a comprehensive beamforming protocol to realize the proposed iterative searching algorithm among WDEVs. As shown in Fig. 5.2, the BS receives a

number of transmission requests in the CAP period of the m^{th} superframe. Then the scheduling decision is made by BS for the $(m+1)^{th}$ CTAP before the $(m+1)^{th}$ beacon period, during which the BS broadcasts the scheduling information to all the WDEVs in the network. The scheduling information indicates the active links in each time slot. The sequence of these active links to conduct beamforming is randomly assigned. Since the BS cannot control the WDEVs during CTAP period, the proposed concurrent beamforming protocol should work distributively. Each time slot of CTAP consists of beamforming period and data transmission period. Thus, the total setup time of concurrent beamforming has significant impact on the resource utilization efficiency. To shorten the total setup time, a concurrent beamforming protocol based on the proposed iterative searching algorithm is presented with four phases, namely, single link beamforming, inter-link notification, iterative searching convergency, and acknowledgement.

The proposed concurrent beamforming protocol is based on the detection of SINR at the receivers. Thus, it can be applied to many mmWave networks to conduct concurrent beamforming without knowledge of channel gain and location information of the WDEVs.

5.6.1 Single Link Beamforming

In the proposed iterative searching algorithm, concurrent beamforming is conducted link by link to reduce the complexity. The existing beamfoming protocols for single link cannot be applied directly due to the following challenges:

1. The interferences from other concurrent links have to be known to determine the best transmission and reception beam patterns for each link. However, other concurrent links cannot conduct beamforming at the same time as the targeted link.

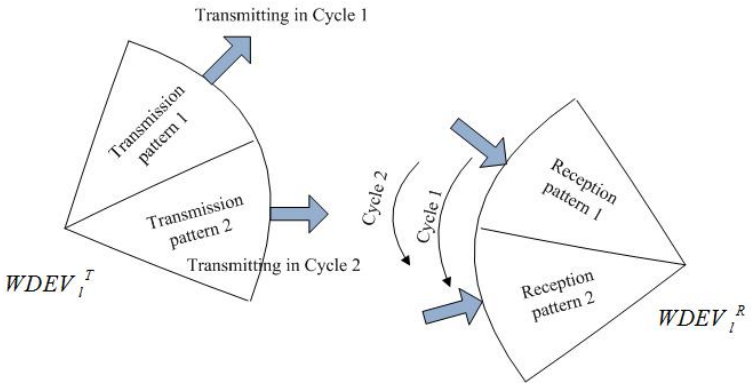

Figure 5.3: An Example for Single Link Beamforming

2. In a distributive manner, it is difficult for a link to know the start/end beamforming time of adjacent links to conduct beamforming sequentially. Since the beamforming duration for each link is quite different, the start time of beamforming for each link cannot be pre-determined even if the time is synchronized.

3. The beam selection metric, i.e., sum rates, is not measurable and each link can not know them by itself.

In BP of the m^{th} superframe, the base station broadcasts the scheduling information to WDEVs to indicate the links to be active in each time slot and the sequence (from 1 to L) for all the active links to conduct beamforming. According to the pre-assigned sequence, each link (e.g., link l) starts beamforming by allowing the transmitter (WDEV_l^T) to send the training sequence with every beam pattern to the receiver (WDEV_l^R) as shown in Fig. (5.3). The transmission beam pattern generates interferences to the receivers of other concurrent links. The SINR values of other concurrent links can be detected at the corresponding receivers by trying each transmission beam pattern of the transmitter WDEV_l^T. Then, the detected SINR values are sent to the receiver WDEV_l^R. For each

transmission beam pattern, the receiver WDEV_l^R switches its beam pattern one by one and records the corresponding received SINR of link l. After traversing every beam pair, WDEV_l^R knows the SINR values of all the links corresponding to each beam pair. WDEV_l^R determines the best beam pair (n_t^l and n_r^l) in terms of sum rates by adding the transmission rates of all the concurrent links together. The training at WDEV_l^R includes N_t cycles and each cycle includes N_r attempts. The training sequence is a long preamble composed of a synchronization sequence and a channel estimation sequence specified in IEEE802.15.3b, generated from Golay code with 32 repetitions of length 128 bits [57].

After determining the best beam pair, WDEV_l^R transmits the feedback information to WDEV_l^T to indicate WDEV_l^T's best transmission beam pattern since WDEV_l^T does not yet know its optimal transmit direction corresponding to WDEV_l^R. Upon the completion of the feedback, the WDEV_l^T activates the selected beam pattern as the interference for beamforming of other links.

5.6.2 Inter-link Notification

Following the Single Link Beamforming phase, Inter-link Notification phase is required to move to next single link beamforming. As indicated in Section 5.6.1, although each link receives the beamforming sequence information during the Beacon period, link $(l+1)$ needs the trigger information to start beamforming right after the completion of link l's beamforming. The transmitter WDEV_l^T activates its optimal beam pattern for SINR considerations of the following links, meanwhile it sends out a notification packet with four fields: header, number of current link (NoC), number of following link (NoF), beam selection identification, as shown in Fig. 5.4.

The information in the Header, NoC and NoF fields, is necessary for other nodes to

Figure 5.4: Format of Each Notification Packet

decide if they should start beamforming currently. After the nodes receive the notification packet, the two nodes forming link $(l+1)$ conduct the single link beamforming triggered by the sequence number $(l+1)$ in the NoF field. The beam selection identification field is used for iterative searching convergency phase as to be described in Section 5.6.3. A short inter-phase spacing (SIPS) time is added for WDEVs to set up beamforming before the single link beamforming phase starts.

Concurrent beamforming is to find the best transmission directions and reception directions to achieve high data rate for mmWave networks supporting bandwidth intensive applications. The notification packet, SINR values, and other management information do not require high transmission rate. It can be supported even with the NLOS transmissions. Thus, it is not necessary to use optimal beam pairs to deliver such information considering the fact that it takes longer time to obtain the optimal beam pairs between WDEVs. With the localization service provided by the mmWave indoor system [58], the BS can obtain the approximate network topology information and broadcast it to WDEVs during CAP. WDEVs can activate the corresponding beam pattern towards each other according to the topology information, to deliver signaling and management information among WDEVs.

5.6.3 Iterative Searching Convergency

With Single Link Beamforming phase and Inter-link Notification phase, beamforming can be conducted link by link to achieve the transmission/reception beam set. To realize the iterative searching algorithm distributively, two main issues need to be considered:

- A Convergency Condition is required to indicate the end of the searching algorithm while determining the accuracy of the obtained solution and the total beamforming time. A strict convergency condition can achieve better solution with more rounds of iterative searching. A loose convergency condition can quickly find the approximate solution.

- Since beamforming is conducted link by link without the control of base station, the convergency condition information should be delivered to WDEVs in a distributive manner.

For concurrent beamforming, the general convergency condition is the iterative searching keeps going until B_T and B_R reach a steady state. Specifically, B_T and B_R converge if n_t^l and n_r^l determined in the n^{th} round are the same as those obtained in the $(n-1)^{th}$ round for all $l \in \{1, 2...L\}$. The above ideal convergency condition can cause a large number of searching rounds and much longer setup time due to the following reasons: 1) A new beam pattern pair (n_t^l and n_r^l) for link l may be found to achieve minor improvement on the sum rates, which brings communication and computation overheads to reduce the system throughput. 2) At system-level beamforming, the new beam pair of one link can lead to the beam pattern re-selection of many other links and more searching rounds because of the mutual interferences. The concurrent beamforming procedure is kind of chain reaction inspired by the the transmission beam re-selection since the transmission beams generate interferences to other links. To significantly shorten the total setup time, a sum rate threshold Th_{rate} in percentages is used to determine the beam pattern re-selection. For example, after conducting single link beamforming for link l, if the new achieved sum rates are within $[1, (1 + Th_{rate}))$ of the sum rates of previous round, the beam patterns of link l remain the same.

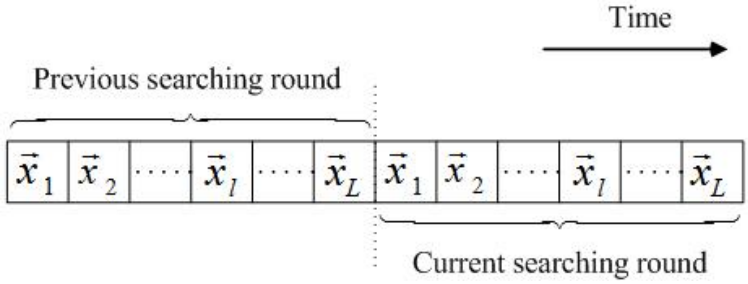

Figure 5.5: Format of Beam Selection Identification in Notification Packet

As shown in Fig. 5.4, Beam Selection Identification Field of the notification packet is used to notify the convergency information among WDEVs. A vector $\overrightarrow{x}_l = [n_t^l, n_r^l]$ is saved in each sub-field of the beam selection identification field to record beam selection result of link l. The format of the beam selection identification field is shown in Fig. 5.5. It records the beam selection results of both the current and the previous searching round for all the L links. If all the $\overrightarrow{x}_l$ $(l = 1, 2, ...l, ..., L)$ of the current round are the same as those of the previous round, it can be concluded that the beam sets B_T and B_R are converged.

5.6.4 Acknowledgement

After a link completes its beamforming for the current searching round, it either waits for the trigger information to start the beamforming for next searching round, or it will stop beamforming after receiving the convergency message and use the selected transmission/reception beam patterns for data transmission. After the receiver of the L^{th} link receives the notification packet and compare the beam selection results of the current and previous searching round, it will broadcast an Acknowledgement (ACK) message to all the WDEVs to indicate the completion of beamforming if the comparison shows that both rounds have the same beam selection results.

The notification packets delivered among WDEVs control the whole process of concurrent beamforming. The following pseudo codes as shown in Algorithm 5 describe the procedure of concurrent beamforming with details on the notification packet transmissions and the corresponding actions done by the WDEVs.

5.7 Performance Analysis

The computational complexity significantly affects the network performance, such as total setup time, transmission efficiency, and energy consumption. Although the proposed iterative searching algorithm reduces the exponential searching space to linear searching space, theoretical analysis is required to show the exact complexity of the proposed algorithm.

To simplified the analysis, the following assumptions are made: a) there is no link blockage between WDEVs and the line-of-sight transmission is always available; b) the channel status does not change during the concurrent beamforming period, thus the beam re-selection of a link can only result from the beam re-selections of other interferers; c) the beam re-selection of a link in the n^{th} searching round is due to the transmission beam re-selections in the $(n-1)^{th}$ round; d) each receiver is randomly located in the area, thus the received power (or interferences) at each receiver is independent. Let X_n be the number of re-selections on transmission beams among L concurrent links in the n^{th} searching round ($X_n \in \{0, 1, 2, ..., L\}$). X_n is a discrete-time Markov chain with transition probability $p_{i,j}$ illustrated in Fig. 5.6. The state transition ends when $X_n = 0$ even if there are some re-selections on reception beams in the n^{th} searching round since the receivers are not interferers.

There are L transmitters and L receivers randomly distributed in the area A. Consider the transmitter WDEV_l^T, receiver WDEV_l^R, and interferer WDEV_k^T shown in Fig. 5.7. The

Algorithm 5 Procedure for Concurrent Beamforming

BEGIN:

Each link l can generate a notification packet
for Each iterative searching round n **do**
  for Link $l = 1$ to L (l is link number index) **do**
    Link l receives a notification packet
    if The $NoF = l$ in the notification packet **then**
      Link l starts beamforming;
      Update $NoC = l$;
      Update $NoF = l + 1$;
      Record the beam selection result $\overrightarrow{x}_l = [n_t^l, n_r^l]$;
      if $l \neq L$ **then**
        Link l sends out updated notification packet;
      else
        The beam selection results of n^{th} round and $(n+1)^{th}$ round are compared;
        if $B_T^n = B_T^{n+1}$ and $B_R^n = B_R^{n+1}$ **then**
          Go to END;
        else
          Update $n = n + 1$;
          Link L sends out message to trigger new round beamforming;
        end if
      end if
    else
      Link l disregards the notification packet
    end if
  end for
end for

END;

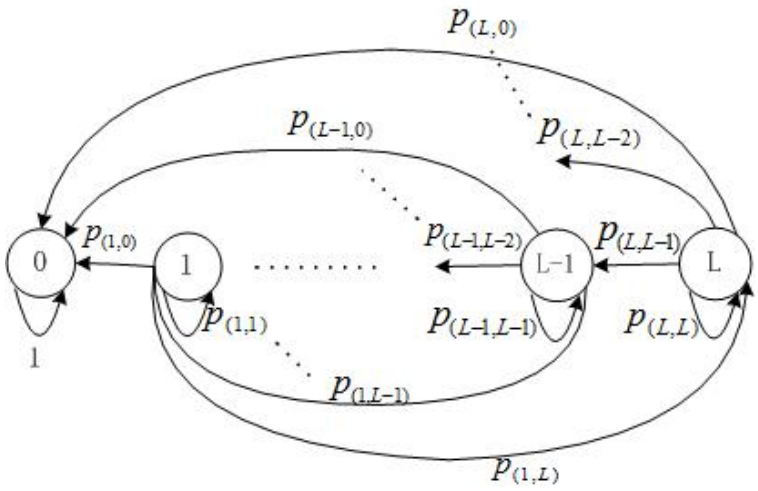

Figure 5.6: Markov Chain for X_n

directional antenna model is described as

$$g(\alpha) = \frac{G(\alpha)}{G_{max}} \text{ where } G_{max} = \max_{\alpha} G(\alpha) \tag{5.7}$$

α describes the horizontal angle in different directions. To make the analysis trackable, the flat-top antenna model is adopted. Specifically,

$$g(\alpha) = \begin{cases} 1, & \mid \alpha \mid \leq \dfrac{\Delta\alpha}{2} \\ 0, & \text{otherwise} \end{cases} \tag{5.8}$$

where $\Delta\alpha$ is the beamwidth. Directional antenna radiates power to all the directions while having focus on specific directions. To reflect this feature, we have $\Delta\alpha > \frac{2\pi}{N}$ where N is the number of beam patterns for each WDEV. The transmitter and the receiver should be within each other's beam in order to communicate with each other. To keep link connectivity after the beam re-selection, only the adjacent beam of the current active beam can be the re-selection candidate based on flat-top antenna model. A receiver can obtain the transmission power from its transmitter after transmission beam re-selection if

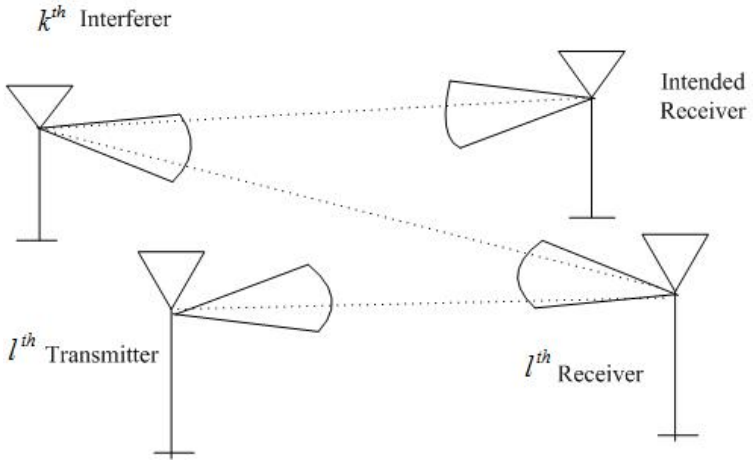

Figure 5.7: Geometry of Directional Interference

the receiver locates within overlap area of the two beams. A sufficient condition to select a new transmission beam for link l to improve the sum rates is that the new beam selection can improve the rate of link l while it does not generate interferences to other receivers.

By standard Friis transmission equation, the received power at WDEV_l^R is

$$P_R^l = P_T G_{max}^2 \frac{\lambda^2}{16\pi^2 d_l^\gamma} e^{-\xi d_l} \tag{5.9}$$

where d_l is the transmission distance, ξ is the attenuation factor due to absorption in the medium, and γ is the the pathloss exponent. Similarly, the interference power from interferer WDEV_k^T is evaluated as:

$$P_{interf}^k = P_T G_{max}^2 f_{k,l} \frac{\lambda^2}{16\pi^2 d_k^\gamma} e^{-\xi d_k} \tag{5.10}$$

where $f_{k,l} = 1$ if WDEV_k^T and WDEV_l^R are within each other's beam; otherwise, $f_{k,l} = 0$. The SINR of link l is

$$SINR = \frac{P_T G_{max}^2 \frac{\lambda^2}{16\pi^2 d_l^\gamma} e^{-\xi d_l}}{WN_0 + \sum_{k \neq l} P_T G_{max}^2 f_{k,l} \frac{\lambda^2}{16\pi^2 d_k^\gamma} e^{-\xi d_k}} \tag{5.11}$$

From (5.11), to improve the rate of link l, the interferences need to be reduced since the transmission beam re-selection does not change the received power if they are still within each other's beam. Three probabilities are defined:

1. The probability that WDEV_1 is located within WDEV_2's beam is $Q_1 = \frac{\Delta\alpha}{2\pi}$;
2. The probability that WDEV_1 is still located within WDEV_2's beam after WDEV_2's beam re-selection is $Q_2 = \frac{(\Delta\alpha - \frac{2\pi}{N})}{2\pi} = \frac{\Delta\alpha}{2\pi} - \frac{1}{N}$;
3. The probability that WDEV_1 moves out of (into) WDEV_2's beam after WDEV_2's beam re-selection is $Q_3 = \frac{\Delta\alpha - (\Delta\alpha - \frac{2\pi}{N})}{2\pi} = \frac{1}{N}$.

In the following, two events are defined: event B is that the transmission beam re-selection on link l can improve the sum rates according to the above sufficient condition, given there are i interferences with transmission beam re-selections in previous searching round while event C is the beam re-selection on the receiver WDEV_l^R. Therefore, we have

$$P(\text{B}) = P(\text{B}, \text{C}) + P(\text{B}, \overline{\text{C}}). \tag{5.12}$$

First, we consider the case that only the transmission beam is re-selected at the transmitter WDEV_l^T to improve the sum rates. The probability that there are h interferers located within the WDEV_l^R's beam among the i interferers is

$$P_h = \binom{i}{h} Q_1^h (1 - Q_1)^{i-h} \triangleq F(i, h, Q_1). \tag{5.13}$$

The probability that there are g interferers whose beams move into or move out of WDEV_l^R's beam by their transmission beam re-selections among the h interferers is

$$P_g = \binom{h}{g} (2Q_3)^g (1 - 2Q_3)^{h-g}. \tag{5.14}$$

Thus, the probability that the total interferences at WDEV_l^R are reduced with g interferers moving into or move out of WDEV_l^R's beam is

$$\begin{aligned} P_1 &= P(\sum_{z=1}^{g} X_z P_T G_{max}^2 \frac{\lambda^2}{16\pi^2 d_z^\gamma} e^{-\xi d_z} < 0) \\ &= P(\sum_{z=1}^{g} X_z \frac{e^{-\xi d_z}}{d_z^\gamma} < 0). \end{aligned} \tag{5.15}$$

where X_z $(z = 1, 2, ..., g)$ are i.i.d. random variables with $P(X_z = 1) = P(X_z = -1) = 1/2$ since receiver has the same probability to move into or move out of the interferer's beam by re-selecting interferer's transmission beam. All the WDEVs are randomly located in the whole area, thus all the d_z are i.i.d. with the same probability density function (pdf) $f(d)$. Let $\vec{X} = \{X_1, X_2, ..., X_g\}$, $\vec{D} = \{d_1, d_2, ..., d_g\}$, and $W(\vec{X}, \vec{D}) = \sum_{z=1}^{g} X_z \frac{e^{-\xi d_z}}{d_z^\gamma}$, then we have

$$\begin{aligned} P_1 &= \sum_{a=1}^{2^g} P(\vec{X} = \vec{X}_a) P(W(\vec{X}_a, \vec{D}) < 0 | \vec{X} = \vec{X}_a) \\ &= \frac{1}{2^g} \sum_{a=1}^{2^g} \int_{W(\vec{X}_a, \vec{D}) < 0} ... \int (\prod_{z=1}^{g} f(d_z)) d\vec{D}. \end{aligned} \tag{5.16}$$

Because $P(X_z = 1) = P(X_z = -1) = 1/2$, all the $P(\vec{X} = \vec{X}_a, a = 1, ..., 2^g)$ have the same probability of $1/2^g$. The probability that i beam re-selections of the transmitters can result in the reduction of total interferences at WDEV_l^R is

$$P_2 = \sum_{h=1}^{i} \sum_{g=1}^{h} P_h P_g P_1 \tag{5.17}$$

According to the sufficient condition to improve sum rates, we have

$$P(\text{B}, \overline{\text{C}}) = Q_2 (1 - Q_1)^{L-1} P_2 \tag{5.18}$$

The probability Q_2 in (5.18) is to let the receiver locate within the overlap area of the adjacent transmission beams such that it can still obtain the power after the transmission

beam is re-selected. The probability $(1-Q_1)^{L-1}$ is to make sure the transmitter does not generate interferences to other $(L-1)$ receivers after transmission beam re-selection.

Similarly, we can obtain the probability $P(\mathrm{B},\mathrm{C})$ as

$$P(\mathrm{B},\mathrm{C}) = \frac{1}{2}Q_2^2(1-Q_1)^{L-1}\sum_{h=1}^{i}\sum_{g=1}^{h}F(i,h,2Q_3)F(h,g,Q_3)P_1 \tag{5.19}$$

Thus the state transition probability is

$$p_{i,j} = \begin{cases} \binom{L}{j}P(B)^j(1-P(B))^{L-j}, \ (1 \le i,j \le L), \\ 0, \qquad\qquad (i=0 \ and \ j \neq 0), \\ 1, \qquad\qquad (i=j=0). \end{cases} \tag{5.20}$$

We can obtain the n-step transition probability matrix as

$$\mathbb{P}^{(n)} = (\mathbb{P})^n = |p_{i,j}^{(n)}|_{(L+1)\times(L+1)} \tag{5.21}$$

The probability that the system enters state $X_n = 0$ from initial state $X_1 = L$ without traversing any state $X_q = 0$ $(1 < q < n)$ is

$$\hat{p}_{L,0}^{(n)} = \sum_{y=1}^{L} p_{L,y}^{(n-1)} p_{y,0} \tag{5.22}$$

The average number of iteration searching rounds from the initial state to the converged beam sets is

$$\mu_n = \sum_{n=1}^{\infty} n \times \hat{p}_{L,0}^{(n)} = \sum_{n=1}^{\infty}\sum_{y=1}^{L} n p_{L,y}^{(n-1)} p_{y,0} \tag{5.23}$$

The searching space of the proposed algorithm, denoted as T_{pro}, is given as

$$T_{pro} = \mu_n L N^2 \tag{5.24}$$

compared with the original searching space with exhaustive search method (denoted as T_{org}) $T_{org} = N^{2L}$.

Given a $\rho \times \rho$ square area A, $f(d)$ is

$$f(d) = \begin{cases} 2\frac{d}{\rho^2}(\frac{d^2}{\rho^2} - 4\frac{d}{\rho} + \pi), & (0 < d \leq \rho), \\ 2\frac{d}{\rho^2}(4\sqrt{\frac{d^2}{\rho^2} - 1} - (\frac{d^2}{\rho^2} + 2 - \pi) \\ -4\arctan(\frac{d^2}{\rho^2} - 1)), & (\rho < d \leq \sqrt{2}\rho). \end{cases} \tag{5.25}$$

With $\rho = 30m$, $L = 10$, $\Delta\alpha = 1.5\frac{2\pi}{N}$, $\gamma = 2$, $\xi = 0$, and $N = 8$, the average number of iteration searching rounds is 15. The searching space can be reduced more than 99%. μ_n provides the average number of iterative searching rounds. For specific cases, the number of searching rounds n depends on the initial search point, the converged point, the convergent rate, and the convergent threshold Th_{rate}.

5.8 Performance Evaluation

In this section, simulation results are provided to demonstrate the performance of the proposed concurrent beamforming protocol, compared with another two beamforming protocols, namely, exhaustive searching concurrent beamforming (ESCB) protocol and isolated sequential beamforming (ISB) protocol. The ESCB protocol finds the best transmission/reception beam sets for all the concurrent links by exhaustive search while the ISB protocol sequentially obtains the transmission/reception beam for each link without consideration of mutual interferences.

We simulate a typical mmWave indoor environment (e.g., large office space) where the BS is placed in the center of the room and all WDEVs are randomly distributed in a $30m \times 30m$ region. The source and destination nodes of each link are randomly selected. The channel model defined in 802.11ad TG [59] is adopted for simulations. The pathloss

Table 5.1: Parameters of shadowing effect

Parameters	**Distribution**
$t_D[s]$	Weibull ($\lambda = 0.59, k = 6.32$)
$t_{decay}[s]$	Weibull ($\lambda = 0.044, k = 2.07$)
$t_{rise}[s]$	Weibull ($\lambda = 0.045, k = 1.76$)
$A_{mean}[dB]$	Gaussian ($\mu = 13.4, \sigma = 2.0$)

Table 5.2: Propagation related parameters

Parameters	**Value**
Central frequency (f)	60 GHz
System bandwidth (W)	1.728 GHz
Transmission power (P_T)	0.1mW
Background noise (N_0)	-134dBm/MHz
Pathloss exponent for LOS (n_{LOS})	2.0
Pathloss exponent for NLOS (n_{NLOS})	1.4
Reference distance (d_{ref})	1.0 m
Pathloss at d_{ref} (PL_0)	68.0 dB
Minimum Rx sensitivity level	-59.3 dBm
Shadow Margin	5.0 dB

Table 5.3: MAC related parameters

Parameters	**Value**
Superframe period	$65535\mu s$
Beacon period	55 μs
Random access period	120 μs
Number of slots in transmission period	817
Slot time	80 μs
Training sequence	3.259 μs
SIFS time (T_{SIFS})	2.5 μs
Notification packet size	2 Kbytes
ACK period T_{ACK}	7.212 μs
Convergency threshold Th_{rate}	0.05
Number of beams N	16

model for conference room environment [59] is given as

$$PL[dB] = A + 20log_{10}(f) + 10\gamma log_{10}(d) \tag{5.26}$$

where γ is the pathloss exponent. $A = 45.5$ and $\gamma = 1.4$ are set for NLOS environment.

To reflect the shadowing effect, the human blockage model defined in 802.11ad channel model document [59] is adopted. Four parameters are used to describe the shadow effect. The duration t_D characterizes the time between the last zero crossing before and the first

zero crossing after the shadowing event. The decay time t_{decay} and the rising time t_{rise} define the time duration between the zero crossings of the signal level and a given threshold of 10 dB in each case. The mean attenuation A_{mean} describes the average attenuation. The shadowing effect model is composed of a series of periods: a linearly decaying period, a period of constant signal level, and a linearly increasing period. The probability distributions of the four parameters are described in Table 5.1. The propagation related and MAC-layer related simulation parameters are shown in Table 5.2 and Table 5.3, respectively.

5.8.1 Total Setup Time

Fig. 5.8 shows the normalized average setup time to conduct concurrent beamforming with respect to the number of concurrent links. For fair comparison, the time to conduct Single Link Beamforming is the same for all the three beamforming protocols. The proposed concurrent beamforming protocol significantly reduces the total setup time compared with ESCB protocol, especially for larger number of concurrent links. For example, with 12 concurrent links, the proposed beamforming protocol can reduce 85% setup time. The total setup time accounts for both the searching time and the time for signaling and management information transmission. Although the proposed concurrent beamforming introduces more communication overheads as described in Sec.5.6.2, the communication overheads are minor in comparison with the total time it saves since it reduces the exponential searching complexity to linear searching complexity.

5.8.2 Network Throughput

Fig. 5.9 shows the normalized system throughput of the three beamforming protocols. The proposed beamforming protocol can achieve much better system throughput than the ISB

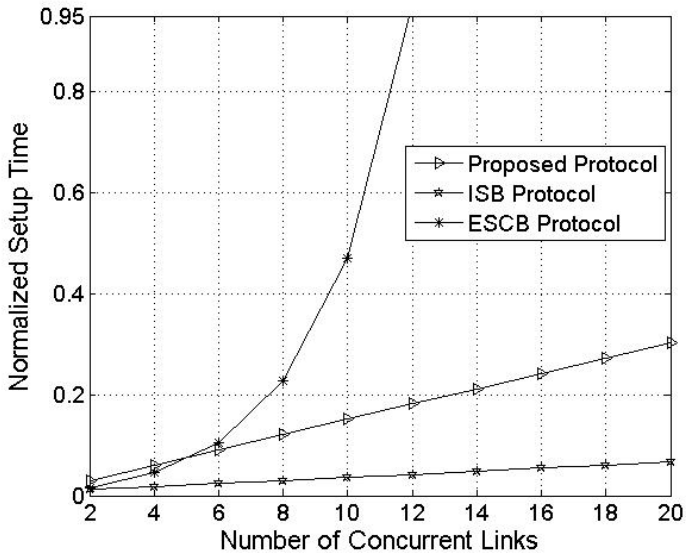

Figure 5.8: Normalized Total Setup Time

protocol, especially in the scenarios with more concurrent links. The ISB protocol conducts beamforming without the consideration of mutual interferences. When the number of concurrent links in the network is small, the mutual interferences have minor impact on the system throughput due to directional transmission and high propagation loss over relatively large transmission distance. For dense networks, the system throughput would be reduced significantly by ISB protocol because of the severe mutual interferences. Additionally, it is demonstrated by Fig. 5.8 and Fig. 5.9, that the proposed beamforming protocol can save significant total setup time with reasonable system throughput reduction compared with ESCB protocol.

The system throughput of the ISB protocol greatly relies on the geographic distribution of the transmitters and receivers. The geographic distribution determines the mutual interferences (not considered in the ISB protocol), and so do the transmission rate and the

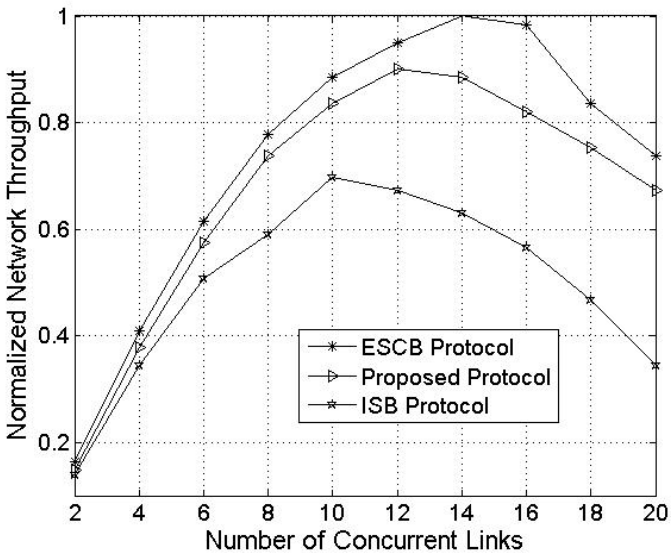

Figure 5.9: Normalized System Throughput

system throughput. Fig. 5.10 shows the system throughputs of the three protocols for 15 concurrent links with 40 random topologies. It is shown that the network throughput of ISB varies much with different topologies while the proposed protocol can achieve more stable system throughput. The multimedia applications operating in mmWave networks require stable or even guaranteed throughput, which the ISB protocol can not provide. The proposed concurrent beamforming protocol determines the beam selection for each link taking into account the interferences and thus it can achieve more stable system throughput.

5.8.3 Energy Consumption

As indicated in the Section 5.6.1, all the interferers (transmitter of other concurrent links) need to be active with the proposed protocol while one link conducts beamforming, in order

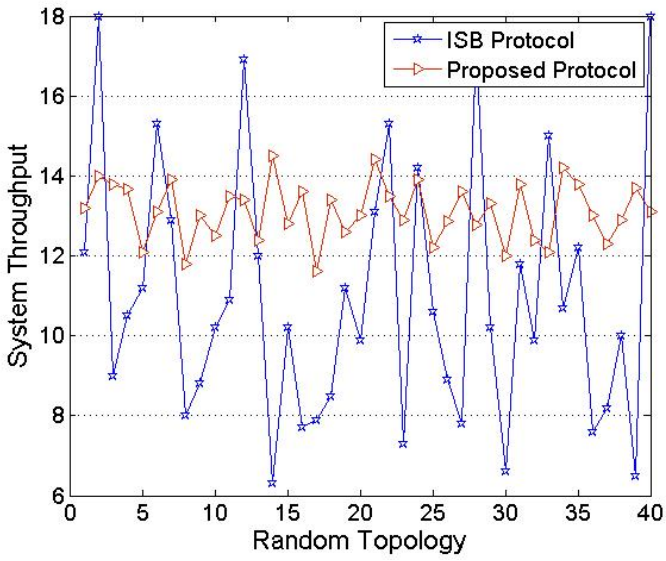

Figure 5.10: Network Throughput Variation with Different Topologies

to consider the interferences. This method introduces extra energy consumption compared with ISB protocol. However, the ESCB protocol consumes much more energy because all the transmitters remain active for the whole beamforming procedure and the ESCB protocol spends much more time to obtain the beamforming results. For fair comparison, we use the total energy consumption (for the transmission of both TSs and communication overheads) divided by achieved system throughput, to show the energy efficiency. As shown in Fig. 5.11, the proposed concurrent beamforming protocol achieves better energy efficiency than ESCB protocol.

5.9 Conclusions of the Chapter

In this chapter, we have proposed a MAC-layer comprehensive beamforming protocol including four phases based on the proposed iterative searching algorithm to setup multiple

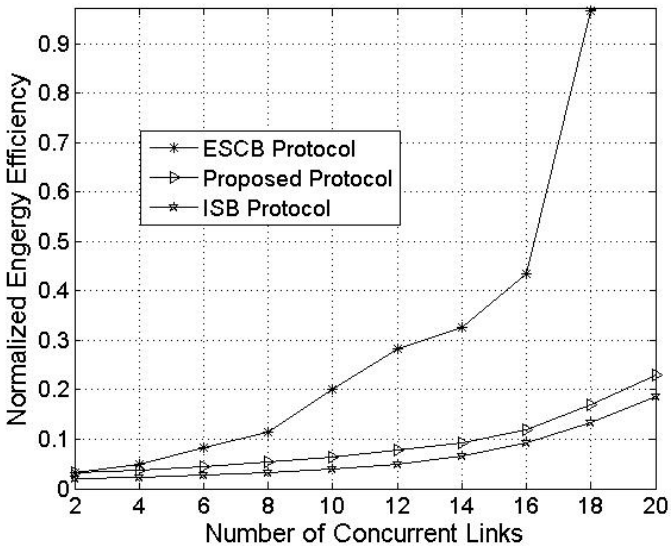

Figure 5.11: Energy Efficiency of the Three Protocols

directional concurrent transmissions. The proposed iterative searching algorithm can provide sub-optimal solution on network throughput to find transmission/reception beam patterns. The theoretical analysis shows the complexity to conduct beamforming can be reduced from exponential searching space to linear searching space. The simulation results have demonstrated significant performance improvements on total setup time and transmission efficiency for dense indoor mmWave networks. The proposed concurrent beamforming protocol has the flexibility to support multiple PHY-layer designs and different antenna configurations for indoor mmWave networks. It should also be useful for other mmWave networks, such as mmWave 5G cellular networks and mmWave based wireless backhaul.

Chapter 6

CSMA/CA-based MAC Protocol Design

Although mmWave communications can achieve instantaneous transmission rate of multi-Gbps, the transmission throughput of each node can not support multimedia applications requiring multi-Gbps throughput, if a large number of nodes contend for the wireless channel [41]. To provide multi-Gbps throughput for each node and support multimedia applications, it is important to improve the network capacity. Increasing the capacity of wireless networks requires increasing the concurrency with which shared channels are accessed or increasing the amount of information sent with each transmission [60]. Multipacket reception consists of the ability of allowing multiple nodes to transmit their packets simultaneously to the same receiver, and the receiver can decode all such packets successfully. Multipacket reception can be implemented by allowing a node to decode multiple concurrent packets using multiuser detection [60,61] or distributed multiple input multiple output (MIMO) techniques. It has been demonstrated that a gain of $\Theta(log(n))$ (n is the number of nodes in the network) can be achieved on network capacity by using multipacket

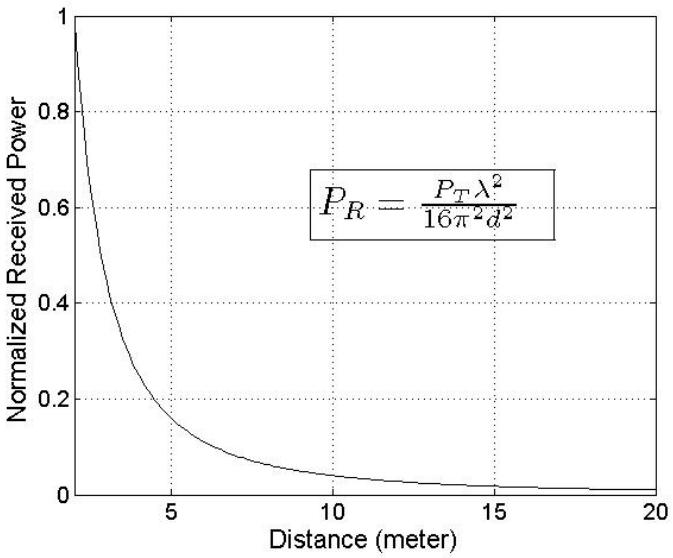

Figure 6.1: Normalized Signal Power over Distance

reception instead of single packet reception [62, 63].

Multipacket reception and the unique features of mmWave communications bring more challenges on mmWave medium access control (MAC) design. Current MAC protocols designed for narrowband systems can not be directly applied to mmWave networks. Indoor mmWave networks are based on the hybrid multiple access of carrier sensing multiple access/collision avoidance (CSMA/CA) and time division multiple access (TDMA) [19,24,25]. Previous work has addressed the MAC design and analysis on the part of TDMA-based MAC in indoor mmWave networks [4,21,41]. In this chapter, we focus on designing efficient CSMA/CA-based MAC protocol for mmWave networks with multipacket reception capability. A basic underlying assumption in the design and evaluation of legacy CSMA/CA-based MAC protocols was that any concurrent transmissions of two or more packets results in collision which leads to a failure in reception of all the packets. The actual situation in

many wireless communication systems is that the packet with the strongest power level can be received successfully (captured) in the presence of contending transmissions. As shown in Fig. 6.1, the mmWave signal power significantly degrades over distance, thus the received signal power of nearby nodes are much stronger than that of distant nodes. Therefore, the channel is always captured by nearby nodes, resulting in serious unfairness. Moreover, for a receiver with multipacket reception capability, much stronger power of the nearest nodes also leads to significant degradation of system throughput. As shown in Fig. 6.2, there are five nodes transmitting to the receiver (network controller) simultaneously. The stronger received signal power from the nearby node results in failure of the other four transmissions from the distant nodes. With multipacket reception capability at the receiver, the system can successfully receive four packets instead of one, if the nearest node does not transmit. Therefore, we propose a backoff mechanism giving higher transmission priority to the distant nodes to achieve better system throughput and fairness, considering the high propagation loss of mmWave communications.

6.1 Related Works

A wide range of MAC-layer protocols and algorithms have been proposed for mmWave networks [4,18,21,40,41,43,44,64,65]. One line of research focuses on transmission scheduling for TDMA-based MAC period to improve the system throughput [4, 18, 21, 41, 43, 44, 65]. The high propagation loss and the utilization of directional antenna result in relatively low multi-user interference (MUI), so that concurrent transmissions can be supported to exploit spatial reuse [4, 18, 21, 43, 44]. In [4], a multi-hop concurrent transmission scheduling (MHCT) scheme is proposed to allow non-interfering transmission links to operate simultaneously over mmWave channels. To further improve the spatial reuse, spatial-time

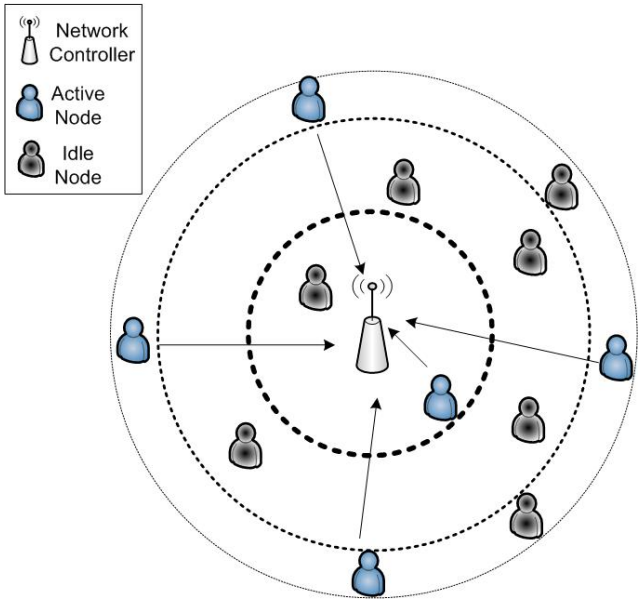

Figure 6.2: Indoor mmWave Network Architecture

division multiple access (STDMA)-based concurrent transmission scheduling schemes are proposed to allow both non-interfering and interfering links to transmit concurrently, either to achieve suboptimal system throughput [41] or to make the accumulated interference in each time slot below a specific threshold [18]. In [21], an exclusive region (ER) based resource management scheme is proposed to exploit the spatial reuse of mmWave WPANs with directional antenna and the optimal ER sizes are analytically derived.

Another line of research on mmWave MAC considers the mutual impact on system throughput of both CSMA/CA-based MAC period and TDMA-based MAC period. A long period of CSMA/CA will cause a low data transmission time in TDMA period, while a short length of access time of CSMA/CA will cause a large number of data transmission collisions in CSMA/CA. In [65], system throughput is optimized by adjusting the access periods of CSMA/CA and TDMA without considering concurrent transmissions.

To the best of our knowledge, there are few works addressing the CSMA/CA-based MAC period for mmWave indoor networks. Most of existing works on mmWave MAC focus on TDMA-based MAC period which is mainly for bandwidth-intensive multimedia applications. Burst-type of applications such as web browsing may not require bandwidth guarantees and use CSMA/CA mechanism to access the channel. To apply applications within much shorter CSMA/CA access period compared with TDMA access period, it is important to improve the network capacity. Multipacket reception is implemented in wireless networks to significantly improve network capacity [66–68]. In [66], the proposed MAC protocol adaptively grants access to the MPR channel to several users to maximize the expected number of successfully received packets in each slot. The system throughput of networks with multipacket reception capability is analyzed in [67] considering spatially distributed nodes. [68] proposes a physical layer multipacket reception technique and the corresponding MAC layer which closely follows the IEEE 802.11 distributed coordination function (DCF) scheme. In this chapter, we propose a new backoff mechanism for CSMA/CA-based MAC for wideband mmWave indoor networks with multipacket reception capability, considering the high propagation loss resulting significant system throughput reduction.

6.2 Packet Capture Model

Since the throughput is one important aspect of the system performance and it is counted by the number of received packets, we describe the packet capture model to decide whether a packet is successfully received. Two packet capture models are adopted, namely, SINR Capture Model and Vulnerability Circle Capture Model. The SINR capture model describes the real situation of packet capture and is used to evaluate the performance of the

proposed backoff mechanism while the vulnerability circle capture model is a simplified model based on the SINR capture model and is used for system throughput analysis.

6.2.1 SINR Capture Model

The channel time is slotted and the transmission duration for packets is one slot long. The propagation model includes pathloss and fading. The received power of a transmission from wireless node i, located at distance d_i from the receiver, is given by

$$P_R(d_i) = P_T K d_i^{-\gamma} F \tag{6.1}$$

where P_T is the transmission power, γ is the pathloss exponent, which is usually determined using a measurement approach (typically in the range of 2 to 6 for indoor environments [28]), K is the attenuation constant, and F is the fading factor. The received signal to interference plus noise ratio (SINR) from node i at the receiver is

$$SINR(i) = \frac{P_R(d_i)}{N_0 W + \sum_{j=1, j\neq i}^{n} P_R(d_j)} \tag{6.2}$$

where N_0 is the one-side power spectral density of noise and W is the system bandwidth. For SINR packet capture model, given a set of simultaneous transmission packets, the packet of node i is received successfully if

$$SINR(i) > h, \tag{6.3}$$

where h is the packet capture threshold ratio. For single packet reception narrowband systems, $1 \leq h \leq 10$, while for wideband multipacket reception systems, such as UWB and mmWave, $h < 1$ [69]. Let M denote the multipacket reception capability, then the

maximum number of packets will be captured if there are M equal received-power packets at the receiver with $M = \lceil 1/h \rceil$ [67, 69], where $\lceil \rceil$ is the ceiling function.

6.2.2 Vulnerability Circle Capture Model

This model was proposed in [70] and studied in [71]. In this model, the node closest to the receiver can capture the channel due to its larger power at the receiver. Specifically, a node i with distance d_i from the receiver captures the channel, if there are no simultaneous transmissions within a disk of radius αd_i ($\alpha > 1$). The parameter α is the vulnerability circle capture ratio. We extend the vulnerability circle capture model to the case of multi-packet reception. Based on SINR capture model, a transmission from a node is successful if

$$SINR(i) > h, (h < 1). \tag{6.4}$$

To achieve this, the received power of the successful packets should be similar to each other. To make the analysis tractable, the successful packets have similar distances to the receiver. Let β denote the number of successful transmission packets. For a receiver with multipacket reception capability, a transmission from a node with distance d_i is successful if there are β ($\beta \leq M$) simultaneous transmissions around the circle with radius d_i and the other simultaneous transmissions (if they exist) are outside of the disk of radius ζd_i, where ζ is the vulnerability circle capture ratio for the case of multipacket reception ($\zeta > 1$). Since mmWave signal power degrades significantly over distance, the value of α and ζ could be relatively smaller than those at lower frequency bands.

6.3 Backoff Mechanism Design

In CSMA/CA-based MAC, the backoff mechanism controls the transmission probability of each node. In a general backoff mechanism, each node i sets an integer backoff counter $B_{i,j}$ randomly generated from a contention window $W_{i,j}$, i.e., $B_{i,j} \in \{0, 1, ..., W_{i,j} - 1\}$. The subscript j represents the backoff stage of node i. The contention window size is re-set after a transmission attempt and node i re-starts to transmit packet after $B_{i,j}$ time slots.

In this section, we first present the design considerations for the proposed backoff mechanism. Then, a detailed backoff mechanism is proposed to adjust the contention window size considering the unique features of indoor mmWave systems, aiming to improve the system throughput and fairness.

6.3.1 Design Considerations

A transmission failure can occur in two independent scenarios: 1) the nearby nodes with stronger received power satisfy the SINR capture model $SINR(i) > h$, thus the transmissions from distant nodes fail; 2) the transmissions arriving at the receiver are only from distant nodes and the received SINR of each node is not larger than the packet capture threshold ratio h in (6.3) if there are too many transmissions. Thus no packet captures the channel. For the first scenario, we can mitigate the channel capture effect of nearby nodes by giving higher transmission probability to distant nodes and the receiver can receive more packets because of its multipacket reception capability. The second scenario indicates that much higher transmission probability of the distant nodes can also result in system throughput reduction. It would make the case worse if the contention window size were further reduced after a transmission failure of distant nodes. Consequently, to achieve better system throughput, the transmission probability needs to be properly adjusted.

The transmission probability controlled by the backoff mechanism has a strong impact on both system throughput and fairness. Giving higher priority to the distant nodes can achieve higher system throughput due to the multipacket reception capability at the receiver. Meanwhile, the nearby nodes would have less chance to access the channel and may even suffer starvation problem. During the contention access period in each superframe, the transmission requests are also sent from the nodes which are active during the next channel time allocation period. The number of received transmission requests has significant impact on the overall system throughput of mmWave indoor networks based on the hybrid multiple access of CSMA/CA and TDMA [65]. Thus, the proposed backoff mechanism needs to avoid starvation to nearby nodes if we give higher transmission priority to distant nodes.

We propose a novel backoff mechanism for CSMA/CA based MAC with the following considerations: 1) Identify different scenarios for packet transmission failure and use different backoff strategies to deal with them in order to increase the number of successful packets. 2) Try to avoid starvation problem in case there is large delay for the transmission requests of those flows operating in channel time allocation period. In addition, the performance of response time would be improved by dealing with the starvation problem. 3) The unique features of mmWave communication should be considered, e.g., the mmWave signal strength is very sensitive to transmission distance due to its high propagation loss.

6.3.2 Proposed Backoff Mechanism

Initially, with neighbor discovery and beamforming technologies, the wireless nodes register with the network controller and train their antenna array in the direction of the receiver to maximize the received signal power at the receiver. With the accurate localization service

provided by the mmWave indoor system [58], the network controller has the valid network topology information. Considering that all the nodes are randomly distributed in a circular region, there are more nodes located in the distant area, thus the transmission probability of distant nodes could have great impact on the number of transmissions arriving at the receiver. To attain more successful transmissions, the contention window size of distant nodes should be adjusted more moderately compared to nearby nodes. Additionally, mmWave signal power degrades significantly over distance and the received signal powers of all the transmissions determine whether a transmission is successful or not. Therefore, we use $d^{-\gamma}$ as the coefficient of the interval to adjust the contention window size. Specifically, the adjustment interval for a node i after a transmission failure is $\lceil d_i^{-\gamma} W_w \rceil$ while the adjustment interval for a node i after a transmission success is $\lceil d_i^{-\gamma} W_r \rceil$, where W_w and W_r denote the basic backoff intervals after a transmission failure and a transmission success, respectively. To give a higher transmission probability to the node with a transmission failure than that with a transmission success, we have $W_r > W_w$.

Both much higher transmission probability and lower transmission probability for distant nodes can result in system throughput reduction; therefore we adjust the contention window after a transmission attempt according to the network congestion status. If the scenario in which transmission fails with no packet successfully received at the receiver occurs frequently, it is very likely that the transmission probability of distant nodes is much higher. Then we enlarge their contention window to reduce the transmission probability. Otherwise, the channel is captured by nearby nodes and we reduce the transmission probability of nearby nodes by increasing their contention window size while the transmission probability of distant nodes needs to be increased by reducing their contention window size.

During the k^{th} slot of the CAP period in the m^{th} superframe, the network controller

Algorithm 6 Backoff Mechanism for CSAM/CA

BEGIN:

1: **for** slot k of CAP in m^{th} superframe **do**
2: Several nodes send packets to network controller
3: For each transmission from node i, network controller determines packet reception according to

$$SINR(i) > h$$

4: **if** Multiple packets arrive but no successful packets **then**
5: Set $f_a(k) = 1$
6: Update $F_{k,m} = \frac{\sum_{l=k-T+1}^{k} f_a(l)}{T}$
7: **if** $F_{k,m} > F_{thr}$ **then**
8: Update contention window $W_i = W_i + \lceil d_i^{-\gamma} W_w \rceil$
9: **else**
10: Update contention window $W_i = W_i - \lceil d_i^{-\gamma} W_w \rceil$
11: **end if**
12: **else**
13: Set $f_a(k) = 0$
14: **if** Transmission of node i is successful **then**
15: Update contention window $W_i = W_i + \lceil d_i^{-\gamma} W_r \rceil$
16: **else**
17: Update contention window $W_i = W_i - \lceil d_i^{-\gamma} W_w \rceil$
18: **end if**
19: **end if**
20: **if** $W_i \geq W_{max}$ **then**
21: Update contention window $W_i = W_{max} - \lceil d_i^{-\gamma} W_w \rceil$
22: **end if**
23: Update $k = k + 1$
24: **if** No more slots in m^{th} CAP **then**
25: FROZEN
26: Until the CAP of next superframe
27: Update $m = m + 1$
28: **end if**
29: Go to line 1
30: **end for**

END;

receives a number of transmission packets and determines which transmission packets are captured, based on the SINR packet capture model in (6.3). Then it broadcasts a feedback packet to all the nodes in the network to indicate the transmission failures and transmission successes. We use packet capture failure frequency ($F_{k,m}$) to determine the contention window size. $F_{k,m}$ is defined as the number of time slots (during the previous T time slots from the current k^{th} time slot in the m^{th} superframe), in which there are transmissions arriving at the receiver but no packet is captured successfully, divided by T. $F_{k,m}$ is given as

$$F_{k,m} = \frac{\sum_{l=k-T+1}^{k} f_a(l)}{T} \tag{6.5}$$

where $f_a(l)$ is

$$f_a(l) = \begin{cases} 1, & M_T(l) > 1 \text{ and } M_R(l) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.6}$$

and $M_T(l)$ and $M_R(l)$ indicate the number of transmissions and successful receptions in the l^{th} time slot, respectively. If $F_{k,m}$ is larger than a specific threshold F_{thr}, the network becomes congested because of the higher transmission probability of distant nodes, thus the contention window of each transmission node i in the current time slot is increased by $\lceil d_i^{-\gamma} W_w \rceil$ to reduce transmission probability. Otherwise, the contention window of node i is increased by $\lceil d_i^{-\gamma} W_r \rceil$ following a transmission success and decreased by $\lceil d_i^{-\gamma} W_w \rceil$ after a transmission failure. The contention window size can be adjusted until it reaches the maximum value W_{max}. To deal with the starvation problem, when the contention window of node i is larger than or equal to W_{max}, it is set as $W_{max} - \lceil d_i^{-\gamma} W_w \rceil$ to increase its transmission probability. When all the transmissions are not successful at the receiver, although the contention window size is increased, it does not mean that the transmission probability after a transmission failure (p_w) is less than that after a transmission success (p_r). In this case, the transmission probability of distant nodes is much higher, thus we

reduce it moderately to achieve better system throughput. It is very likely that p_w is still larger than p_r after the contention window size is increased for the nodes with transmission failure.

The contention access period is not continuous in different superframes. Therefore, the backoff status of each node is frozen at the end of each contention access period and restarts at the beginning of next contention access period in the following superframe. The detailed procedure of the proposed backoff mechanism is described by the pseudocode in Algorithm 6.

6.4 System Throughput Analysis

In this section, we analyze the system throughput for indoor mmWave networks with multipacket reception capability. The packet capture model is based on the vulnerability circle capture model described above. The performance analysis of this section is to theoretically demonstrate that the transmission probability adjustment in the proposed backoff mechanism can significantly improve the system throughput, compared with traditional backoff mechanisms.

The transmission states of a wireless node can be described as the Markov state diagram shown in Fig. 6.3. A node can be in two states: AS (after success) and AF (after failure). State transition may take place after a transmission attempt. A node moves into the AS (AF) state after a successful (failed) transmission of a packet. Since transitions take place after an attempt, a node does not change its state following an idle slot. A node in the AS state transmits with probability p_r at each slot, disregarding the status of the channel while the transmission probability is p_w for a node in the AF state. The values of p_r and p_w represent the size of contention window in the backoff mechanism.

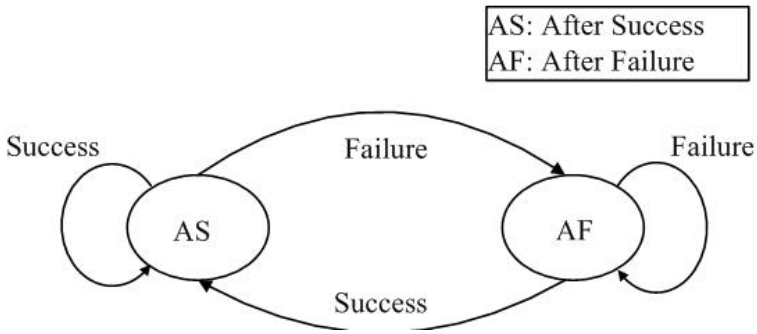

Figure 6.3: The State Transition Diagram

For example, a high value of p_r corresponds to maintaining a small contention window following a successful transmission in the traditional backoff mechanisms as the one used in IEEE 802.11. Similarly, a low p_w value corresponds to maintaining a large contention window following a transmission failure. Therefore, traditional backoff mechanisms have a large value of p_r and a small value of p_w, which corresponds to the event the contention window is increased after a transmission failure and the contention window is reduced after a successful transmission.

Generally, when a node transmits a packet, the loss probability due to collisions depends on other transmissions during that slot. However, in most work on performance analysis, the loss probability of a packet transmitted by a node at distance d, defined as $PF(d)$, is assumed to be independent of the number of retransmissions suffered. The validity and accuracy of the assumption have been recently verified. The state transition probability from state AS to state AF is $p_r PF(d)$. Similarly, the state transition probability from state AF to AS is $p_w(1 - PF(d))$. Hence, the transition probability matrix is given as

$$\mathbb{P} = \begin{bmatrix} 1 - p_r PF(d), & p_r PF(d) \\ p_w(1 - PF(d)), & 1 - p_w + p_w PF(d)) \end{bmatrix} \tag{6.7}$$

With the transition probability matrix, steady-state probabilities of states AS (π_{AS}) and AF (π_{AF}) for a node at distance d can be obtained by solving the following equations:

$$\begin{cases} \pi_{AS} + \pi_{AF} = 1 \\ \overline{\pi}\mathbb{P} = \overline{\pi} \end{cases} \tag{6.8}$$

where $\overline{\pi} = \{\pi_{AS}, \pi_{AF}\}$. Then, we have

$$\begin{aligned} \pi_{AS} &= \frac{p_w(1 - PF(d))}{p_r PF_d + p_w(1 - PF(d))} \\ \pi_{AF} &= \frac{p_r PF(d)}{p_r PF_d + p_w(1 - PF(d))} \end{aligned} \tag{6.9}$$

The total transmission probability of a node at distance d is

$$\begin{aligned} p(d) &= \pi_{AS} p_r + \pi_{AF} p_w \\ &= \frac{p_r p_w}{p_w - p_w PF(d) + p_r PF(d)} \end{aligned} \tag{6.10}$$

According to the vulnerability circle capture model, for a receiver with single packet reception capability, a transmission from a node at distance d succeeds if there is no simultaneous transmission among the other $(n-1)$ nodes within a disk with radius αd around the receiver, where n is the total number of transmission nodes in the network. For a receiver with multipacket reception capability M, a transmission of a node at distance d becomes successful if there is no simultaneous transmission among the $(n-\beta)$ $(2 \leq \beta \leq M)$ nodes within a disk of radius ζd around the receiver; meanwhile the β simultaneous transmissions are distributed around the circle with radius d. β is the number of simultaneous transmissions received successfully. The probability that all the $(n-\beta)$ nodes do not transmit packets within the disk of radius ζd is given as

$$\mathcal{P}' = [1 - \int_0^{\zeta d} p(y) f(y) dy]^{n-\beta} \tag{6.11}$$

where $f(d)$ is the probability density function of the distance from a node to the receiver, and $p(d)$ is transmission probability of a node at distance d. To achieve successful reception of β packets simultaneously, the transmitting nodes of β packets need to be distributed around the circle of radius d. For example, the packet transmission of a node located within the disk of radius d/ζ can result in the transmission failure of a node at distance d according to the vulnerability circle capture model. To make the analysis tractable, it is assumed that the other $(\beta-1)$ simultaneous transmitting nodes are distributed within the area between the circle of radius $d(1+\varepsilon)$ and the circle of radius $d(1-\varepsilon)$. ε is a parameter determining the location area for the $(\beta-1)$ transmitting nodes. Thus, the corresponding probability is

$$\mathcal{P}'' = [\int_{d(1-\varepsilon)}^{d(1+\varepsilon)} f(y)dy]^{\beta-1} \tag{6.12}$$

Therefore, the transmission failure probability for a node at distance d for multipacket reception case is

$$PF(d) = 1 - \binom{n-1}{\beta-1}\mathcal{P}'\mathcal{P}'' \tag{6.13}$$

Substituting (6.11) and (6.12) into (6.13), we have

$$PF(d) = 1 - \binom{n-1}{\beta-1}[\int_{d(1-\varepsilon)}^{d(1+\varepsilon)} f(y)dy]^{\beta-1}[1 - \int_0^{\zeta d} p(y)f(y)dy]^{n-\beta} \tag{6.14}$$

The throughput of a node at distance d from the receiver is

$$C(d) = p(d)(1 - PF(d)) \tag{6.15}$$

and the total system throughput is given as

$$\begin{aligned} C &= n\int_0^D C(y)f(y)dy \\ &= n\int_0^D f(y)p(y)(1-PF(y))dy \end{aligned} \tag{6.16}$$

where D is the radius of the whole communication area. Given the probability density function $f(d)$, the transmission probability of a node at AS state p_r, and the transmission probability of a node at AF state p_w, we can obtain the system throughput C numerically. If all the nodes are uniformly distributed in the disk of radius D with the network controller in the center, we have

$$f(d) = \begin{cases} \dfrac{2d}{D^2}, & (0 < d \leq D) \\ 0, & \text{otherwise} \end{cases} \tag{6.17}$$

With multipacket reception capability for $M = \beta = 4$, Fig. 6.4 shows the numerical results of system throughput for different combinations of p_r and p_w. In traditional CS-MAC/CA, the transmission probability after transmission failure p_w is less than that after transmission success p_r, in order to reduce the transmission collision. In Fig. 6.4, we show the system throughput for the cases of both $p_w < p_r$ and $p_w > p_r$. It can be seen that the network with multipacket reception capability can achieve much higher system throughput with the case of $p_w > p_r$, compared with the traditional case of $p_w < p_r$. Initially, the system throughput increases as more nodes transmit packets in the network, and then it decreases since more collisions occur due to larger number of nodes involved in packet transmission. For the case of $p_w > p_r$, much larger p_w can result in fast degradation of system throughput.

In summary, the above analysis demonstrates that, in the case of multipacket reception, the system throughput can be improved by giving a higher transmission probability to the node after a transmission failure than that after a transmission success. Meanwhile, much higher transmission probability after transmission failure can also result in system throughput reduction. The above analysis theoretically verifies the performance of the proposed backoff mechanism.

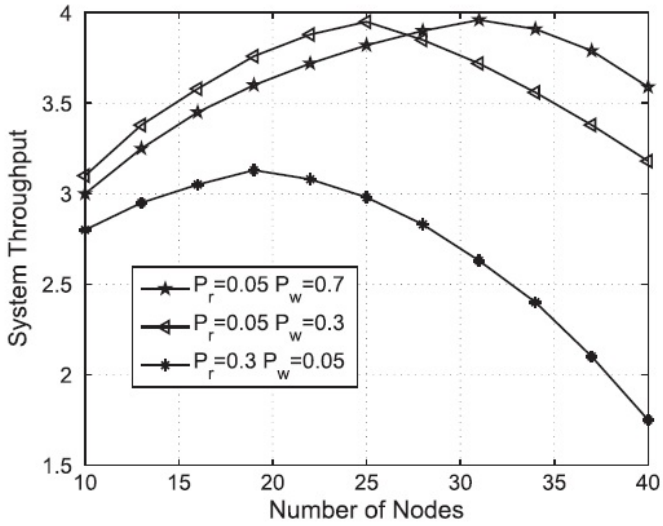

Figure 6.4: System Throughput of Different Combinations of Transmission Probabilities

6.5 Performance Evaluation

In this section, we describe the performance evaluation settings and present the simulation results for the proposed backoff mechanism compared with two other exponential backoff mechanisms.

We evaluate the performance of the proposed backoff mechanism in terms of system throughput, fairness, and power consumption in a typical mmWave indoor environment (i.e., large office space). The network controller is placed in the center of the room and all wireless nodes are randomly distributed in the circular region with a radius of 20 meters. Each node is equipped with a directional antenna with a beamwidth of 60°, corresponding to six beams at each node. Each packet is one-slot long and the packet capture is based on SINR packet capture model with packet capture threshold ratio $h = 0.25$ (i.e., multipacket reception capability is 4). The signal propagation model is based on the free space Friis

Table 6.1: SIMULATION PARAMETERS

Parameters	**Symbol**	**Value**
System bandwidth	W	1200 MHz
Transmission power	P_T	0.1mW
Background noise	N_0	-134dBm/MHz
Reference distance	d_{ref}	1.5m
Pathloss at d_{ref}	PL_0	71.5 dB
Slot time	ΔT	$10\mu s$
Beacon period	T_{BEA}	50 μs
Random access period	T_{RAP}	80 ms
Channel time allocation period	T_{CTAP}	500 ms
Maximum contention window	W_{max}	10000
Basic success backoff interval	W_r	300
Basic failure backoff interval	W_w	200
Congestion threshold	F_{thr}	0.3
Congestion concern duration	T	20

model with pathloss exponent $\gamma = 2$. The reference distance is set to 1.5 meters, which is also used to bound the adjustment interval since it is proportional to $1/d^{\gamma}$. The main parameters used in our simulations are listed in Table 6.1.

We compare the proposed backoff mechanism with two other exponential backoff mechanisms, namely, traditional exponential backoff mechanism and alternative exponential backoff mechanism. In traditional exponential backoff mechanism, after every transmission failure, the contention window size is doubled while it is reduced half after a transmission success. The alternative exponential backoff mechanism doubles the contention window size after a transmission success and reduces the contention window by half following a transmission failure. We use the performance of traditional exponential backoff mechanism as the baseline for comparison. The alternative exponential backoff mechanism gives higher transmission probability after a transmission failure. However, its exponential backoff would give much higher transmission probability after transmission failure and results in congestions.

Fig. 6.5 shows the normalized system throughput of the three backoff mechanisms as a function of the number of nodes in the network. The alternative exponential backoff mechanism and proposed backoff mechanism can achieve higher system throughput compared with traditional exponential backoff mechanism since it gives higher transmission probability after a transmission failure. As more nodes are involved in the network, the proposed backoff mechanism adjusts the contention window considering the network congestion status, thus achieves higher system throughput in comparison with the alternative exponential backoff mechanism.

Fig. 6.6 shows the average throughput per node at different distances to indicate the fairness. We run the simulation 100 times with 30 nodes randomly distributed in the network. With the proposed backoff mechanism, the throughput of each node does not

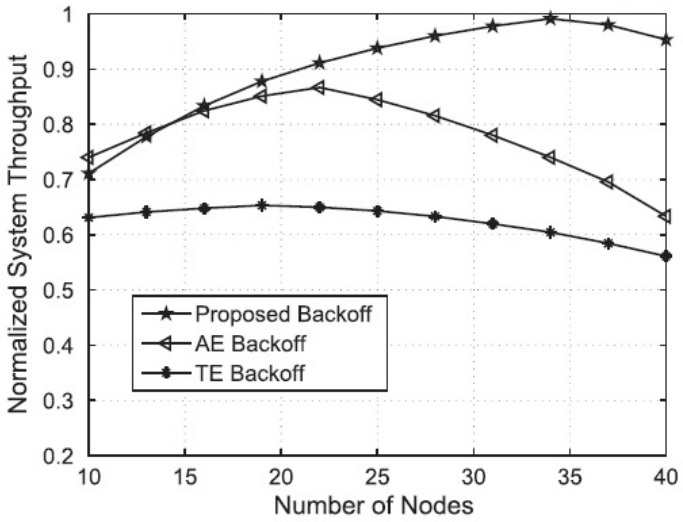

Figure 6.5: Normalized System Throughput of Three Backoff Mechanisms

decrease much with the distance, while with traditional exponential backoff mechanism, the throughput of distant nodes is much less than that of nearby nodes. The proposed backoff mechanism considers both the network congestion status and the transmission status of each node to adjust the transmission probability, and achieves better fairness. Although the achieved gain on the throughput of each node is not that much, the overall system throughput would be improved significantly since there are many nodes in the network.

Transmission failure probability is the probability that a transmission attempt experiences a failure. Note that a packet can suffer multiple transmission failures before it is successfully received. Fig. 6.7 shows the transmission failure probabilities of the three backoff mechanisms. They have similar performance on transmission failure probability if there are not many nodes in the network. The alternative exponential backoff mechanism gives much higher transmission probability to distant nodes and more nodes are distributed in the distant area. Thus, there are many transmissions coming from distant area and the

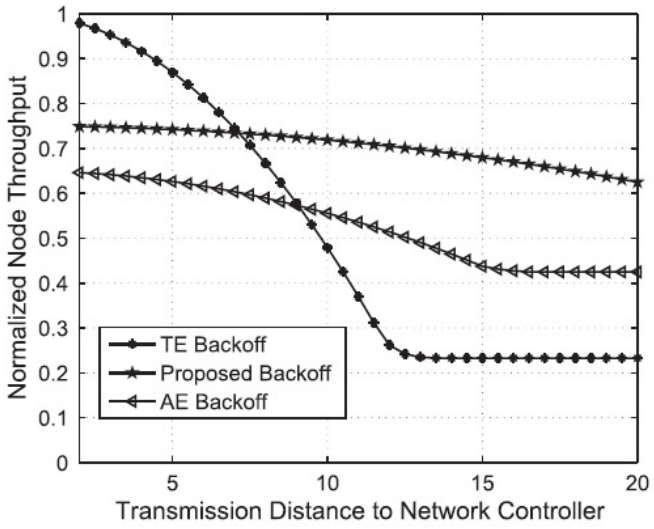

Figure 6.6: Average Node Throughput with Different Distance for Fairness

SINR of each can not exceed the packet capture threshold ratio. As a result, the transmission failure probability of the alternative exponential backoff mechanism increases rapidly with the increased number of nodes. Although the transmission failure probability of the alternative exponential backoff mechanism exceeds that of the others for a dense network, the network throughput of it is still more than that of the traditional exponential backoff mechanism since there are more transmission attempts with the alternative exponential backoff mechanism due to its higher transmission probability for distant nodes.

The normalized average packet delays of the three mechanisms are shown in Fig. 6.8. The packet delay is defined as the time duration from the time the packet is transmitted to the time the packet is received successfully. As the number of nodes increases, the network becomes more congested and the packet delay increases. The proposed backoff mechanism has shorter packet delay corresponding to less congestions in the network.

We then compare the energy consumption of the proposed backoff mechanism with that

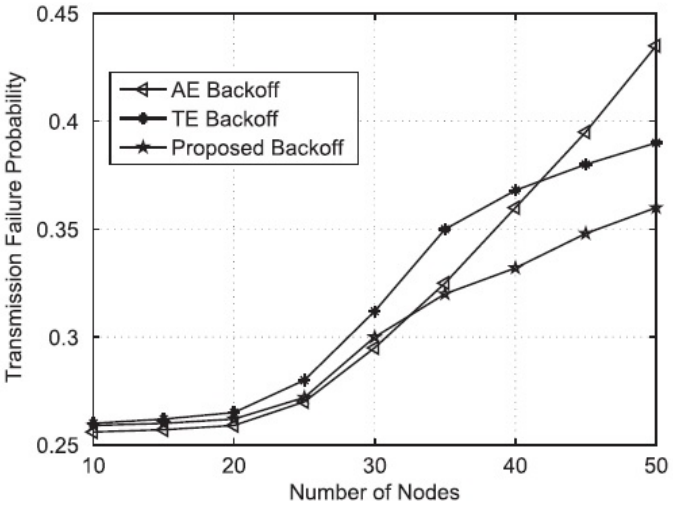

Figure 6.7: Transmission Failure Probability

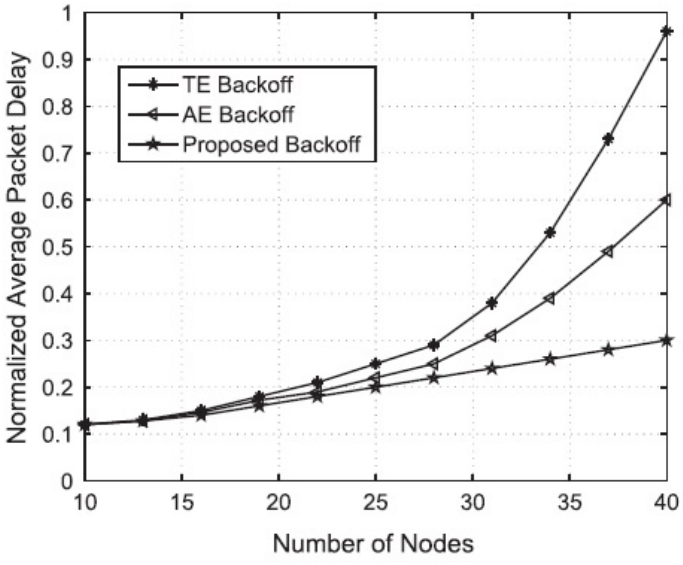

Figure 6.8: Average Packet Delay

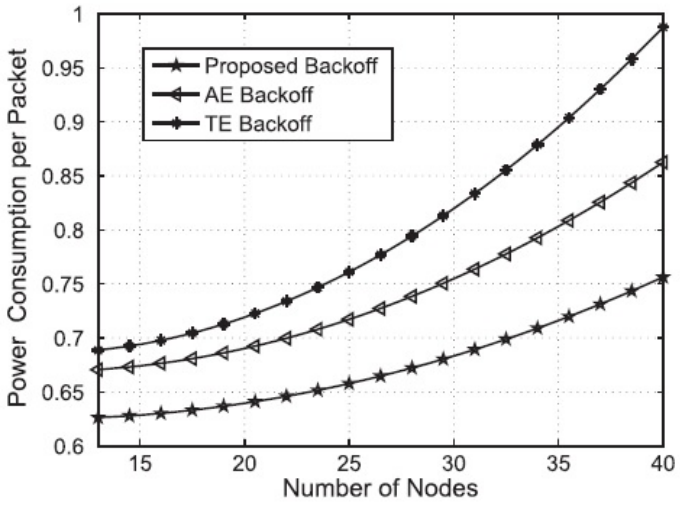

Figure 6.9: Normalized Power Consumption per Packet

of the other two exponential backoff mechanisms. For fair comparison, we use the total transmission energy (consumed by both successful and unsuccessful packets) divided by the total number of successful packets, to obtain the energy consumption per successful packet. From Fig. 6.9, we can see that our proposed backoff mechanism is more energy-efficient than the other two exponential backoff mechanisms. By adjusting the contention window, the proposed backoff mechanism can reduce the number of transmission failures (energy waste) and make the receiver accept more packets.

The proposed backoff mechanism can achieve better performance on system throughput, fairness, average packet delay and energy consumption. Meanwhile, it introduces communication overheads and computational overheads as indicated in the proposed mechanism. The main communication overheads are the feedback packets from the network controller to other nodes in the network. However, in traditional backoff mechanisms in CSMA/CA, there are still ACK packets from the receiver to confirm the packet reception success. In

the proposed backoff mechanism, the nodes need to compute packet capture failure frequency ($F_{k,m}$) and compare it with the threshold. It is a simple computation and does not consume much computational power. Therefore, the proposed backoff mechanism can significantly improve the network performance with limited extra cost.

6.6 Conclusions of the Chapter

In this chapter, we have proposed a backoff mechanism for CSMA/CA-based MAC for mmWave indoor networks with multipacket reception capability. Generally, the proposed backoff mechanism gives higher transmission probability to distant nodes to take the advantage of multipacket reception capability at the receiver. The transmission probability can be adjusted by changing the contention window size according to the node's transmission status (failure or success) and network congestion status. With the proposed backoff mechanism, the mmWave indoor networks can achieve higher system throughput and better fairness.

To the best of our knowledge, this work is one of the first attempts to design CSMA/CA-based MAC for mmWave networks with multipacket reception capability, considering the severe signal power degradation over distance in mmWave band. Since different parameter settings can impact the system performance, we intend to develop mechanisms selecting appropriate parameter values to achieve optimal system performance, e.g., system throughput. Moreover, mmWave links are highly susceptible to blockage because of the limited ability to diffract around obstacles such as moving people and furniture in indoor environment. Therefore, designing efficient CSMA/CA-based MAC for multi-hop scenario is an interesting and challenging problem in mmWave indoor networks.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

mmWave communication with huge available bandwidth provides a promising solution for 5G cellular networks to achieve high capacity. In this thesis, we developed MAC-layer protocols to effectively and efficiently exploit the large available mmWave bandwidth to satisfy high capacity demands of 5G cellular networks while overcoming the challenges on QoS provisioning, brought by mmWave propagation characteristics, to support various kinds of applications. This thesis mainly focuses on MAC-layer modeling, design and analysis of mmWave network to provide efficient MAC protocol, required capacity, and required QoS to various kinds of applications in 5G cellular networks. Multi-user beamforming based on codebook is conducted to determine best transmission/reception beams to increase network capacity. Concurrent transmission scheduling are enabled to exploit spatial reuse and improve network capacity. Link outage problem of mmWave communication is addressed by enabling multi-hop transmissions (for low-mobility scenarios) and buffer design with bandwidth allocation (for high-mobility scenarios). Finally, a novel backoff mechanism for

CSMA/CA is proposed to achieve better fairness and higher network capacity.

7.2 Further Research

My long-term research objective is to establish the overall cross-layer architecture of 5G cellular networks, which incorporates 4G cellular networks and small-cell mmWave networks to provide extremely high capacity and satisfying QoS of various kinds of emerging applications, with perfect solutions to numerous challenges, in a more efficient, reliable, sustainable, and secure way. My mid-term research objective is to develop MAC-layer protocols and algorithms to exploit the large available bandwidth while addressing the challenges brought by unique propagation features of mmWave communication. New stochastic modeling and optimization techniques will be developed for fast directional neighbor discovery in frequent handoff; network capacity will be analyzed with wireless backhaul constraint and considering emerging communication modes; and mmWave base station sleep scheduling based on burst multimedia traffic will be conducted to achieve energy efficiency.

7.2.1 Fast Directional Neighbor Discovery

Severe propagation loss confines mmWave communication range. Thus, mmWave base stations are densely deployed in small cell to satisfy the high capacity demands. Users with high mobility would suffer from frequent handoff due to the smaller cell size. Neighbor discovery plays an important role for mobile users with frequent handoffs to detect the base station to be connected. Highly directional transmission and reception make mobile user and base station difficult to detect each other due to deafness problem. Algorithms and protocols to achieve fast directional neighbor discovery is really necessary in frequent

bandoff to maintain mobile users' connectivity. The neighbor discovery algorithm would be a sub-optimal solution to achieve relatively high discovery probability and significantly reduced discovery time required by the frequent handoff.

7.2.2 Network Capacity Analysis

Next generation wireless networks support applications with quite different transmission rates. Thus, hybrid medium access control of CSMA/CA+TDMA would be promising for next generation wireless networks with CSMA/CA for low rate applications and TDMA for high rate applications requiring guaranteed performance. The CSMA/CA period and TDMA period are highly correlated and determine the system capacity together, e.g., larger CSMA/CA period can collect more transmission requests which would be scheduled in TDMA period to improve network capacity while larger CSMA/CA period would reduce network capacity since it cannot transmit as much data as TDMA period. Optimization techniques can be used to balance these periods to achieve optimal network capacity constrained by the directional wireless backhaul, while considering the emerging communication modes, such as directional multiple D2Ds.

7.2.3 Power-efficient Sleep Scheduling for mmWave Base Stations

mmWave networks mainly target for burst type applications with high data rate such as video streaming. Thus, it is possible that during a specific period, some mmWave base stations are idle if there are no users within their cells requiring high rate mmWave connections. Since mmWave base stations are densely deployed, there could be lots of opportunities to let idle mmWave base stations sleep to greatly save the energy. Sleep

scheduling for mmWave base stations needs to consider the whole system since mobile users go through different mmWave base stations for mmWave connection. Sleep scheduling for single mmWave base station would introduce large cost of frequently turning on/off and also results in intermittent connection for the mobile users because mmWave users would come from the neighboring mmWave base stations. Global optimization can be used to generate sleep scheduling for all the mmWave base stations in the system to achieve minimum energy consumption while satisfying connectivity requirement.

Abbreviations

ITU	International Telecommunication Union
IMT	International Mobile Telecommunication
FCC	Federal Communications Commission
ETSI	European Telecommunications Standards Institute
GSM	Global System for Mobile Communications
3GPP	the Third Generation Partnership Project
LTE-A	Long Term Evolution - Advanced
UMTS	Universal Mobile Telecommunications System
WIMAX	Worldwide Interoperability for Microwave Access
QoS	Quality of Service
HSPA	High Speed Packet Access
HSDPA	High Speed Downlink Packet Access
HSUPA	High Speed Uplink Packet Access
CSP	Communication Service Provider
AP	Access Point
BS	Base Station
VNI	Visual Network Index

CMOS	Complementary Metaloxidesemiconductor
RFIC	Radio Frequency Integrated Circuits
SINR	Signal to Interference plus Noise Ratio
EIRP	Equivalent Isotropically Radiated Power
MAC	Medium Access Control
PHY	Physical Layer
CDMA	Code Division Multiple Access
TDMA	Time Division Multiple Access
STDMA	Spatial-Time Division Multiple Access
OFDM	Orthogonal Frequency-Division multiplexing
MIMO	Multi-Input-Multi-Output
CoMP	Coordinated Multi-Point Transmission
HetNet	Heterogeneous Networks
D2D	Device to Device
UWB	Ultra-Wideband
WBAN	Wireless Body Area Network
WPAN	Wireless Personal Area Network
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Network
LOS	Line-of-sight
NLOS	Non-line-of-sight
MUI	Multi-user Interference
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance

BP	Beacon Period
CAP	Contention Access Period
CTAP	Channel Time Allocation Period
AWGN	Additive White Gaussian Noise
MHCT	Multi-hop Concurrent Transmission
SHCT	Single-hop Concurrent Transmission
SHT	Single-hop Transmission
VBR	Variable Bit Rate
CBR	Constant Bit Rate
MDP	Markov Decision Process
AOA	Angle of Arrival
AOD	Angle of Departure
CSI	Channel State Information
RF	Radio Frequency
DAC	Digital-to-analog Converter
DFT	Discrete Fourier Transform
WDEV	Wireless Device
SIPS	Short Inter-phase Spacing

Symbols

R_{total}	Total data rate of an area
W_{sys}	System bandwidth
$PL(d)$	Pathloss at distance d
PL_{thr}	Pathloss threshold
P_R	Received signal power
G_T	Transmitter antenna gain
G_R	Receiver antenna gain
$d(A, B)$	Link length
N	Number of mobile users
B_m	mmWave base station
C_m	mmWave base station capacity
R_s	Instantaneous transmission rate
U_n	Mobile user n
q	User moving probability
G	Buffer size
$G(i, a)$	Stage cost
T_{BP}	Time duration for beacon period

T_{CAP}	Time duration for contention access period
L	Number of concurrent links in each time slot
B_T	Transmission beam set
B_R	Reception beam set
N_t	Number of transmission beams
N_r	Number of reception beams
N_0	Background noise
d_{ref}	Reference distance
N_{BS}	Number of base stations
N_0	One-side power spectral density
G_{max}	Maximum antenna gain over angle
$w(A, B)$	Link weight
E_F	Average node traffic load
S_{R1}	Beam overlap area
x_a	Minimum mumber of time slots in each group
x_b	Maximum mumber of time slots in each group
$f(d)$	Pdf of link length
K	Total number of time slots in each superframe

Publication List

1. J. Qiao, X. Shen, J. W. Mark, and L. Lei, "Video Quality Provisioning for Millimeter Wave 5G Cellular Networks with Link Outage", under review in IEEE Transactions on Wireless Communications.

2. J. Qiao, X. Shen, J. W. Mark, Q. Shen, Y. He, and L. Lei, "Enabling Device-to-Device Communication in Millimeter Wave 5G Cellular Networks", IEEE Communications Magazine, vol. 53, no. 1, pp. 209-215, Jan. 2015.

3. J. Qiao, X. Shen, J. W. Mark, and Y. He, "MAC-layer Concurrent Beamforming Protocol for Indoor Millimeter Wave Networks", IEEE Transactions on Vehicular Technology, vol. 64, no. 1, pp. 327-338, Jan. 2015.

4. J. Qiao, X. Shen, J. W. Mark, B. Cao, Z. Shi, and K. Zhang, "CSMA/CA-based Medium Access Control for Indoor Millimeter Wave Networks", to appear in Wiley Wireless Communications and Mobile Computing, early version available online.

5. J. Qiao, L. X. 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.

6. Z. Shi, R. Sun, R. Lu, J. Qiao, J. Chen, and X. Shen, "A Wormhole Attack Resistant Neighbor Discovery Scheme with RDMA Protocol for 60 GHz Directional Network", IEEE Transactions on Emerging Topics in Computing, vol. 1, no. 2, pp. 341-352, Dec. 2013.

7. L. Bin, L. Chuang, J. Qiao, J. He, and U. Peter, "A NetFlow Based Flow Analysis and Monitoring System in Enterprise Networks", Elsevier Computer Networks, vol. 52, issue 5, pp. 1074-1092, April 2008.

8. Y. He, Z. Pan, X. Cheng, J. Qiao, and M. Tentzeris, "A Dual-Band, Dual-Polarized, Miniaturized and Beautified Base Station Antenna for FDD Systems", 2nd-round review in IEEE Transactions on Antennas and Propagation.

9. K. Zheng, H. Li, W. Xiang, J. Qiao, and X. Shen, "Design and Implementation of an Energy-Efficient Localization and Tracking System using ZigBee", submitted to IEEE Transactions on Industrial Informatics.

10. L. Lei, Y. Kuang, J. Qiao, X. Shen, and Z. Zhong, "Efficient Energy Management for Reliable Real-Time Energy Harvesting in Industrial Wireless Sensor Networks", submitted to IEEE Transactions on Industrial Informatics.

11. J. Qiao, X. Shen, J. W. Mark, Z. Shi, and N. Mohammadizadeh, "MAC-Layer Integration of Multiple Radio Bands in Indoor Millimeter Wave Networks", in Proc. IEEE WCNC, pp. 889-894, April 2013.

12. J. Qiao, B. Cao, X. Zhang, X. Shen, and J. W. Mark, "Efficient Concurrent Transmission Scheduling for Cooperative Millimeter Wave Systems", in Proc. IEEE GLOBECOM, pp. 4187-4192, Dec. 2012.

13. J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "STDMA-based Scheduling Algorithm for Concurrent Transmissions in Directional Millimeter Wave Networks", in Proc. IEEE ICC, pp. 5221-5225, June, 2012.

14. J. Qiao, L. X. Cai, and X. Shen, "Multi-Hop Concurrent Transmission in Millimeter Wave WPANs with Directional Antenna", in Proc. IEEE ICC, pp. 1-5, May 2010.

15. Z. Shi, R. Lu, X. Liang, J. Qiao, and X. Shen, "SND: Secure Neighbor Discovery for 60 GHz Network with Directional Antenna", in Proc. IEEE WCNC, pp. 4712-4717, April

2013.

16. K. Zhang, R. Lu, J. Qiao, and X. Shen, "PARK: A Privacy-preserving Aggregation Scheme with Adaptive Key Management for Smart Grid", in Proc. IEEE ICCC, pp. 236-241, August 2013.

Bibliography

[1] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–449, May 2013. 1, 2, 7, 16

[2] Z. Pi and F. Khan, "An Introduction to Millimeter-wave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, June 2011. 1, 12, 19, 35

[3] 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 J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014. 1, 7, 35

[4] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "Enabling Multi-Hop Concurrent Transmissions in 60 GHz Wireless Personal Area Networks," *IEEE Trans. on Wireless Commun.*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011. 2, 51, 126, 127

[5] S. Singh, R. Mudumbai, and U. Madhow, "Distributed Coordination with Deaf Neighbors: Efficient Medium Access for 60 GHz Mesh Networks," in *Proc. IEEE INFOCOM*, May 2010, pp. 1–5. 2, 13, 32

[6] S. Hwang, D. Lyu, and K. Chang, "4G Vision and Technology Development in Korea," in *Proc. IEEE Int. Conf. Commun. Technol.*, April 2003, pp. 26–31. 4

[7] S. Rajagopal, S. Abu-Surra, Z. Pi, and F. Khan, "Antenna Array Design for Multi-Gbps mmwave Mobile Broadband Communication," in *Proc. IEEE GLOBECOM*, Dec. 2011, pp. 1–6. 4

[8] Cisco, "Visual Networking Index," http://www.cisco.com, Feb. 2014. 5

[9] "FP7 European Project 317669 METIS (Mobile and Wireless Communications Enablers for the Twenty-Twenty Information Society)," https://www.metis2020.com/, 2012. 7

[10] H. C. Minh, "5G Vision and Requirements of 5G Forum, Korea," in *ITU-R 2020 Vision Workshop 5G Forum Korea*, Feb. 2014, pp. 1–6. 7

[11] F. Gutierrez, S. Agarwal, K. Parrish, and T. S. Rappaport, "On-Chip Integrated Antenna Structures in CMOS for 60 GHz WPAN Systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1367–1378, Oct. 2009. 9, 13

[12] S. Emami, C. Doan, A. Niknejad, and R. Brodersen, "A Highly Integrated 60GHz CMOS Front-End Receiver," in *Proc. IEEE International Solid-State Circuits Conference*, Feb. 2007, pp. 190–192. 9

[13] T. Marzetta, "Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Sept. 2010. 10

[14] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, Future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012. 11, 12

[15] R. Heath, S. Peters, Y.Wang, and J. Zhang, "A Current Perspective on Distributed Antenna Systems for the Downlink of Cellular Systems," *IEEE J. Sel. Areas Commun.*, vol. 51, no. 4, pp. 161–167, Apr. 2013. 11

[16] P. Smulders, "Exploiting the 60 GHz Band for Local Wireless Multimedia Access: Prospects and Future Directions," *IEEE Commun. Mag.*, vol. 40, no. 1, pp. 140–147, Jan. 2002. 13

[17] R. Daniels and R. Heath, "60 GHz Wireless Communications: Emerging Requirements and Design Recommendations," *IEEE Veh. Technol. Mag.*, vol. 2, no. 3, pp. 41–50, Sept. 2007. 13

[18] C. Sum, Z. Lan, R. Funada, J. Wang, T. Baykas, M. Rahman, and H. Harada, "Virtual Time-Slot Allocation Scheme for Throughput Enhancement in a Millimeter-Wave Multi-Gbps WPAN System," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1379–1389, Oct. 2009. 13, 51, 83, 93, 127, 128

[19] J. Foerster, J. Lansford, J. Laskar, T. Rappaport, and S. Kato, "Realizing Gbps Wireless Personal Area Networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1313–1317, Oct. 2009. 13, 39, 126

[20] S. Y. Geng, J. Kivinen, X. W. Zhao, and P. Vainikainen, "Millimeter-Wave Propagation Channel Characterization for Short-Range Wireless Communications," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 3–13, Jan. 2009. 13

[21] L. X. Cai, L. Cai, X. Shen, and J. W. Mark, "REX: a Randomized EXclusive Region based Scheduling Scheme for mmWave WPANs with Directional Antenna," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 113–121, Jan. 2010. 13, 39, 80, 83, 126, 127, 128

[22] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "Enabling Multi-Hop Concurrent Transmissions in 60 GHz Wireless Personal Area Networks," *IEEE Trans. on Wireless Commun.*, vol. 10, no. 11, pp. 3824–3833, Nov. 2011. 13, 49

[23] "ECMA International TC48 High Rate Short Range Wireless Communications," http://www.ecma-international.org/memento/TC48-M.htm, 2010. 14

[24] "IEEE 802.15 WPAN Millimeter Wave Alternative PHY Task Group 3c(TG3c)," http://www.ieee802.org/15/pub/TG3c.html, 2010. 14, 126

[25] "IEEE 802.11 VHT Study Group," http://www.ieee802.org/11/Reports/vht_update.htm, 2010. 14, 126

[26] Q. Zhao and J. Li, "Rain Attenuation in Millimeter Wave Ranges," in *IEEE Int. Symp. Antennas, Propag. EM Theory*, Oct. 2006, pp. 1–4. 17, 18

[27] T. S. Rappaport, J. N. Murdock, and F. Gutierrez, "State of the Art in 60 GHz Integrated Circuits and Systems for Wireless Communications," *Proc. IEEE*, vol. 99, no. 8, pp. 1390–1436, Aug. 2011. 17

[28] S. Geng, J. Kivinen, X. Zhao, and P. Vainikainen, "Millimeter-Wave Propagation Channel Characterization for Short-Range Wireless Communications," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 3–13, Jan. 2009. 18, 35, 100, 130

[29] A. I. Sulyman, A. T. Nassar, M. K. Samimi, G. R. MacCartney, T. S. Rappaport, and A. Alsanie, "Radio Propagation Path Loss Models for 5G Cellular Networks in the 28 GHZ and 38 GHZ Millimeter-wave Bands," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 23–30, Sept. 2014. 19, 35

[30] A. Maltsev, "Channel Models for 60 GHz WLAN Systems," IEEE 802.11-09/0334r8, May 2010. 32

[31] R. Mudumbai, S. Singh, and U. Madhow, "Medium Access Control for 60 GHz Outdoor Mesh Networks with Highly Directional Links," in *Proc. IEEE INFOCOM*, April, pp. 2871–2875. 34

[32] "Disk Line Picking," http://mathworld.wolfram.com/DiskLinePicking.html, 2013. 46

[33] X. Pei, T. Jiang, D. Qu, G. Zhu, and J. Liu, "Radio Resource Management and Access Control Mechanism Based on a Novel Economic Model in Heterogeneous Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3047–3056, July 2010. 48

[34] W. Song, H. Jiang, and W. Zhuang, "Performance Analysis of the WLAN-First Scheme in Cellular/WLAN Interworking," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 5, pp. 1932–1943, May 2007. 48

[35] W. Song, Y. Chen, and W. Zhuang, "Improving Voice and Data Services in Cellular/WLAN Integrated Network by Admission Control," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 11, pp. 4015–4037, Nov. 2007. 48

[36] M. Ismail and W. Zhuang, "A Distributed Multi-Service Resource Allocation Algorithm in Heterogeneous Wireless Access Medium," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 425–432, Feb. 2012. 48, 49

[37] D. Cavalcanti, D. Agrawal, C. Cordeiro, B. Xie, and A. Kumar, "Issues in Integrating Cellular Networks, WLANs, and MANETs: A Futuristic Heterogeneous Wireless Network," *IEEE Wireless Commun.*, vol. 12, no. 3, pp. 30–41, Jun. 2005. 48, 49

[38] A. Alshamrani, X. Shen, and L. Xie, "QoS Provisioning for Heterogeneous Services in Cooperative Cognitive Radio Networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 819–830, April 2011. 48

[39] D. Niyato and E. Hossain, "Bandwidth Allocation in 4G Heterogeneous Wireless Access Networks: A Noncooperative Game Theoretical Approach," in *IEEE GLOBECOM*, Nov. 2006, pp. 1–5. 49

[40] S. Singh, F. Ziliotto, U. Madhow, E. M. Belding, and M. J. W. Rodwell, "Millimeter Wave WPAN: Cross-Layer Modeling and Multihop Architecture," in *Proc. IEEE INFOCOM*, May 2007, pp. 2336–2340. 49, 82, 127

[41] J. Qiao, L. X. Cai, X. Shen, and J. W. Mark, "STDMA-based Scheduling Algorithm for Concurrent Transmissions in Directional Millimeter Wave Networks," in *Proc. IEEE ICC*, June 2012, pp. 1–5. 51, 125, 126, 127, 128

[42] A. Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queueing Algorithm," in *Applications, Technologies, Architectures, and Protocols for Computer Communication*, 1989, pp. 1–12. 62, 65

[43] J. Qiao, L. X. Cai, and X. Shen, "Multi-Hop Concurrent Transmission in Millimeter Wave WPANs with Directional Antenna," in *Proc. IEEE ICC*, May 2010, pp. 23–27. 83, 90, 91, 127

[44] Z. Yang, L. Cai, and W. Lu, "Practical Concurrent Transmission Scheduling Algorithms for Rate-adaptive Wireless Networks," in *Proc. IEEE INFOCOM*, March 2010, pp. 230–238. 83, 87, 127

[45] P. Dutta, V. Mhatre, D. Panigrahi, and R. Rastogi, "Joint Routing and Scheduling in Multi-hop Wireless Networks with Directional Antennas," in *Proc. IEEE INFOCOM*, March 2010, pp. 389–397. 83

[46] D. Pisinger, "Where are the hard knapsack problems?" in *Computers and Operations Research*, March 2005, pp. 2271–2284. 84, 100

[47] J. Wang, Z. Lan, C. W. Pyo, T. Baykas, C. S. Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, and S. Kato, "Beam Codebook Based Beamforming Protocol for Multi-Gbps Millimeter-Wave WPAN Systems," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1390–1399, Oct. 2009. 93, 95, 96, 97, 98, 102

[48] H. H. Lee and Y. C. Ko, "Low Complexity Codebook-Based Beamforming for MIMO-OFDM Systems in Millimeter-Wave WPAN," *IEEE Trans. on Wireless Commun.*, vol. 10, no. 11, pp. 3607–3612, Nov. 2011. 93, 95, 97, 102

[49] R. Nelson and L. Kleinrock, "Spatial-TDMA: A Collision-free Multihop Channel Access Protocol," *IEEE Trans. Commun.*, vol. 33, no. 9, pp. 934–944, Sept. 1985. 93

[50] M. S. Choi, G. Grosskopf, D. Rohde, B. Kuhlow, G. Przyrembel, and H. Ehlers, "Experiments on DOA-Estimation and Beamforming for 60 GHz Smart Antennas," in *Proc. IEEE VTC-Spring*, April 2003, pp. 1041–1045. 95

[51] X. An, C. Sum, R. V. Prasad, J. Wang, Z. Lan, J. Wang, H. H. R. Hekmat, and I. Niemegeers, "Beam Switching Support to Resolve Link-Blockage Problem in 60 GHz WPANs," in *Proc. IEEE PIMRC*, Sept. 2009, pp. 390–394. 95

[52] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Multilevel Millimeter Wave Beamforming for Wireless Backhaul," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2012, pp. 253–257. 95, 97

[53] L. Zhou and Y. Ohashi, "Efficient Codebook-Based MIMO Beamforming for Millimeter-Wave WLANs," in *Proc. IEEE PIMRC*, Sept. 2012, pp. 1885–1889. 95

[54] G. Grosskopf, R. Eggemann, and H. Ehlers, "Maximum Directivity Beam-former at 60 GHz with Optical Feeder," *IEEE Trans. on Antennas and Propagation*, vol. 51, no. 11, pp. 3040–3046, Nov. 2003. 95

[55] N. Celik, M. F. Iskander, R. Emrick, S. J. Franson, and J. Holmes, "Implementation and Experimental Verification of a Smart Antenna System Operating at 60 GHz Band," *IEEE Trans. on Antennas and Propagation*, vol. 56, no. 9, pp. 2790–2800, Sept. 2008. 95

[56] L. Wan, X. Zhong, Y. Zheng, and S. Mei, "Adaptive Codebook for Limited Feedback MIMO System," in *Proc. WOCN*, April 2009, pp. 1–5. 98

[57] "IEEE 802.11n-2009Amendment 5: Enhancements for Higher Throughput," , 2009. 106

[58] F. Winkler, E. Fischer, E. Grass, and P. Langendorfer, "An Indoor Localization System Based on DTDOA for Different Wireless LAN Systems," in *Proc. WPNC*, June 2006, pp. 117–122. 107, 134

[59] A. Maltsev, "Channel Models for 60 GHz WLAN Systems IEEE 802.11-09-0344-07ad," , March 2010. 117, 119

[60] F. Winkler, E. Fischer, E. Grass, and P. Langendorfer, "On the Capacity Improvement of Ad Hoc Wireless Networks Using Directional Antennas," in *Proc. ACM Mobihoc'03*, June 2003, pp. 117–125. 125

[61] C. Peraki and S. Servetto, "On the Maximum Stable Throughput Problem in Random Networks with Directional Antennas," in *Proc. ACM MobiHoc*, June 2003. 125

[62] Q. Zhao and L. Tong, "A Multiqueue Service Room MAC Protocol for Wireless Networks with Multipacket Reception," *IEEE/ACM Trans. Networking*, vol. 11, no. 1, pp. 125–137, Feb. 2003. 126

[63] G. D. Celik, Z. G, W. F. Khan, and E. Modiano, "MAC for Networks with Multipacket Reception Capability and Spatially Distributed Nodes," *IEEE/ACM Trans. Mobile Computing*, vol. 9, no. 2, pp. 226–240, Feb. 2010. 126

[64] S. Singh, F. Ziliotto, U. Madhow, E. M. Belding, and M. Rodwell, "Blockage and Directivity in 60 GHz Wireless Personal Area Networks: From Cross-Layer Model to Multihop MAC Design," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1400–1413, Oct. 2009. 127

[65] C. W. Pyo and H. Harada, "Throughput Analysis and Improvement of Hybrid Multiple Access in IEEE 802.15.3c mm-Wave WPAN," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1414–1424, Oct. 2009. 127, 128, 133

[66] Q. Zhao and L. Tong, "A Multiqueue Service Room MAC Protocol for Wireless Networks with Multipacket Reception," *IEEE/ACM Trans. on Netw.*, vol. 11, no. 1, pp. 125–137, Feb. 2003. 129

[67] G. D. Celik, Z. Gil, W. F. Khan, and E. Modiano, "MAC for Networks with Multipacket Reception Capability and Spatially Distributed Nodes," *IEEE/ACM Trans. Mobile Computing*, vol. 9, no. 2, pp. 226–240, Feb. 2013. 129, 131

[68] P. Zheng, Y. Zhang, and S. Liew, "Multipacket Reception in Wireless Local Area Networks," in *Proc. IEEE ICC*, June 2006, pp. 112–116. 129

[69] G. Nguyen, A. Ephremides, and J. Wieselthier, "Comments on "Capture and Retransmission Control in Mobile Radio," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 12, pp. 2340–2351, Dec. 2006. 130, 131

[70] L. G. Roberts, "Aloha Packet System with and without Slots and Capture Broadcast Channels," *Computer Communicationr Review*, vol. 2, no. 5, pp. 28–42, April 1975. 131

[71] C. T. Lau and C. Leung, "Capture Models for Mobile Packet Radio Networks," *IEEE Trans. Commun.*, vol. 40, no. 5, pp. 917–925, May 1992. 131