ROYAL INSTITUTE
OF TECHNOLOGY

Throughput and Latency of Millimeter-Wave Networks: Performance Analyses and Design Principles

GUANG YANG

Doctoral Thesis
Stockholm, Sweden 2018

Information Science and Engineering
Electrical Engineering and Computer Science
KTH Royal Institute of Technology
SE-100 44 Stockholm
SWEDEN

TRITA-EECS-AVL-2018:28
ISBN 978-91-7729-731-4

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i electro- och systemteknik torsdag den 3 maj 2018 klockan 9.30 i Kollegiesalen, Brinellvägen 8, Stockholm.

Tryck: Universitetsservice US AB

The important thing in life is not the triumph, but the fight;

the essential thing is not to have won, but to have fought well.

— Pierre de Coubertin

Abstract

Nowadays, the ever-increasing demands on higher data rates and better service performance have posed extremely huge challenges to the existing wireless communications within sub-6 GHz bands, mainly due to the spectrum scarcity in lower frequency bands. In recent years, the millimeter-wave (mm-wave) technology, as a promising candidate to meet the aforementioned demands, have attracted extensive research attention, and has been regarded as one of the key enablers for the forthcoming the 5^{th} generation (5G) mobile communications. The main features of mm-wave communications include: abundant spectral resources, high penetration loss, severe path loss, weak multi-path effects, and narrow antenna beams, and these particular features make the potential challenges and solutions with mm-wave differ a lot from those in the conventional 6-GHz systems.

It is known that the high throughput and the low latency are two critical quality-of-service (QoS) aspects in future mobile networks, while the related research with mm-wave are fairly recent and insufficient in the past few years. Motived by the urgent needs for further development and the blanks remained in previous works, in this doctoral thesis, we investigate the throughput and the latency in mm-wave networks through conducting performance analyses and identifying design principles, with the objective of seeking clues for improving the QoS of mm-wave wireless communications in practice.

Our main research regarding throughput and latency in mm-wave networks that are included in this doctoral thesis can be categorized from the following three aspects:

(i) *Throughput of mm-wave relay networks*: For indoor scenarios, we study the half-duplex (HD) relaying with mm-wave in the presence of random link blockages, where a distance-based routing algorithm is proposed to maximize the throughput. For outdoor scenarios, focusing on a two-hop amplify-and-forward (AF) relay network in the HD or the full-duplex (FD) mode, we investigate the impacts of beamwidth, ground reflections, and self-interference coefficient on the throughput, where Gaussian-type directional antenna model and two-ray channel model are jointly adopted.

(ii) *Latency analysis via stochastic network calculus*: With the aid of stochastic network calculus, we derive upper bounds for the probabilistic delay to keep track of the latency performance of buffer-aided mm-wave networks.

We mainly study mm-wave systems designed in tandem or parallel manners, and also consider a hybrid design that combines the tandem and parallel schemes in a flexible manner. Moreover, the capability of achieving low-latency mm-wave communications is characterized and investigated in terms of effective capacity, and the comparison among different transmission schemes is conducted to identify the respective strengths and proper conditions for their applications.

(iii) *Traffic allocation for low-latency mm-wave systems*: Traffic allocation schemes for low latency in buffer-aided mm-wave networks are investigated. Due to the use of buffers, the delay optimization problem hereby differs from those without buffers, where the conventional graph-based network optimization techniques become intractable. We demonstrate the impacts of different traffic allocation schemes on the latency. For multi-hop networks with multiple parallel channels in each hop, we consider both local and global traffic allocation schemes, quantify their resulting end-to-end (E2E) latencies, and analyze the respective strengths and weaknesses.

Sammanfattning

Dagens konstant ökande efterfrågan på högre datahastigheter och bättre prestanda har skapat extremt stora utmaningar för existerande trådlös kommunikation i band under 6 GHz, framförallt på grund av brist på spektrum i de lägre banden. De senaste åren har millimetervågs (mm-våg) teknologier, som anses vara en lovande metod för att uppfylla ovan nämnda krav, fått stort forskningsgenomslag och ses därför som en nyckel i realisationen av kommande 5:e generationens (5G) mobila kommunikationssystem. Några av de utmärkande egenskaperna för mm-vågskommunikation är: kraftigt ökade spektrala resurser, höga överförings- samt penetrations-förluster, svaga effekter från flervägsutbredning och smala antennstrålar. Dessa egenskaper skapar unika utmaningar som skiljer sig markant från de som traditionella 6 GHz system ställts inför.

Det är välkänt att hög överföringshastighet och låg latens är två kritiska kvalitetsmarkörer i framtida mobila nätverk, men relaterad forskning inom mm-vågs kommunikation har varit otillräcklig under de senaste åren. I denna doktorsavhandling, motiverad av kraven på utveckling samt frågor som står obesvarade inom detta område, utforskar vi överföringshastighet och latens i mm-vågs-nätverk genom att utvärdera prestanda samt utveckla designprinciper. Motivet är att söka ledtrådarna för att i praktiken förbättra upplevd kvalitet i trådlös mm-vågs-kommunikation.

Den forskning relaterad till överföringshastighet samt latens i mm-vågs-nätverk som presenteras i denna avhandling kan delas in enligt följande tre kategorier:

(i) *Överföringshastighet i mm-vågs relänätverk:* I inomhusscenarier studerar vi halvduplex (HD) reläöverföring med mm-våg då nätverkslänkar blockeras slumpmässigt. En distans-baserad routing-algoritm som maximerar överföringshastigheten presenteras. I utomhusscenarier fokuserar vi på ett två-stegs förstärk-och-vidarebefodra relänätverk i HD eller full duplex (FD) läge. Vi utforskar effekterna av strålbredd, markreflektioner samt självinterferens på överföringshastigheten, där en direktiv antennmodell av Gaussisk typ samt en tvåvägskanalmodell appliceras.

(ii) *Latensanalys via stokastisk nätverksanalys:* Med hjälp av stokastisk nätverksanalys härleder vi övre gränsvärden för probabilistiska fördröjningar med syftet att utvärdera latens-prestandan i buffrade mm-vågs-nätverk. Vi studerar serie- samt parallellkopplade mm-vågs-system på ett flexibelt sätt, samt utvärderar även en hybrid design där de två kombineras. Utöver detta används

effektiv kapacitet för att studera förmågan att uppnå låg latens i mm-vågskommunikation, och en jämförelse mellan olika överföringstekniker utförs för att identifiera styrkor samt lämpliga förutsättningar för deras applicering.

(iii) *Trafikallokering för mm-vågs-system med låg latens:* Trafikallokering schemea för låg latens i buffrade mm-vågs-nätverk utvärderas. Minimering av fördröjningar skiljer sig som optimeringsproblem i nätverk med och utan buffrar, där traditionella graf-baserade nätverksoptimerings-tekniker inte längre fungerar i den förstnämnde. Vi demonstrerar implikationerna av traffikallokeriong på latensen. I multi-hop nätverk med flera parallella kanaler i varje steg undersöker vi både lokala samt globala allokerings-tekniker, kvantifierar deras resulterande totala fördröjning samt analyserar deras respektive styrkor och svagheter.

List of Papers

J: Journal Publication

C: Conference Publication

The thesis is based on the following papers:

[J1] **Guang Yang**, Jinfeng Du, and Ming Xiao, “Maximum Throughput Path Selection with Random Blockage for Indoor 60 GHz Relay Networks,” *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3511-3524, October, 2015. [YDX15]

[J2] **Guang Yang** and Ming Xiao, “Performance Analysis of Millimeter-Wave Relaying: Impacts of Beamwidth and Self-Interference”, *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 589-600, February, 2018. [YX18]

[J3] **Guang Yang**, Ming Xiao, Hussein Mohammed Al-Zubaidy, Yongming Huang, and James Gross, “Analysis of Millimeter-Wave Multi-Hop Networks with Full-Duplex Buffered Relays,” *IEEE/ACM Transactions on Networking*, vol. 26, no. 1, pp. 576-590, February, 2018. [YXAZ$^+$18]

[J4] **Guang Yang**, Ming Xiao, and H. Vincent Poor, “Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification?” to appear in *IEEE Transactions on Communications*, 2018. [YXP18]

[J5] **Guang Yang**, Ming Xiao, Muhammad Alam, and Yongming Huang, “Low-Latency Heterogeneous Networks with Millimeter-Wave Communications,” to appear in *IEEE Communications Magazine*, 2018. [YXAH18]

[J6] **Guang Yang**, Martin Haenggi, and Ming Xiao, “Traffic Allocation for Low-Latency Multi-Hop Networks with Buffers,” submitted to *IEEE Transactions on Communications*, 2018. (Major Revision) [YHX18]

In addition, the following papers have also been (co)-authored by the author of this thesis:

[J8] **Guang Yang** and Ming Xiao, “Blockage Robust Millimeter-Wave Networks,” Science China Information Science, vol. 60, no. 8, pp. 080307:1 - 3, 2017. [YX17a]

[J9] Zhenquan Zhang, **Guang Yang**, Zheng Ma, Ming Xiao, Zhiguo Ding, and Pingzhi Fan, “Coordination for Virtualized Heterogeneous Ultra-Dense Networks with NOMA,” to appear in *IEEE Vehicular Technology Magazine*, 2018.

[J10] Bing Li, Daniel Månsson, and **Guang Yang**, “An Efficient Method for Solving Frequency Responses of Power-Line Networks, ” *Progress in Electromagnetics Research B*, vol. 62, no. 1, pp. 202-317, 2015. [BMY15]

[C1] **Guang Yang** and Ming Xiao, “Interference Statistics of Regular Ring Structured Networks with 60 GHz Directional Antennas,” in *Proc. IEEE International Conference on Communications (ICC)*, Paris, France, May 2017. [YX17b]

[C2] **Guang Yang**, Ming Xiao, James Gross, Hussein Mohammed Al-Zubaidy, and Yongming Huang, “Delay and Backlog Analysis for 60 GHz Wireless Networks,” in *Proc. IEEE Global Communications Conference(Globecom)*, Washington D.C., USA, Dec. 2016. [YXG^{+}16]

[C3] **Guang Yang**, Ming Xiao, and Zhibo Pang, “Delay Analysis of Traffic Dispersion with Nakagami-m Fading in Millimeter-Wave Bands,” in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Spain, Apr. 2018.

[C4] Yu Ye, Zhenquan Zhang, **Guang Yang**, and Ming Xiao, “Minimum Cost Based Clustering Scheme for Cooperative Wireless Caching Network with Heterogeneous File Preference,” in *Proc. IEEE International Conference on Communications (ICC)*, Paris, France, May 2017. [YZYX17]

Acknowledgments

First and foremost, I would like to thank my main advisor Assoc. Prof. Ming Xiao. It is my great fortune to meet him at my early stage entering the field of wireless communications and to work with him till now. Ming is not only a nice supervisor, but also a close friend. I have harvested a lot from him in various aspects, from methodologies in academic research to philosophies in life. Without his guidance, I do not think I can improve and become the me of today. Besides, I want to thank Prof. Mikael Skoglund for being my co-advisor during my study. Mikael offered me the opportunity to join Department of Information Science and Engineering at KTH Royal Institute of Technology, and he is always kind and supportive.

I would like to express my sincere gratitude to Assoc. Prof. Mehdi Bennis from University of Oulu for acting as the opponent, and to the grading board members: Dr. Angela Sara Cacciapuoti from University of Naples Federico II, Dr. Shahid Mumtaz from Instituto de Telecomunicações Aveiro, and Assoc. Prof. Jiajia Chen from KTH Royal Institute of Technology. I am also grateful to Prof. Mats Bengtsson from KTH Royal Institute of Technology for being the advance reviewer of my doctoral thesis.

I would like to acknowledge Dr. Hussein Mohammed Al-Zubaidy and Assoc. Prof. James Gross. They provided an excellent series of lectures on network calculus and brought me to a different world full of fun. It is a pleasant experience to have discussions with them and to co-author one publication on IEEE/ACM Transactions on Networking. I would like to thank Dr. Jinfeng Du for giving me many beneficial research suggestions and thinking manners, which helped me a lot especially at the first few years of my Ph.D study.

I am deeply grateful to Prof. Martin Haenggi for accepting me as a two-month visiting student at University of Notre Dame, Indiana. Martin generously shared his rich expertise and invaluable techniques in research, which tremendously help me in the following studies. I am also thankful to Dr. Sanket Kalamkar and Ms. Debrah Gillean for arranging a plenty of logistic issues to make my life and work at Notre Dame much easier.

I would like to express my thanks to other senior faculties for their every effort to make our department fantastic: Prof. Lars Kildehøj, Prof. Mats Bengtsson, Prof. Magnus Jansson, Assoc. Tobias Oechtering, Assoc. Prof. Ragnar Thobaben, Assoc. Prof. Markus Flierl, Asst. Saikat Chatterjee. I also would like to thank all my past and present colleagues for all the good times we shared together in the workplace:

Dr. Ali Zaidi, Dr. Frédéric Gabry, Dr. Hamed Farhadi, Dr. Maksym Girnyk, Dr. Iqbal Hussain, Dr. Sheng Huang, Dr. Zhao Wang, Dr. Tai Do, Dr. Haopeng Li, Dr. Hadi Ghauch, Dr. Majid Gerami, Dr. Alla Tarighati, Dr. Nima Najari Moghadam, Dr. Germán Bassi, Dr. Rami Mochaourab, Dr. Antonios Pitarokoilis, Dr. Qiwen Wang, Dr. Zuxing Li, Dr. Jing Yue, Dr. Nan Qi, Dr. Du Liu, Dr. Lin Zhang, Dr. Nan Li, Le Phuong Cao, Peter Larsson, Sebastian Schiessl, Ahmed Zaki, Michail Mylonakis, Yu Ye, Zhengquan Zhang, Hanwei Wu, Dong Liu, Yang You, and Shaocheng Huang. Here, I greatly appreciate the help from Henrik Forssell for translating the abstract of this thesis into Swedish, and particularly thank Le Phuong Cao and Yu Ye for the thesis proofreading. I also have my special thanks to Ms. Raine Tiivel and Ms. Eleni Nylén for providing administrative supports and thanks to Niclas Horney and Pontus Friberg for IT supports.

I would like to thank all my friends met in Stockholm who make my life joyful and unforgettable: Dr. Jingqi Zhang, Dr. Yubao Song, Dr. Li Zhang, Yingchen Zhang, Dr. Xuge Fan, Dr. Lipeng Liu, Dr. Xin Xu, Hongyu Jin, Ye Tian, Dr. Jin Yang, Dr. Zhili Pang, Dr. Zhou Li, Dr. Minyu Zuo, Dr. Yang Yang, Rong Du, Muhammad Taha Ali, and Dr. Xinhai Zhang. I am also thankful to all the workout buddies and instructors in KTH-Hallen, since their enthusiasm and encouragement make me energetic, happy, and strong.

Finally, I want to express my deepest and sincerest gratitude to my family for their continuous and unparalleled love and support. My special acknowledgment is reserved to my wife, Bing, for every splendid and memorable moment we shared around the world. Her encouragement and passion light me up when I was feeling frustrated and depressed. I would like to especially thank my parents for their love, care and encouragement. This doctoral thesis is dedicated to my family with love, and I wish you all the best in everything!

Guang Yang

Stockholm, April 2018

Contents

List of Acronyms

3G	the 3$^{\text{rd}}$ generation
4G	the 4$^{\text{th}}$ generation
5G	the 5$^{\text{th}}$ generation
AF	amplify-and-forward
c.d.f.	cumulative distribution function
DAS	distributed antenna system
DF	decode-and-forward
E2E	end-to-end
eMBB	enhanced mobile broadband
FIFO	first-in first-out
FD	full-duplex
Gbps	gigabits per second
HD	half-duplex
HetNet	heterogeneous network
i.i.d.	independent and identically distributed
LoS	line-of-sight
mm-wave	millimeter-wave
mMTC	massive machine-type communciations
MeNB	macro-cell evolved NodeB
MTPS	maximum throughput path selection
MHCT	multi-hop concurrent transmission

MGF	moment generating function
MAC	medium access control
MIMO	multiple-input multiple-output
NLoS	non-line-of-sight
p.d.f.	probability density function
QoS	quality-of-service
RRH	remote radio head
RPR	relay-prioritized region
SeNB	small-cell evolved NodeB
SNR	signal-to-noise ratio
SINR	signal-to-interference-plus-noise ratio
UE	user equipment
URLLC	ultra-reliable and low-latency communication
WPAN	wireless personal area networks
WLAN	wireless local area networks
WiGig	wireless gigabit alliance

Mathematical Notations

$\mathbb{C}$	set of complex numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of positive real numbers
$\mathbb{R}_+^d$	set of d-dimensional positive real vectors
$\mathbb{N}$	set of non-negative integers
$\mathcal{O}\left(\cdot\right)$	big-$\mathcal{O}$ notation
$\otimes$	$(\min,+)$-algebra convolution
$\oslash$	$(\min,+)$-algebra de-convolution
$\mathbb{M}_X\left(\theta\right)$	moment generating function of random variable X (w.r.t. $\theta > 0$)
$\overline{\mathbb{M}}_X\left(\theta\right)$	moment generating function of random variable X (w.r.t. $\theta < 0$)
$\mathbb{E}\left[X\right]$	expectation of random variable X
$\lvert x\rvert$	absolute value or magnitude, for $\forall x \in \mathbb{C}$
$\lVert\mathbf{x}\rVert_1$	ℓ_1-norm of vector $\mathbf{x}$
$\left[x\right]^+$	maximum between x and 0, for $\forall x \in \mathbb{R}$
$[n]$	set of integers from 1 to n, for $\forall n \in \mathbb{N}$
$\overset{\ell}{=}$	equality in law, i.e., distribution
$\triangleq$	definition

Part I

Thesis Overview

1

Introduction

1.1 Backgrounds

Due to the rapid development of electronic technologies and computer science, the conventional wireless communications within lower frequency bands, i.e., sub 6 GHz, become insufficient to support the ever-increasing demands on a higher throughput and better network performance for various emerging applications. Thus, the capability of the existing the 3$^{\mathrm{rd}}$ generation (3G) or the 4$^{\mathrm{th}}$ generation (4G) mobile networks for meeting the requirements above are seriously challenged. To enable a broad classes of potential applications and scenarios, i.e., enhanced mobile broadband (eMBB), massive machine-type communciations (mMTC), and ultra-reliable and low-latency communication (URLLC) [3GP17], the unprecedentedly high throughput and low latency are the two key characteristics to be fulfilled in future mobile networks, i.e., 5G and beyond.

As one of the key enablers of 5G communications, mm-wave technology exhibits appealing potential to remarkably improve the network performance from many aspects, e.g., data rate, latency, and energy consumption [ABC$^+$14]. These attractive benefits have attracted massive attention for standardization, e.g., IEEE 802.15.3c, IEEE 802.11ad, WirelessHD specification, and wireless gigabit alliance (WiGig). The mm-wave radios in 5G mobile networks usually refer to the electromagnetic waves with frequency roughly among 24 GHz to 300 GHz [XMH$^+$17]. In contrast to radios in lower frequency bands, i.e., ≤ 6 GHz, mm-wave signals encounter much severer atmospheric attenuations (see Fig. 1.1). Therefore, mm-wave technologies are recommended for short-range communication scenarios, e.g., wireless personal area networks (WPAN) and wireless local area networks (WLAN).

Compared to the existing radio technologies in sub-6 GHz bands, there are two major advantages of adopting mm-wave in future mobile networks, i.e., abundant spectral resources and short wavelength. Specifically:

- The available bandwidth in total for the current mobile networks is roughly 780 MHz [RSM$^+$13], while the potentially available bandwidth of mm-wave in total is at least 150 GHz (even excluding those higher atmospheric absorption

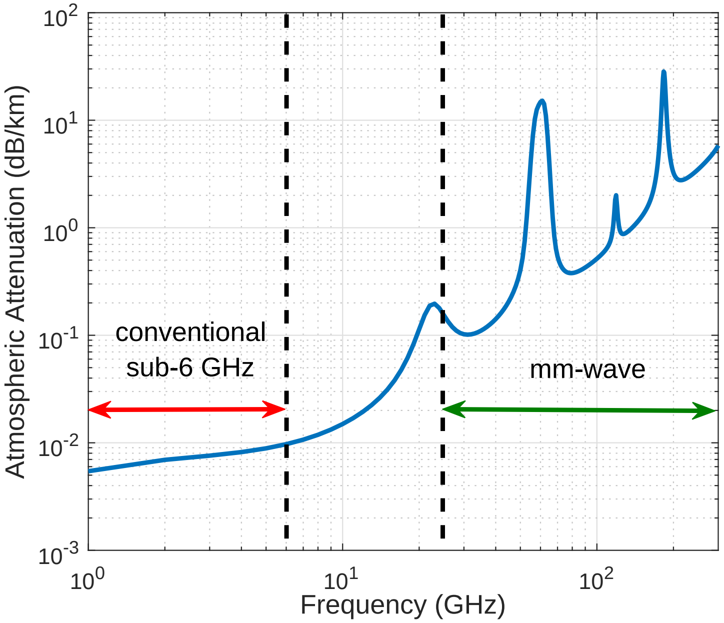


Figure 1.1: Attenuation by atmospheric absorption in differ frequency bands.

bands shown in Fig. 1.1). The abundance of spectral resources in mm-wave allows ultra-high data rates up to gigabits per second (Gbps).

- We know that mm-wave radios have much shorter wavelengths than those in sub-6 GHz bands. Thus, tens-to-hundreds of antenna elements can be integrated onto small-size chips in an ultra-dense fashion, thereby producing highly directional antennas. The high directivity not only indicates the high antenna gain for throughput improvement, but also enhances the spatial reuse and mitigates the interference in the presence of concurrent transmissions.

In spite of the huge benefits mentioned above, mm-wave communications however also have several critical technical challenges that need to be addressed as below:

- From the Friis transmission formula, it is known that the free-space path loss increases with the square of the carrier frequency, indicating that mm-wave radios suffer much higher path loss than sub-6 GHz radios. Paired with the higher signal attenuation induced by atmospheric absorption, the propagation attenuation in mm-wave bands is rather severe.

- Due to the short wavelength, the multi-path effects in wireless communications, e.g., diffraction and higher-order reflection, are rather weak in mm-wave, and the penetration loss of mm-wave signals is significantly higher than the sub-6 GHz counterpart. Hence, it is more desirable to fulfill mm-wave communications in line-of-sight (LoS) scenarios, since the link connectivity cannot be guaranteed in the presence of non-line-of-sight (NLoS). This particular propagation characteristic makes mm-wave links rather vulnerable to blockage caused by objects emerging between the transmitter and the receiver.

Fortunately, the clues to solve the difficulties above can be found from the advanced mm-wave technologies or the existing schemes in conventional mobile networks. For instance, highly directional mm-wave antennas or beamforming techniques in massive multiple-input multiple-output (MIMO) systems are usually employed to combat the propagation attenuation. Besides, for link blockage, proper relaying or cooperative strategies along with networking designs are capable of preserving the connectivity and ensuring the network coverage.

As aforementioned, the high throughput and the low latency, as two of the critical features of 5G mobile networks, have raised extensive research interest, and numerous efforts have been dedicated to these two topics with the past few years. Due to the fact that studies towards those aspects are fairly recent, our research is motivated by the blanks in existing works and hence conducted to explore the further potential of improving the performance for networks with mm-wave. In what follows, we will briefly present our research regarding the throughput and latency of mm-wave networks, where performance analyses are conducted and potential design principles are identified.

1.2 Our Research in Brief

In this section, we will discuss our contributions in the study regarding the performance of mm-wave communications from the perspective of throughput and latency. Specifically, in the first part of this section, we focus on the problem of throughput in mm-wave relay networks, where the impacts of a few important factors, i.e., random link blockages, beamwidth of directional antennas, and ground reflections, will be investigated. We have studied this problem in [YDX15, YX18], which are labeled as Paper J1 and Paper J2 in this thesis, respectively. In the second part, we focus on the problem of latency in buffer-aided mm-wave networks, where stochastic network calculus is employed for analysis. We have studied this problem in [YXAZ+18, YXP18], which are labeled as Paper J3 and Paper J4 in this thesis, respectively. Finally, we focus on the problem of traffic allocation to achieve low latency in mm-wave networks. We have studied this problem in [YXAH18, YHX18], which are labeled as Paper J5 and Paper J6 in this thesis, respectively.

For each paper, we will briefly discuss the previous related works, present the system model and methodology used in the paper, demonstrate the central results, and summarize the concluding remarks.

Contributions by the author

The contributions of the author included in papers come from the author's own research, in collaboration with co-authors as listed. The author of this thesis is the main contributor in all included papers, especially regarding theoretical analysis, computer simulations, and paper writing.

1.2.1 Throughput of mm-wave relay networks

J1: Maximum Throughput Path Selection with Random Blockage for Indoor 60 GHz Relay Networks [YDX15]

In this paper, we have studied the throughput maximization problem of 60 GHz indoor communications in the presence of random link blockages. Given sufficiently precise network topology information, we proposed a generic algorithm, namely maximum throughput path selection (MTPS), to select the optimal path that maximizes the throughput for networks with a central coordinator and multiple relays.

Backgrounds In the past decade, many efforts have been devoted to the study of link blockage in 60 GHz indoor scenarios [MEP+10, GT12, JPK+13]. To combat link blockages, commonly, there are two methods to maintain the link connectivity, i.e., via relay(s) [LLNS04, GOON10] or reflection(s) [GKZV09, GRON10]. Then the problem can be interpreted as the optimal routing in 60 GHz networks, which has been widely investigated. For instance, a geographic routing protocol for multi-hop LoS transmissions was proposed in [CHS+09], and a link scheduling scheme for multi-channel wireless mesh networks was studied in [SZ09]. To find the optimal relay path that maximizes the throughput, a routing algorithm regarding relay with deflection was proposed in [LSW+09].

In spite of numerous achievements in the field of 60 GHz networking, there still remain several open problems that are outlined from the following aspects:

- In [CHS+09, SZ09, LSW+09, CCSM10, QCSM11], link blockage is not incorporated in the proposed routing schemes.

- Although 60 GHz relaying are studied in [LLNS04, GOON10, LPCF12], the network is restricted to two-hop networks, and it is not clear how to optimize the relaying path.

- For preserving the link connectivity, measurements have shown that reflection signals are viable [DLZ12, SZM+09, GRON10, ASP+09, SLC11, Fan08], while related closed-form analysis is missing.

- The multi-hop concurrent transmission (MHCT) by [CCSM10, QCSM11] is limited to linear networks, and its performance in networks with arbitrarily placed nodes or link blockage is not clear.

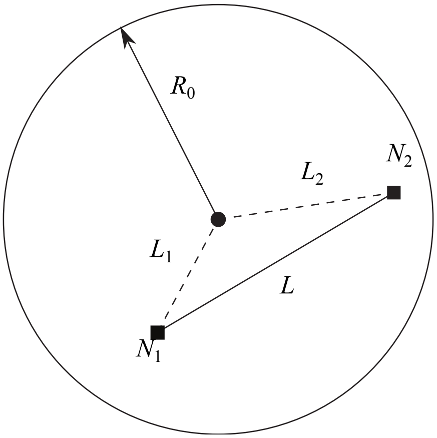


Figure 1.2: Randomly distributed nodes in a circular space, where the relay node is placed at the center and two communicating user nodes N_1 and N_2 are randomly located in the hall with radius R_0.

Motivation & Contributions The motivation of our research stems from the unsolved problems summarized above, and the corresponding outcomes build up our contributions. Specifically, in our work, we take into account the link blockage and develop a generic routing scheme, where relay selection and reflection utilization are both incorporated to maximize the throughput. The benefit of the proposed MTPS mainly lies in its low computational complexity, i.e., $\mathcal{O}\left(n^2\right)$ with respect to n relays, in contrast to $\mathcal{O}\left(n \cdot n!\right)$ via a brute-force manner. In addition, an analytical model for reflected signals is provided, and the placement of communicating nodes can be arbitrary. These contributions address the limitations in previous research works.

Model, Methodology & Analysis As an example, we consider a 60 GHz networks consisting of a pair of communication nodes N_1 and N_2 (randomly placed within a circular space) and one HD relay (located at the center C), as shown in Fig. 1.2. Supposing N_1 and N_2 are uniformly distributed in the circular area, and their distances to the center C are denoted by L_1 and L_2, respectively. Throughout our study, decode-and-forward (DF) relaying is assumed for analyzes. The LoS link between any two nodes encounter blockage caused by an obstacle (e.g., human body) in probability. We assume that the reflection path between any two nodes is always available [MEP$^+$10]. In light of above settings, the average throughput is investigated via considering the following for cases:

- Case I (LoS): only LoS link is available,
- Case II (LoS, Relay): both LoS and relay paths are available,
- Case III (LoS, Reflection): both LoS and reflection paths are available,

- Case IV (LoS, Relay, Reflection): LoS, relay, and reflection are available.

Regarding the random link blockage model, we assume the probability that an obstacle emerges outsides all potential links is $1-p$ for $p \in (0, 1)$. In other words, the blockage probability p denotes the probability that any one of all potential links encounters obstruction. In this work, we consider two types of random blockage models, namely *topology-independent model* and *topology-dependent model.* More specifically, given M links, labeled as $k \in [M]$, if a link blockage event has occurred, then the probability that link k is blocked, denoted by p_k, is given by $p_k = p\tau_k$, where the distribution coefficient τ_k differs for two distinct models, i.e.,

(a) $\tau_k = M^{-1}$ for topology-independent model,

(b) $\tau_k = l_k \cdot \sum_{i=1}^{M} l_i$ for topology-dependent model, with respect to the length l_i for link i.

Here, the topology-dependent model stems from the fact that longer links are more likely to be blocked.

In our study, a central coordinator is assumed for the indoor communications, to collect and update necessary information, manage the path selection process, and coordinate the scheduled communications. A practical example of such coordinator-aided systems is the high rate WPAN with a piconet coordinator or indoor network with an access point. With the assistance of central coordinator, the communication protocol for routing is outlined as follows:

(i) the coordinator monitors the network and updates a modified adjacency matrix that includes the location information of all nodes,

(ii) upon the request from a source-destination pair, the coordinator selects the optimized path that gives the maximum throughput,

(iii) the coordinator sends out the communication schedule to the source-destination pair and the selected relay nodes.

The first two steps are accomplished via performing the weighted distance calculation and the maximum throughput path selection, respectively (refer to Algorithm 2 and Algorithm 3 in Paper J1 for details).

We note that distance between nodes plays an important role in our algorithms. Thus, we introduce two notions, namely *relay-prioritized region* and *critical distance*, to decide if a relay should be incorporated in the routing.

Definition 1.2.1 (Relay-prioritized Region)**.** We define relay-prioritized region (RPR) as the area in which relaying can provide a higher transmission rate than the LoS path.

Specifically, with the HD relaying, let l, l_1, l_2 denote the length of the LoS path and the two relaying paths, respectively, the RPR exists if and only if

$$\frac{R\left(l_1\right) R\left(l_2\right)}{R\left(l_1\right)+R\left(l_2\right)}>R(l), \tag{1}$$

for some $l_1, l_2 > 0$ and $l_1 + l_2 \geq l$, where $R(\cdot)$ denotes of the rate with respect to the link length. It is worth noting that the RPR may not always exist.

Definition 1.2.2 (Critical Distance)**.** We define critical distance as the minimum distance between a pair of nodes for the existence of the RPR.

Proposition 1.2.1. *Given the signal-to-noise ratio (SNR) α and the path loss exponent n, the RPR exists if and only if*

$$d \geq d^* \triangleq \sqrt[n]{\frac{\alpha}{2^n-2}}, \tag{2}$$

where d is the communication distance and d^* is the critical distance.

We can see that using relays is more advantageous in the presence of higher path loss, since critical distance ld^* decays as n grows.

With the RPR and the critical distance, the optimal routing can be determined via performing Algorithm 2 and Algorithm 3 in Paper II. Applying the following Proposition 1.2.2, the throughput of HD relaying network with respect to the optimized path can be obtained.

Proposition 1.2.2. *For an L-hop channel, let R_i be the capacity of each individual channel i, $i = 1, 2, \ldots, L$, the maximum achievable rate R^* over the cascaded channel under the HD constraint is given by:*

$$R^*=\min \left\{\frac{R_1 R_2}{R_1+R_2}, \frac{R_2 R_3}{R_2+R_3}, \cdots, \frac{R_{L-1} R_L}{R_{L-1}+R_L}\right\}. \tag{3}$$

Performance Evaluation In performance evaluation, we consider a scenario where $N = 20$ obstacles are randomly distributed within a circular hall of radius $R_0 = 15$ meters and the HD relay node is deployed at the center of the hall. The default transmit power is $P_t = 0.1$mW with bandwidth $W = 1200$ MHz.

The average throughput of using relaying and/or reflection with the topology-independent random blockage model is shown in Fig. 1.3. In the presence of many obstacles ($N = 20$) with $p \in (0.1, 0.2)$, the average throughput is improved by roughly 10 times via introducing an HD relay node. Resorting to reflection paths is not significantly advantages for small p (< 0.05), while it outperforms that of relaying for $p > 0.35$. Without reflection, the average throughput for the LoS and (LoS, relay) scenarios both encounter drastic decrease, which however can be significantly mitigated by adding a relay node. When reflection is incorporated, the

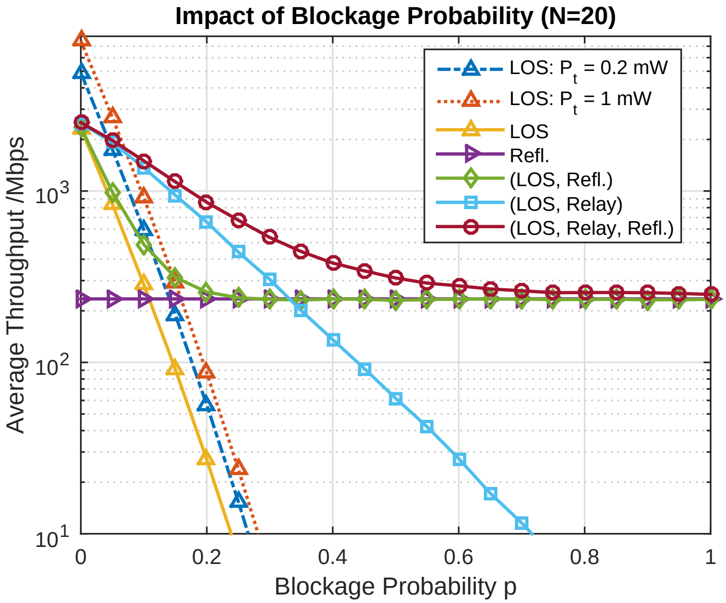


Figure 1.3: Average throughput [Mbps] versus link blockage probability p for topology-independent models.

average throughput of roughly 290 Mbps is still available even if the link blockage probability is very high. Besides, we notice that although increasing the transmit power can improve the average throughput, it is not an effective way to combat link blockage (see the resulting average throughput by adopting $P_t = 0.2$ mW and $P_t = 1$ mW).

In Fig. 1.4 we show the performance difference via applying the two random link blockage models (topology-independent and topology-dependent). For the (LoS, Relay) scenario, the difference can be remarkable when p is high. However, the gap is much smaller when reflection is taken into account. This is because the reflection paths are assumed to be always available and therefore not subject to link blockage.

Concluding Remarks We have studied the throughput of indoor 60 GHz communications over relaying networks, where the communication between two nodes can be established either by the LoS path, via an HD relay, or via the reflection path. To maximize the throughput in the presence of random link blockages, we propose the MTPS algorithm to select the optimal transmission path. Given n relay nodes, we show that the complexity of the proposed algorithm is $\mathcal{O}(n^2)$, in contrast to $\mathcal{O}(n \cdot n!)$ by the brute-force approach. Results have demonstrated that:

- The average throughput can be remarkably improved by using more relays.

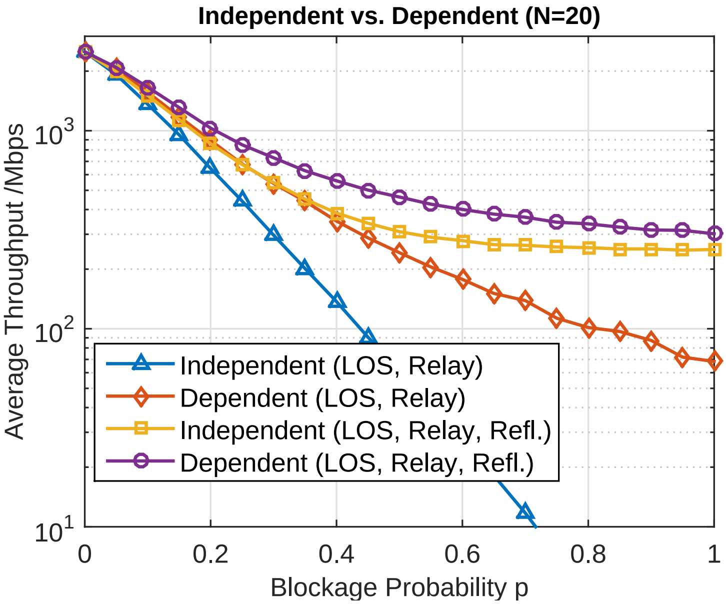


Figure 1.4: Impact of blockage probability p on average throughput [Mbps] in the topology-dependent or topology-independent model.

- Reflection path is quite beneficial to preserve the link connectivity, especially when the link blockage probability is high.
- Distinct random blockage models may result in remarkably different performance in terms of average throughput.

J2: Performance Analysis of Millimeter-Wave Relaying: Impacts of Beamwidth and Self-Interference [YX18]

In this paper, we have studied the maximum achievable rate of a two-hop AF relaying system with mm-wave. Specifically, with Gaussian-type directional antenna and two-ray mm-wave channel incorporated, we have investigated two AF relaying schemes, namely HD and FD, where the impacts of beamwidth, self-interference and ground reflections are studied.

Backgrounds It is well known that the severe path loss is a typical challenge in mm-wave communications, and one common solution is to use directional antennas, such that the power can be concentrated along the desired orientation for transmission [RRS$^+$05, YYMZ06, Epp06, WNLa$^+$14]. Usually, an idealized radiation pattern (often referred to as the “flat-top” model, which consists of a large constant antenna gain within the narrow main-lobe and zero elsewhere) was widely used for performance analysis [WNE02,KPL03,SMM11]. Note that mm-wave signals suffer higher penetration loss, weaker diffusion, diffraction, and high-order reflections [GKZV09], which restrict mm-wave to short-range wireless communications. Thus, relays may be used for mm-wave systems to achieve higher coverage and robustness. Numerous efforts have been devoted to exploring potentials of performance enhancement via relays in mm-wave communications [GOON10, DSZC13, RST14, LJT14].

However, for mm-wave relaying systems with directional antennas, there are two major limitations in the existing works:

- Although the sectorized model (refer to [WNLa$^+$14,SGFF$^+$15,BH15], consisting of only two constant gains for main and side lobes without any transition) is extensively used for modeling the radiation pattern of directional antennas, a crucial "roll-off" feature (a gradual decay from the main lobe to the side lobe) of the real-world radiation pattern for directional antennas is missing, thereby potentially causing discontinuity in performance assessments.

- Most of the preceding works (e.g., [GKZV09, SGFF$^+$15, SKGA15, TBHJ16]) do not consider the impact of ground reflections, which is commonly regarded as first-order reflections that however have been shown not negligible in mm-wave communications [GRON10, YDX15, LvdBS$^+$16, VH16]. The conventional mm-wave channel model usually considers the LoS path only, potentially leading to the imprecision in performance evaluations when omitting the non-negligible effects of reflections.

Motivation & Contributions To address the aforementioned two concerns, we jointly consider the *Gaussian-type directional antenna model* [Gag12,MiW14,TH15] and the *two-ray channel model* [Gol05, RJLMPG17] in our study. Under these two models, the objective of our work is to investigate the maximum achievable rates of a two-hop AF relaying system with mm-wave, where both HD and FD modes with a sum-power constraint are considered.

Main contributions of our work can be briefly presented as below:

- With a two-ray channel model, the impact of main-lobe beamwidth θ_m on the rate of mm-wave relaying systems scales as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$.

- The rate of mm-wave relaying systems with an FD relay scales as $\mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$, where $\mu \in (0,1)$ characterize the self-interference cancellation.

- The constructive or destructive contribution of ground reflections is indeed nontrivial, which contradicts the conventional belief that the effect of ground reflections is negligible in mm-wave communications.

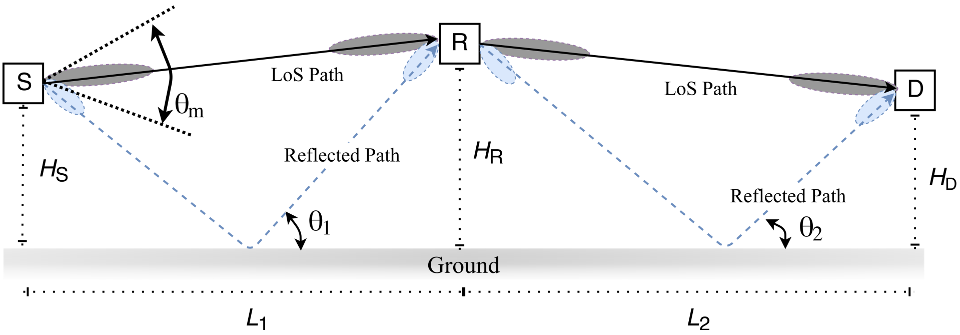


Figure 1.5: Two-hop AF relaying system: two-ray channel and directional antennas.

Model, Methodology & Analysis An illustration of two-hop AF relaying system is given in Fig. 1.5, where the source node, the destination node and the relay node are denoted by S, D and R (in the HD or FD mode), respectively[1]. Directional antennas are employed at all nodes. Regarding mm-wave channels, we denote by $h_i \in \mathbb{C}$, $i \in \{1,2\}$ the channel coefficients of links $S-R$ and $R-D$, respectively. For simplicity, small-scale fading and shadowing effect are not considered in this paper.

In the two-ray channel model, h_i for $i \in \{1,2\}$ are given as

$$\begin{cases} h_1 = \dfrac{\lambda\left(G(0) + G(\theta_1)\,\Gamma(\theta_1)\cos(\theta_1)\,e^{-j\Delta\varphi_1}\right)}{4\pi\sqrt{(H_S - H_R)^2 + L_1^2}} \\ h_2 = \dfrac{\lambda\left(G(0) + G(\theta_2)\,\Gamma(\theta_2)\cos(\theta_2)\,e^{-j\Delta\varphi_2}\right)}{4\pi\sqrt{(H_R - H_D)^2 + L_2^2}} \end{cases}, \tag{4}$$

where λ denotes the wavelength, and θ_i for $i \in \{1,2\}$ denote the reflection angles relative to the ground plane:

$$\theta_1 = \arctan\left(\frac{H_S + H_R}{L_1}\right) \quad \text{and} \quad \theta_2 = \arctan\left(\frac{H_R + H_D}{L_2}\right). \tag{5}$$

Furthermore, the phase differences $\Delta\varphi_i$ are expressed as

$$\begin{cases} \Delta\varphi_1 = \dfrac{2\pi}{\lambda}\left(\sqrt{(H_S{+}H_R)^2{+}L_1^2} - \sqrt{(H_S{-}H_R)^2{+}L_1^2}\right) \\ \Delta\varphi_2 = \dfrac{2\pi}{\lambda}\left(\sqrt{(H_R{+}H_D)^2{+}L_2^2} - \sqrt{(H_R{-}H_D)^2{+}L_1^2}\right) \end{cases}. \tag{6}$$

[1] We assume that there exists no direct link between S and D in our scenario.

The reflection coefficient $\Gamma(\theta_i)$ with reflection angle θ_i [Gol05] is given as

$$\Gamma(\theta_i) = \frac{\omega \sin\theta_i - \sqrt{\omega - \cos^2(\theta_i)}}{\omega \sin\theta_i + \sqrt{\omega - \cos^2(\theta_i)}}. \tag{7}$$

where ω denotes the dielectric constant of ground. In our work, the radiation pattern $G(\cdot)$ for Gaussian-type directional antenna is given as

$$G(\phi) = \frac{2\pi}{2\pi + 42.64443\theta_m} \cdot 10^{2.028\left[1-\left(\frac{2\phi}{\theta_m}\right)^2\right]^+}. \tag{8}$$

For notational simplicity, we define $g_i \triangleq |h_i|^2$ and $C(s) \triangleq \log_2(1+s)$ in what follows. Besides, ξ_i with $i \in \{1,2\}$ denotes the transmit powers on the i^{th} hop, and ξ denotes the sum-power constraint. The maximum achievable rates with respect to different duplex modes are given as below.

For two-hop AF relaying with HD, the maximum achievable rate of via time sharing and power allocation is formulated in the following proposition.

Proposition 1.2.3. *For two-hop HD-AF relaying, given sum-power constraint $\beta\xi_1 + (1-\beta)\xi_2 = \xi$, the maximum achievable rate is formulated as*

$$\eta^*_{\text{HD}} = \max_{\substack{0\le\beta\le1 \\ \xi_1,\xi_2\ge0 \\ \beta\xi_1+(1-\beta)\xi_2=\xi}} \min \left\{ \begin{array}{c} \beta C(g_1\xi_1), \\ (1-\beta)\, C\left(\dfrac{g_1 g_2 \xi_1 \xi_2}{1+g_1\xi_1+g_2\xi_2}\right) \end{array} \right\}. \tag{9}$$

However, it is non-trivial to solve the optimization problem formulated in Proposition 1.2.3, since the selection of the optimal time-sharing parameter and power allocation are intertwined with each other (refer to the sum-power constraint $\beta\xi_1 + (1-\beta)\xi_2 = \xi$). In this case, we are unable to decouple the constraints straightforwardly, such that it is intractable to obtain a closed-form solution for η^*_{HD}.

For two-hop AF relaying with FD, the maximum achievable rate of by applying power allocation is given in the following proposition.

Proposition 1.2.4. *For two-hop FD-AF relaying, given the sum-power constraint $\xi_1 + \xi_2 = \xi$, the maximum achievable rate is*

$$\eta^*_{\text{FD}} = C\left(\frac{g_1 g_2 \xi^2}{2+(g_1+g_2+\mu)\,\xi+2\sqrt{(1+g_1\xi)\,(1+g_2\xi)\,(1+\mu\xi)}}\right), \tag{10}$$

where the optimal power allocation is given as

$$\begin{cases} \xi_1^* = \dfrac{\xi\sqrt{1+(\mu+g_2)\,\xi+\mu g_2\xi^2}}{\sqrt{1+g_1\xi}+\sqrt{1+(\mu+g_2)\,\xi+\mu g_2\xi^2}} \\ \xi_2^* = \dfrac{\xi\sqrt{1+\xi g_1}}{\sqrt{1+\xi g_1}+\sqrt{1+(\mu+g_2)\,\xi+\mu g_2\xi^2}} \end{cases}. \tag{11}$$

With the maximum achievable rates given in Proposition 1.2.3 and Proposition 1.2.4, we have following theorems regarding the impacts of beamwidth and self-interference coefficient on the maximum achievable rate.

Theorem 1.2.1. *With respect to beamwidth θ_m, we have $\eta^*_{\mathrm{HD}} \in \mathcal{O}\left(\min\left(\theta_m^{-1}, \theta_m^{-2}\right)\right)$ and $\eta^*_{\mathrm{FD}} \in \mathcal{O}\left(\min\left(\theta_m^{-1}, \theta_m^{-2}\right)\right)$.*

The unsurprising result in Theorem 1.2.1 is in line with the fact that a narrower beamwidth enables a higher achievable rate.

Theorem 1.2.2. *η^*_{FD} is strictly convex and monotonically decreasing with respect to self-interference coefficient $\mu \in (0, 1)$, and $\eta^*_{\mathrm{FD}} \in \mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$.*

Theorem 1.2.2 shows the convexity and monotonicity of η^*_{FD} with respect to μ, indicating the critical role of the self-interference coefficient in improving the performance of FD-AF mm-wave relaying.

Performance Evaluation In performance evaluation, for simplicity, we consider a linear network, i.e., $L_1 + L_2 = L$, where all nodes are deployed with the identical height, i.e., $H_S = H_R = H_D = H$.

Maximum achievable rates of different mm-wave systems are illustrated in Fig. 1.6. For HD-AF relaying, a performance gain, i.e., roughly 5 dB, is achieved when using the optimal time-sharing scheme, compared to $\beta = \frac{1}{2}$. For FD-AF relaying, η^*_{FD} substantially increases when μ reduces. We note that a smaller μ is important for FD-AF relaying with mm-wave, particularly given a medium or high ξ, i.e., $\xi \geq 100$ dB, while the impacts of μ are relative smaller for $\xi \leq 90$ dB. It can be seen that, if a higher ξ is given, the direct transmission is the best choice. For relaying schemes, FD-AF relaying with a small μ should be adopted for scenarios with medium ξ, while HD-AF relaying is more suitable for scenarios with low ξ.

The impacts of beamwidth are shown in Fig. 1.7. The performance of all schemes suffer degradation as θ_m increases, where direct transmission encounters the fastest decay, and FD-AF follows. The tendency that the maximum achievable rate degrades as θ_m as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$ is reflected from all curves. Compared to η^*_{HD}, we see that η^*_{FD} dramatically decreases when θ_m grows, indicating that FD-AF mm-wave relaying is more sensitive to the variation of beamwidth. A higher η^*_{FD} can be achieved if a smaller μ (e.g., $\mu \leq -90$ dB) and θ_m (e.g., $\theta_m \leq \frac{\pi}{6}$) are given. Otherwise, HD-AF or direct transmission strategy gives better performance. Given $\xi = 100$ dB, we figure out that the direct transmission outperforms HD-AF when sharp beams are used.

Fig. 1.8 depicts the contribution of ground reflections, where curves labeled with "1-ray" are taken as references, corresponding to the conventional modeling method based on LoS only. Evidently, rates with the two-ray model (labeled with "2-ray") heavily rely on L_1 and L_2. Also, the performance fluctuations resulted by ground reflections, which become significant as θ_m grows, are not as minor as commonly

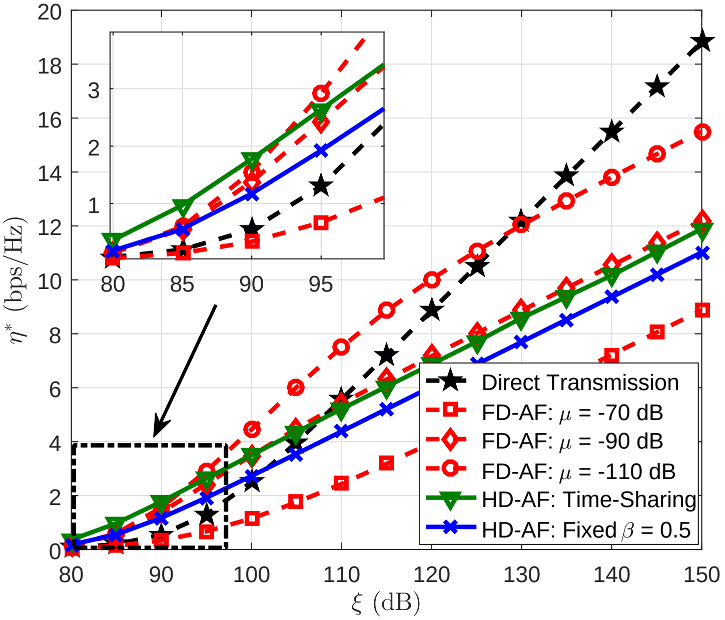


Figure 1.6: Maximum rates of direct transmission and both AF relaying schemes under different sum-power constraints, where $L_1 = 80$ m and $\theta_m = \frac{\pi}{6}$.

believed. The observations conveys an insight that, if the beamwidth is not very small, a proper deployment of relay can exploit constructive ground reflections, thereby gaining considerable improvements in rates. When θ_m gets smaller, the rate with the "2-ray" model converges to that with the "1-ray" model, indicating that the "1-ray" model only fits scenarios with very sharp beams.

The impact of μ on the rate of the FD-AF mm-wave relaying is shown in Fig. 1.9. As aforementioned in Theorem 1.2.2, η^*_{FD} varies in the convex and decreasing manner with respect to μ. Hence, slightly reducing μ can achieve a remarkable performance improvement, and the gain becomes significant at lower μ and higher ξ.

Concluding Remarks We have studied the maximum achievable rate of the two-hop AF relaying system with mm-wave, where both HD and FD are discussed. With the joint treatment of the two-ray mm-wave channel and Gaussian-type directional antenna, we have investigated the impact of the beamwidth and the self-interference coefficient. Under sum-power constraint, results have shown that:

- FD outperforms HD only if both the beamwidth θ_m and the self-interference coefficient μ are small.

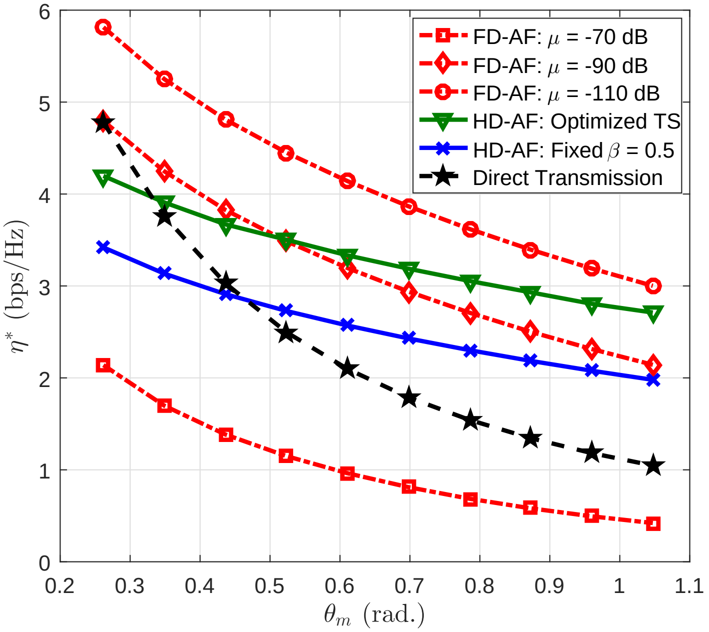


Figure 1.7: Maximum rates of AF relaying schemes (FD-AF, HD-AF with time-sharing, and HD-AF with $\beta = \frac{1}{2}$) with different beamwidth under a sum-power constraint $\xi = 100$ dB are applied, and $L_1 = 80$ m.

- If the sum-power budget ξ is sufficiently high or the beamwidth θ_m is sufficiently small, the best option would be the direct transmission.
- For both relaying schemes, the rate scales as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$, and the rate for FD-AF scales as $\mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$.
- Ground reflections may significantly affect the performance of mm-wave communications in a constructive or destructive manner.

1.2.2 Latency analysis via stochastic network calculus

J3: Analysis of Millimeter-Wave Multi-Hop Networks with Full-Duplex Buffered Relays [YXAZ+18]

In this paper, using stochastic network calculus based on moment generating function (MGF), we have analyzed the two critical performance guarantees, namely the total backlog and the E2E delay, where buffer-aided relays are incorporated in multi-hop mm-wave networks. We have studied the effect of self-interference at

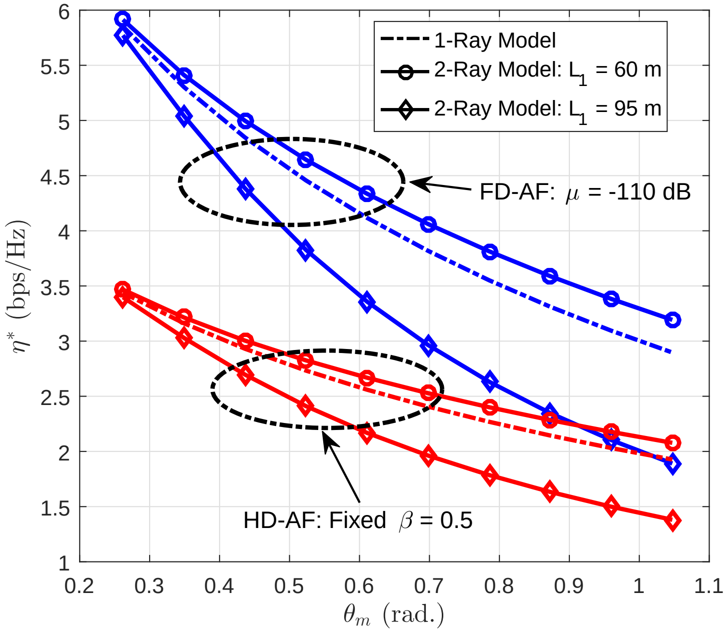


Figure 1.8: Contribution of ground reflections to rates with respect to different beamwidth, where $\xi = 100$ dB.

FD relays on the network performance and proposed an optimal power allocation scheme to improve network performance.

Backgrounds In mm-wave communications, multi-hop buffered network is a promising solution to deal with the unprecedented data volumes and to mitigate the effects of serious path loss over long distances and/or the effect of NLoS, while maintaining the traffic flows' QoS requirements. Also, in order to improve the network throughput, FD relays may be used since they allow simultaneous transmission and reception. In spite of self-interference, FD relaying is still promising in enhancing the network throughput through interference cancellations [DS10, JCK$^+$11, CJS$^+$10]. Numerous efforts dedicated to FD relaying with mm-wave can be found in [LJT14, MKJ$^+$16, WZS$^+$16a, AH16, DCK16, WZSH16].

Stochastic network calculus is extensively employed for assessing network performance in the presence of stochastic arrivals and/or service process. In the pioneering literature [Cha00, CCS01, LBL07], MGF-based characterizations were developed to model traffic and service processes with independent increments and to utilize the independence among multiplexed flows, the MGF-based network calculus was proposed [Fid10]. However, the conventional MGF-based network calculus cannot be

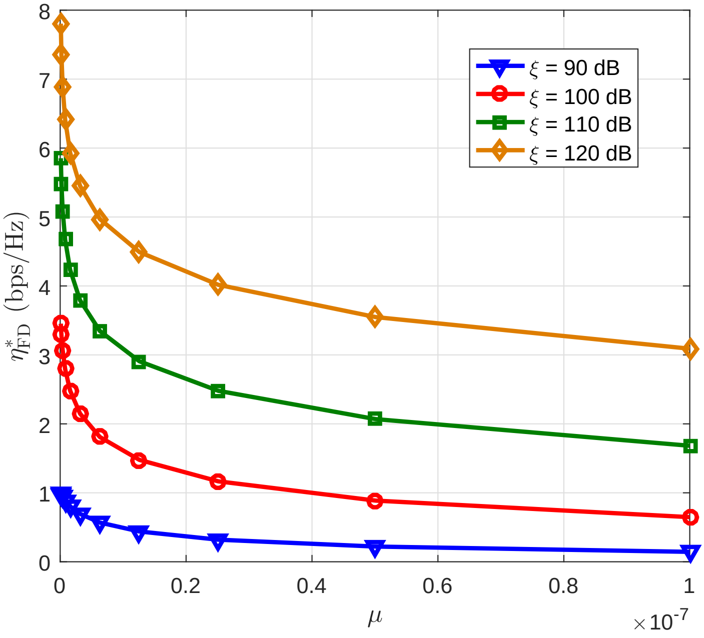


Figure 1.9: Impact of μ on η^*_{FD}, where $\theta_m = \frac{\pi}{4}$ and $L_1 = 100$ m.

extended to analyzing wireless networks, due to the existence of logarithm operator for channel capacity (often treated as instantaneous service process). To address this intractability, in the recent work [AZLB16], the $(\min, \times)$ network calculus approach was proposed, which is capable of giving probabilistic performance bounds directly in terms of the fading channel parameters via the transfer between the "bit domain" and the "SNR domain". To apply the $(\min, \times)$ network calculus to non-identically distributed multi-hop wireless networks, a recursive formula for delay bound computation was developed in [PAZKG15].

However, the following two major limitations are still not addressed in the existing literature:

- Although network calculus has developed for decades, its application to wireless networks analysis is fairly few, and the self-interference factor was not taken into account in most of previous works that are associated with FD relays.

- The performance guarantees of mm-wave multi-hop wireless networks considering heterogeneous self-interfered channels have not yet been studied before.

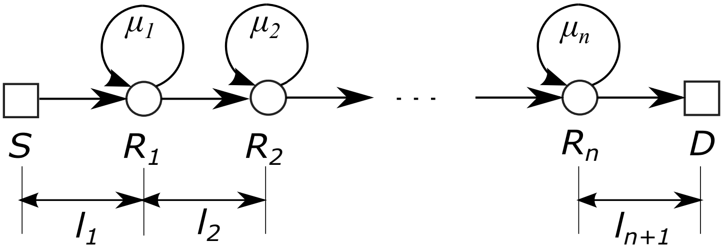


Figure 1.10: A multi-hop wireless network with n full-duplex relays.

Motivation & Contributions Our work is motivated by the importance and the potential of mm-wave buffered networks in future mobile communications along with the aforementioned unsolved limitations in preceding research. Specifically, our objective is to study the total backlog and delay performance, in the presence of self-inferred mm-wave channels, constrained sum-power budget and given QoS requirement. Main contributions of our work can be outlined as below:

- In contrast to the recursive method developed in [PAZKG15], we derive performance bounds for homogeneous and heterogeneous wireless networks in a straightforward manner, which make a general contribution to the theory of stochastic network calculus.
- In the presence of self-interference and sum-power constraint, a power allocation scheme is developed to optimize the network performance, and the impacts of self-interference coefficient on the two performance guarantees are studied.

Model, Methodology & Analysis An illustration of multi-hop mm-wave network is given in Fig. 1.10, which consists of a source S, n ($n \geq 1$) FD relays R_i, $i \in [n]$ and a destination D. For simplifying illustration, the labels 0, 1, …, $n+1$ are assigned to the ordered nodes, where S and D correspond to nodes 0 and $(n+1)$, respectively. We label the channel between nodes $(i-1)$ and i as the i^{th} hop in the set of hops $\mathcal{I}_{\mathcal{H}}$, i.e., $i \in \mathcal{I}_{\mathcal{H}} = [n+1]$, and the distance between adjacent nodes by l_i. We denote the channel gain coefficient of the i^{th} hop by g_i. In this work, we mainly consider large-scale fading, e.g., shadowing, in mm-wave bands. Given the separation distance l_i, a generalized expression of g_i based on path-loss model is given as [RRE14, RMSS15]

$$g_i[\text{dB}] = -\left(\alpha + 10\beta \log_{10}(l_i) + \xi_i\right), \tag{12}$$

where α and β are the least square fits of floating intercept and slope of the best fit, and $\xi_i \sim \mathcal{N}(0, v_i^2)$ corresponds to the log-normal shadowing with variance v_i^2.

A self-interference coefficient $\mu \in (0,1)$ is introduced to characterize the self-interference residing at FD relays [DS10]. In the presence of self-interference, the

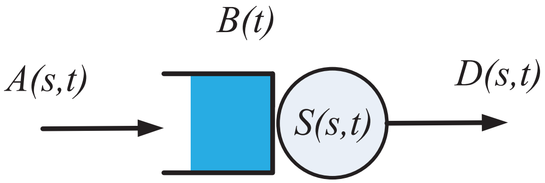


Figure 1.11: A queuing model for a store-and-forward node.

signal-to-interference-plus-noise ratio (SINR) in the i^{th} hop, denoted by γ_i, for the described channel is $\gamma_i = \kappa \cdot \omega_i \cdot g_i$ with ω_i given as

$$\omega_i = \begin{cases} \dfrac{\lambda_{i-1}}{1+\mu_i\lambda_i}, & i \in [n] \\ \lambda_{i-1}, & i = n+1 \end{cases}, \tag{13}$$

where κ is a scalar determined by system configurations, $\lambda_i \triangleq \frac{P_i}{N_0}$ is the transmitted SNR at node i defined by the transmit power P_i and background noise power N_0. We also assume a sum-power constraint P_{tot} throughout our work, i.e., $\sum_{i=0}^{n} P_i = P_{\text{tot}}$. Equivalently, given a constant background noise power N_0 for all hops, the sum-power constraint can be reformulated as $\sum_{i=0}^{n} \lambda_i = \lambda_{\text{tot}} \triangleq \frac{P_{\text{tot}}}{N_0}$.

We assume the stochastic process of each hop to be stationary and independent in time. That is, we can use a series of independent random variables γ_i to characterize the multi-hop channels, namely, $\gamma_i^{(k)} \overset{\ell}{=} \gamma_i$ in all time slot k, where $\overset{\ell}{=}$ denotes equality in law (i.e., in distribution)[2]. We decompose the set of hops, $\mathcal{I}_{\mathcal{H}}$, into m subsets, $\mathcal{X}_k, k \in \mathcal{I}_{\mathcal{M}} = [m]$, where, $\mathcal{I}_{\mathcal{H}} = \bigcup_{k=1}^{m} \mathcal{X}_k$, with $\mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset$ for all $i, j \in \mathcal{I}_{\mathcal{M}}$ such that $i \neq j$, where $\mathcal{X}_k$ is defined as the set of indices such that

$$\mathcal{X}_k = \{j \in \mathcal{I}_{\mathcal{H}}, k \in \mathcal{I}_{\mathcal{M}} : F_{\gamma_j}(x) = F^{\langle k \rangle}(x)\}, \tag{14}$$

where $F_X(x)$ is the probability density function (p.d.f.) of the random variable X, $F^{\langle k \rangle}(x)$ represents a unique distribution function corresponding to the subset of i.i.d. hops denoted by the index $k \in \mathcal{I}_{\mathcal{M}}$. We emphasize that $|\mathcal{I}_{\mathcal{H}}| \geq |\mathcal{I}_{\mathcal{M}}|$, where $|\mathcal{Y}|$ represents the cardinality of the set $\mathcal{Y}$, and the equality is attained when the multi-hop network is fully heterogeneous.

An illustration of traffic and service process is given in Fig. 1.11, where $A(s,t)$, $S(s,t)$ and $D(s,t)$ denote the cumulative arrival process, service process, and departure process, respectively. For notational simplicity in analyses below, we define $p_a^{t-s}(\theta) \triangleq \exp(-\theta\sigma(\theta)) \cdot \mathbb{M}_A(\theta, s, t)$ and $q^{t-s}(-\theta) \triangleq \overline{\mathbb{M}}_S(\theta, s, t)$ regarding

[2] The log-normal shadowing resulted by objects obstructing the propagation of mm-wave radios is incorporated in the channel gain coefficient. Normally, shadowing is not spatially independent. However, considering the fact that highly directional antennas are commonly used for mm-wave communications, it is safe to assume the independence across hops.

the MGF-based characterizations of arrival and service, respectively, where $\sigma(\theta)$ characterizes the burstiness of stochastic arrivals. Solving $b^{\varepsilon'}$ for probabilistic backlog requirement $\Pr\left(B(t) > b^{\varepsilon'}\right) \leq \varepsilon'$ and $w^{\varepsilon'}$ for probabilistic delay requirement $\Pr\left(W(t) > w^{\varepsilon''}\right) \leq \varepsilon''$, we have [Fid06b, AZLB16]

$$b^{\varepsilon'} = \inf_{\theta>0} \left\{ \frac{1}{\theta} \left(\log \mathsf{M}(\theta, t, t) - \log \varepsilon' \right) \right\} \tag{15}$$

and

$$w^{\varepsilon''} = \inf \left\{ w : \inf_{\theta>0} \left\{ \mathsf{M}(\theta, t+w, t) \right\} \leq \varepsilon'' \right\}, \tag{16}$$

where $\mathsf{M}(\theta, s, t)$ is defined as

$$\mathsf{M}(\theta, s, t) \triangleq \exp\left(\theta \sigma(\theta)\right) \cdot \sum_{u=0}^{\min(s,t)} p_a^{t-u}(\theta) \cdot q^{s-u}(-\theta). \tag{17}$$

We notice that, with the specific fading characteristic of mm-wave channels in our study, i.e., log-normal shadowing, it is intractable to give the MGF of cumulative service process in closed-form, since the *shifted log-normal random variable* has no closed-form expression for the inverse moment. Instead, in the following theorem, we present an upper bound for $q(-\theta)$ to characterize the MGF of Shannon-type service processes $S(s,t)$.

Theorem 1.2.3. *An upper bound for $q(-\theta)$ is given as*

$$q(-\theta) \leq \min_{u \geq 0} \left\{ \left(1 + \delta N_\delta(u)\right)^{-\eta\theta} + \sum_{i=1}^{N_\delta(u)} a_{\eta\theta,\delta}(i) F_\gamma(i\delta) \right\},$$

where $F_\gamma(\cdot)$ denotes the cumulative distribution function (c.d.f.) of the log-normally distributed SINR γ, and $N_\delta(u)$ and $a_{\theta,\delta}(i)$ are respectively given by $N_\delta(u) = \lfloor \frac{u}{\delta} \rfloor$ and $a_{\theta,\delta}(i) = \left(1 + (i-1)\delta\right)^{-\theta} - \left(1 + i\delta\right)^{-\theta}$.

Note that the tightness of the bound for $q(-\theta)$ depends on the parameter δ, the discretization step size. A smaller step size yields a tighter upper bound while leading to higher computational costs. More details can be referred to in [YXG$^+$16].

Recalling (15) and (16), it can be seen that $\mathsf{M}(\theta, s, t)$ plays an important role in obtaining the probabilistic backlog $b^{\varepsilon'}$ and probabilistic delay $w^{\varepsilon''}$. Hence, in what follows, we focus on computing $\mathsf{M}(\theta, s, t)$ for multi-hop wireless networks. Associated with the channel classifications in terms of $\mathcal{X}_i$ for $i \in [m]$, we define $\hat{q}_i(-\theta)$ to uniquely characterize the MGF of the cumulative service process for channels in the subset $\mathcal{X}_i$, such that $\hat{q}_i(-\theta) \neq \hat{q}_j(-\theta)$ for all $i \neq j$.

Regarding multi-hop wireless networks with heterogeneous channels in general, an upper bound for $\mathsf{M}(\theta, s, t)$ is provided in the following Theorem 1.2.4.

Theorem 1.2.4. *Let m be the number of subsets of identically distributed channels, $\tau \triangleq \max(s-t,0)$, and $V_i(\theta) \triangleq p_a(\theta) \cdot \hat{q}_i(-\theta)$, a upper bound for $\mathsf{M}(\theta,s,t)$, $\theta > 0$, for the $(n+1)$-hop wireless network is given by*

$$\mathsf{M}(\theta,s,t) \leq \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \cdot \sum_{i=1}^{m} \psi_i(\theta) \cdot V_i^{m-1}(\theta) \cdot \mathcal{K}_{\tau,n,m}\left(V_i(\theta)\right), \tag{18}$$

whenever the stability condition, $\max\limits_{i \in \mathcal{I}_{\mathcal{M}}} \{V_i(\theta)\} < 1$, is satisfied. Here, $\psi_i(\theta)$ for all $i \in \{1,\ldots,m\}$ is defined as

$$\psi_i(\theta) \triangleq \begin{cases} \prod\limits_{j \neq i} \left(V_i(\theta) - V_j(\theta)\right)^{-1}, & m \geq 2 \\ 1, & m = 1 \end{cases}, \tag{19}$$

and $\mathcal{K}_{\tau,n,m}(x)$ is defined as

$$\mathcal{K}_{\tau,n,m}(x) \triangleq x^{\tau} \cdot \binom{n+1-m+\tau}{n+1-m} \cdot {}_2F_1\left(1, n+2-m+\tau; \tau+1; x\right),$$

where ${}_pF_q\left(\underline{a};\underline{b};x\right)$ denotes the Generalized Hypergeometric Function *with vectors $\underline{a} = [a_1,\ldots,a_p]$ and $\underline{b} = [b_1,\ldots,b_q]$.*

If the multi-hop network is homogeneous, then applying the result from Theorem 1.2.4 for computation would be rather complex, due to the generalized hypergeometric function. The following Theorem 1.2.5 is proposed to simplify the computation with homogeneity, where $\hat{q}(\theta)$ characterizes the MGF of service process for all channels.

Theorem 1.2.5. *For homogeneous $(n+1)$-hop wireless networks characterized by the MGF service bound $\hat{q}(\theta)$, and for any $\theta > 0$, given $p_a(\theta)$ we have*

$$\mathsf{M}(\theta,s,t) \leq \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \cdot \mathcal{G}_{\tau,n}\left(p_a(\theta)\hat{q}(-\theta)\right), \tag{20}$$

where $\tau \triangleq \max(s-t,0)$, whenever the stability condition, $p_a(\theta)\hat{q}(-\theta) < 1$, holds. Here, $\mathcal{G}_{\tau,n}(x)$ is defined as

$$\mathcal{G}_{\tau,n}(x) \triangleq \min\left(\frac{\min\left(1, x^{\tau}\binom{n+\tau}{n}\right)}{(1-x)^{n+1}}, \frac{1}{(1-x)^{n+1}} - \binom{n+\tau}{n+1}x^{\tau-1}\right). \tag{21}$$

In what follows, we investigate the power allocation scheme in the presence of the self-interference, which enables the minimum E2E delay under sum-power constraint. We denote by $\Xi_n \subset \mathbb{R}_+^{n+1}$ the set of feasible power allocation schemes with respect to n intermediate nodes. Besides, the sum of all allocated power should be constrained by the total power budget, i.e.,

$$P_{\Sigma} \triangleq \sum_{i\in\mathcal{I}_{\mathcal{H}}} P_i \leq P_{\text{tot}}, \tag{22}$$

where P_{tot} is the total power budget. A power allocation vector $\mathbf{P} \triangleq \{P_i\}_{i\in\mathcal{I}_{\mathcal{H}}} \in \Xi_n$ is called as a *feasible* scheme if the power constraint above can be satisfied.

Prior to developing the optimal power allocation, the following two lemmas are given as preliminaries. Specifically, Lemma 1.2.1 tells that, given a feasible power allocation, the maximal network service capability is achieved only when the SINRs for all hops are identically distributed, and Lemma 1.2.2 indicates the existence and uniqueness of the optimal power allocation vector $\mathbf{P}^*$ when $P_{\Sigma} = P_{\text{tot}}$.

Lemma 1.2.1 (Sufficiency). *Given the sum transmit power budget P_{tot} for a $(n+1)$-hop wireless network, a feasible power allocation $\mathbf{P}^*$, where $P_{\Sigma} = P_{\text{tot}}$, that results in identically distributed SINR over all hops maximizes the lower bound on network service process whenever such $\mathbf{P}^*$ exists.*

Lemma 1.2.2 (Existence). *Given a $(n+1)$-hop wireless network operating under transmit power budget P_{tot}, there always exists a unique optimal power allocation $\mathbf{P}^*$ such that $P_{\Sigma} = P_{\text{tot}}$, among all feasible $\mathbf{P} \in \Xi_n$ in terms of maximizing a lower bound on network service process.*

Based on the above two lemmas, an optimal power allocation incorporating self-interference is elaborated in the following Theorem 1.2.6. For tractability, we assume the shadowing variance ν_i^2 for all $n+1$ channels are identical, while the self-interference coefficient μ_i at the i^{th} FD relay and the hop length l_i of the i^{th} channel can be distinct.

Theorem 1.2.6. *Given the total power budget P_{tot}, i.e., $P_{\Sigma} \leq P_{\text{tot}}$, and the background noise power N_0, for the mm-wave channel described in* (12)*, and let x^* denote the positive solution for the algebraic equation*

$$\sum_{k=1}^{n+1} \left(\sum_{i=0}^{n+1-k} \nu_{i,k} \right) x^k = \frac{P_{\text{tot}}}{N_0}, \tag{23}$$

with $\nu_{i,k}$ given by

$$\nu_{i,k} = \mu_{i+k}^{-1} \cdot \prod_{u=1}^{k} \mu_{i+u} \cdot l_{i+u}^{\beta}. \tag{24}$$

Then there exists a unique optimal power allocation strategy $\mathbf{P}^* \in \Xi_n$, such that

$$P_i^* = N_0 \sum_{k=1}^{n-i+1} \nu_{i,k} \cdot (x^*)^k, \quad \textit{for } i \in \mathcal{I}_{\mathcal{H}}.$$

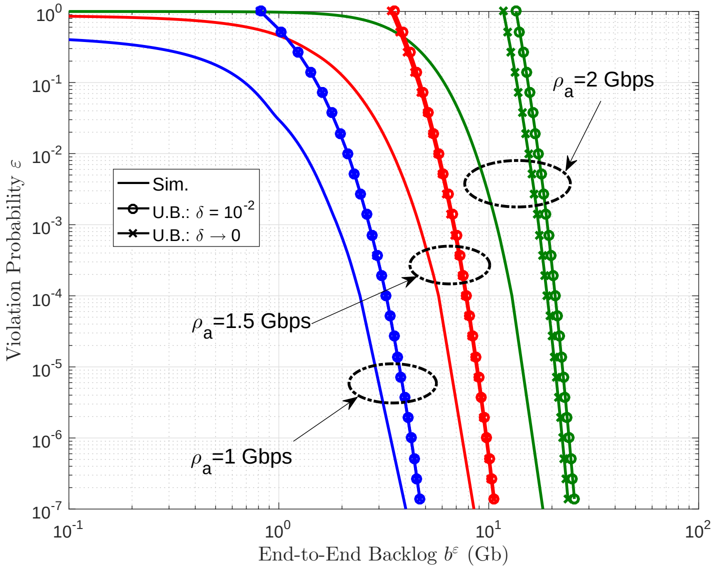


Figure 1.12: Violation probability ε vs. targeted theoretical backlog bounds b^{ε}, compared to simulations for different $\rho_a = 1$, 1.5 and 2 Gbps, with $n = 10$, and $\delta = 10^{-2}$ and $\delta \to 0$, respectively.

Performance Evaluation In performance evaluation, we assume the following configurations regarding the multi-hop buffer-aided network: (i) constant-rate arrivals without burstiness, i.e., $\sigma(\theta) = 0$ and $\rho(\theta) = \rho_a$; (ii) identical μ for all relays, and uniform and linear relay deployment; (iii) infinite buffer at each relay, i.e., no overflow; (iv) time-slotted system with time intervals of 1 second. For the sake of simplicity, in what follows the analytic bounds for homogeneous scenarios are all illustrated by Theorem 1.2.5, while heterogeneous counterparts are provided by applying Theorem 1.2.4.

Considering a tandem 60 GHz network consisting of $n = 10$ relays that have identical self-interference coefficient $\mu = -80$ dB and the total power budget $\lambda_{\text{tot}} = 134$ dB, we show the total backlog and the E2E delay in Fig. 1.12 and Fig. 1.13, respectively, where Theorem 1.2.5 is applied for computation since the optimal power allocation in Theorem 1.2.6 enables the homogeneity. We find that the bound accurately predicts the slope of the simulation curve, and the gap diminishes asymptotically. It can be seen that the performance with $\delta = 10^{-2}$ is close to that by $\delta \to 0$, while a smaller δ may be required as the system utilization becomes higher (i.e., higher arrival rate ρ_a). Two figures above indicate the availability of the derived bounds for asymptotic performance analysis.

Fig. 1.14 illustrates benefits of using the optimal power allocation on reduc-

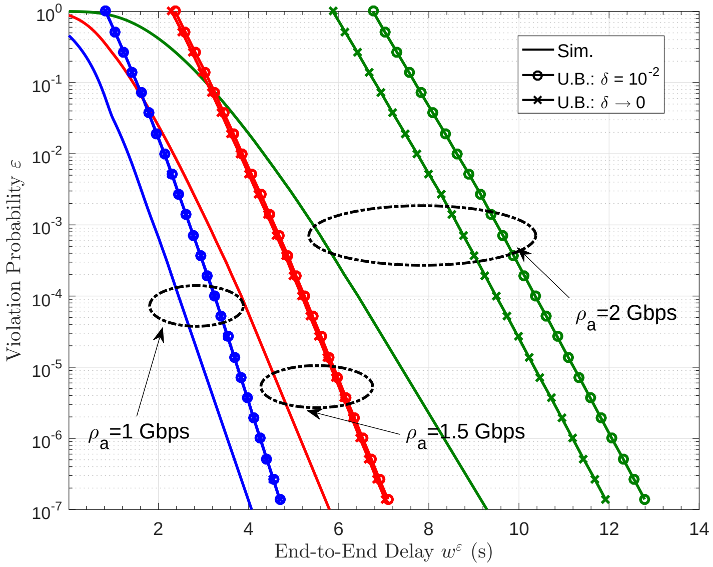


Figure 1.13: Violation probability ε vs. targeted theoretical delay bounds w^ε, compared to simulations for different $\rho_a = 1, 1.5$ and 2 Gbps, with $n = 10$, and $\delta = 10^{-2}$ and $\delta \to 0$, respectively.

ing the E2E delay. A uniform power allocation, i.e., $\{P_i\}_{i=0}^{n} = \frac{50}{11}$W, is taken as a reference. The upper bound associated with the optimal power allocation (see Theorem 1.2.5 for the homogeneous case) is asymptotically tight, while its counterpart (see Theorem 1.2.4 for heterogeneity with $m \geq 2$ due to the non-optimal power allocation) is not. The slackness of bounds for the heterogeneous scenario comes from producing a binomial coefficient in Theorem 1.2.4, which is a relaxed coefficient for the simplified and unified purpose. In this sense, the recursive approach by [PAZKG15] is able to give a tighter bound at the expense of higher computational complexity for the heterogeneous cases. Observing the slope of the probabilistic E2E delay, we also note that the network without the optimal power allocation suffers severer performance degradation.

Varying self-interference coefficient μ or the number of relays n, the probabilistic E2E delay bound w^ε is shown in Fig. 1.15. For noise-limited scenarios, the delay bound is less sensitive to the variation of relay numbers (actually a higher relay density constructively contribute to decreasing the delay, which however is not displayed here), while for interference-limited scenarios, either a lower or a higher relay density outperforms that in between significantly.

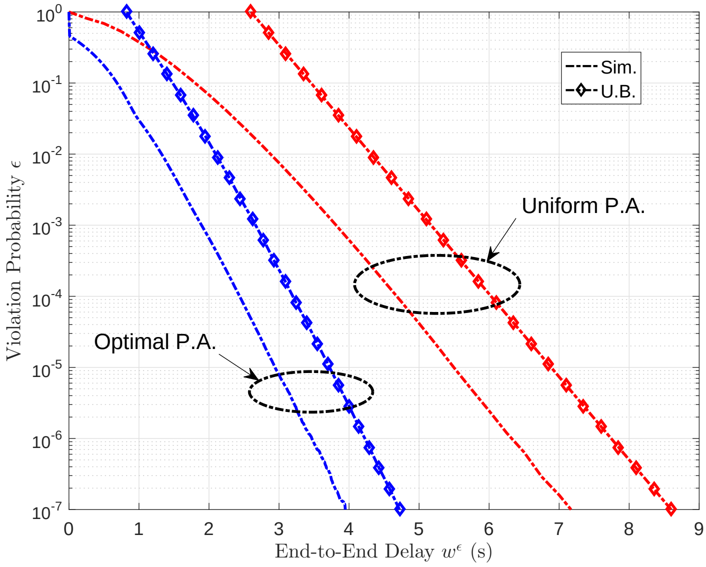


Figure 1.14: Violation probability ε vs. targeted theoretical delay bounds w^{ε}, compared to simulations for two power allocation strategies, with $\mu = -80$ dB, $n = 10$, $\rho_a = 1$ Gbps and $\delta = 10^{-2}$.

Concluding Remarks We have investigated the probabilistic total backlog and the E2E delay for multi-hop mm-wave networks with FD buffered relays, via MGF-based stochastic network calculus. Results showed that:

- The derived bounds for multi-hop networks are asymptotically tight, indicating the capability of keeping the track of probabilistic backlog and delay performance in a low-complexity and straightforward manner.
- The optimal power allocation in the presence of self-interference plays a crucial role in improving network performance, and it is beneficial to reduce the self-interference coefficient.
- Given a sum-power constraint, a higher relay density does not always guarantee a better network performance unless the self-interference coefficient is sufficiently small.

J4: Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification? [YXP18]

In this paper, we have considered the distinct strategies, namely traffic dispersion and network densification, which are potentially used in future mobile networks

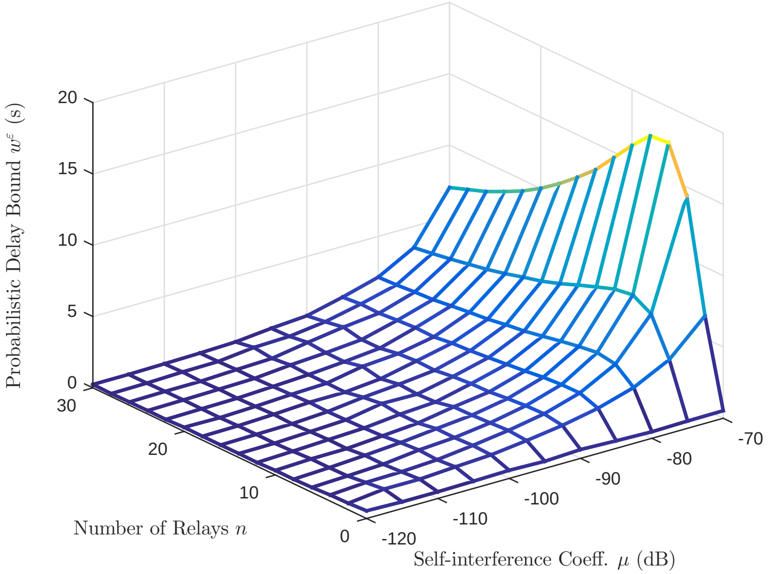


Figure 1.15: Probabilistic delay bound w^{ε} vs. n and μ jointly, with $\varepsilon = 10^{-6}$.

for low latency, and we have investigated their latency performance in terms of probabilistic delay and effective capacity. A hybrid scheme that combines these two strategies has also been studied. The respective benefits of traffic dispersion, network discussion and the hybrid scheme have been discussed.

Backgrounds Latency as an important metric for QoS plays a crucial role in the 5G mobile communications [FZM+17, SMS+17, OBB+14]. Particularly, for delay-sensitive applications, the queuing delay has attracted massive research attention, since it is the major contributor to the total delay in buffer-aided wireless networks. In [FZM+17], critical challenges and possible solutions for delivering E2E low-latency services in mm-wave cellular systems were surveyed from various perspectives, e.g., protocols at the medium access control (MAC) layer, congestion control, and core network architecture. By applying the Lyapunov technique for the utility-delay control, the problem of ultra-reliable and low-latency in mm-wave-enabled massive MIMO networks was studied in [VLB+17]. For systems with buffers at transceivers, the probabilistic delay for point-to-point mm-wave communications is analyzed in [YXG+16], where the delay bound was derived based on network calculus theory.

In queuing theory, the key idea to reduce the queuing delay is to keep low service utilization. Commonly, low service utilization is achieved via offloading the arrival traffic or improving the service capability. In wireless networks, offloading the arrival traffic can be realized by adopting the traffic dispersion scheme (stemming from the distributed antenna system (DAS) or remote radio head (RRH) in mm-wave communications), and service enhancement can be realized by adopting the network densification scheme (motivated from the fashion of dense facility deployments). Both traffic dispersion and network densification are promising and competitive candidates for low-latency mm-wave communications. In the past decade, numerous efforts have been dedicated to the research regarding the dispersion scheme (e.g., [MPH06, PFLZ12, RFC12, HCCP16, FJ16, RPC16]) and multi-hop relaying scheme (e.g., [CBL06, LH08, BA09, AZLB16]).

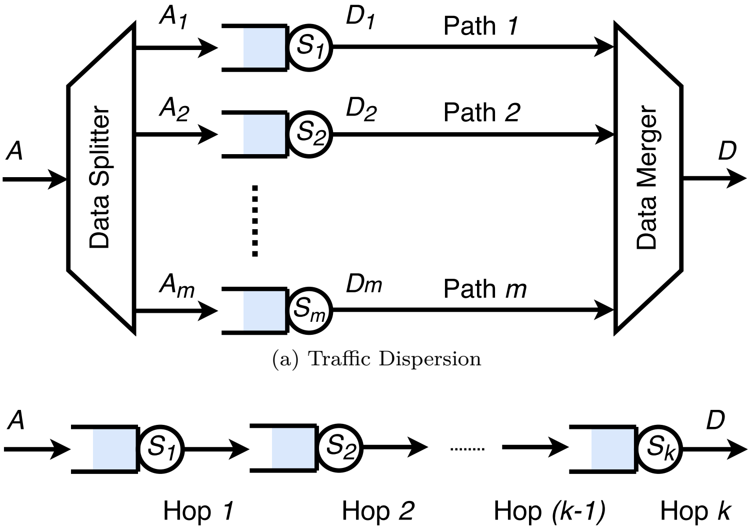
Motivation & Contributions The motivation of our research comes from the following two facts: (i) For the mentioned traffic dispersion and network densification schemes, their potential in achieving low-latency communications is not clear yet, although many related works have been done. (ii) A combination of traffic dispersion and network densification should work well in certain scenarios, thereby deserving to be thoroughly studied.

The objective of our work is to investigate the capabilities of traffic dispersion, network densification, and the hybrid scheme in reducing the E2E latency of buffer-aided networks. Main contributions are summarized as below:

- We comprehensively investigate the respective strengths of above two strategies and propose a generic hybrid scheme, i.e., a flexible combination of traffic dispersion and network densification. Probabilistic delay bounds for traffic dispersion, network densification, and the hybrid scheme are derived, respectively, with respect to networks with heterogeneous settings.
- With Nakagami-m fading in mm-wave channels, the MGF of the service process is provided in closed-form, and the availability of using $(\min, +)$-algebra to address problems in wireless communications indicates an alternative parallel to the $(\min, \times)$-based methodology.
- For traffic dispersion, we show the maximum effective capacity and identify the condition for achieving the optimum. Also, for network densification and the hybrid scheme, we derive lower and upper bounds to characterize their actual effective capacity.

Model, Methodology & Analysis For simplicity, we assume constant arrival rate ρ for the incoming data traffic, i.e., $A(s,t) = \rho \cdot (t-s)$ for any $0 \leq s \leq t$. The mechanisms of the two schemes illustrated in Fig. 1.16 are respectively described as follows:

- For traffic dispersion (see Fig. 1.16a), the original arrival traffic is partitioned into multiple sub-streams by the data splitter, where a set deterministic splitting coefficients $(z_1, z_2, \cdots, z_n)$ satisfying $\sum_{i=1}^{n} z_i = 1$ is adopted to gener-



(a) Traffic Dispersion

(b) Network Densification

Figure 1.16: Illustrations of two schemes for reducing latency for mm-wave communications: (a) traffic dispersion, and (b) network densification.

ate multiple sub-streams. Each sub-stream subsequently is served and delivered towards the receiver through the given path, independently, and all sub-streams are combined through the data merger from different paths, thereby forming the output traffic.

- For network densification (see Fig. 1.16b), multiple relays[3] as servers are deployed along the source-destination transmission path. Due to the concatenation of relying nodes, the output traffic departing from one node is the input traffic for the next node.

Paired traffic dispersion with network densification, a hybrid scheme is established, as shown in Fig. 1.17.

Different from the large-scale fading model in [YXAZ+18], we hereby consider the small-scale fading in mm-wave bands, which is commonly modeled as a Nakagami-m random variable [BH15, YZHL17]. Assuming the small-scale fading channel with Nakagami-m distribution is independent and identically distributed (i.i.d.) over time in terms of blocks, i.e., i.i.d. block fading, the capacity of mm-wave channel is

$$C = B \log_2 \left(1 + \xi \gamma l^{-\alpha}\right), \tag{25}$$

[3] FD relay nodes are used, and the self interference is not taken into account for simplicity.

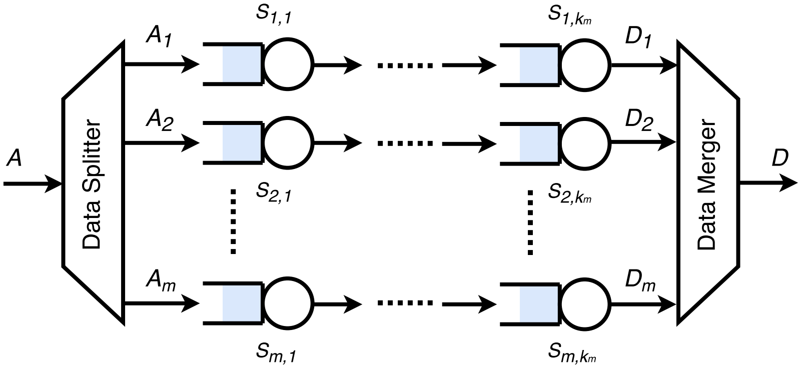


Figure 1.17: Hybrid scheme for low-latency mm-wave communications.

where l, α, γ, and B denote the separation distance, the path-loss exponent, and the transmit power normalized by the background noise, and the bandwidth, respectively. The random variable ξ follows the normalized Gamma distribution, i.e., $\xi \sim \Gamma\left(M, M^{-1}\right)$, with respect to Nakagami parameter M, and its p.d.f. is

$$f\left(x; M, M^{-1}\right) \triangleq \frac{x^{M-1}\exp\left(-Mx\right)}{M^{-M}\Gamma\left(M\right)},$$

where $\Gamma\left(z\right) \triangleq \int_0^\infty z^{t-1}\exp\left(-t\right)dt$ is the gamma function for $\Re\left(z\right) > 0$.

In what follows, regarding the MGF-based characterizations for arrival traffic $A\left(s,t\right)$ and service process $S\left(s,t\right)$, we respectively adopt

$$\mathbb{M}_A\left(\theta, s, t\right) = \exp\left(\theta \cdot \rho \cdot \left(t-s\right)\right) \triangleq \mu^{t-s}\left(\theta\right) \tag{26}$$

and

$$\overline{\mathbb{M}}_S\left(\theta, s, t\right) = \exp\left(-\theta \cdot \sum_{i=s}^{t-1} C^{(i)}\right) \triangleq \mathcal{U}_C^{t-s}\left(\eta\theta\right), \tag{27}$$

with $\mathcal{U}_C\left(x\right)$ for $x > 0$ defined as

$$\mathcal{U}_C\left(x\right) \triangleq \left(\frac{Ml^\alpha}{\gamma}\right)^M \Gamma\left(M\right) \int_0^\infty \exp\left(-\frac{tMl^\alpha}{\gamma}\right) t^{M-1}\left(1+t\right)^{-x} dt. \tag{28}$$

To distinguish the MGFs of service process in different schemes, we use the following distinct notations:

- for traffic dispersion, we use $\overline{\mathbb{M}}_{S_i}\left(\theta, s, t\right) \triangleq \psi_i^{t-s}\left(\theta\right)$ for $i \in [m]$;
- for network densification, we use $\overline{\mathbb{M}}_{S_i}\left(\theta, s, t\right) \triangleq \phi_i^{t-s}\left(\theta\right)$ for $i \in [k]$;
- for the hybrid scheme, we use $\overline{\mathbb{M}}_{S_{i,j}}\left(\theta, s, t\right) \triangleq \varphi_{i,j}^{t-s}\left(\theta\right)$ for $i \in [m]$ and $j \in [k]$.

In Theorem 1.2.7, we give an upper bound for the probabilistic delay of traffic dispersion, where $\mu_i(\theta) \triangleq \exp(\theta\rho_i)$ is defined subject to $\sum_{i=1}^{m} \rho_i = \rho$.

Theorem 1.2.7. *Let $W(t) \triangleq \max\{W_1(t), W_2(t), \cdots, W_m(t)\}$ be the delay for the traffic dispersion scheme with m independent paths, where $W_i(t)$ denotes the delay on the i^{th} path. Then, for any $w \geq 0$, the probabilistic delay is bounded as follows:*

$$\Pr\left(W(t) \geq w\right) \leq 1 - \prod_{i=1}^{m}\left[1 - \inf_{\theta_i > 0}\left\{\frac{\psi_i^w(\theta_i)}{1 - \mu_i(\theta_i)\psi_i(\theta_i)}\right\}\right]^{+},$$

whenever the stability condition $\mu_i(\theta_i)\psi_i(\theta_i) < 1$ holds for some $\theta_i > 0$ and all $i \in [m]$, where $[x]^+ \triangleq \max\{x, 0\}$ for $x \in \mathbb{R}$.

An upper bound for the probabilistic delay of network densification is given in the following Theorem 1.2.8.

Theorem 1.2.8. *Let $W(t)$ be the E2E delay of a k-hop network. Then, for any $w \geq 0$, the probabilistic delay is bounded as follows:*

$$\Pr\left(W(t) \geq w\right) \leq \inf_{\theta > 0}\left\{\mu^{-w}(\theta)\sum_{v=w}^{\infty}\sum_{\sum_{i=1}^{k}\pi_i = v}\prod_{i=1}^{k}\left(\mu(\theta)\phi_i(\theta)\right)^{\pi_i}\right\},$$

whenever the stability condition $\mu(\theta)\phi_i(\theta) < 1$ holds for some $\theta > 0$ and all $i \in [k]$.

In light of Theorem 1.2.7 and Theorem 1.2.8, an upper bound for probabilistic delay of the hybrid scheme is given in the following Theorem 1.2.9.

Theorem 1.2.9. *Assuming m ($m \geq 1$) independent paths for the hybrid scheme system, with k_i hops on the i^{th} path for $1 \leq i \leq m$, we define $\hat{p}_i$ for any given $w \geq 0$ as*

$$\hat{p}_i \triangleq \inf_{\theta_i > 0}\left\{\mu_i^{-w}(\theta_i)\sum_{v=w}^{\infty}\sum_{\sum_{j=1}^{k_i}\pi_j = v}\prod_{j=1}^{k_i}\left(\mu_i(\theta_i)\varphi_{i,j}(\theta_i)\right)^{\pi_j}\right\}.$$

Then, the E2E probabilistic delay is upper bounded as

$$\Pr\left(W(t) \geq w\right) \leq 1 - \prod_{i=1}^{m}\left[1 - \hat{p}_i\right]^{+},$$

whenever the stability condition $\mu_i(\theta_i)\varphi_{i,j}(\theta_i) < 1$ holds for some $\theta_i > 0$ and all $i \in [m]$ and $j \in [k_i]$.

Subsequently, effective capacity is adopted to analyze the performance of the three different scheme, which is defined as

$$\mathcal{C}\left(-\theta\right) \triangleq \lim_{t\to\infty} \frac{\log \overline{\mathbb{M}}_S\left(\theta,0,t\right)}{-\theta t}. \tag{29}$$

We hereafter mainly consider homogeneous settings for three schemes, aiming at obtaining closed-form expressions for fair comparisons without loss of tractability. We denote by L the E2E distance between the source and the destination (in network densification and the hybrid scheme, L is assumed to each independent path). Also, we assume that all transmitters are subject to sum-power constraint γ.

An upper bound for the effective capacity of traffic dispersion is given in Theorem 1.2.10.

Theorem 1.2.10. *Given sum power constraint $\sum_{i=1}^{m}\gamma_i = \gamma$, the effective capacity $\mathcal{C}\left(-\theta\right)$ is upper bounded as*

$$\mathcal{C}\left(-\theta\right) \leq \frac{m\log\left(\left(\frac{MmL^\alpha}{\gamma}\right)^M U\left(M,1+M-\eta\theta,\frac{MmL^\alpha}{\gamma}\right)\right)}{-\theta},$$

where the equality holds if $\gamma_i = m^{-1}\gamma$ for all $i\in[m]$.

For network densification with homogeneity, upper and lower bounds for the effective capacity are derived in the following Theorem 1.2.11.

Theorem 1.2.11. *For homogeneous network densification with k independent hops, given $\theta>0$, the effective capacity is upper bounded as*

$$\mathcal{C}\left(-\theta\right) \leq -\frac{k}{\theta}\log\left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M,1+M-\frac{\eta\theta}{k},\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right),$$

and it is lower bounded as

$$\mathcal{C}\left(-\theta\right) \geq -\frac{1}{\theta}\log\left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M,1+M-\eta\theta,\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right).$$

Combining the results in Theorem 1.2.10 and Theorem 1.2.11, upper and lower bounds for the effective capacity of the hybrid scheme are given in the following Theorem 1.2.12.

Theorem 1.2.12. *For the homogeneous hybrid scheme with m independent paths and $k=n/m$ relay nodes per path, given $\theta>0$, the effective capacity is upper bounded as*

$$\mathcal{C}\left(-\theta\right) \leq -\frac{n}{\theta}\cdot\log\left(\left(\frac{M\left(mL\right)^\alpha}{\gamma n^{\alpha-1}}\right)^M U\left(M,1+M-\frac{\eta\theta}{k},\frac{M\left(mL\right)^\alpha}{\gamma n^{\alpha-1}}\right)\right),$$

and it is lower bounded as

$$\mathcal{C}\left(-\theta\right) \geq -\frac{m}{\theta} \cdot \log \left(\left(\frac{M\left(mL\right)^{\alpha}}{\gamma n^{\alpha-1}} \right)^{M} U\left(M, 1+M-\eta\theta, \frac{M\left(mL\right)^{\alpha}}{\gamma n^{\alpha-1}}\right) \right).$$

Performance Evaluation In performance evaluation, we adopt the homogeneous settings for the sake of fairness and simplicity. The general system configurations are summarized as follows: the bandwidth is allocated with $B = 500$ MHz, the path loss exponent $\alpha = 2.45$, Nakagami-m parameter $M = 3$, and the source-destination distance $L = 1$ km.

Fig. 1.18 shows the violation probabilities of delay for traffic dispersion and network densification. In both Fig. 1.18a and Fig. 1.18b, the numerical results for probabilistic delay bound accurately characterize the slope of the simulated result. Given sum transmit power and arrival rate, we note that the gain by increasing n is more significant to network densification than to traffic dispersion.

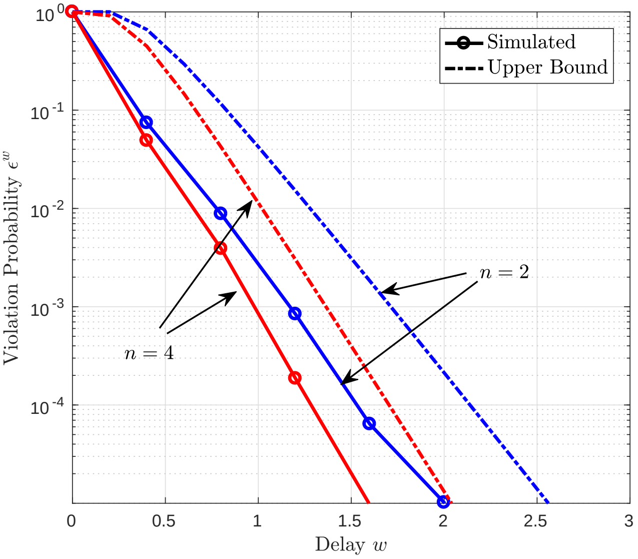
Fig. 1.19 illustrates the effective capacity for traffic dispersion and network densification. The minor gap between the derived lower and upper bounds for network densification indicates the feasibility of using bound to keep track of the actual effective capacity. Traffic dispersion is substantially remarkable when the sum power is high, while network densification performs better for lower sum power. We find that the advantage of traffic dispersion diminishes when the sum transmit power or the number of independent paths decreases.

The effective capacity of the hybrid scheme with $1 \leq m \leq n$ for given $n = 12$ is illustrated in Fig. 1.20, where the actual effective capacity is characterized by lower and upper bounds. With a lower sum power, e.g., $\gamma = 70$ dB, $\mathcal{C}\left(-\theta\right)$ decays as m grows. For higher γ, e.g., $\gamma = 85$ dB, $\mathcal{C}\left(-\theta\right)$ first goes up and falls down afterwards. In this sense, arranging the number of independent paths and the number of hops per path plays an important role in maximizing the effective capacity.

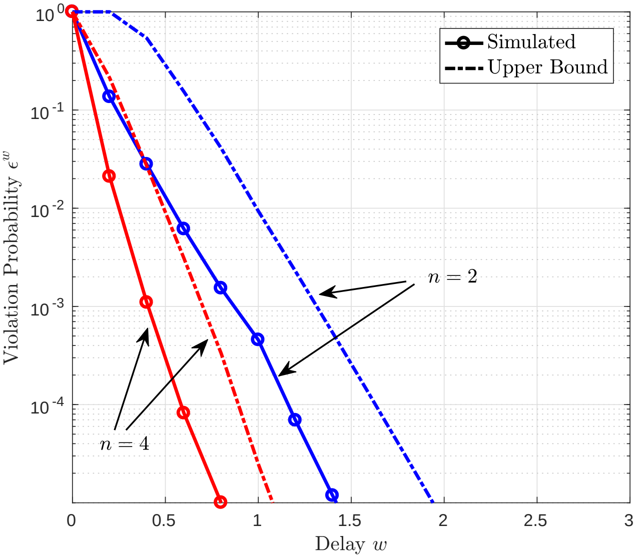
Concluding Remarks We have considered three transmission schemes, namely, traffic dispersion, network densification, and the hybrid scheme, and have investigated their E2E performance in terms of probabilistic delay and effective capacity. Results have demonstrated that:

- Traffic dispersion, network densification, and the hybrid scheme show own strengths, with respect to high, lower and medium sum transmit powers, respectively.

- More independent paths or higher relay density is always advantageous for reducing the E2E communication delay, while the performance gain heavily relies on the arrival rate and the sum power budget, jointly.



(a) traffic dispersion



(b) network densification

Figure 1.18: Violation probability ϵ^w vs. targeted delay bound w for two different schemes, where $\rho = 2$ Gbps and $\gamma = 85$ dB.

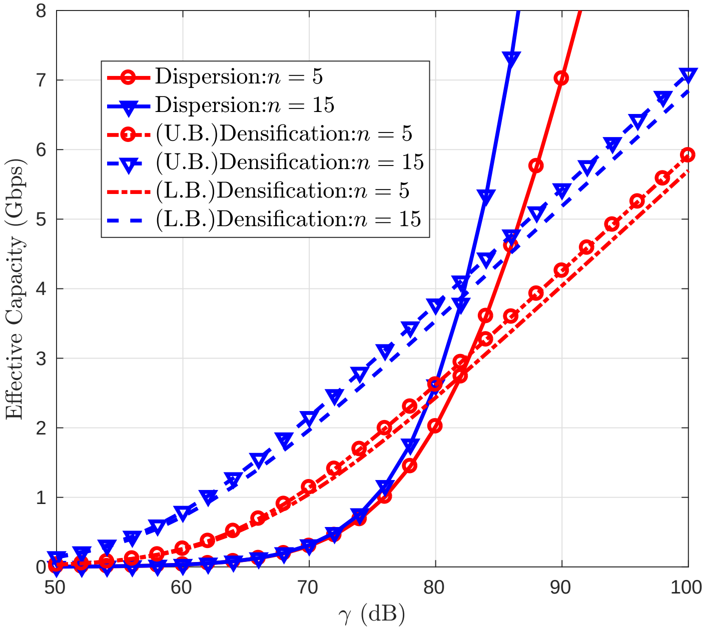


Figure 1.19: Effective Capacity $\mathcal{C}(-\theta)$ vs. sum transmit power γ for two transmission schemes, where QoS exponent $\theta = 2$. Here, U.B. and L.B. stand for "upper bound" and "lower bound", respectively.

1.2.3 Traffic allocation for low-latency mm-wave systems

J5: Low-Latency Heterogeneous Networks with Millimeter-Wave Communications [YXAH18]

In this paper, we have considered a heterogeneous network (HetNet) with mm-wave that consists of one macro-cell evolved NodeB (MeNB), two small-cell evolved NodeB (SeNB)s and one user equipment (UE), where buffers are adopted both at the MeNB and the SeNB. We have discussed several transmission schemes and demonstrated the potential of collaborative networking in reducing the latency.

Backgrounds Many recent efforts have been dedicated to the study regarding latency in 5G communications. [TS15] discussed the technical challenge and possible solution of point-to-multipoint mm-wave backhaul for 5G networks, and [FZM^{+}17] surveyed the main challenges and potential solutions for ultra-low latency 5G cellular networks from the perspective of network architecture, protocols at the MAC layer, and congestion control policy. In [YXG^{+}16], an upper bound on the proba-

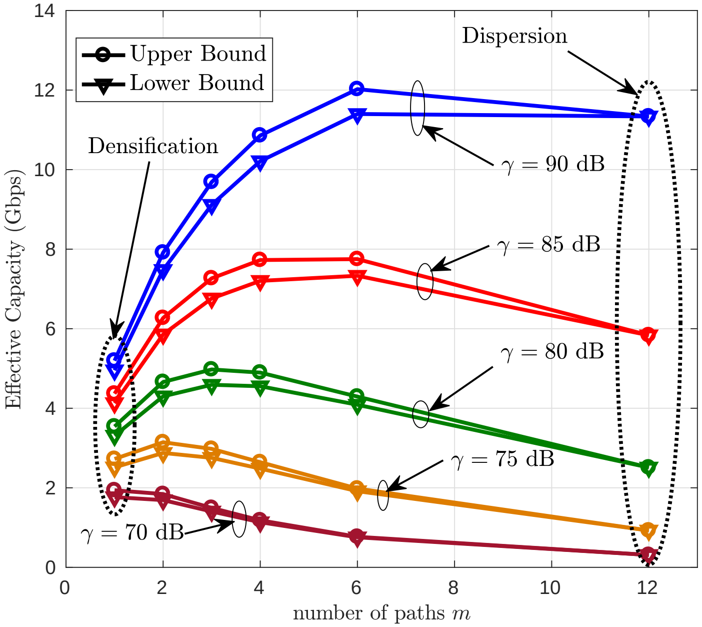


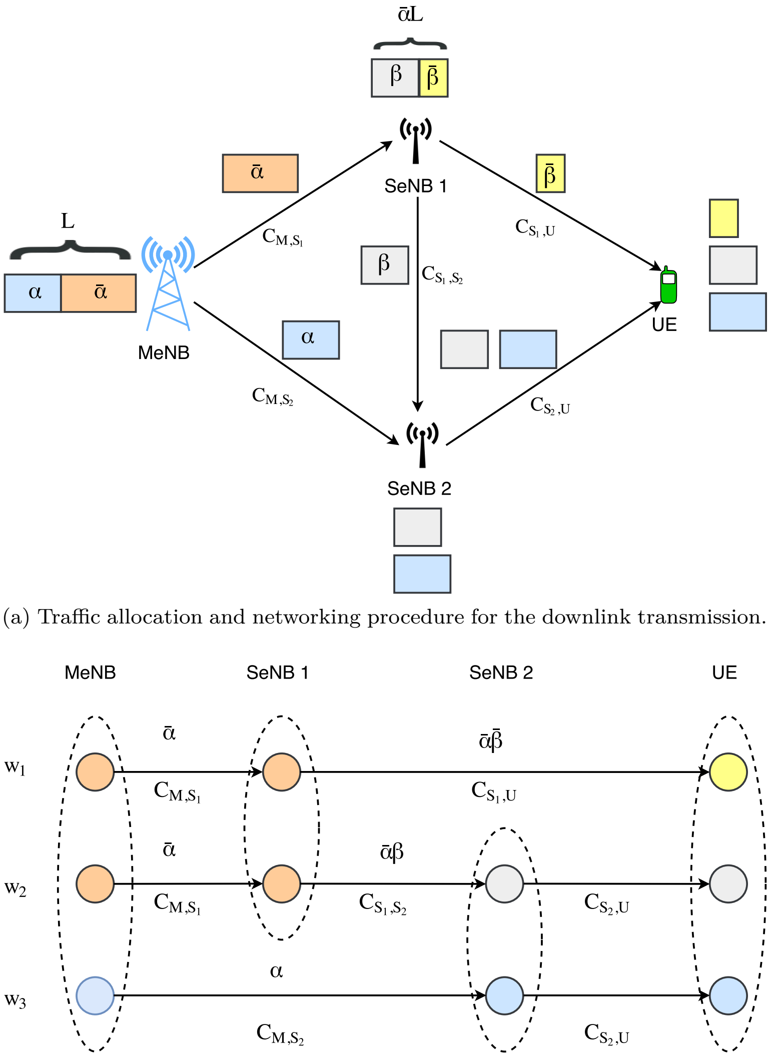
Figure 1.20: Effective capacity $\mathcal{C}(-\theta)$ vs. number of independent paths m for the hybrid scheme with respect to $\theta = 2$ and $n = 12$, where the number of independent paths is $m = 1$, 2, 3, 4, 6 or 12.

bilistic delay was derived to assess the latency performance of point-to-point buffer-aided systems with mm-wave.

However, it is rather challenging to achieve low-latency communications in 5G mm-wave HetNets. The main difficulty stems from the adoption of buffers, since the resulting problem is different from the conventional problems regarding latency minimization. Specifically, in buffer-aided networks, the E2E delay is determined by both link capacities and queuing delays in the buffer, and therefore it cannot be simply formulated as a graph-based network optimization problem [Ber98], e.g., shortest path problem, max-flow problem or min-cost flow problem.

Motivation & Contributions To the best of our knowledge, the study regarding the latency minimization problem of HetNets with buffers is rather limited, and this open problem motivates our study mainly due to its great importance and particular difficulty. Main contributions of our research are outlined as follows:

- We demonstrate the benefits of cooperative networking in mm-wave HetNets with buffers, which is capable of significantly reducing the downlink trans-



(a) Traffic allocation and networking procedure for the downlink transmission.

(b) Abstraction of traffic allocation and networking.

Figure 1.21: Illustration of traffic allocation and networking procedure.

mission latency.

- Proper traffic (re)allocation setups in cooperative networking are critical in minimizing the E2E latency.

Model, Methodology & Analysis Considering one MeNB, two SeNBs and one UE for the mm-wave HetNet as an example, we show the strategy of traffic allocation for cooperative networking in Fig. 1.21.

We denote by α and β the fractions of traffic allocated onto MeNB–SeNB 2 and SeNB 1–SeNB 2 mm-wave backhauls, respectively (here SeNB 2 is taken as the coordinator in Fig. 1.21), and denote by L units the size of file for the downlink transmission from the MeNB to the UE. The traffic allocation is fulfilled via the following two phases:

(i) At MeNB, αL and $\bar{\alpha} L$ units are pushed onto the mm-wave backhauls MeNB–SeNB 2 and MeNB–SeNB 1, respectively.

(ii) At SeNB 1, the received $\bar{\alpha} L$ units are chopped into two parts, i.e., $\bar{\alpha}\beta L$ units and $\bar{\alpha}\bar{\beta} L$ units, for the mm-wave backhaul SeNB 1–SeNB 2 and the mm-wave access SeNB 1–UE, respectively.

Here, $\bar{\alpha} \triangleq 1 - \alpha$ and $\bar{\beta} \triangleq 1 - \beta$. SeNB 2 receives αL and $\bar{\alpha}\beta L$ units from both MeNB and SeNB 1, and buffers them in the queue. The downlink transmission is not completed until all units reach the UE. An abstraction of cooperative networking is depicted in Fig. 1.21b, and the E2E latency is $\max_{1 \leq i \leq 3} w_i$, where w_1, w_2 and w_3 denote component delays on the three traversing paths.

The feasible sets of traffic allocations are denoted by $\mathcal{A} = \{\alpha : 0 \leq \alpha \leq 1\}$ and $\mathcal{B} = \{\beta : 0 \leq \beta \leq 1\}$, respectively. For $i \in \{1, 2\}$, we denote the channel capacity of mm-wave backhaul between the MeNB and SeNB i by C_{M,S_i}, the channel capacity of mm-wave access between SeNB i and the UE by $C_{S_i,U}$, and the channel capacity of the mm-wave backhaul between two SeNBs by C_{S_1,S_2}. Clearly, optimizing α and β is critical for minimizing the E2E latency. With respect to first-in first-out (FIFO) queuing, the earlier arrival will be pushed onto mm-wave access SeNB 2–UE first, and the later one may have to wait in the queue until the earlier comer completely departs from the buffer. Taking into account the arriving order of different file fractions, we then are able to formulate component delays w_2 and w_3. For instance, we have

$$w_2 \leftarrow \frac{\bar{\alpha} L}{C_{M,S_1}} + \frac{\bar{\alpha}\beta L}{C_{S_1,S_2}} + \frac{\bar{\alpha}\beta L}{C_{S_2,U}}$$

and

$$w_3 \leftarrow \max\left\{w_2, \frac{\alpha L}{C_{M,S_2}}\right\} + \frac{\alpha L}{C_{S_2,U}},$$

if SeNB 2 first receives the file fraction on the second path. Otherwise,

$$w_3 \leftarrow \frac{\alpha L}{C_{M,S_2}} + \frac{\alpha L}{C_{S_2,U}}$$

and

$$w_2 \leftarrow \max\left\{w_3, \frac{\bar{\alpha} L}{C_{M,S_1}} + \frac{\bar{\alpha}\beta L}{C_{S_1,S_2}}\right\} + \frac{\bar{\alpha}\beta L}{C_{S_2,U}}.$$

Finally, to achieve $\tau_1^* \leftarrow \min\max\{w_1, w_2, w_3\}$, the MeNB traverses all feasible $(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$ to identify the optimal traffic allocation, i.e., (α_1^*, β_1^*), which enables the minimal E2E latency when treating SeNB 2 as the coordinator.

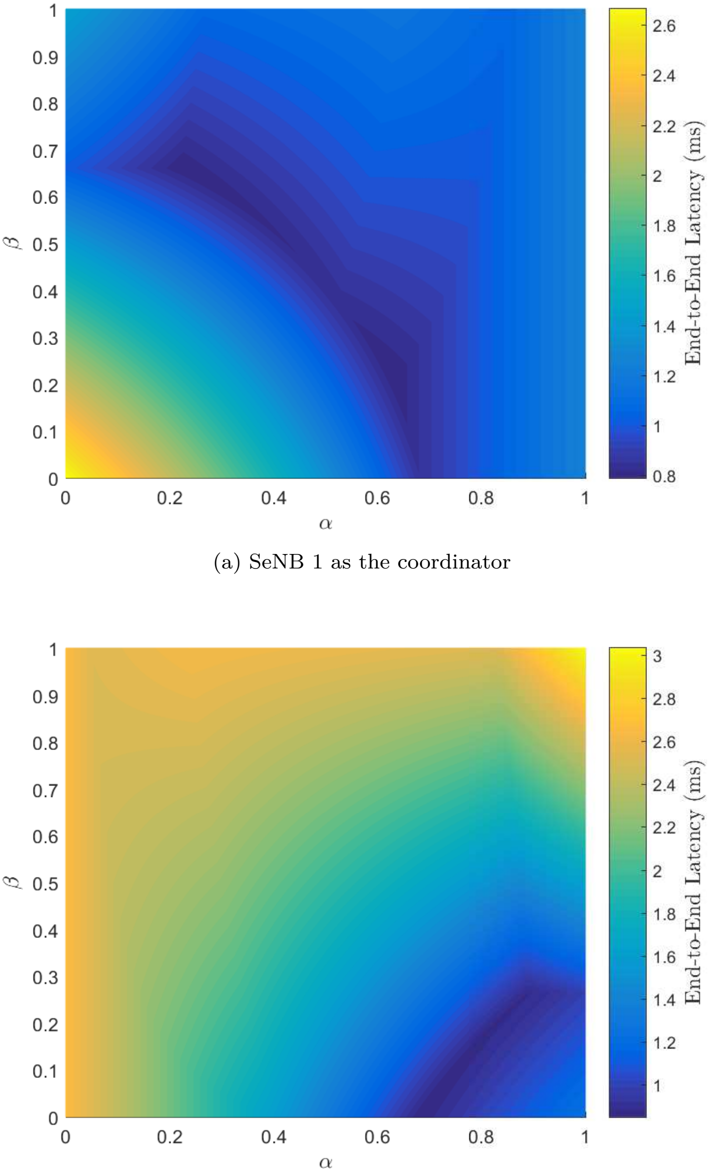
The procedure above is depicted by the following Algorithm 1.

Algorithm 1: Optimal Allocation for Cooperative Networking

```
Input: $C_{M,S_1}$, $C_{M,S_2}$, $C_{S_1,S_2}$, $C_{S_1,U}$, $C_{S_2,U}$, $\mathcal{A}$, $\mathcal{B}$
$\tau^* \leftarrow \infty$
foreach $(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$ do
    $\tau_1 \leftarrow \frac{\bar{\alpha}}{C_{M,S_1}} + \frac{\bar{\alpha}\bar{\beta}}{C_{S_1,U}}$
    if $\frac{\bar{\alpha}}{C_{M,S_1}} + \frac{\bar{\alpha}\beta}{C_{S_1,S_2}} \leq \frac{\alpha}{C_{M,S_2}}$ then
        $\tau_2 \leftarrow \frac{\bar{\alpha}}{C_{M,S_1}} + \frac{\bar{\alpha}\beta}{C_{S_1,S_2}} + \frac{\bar{\alpha}\beta}{C_{S_2,U}}$
        $\tau_3 \leftarrow \max\left\{\tau_2, \frac{\alpha}{C_{M,S_2}}\right\} + \frac{\alpha}{C_{S_2,U}}$
    else
        $\tau_3 \leftarrow \frac{\alpha}{C_{M,S_2}} + \frac{\alpha}{C_{S_2,U}}$
        $\tau_2 \leftarrow \max\left\{\tau_3, \frac{\bar{\alpha}}{C_{M,S_1}} + \frac{\bar{\alpha}\beta}{C_{S_1,S_2}}\right\} + \frac{\bar{\alpha}\beta}{C_{S_2,U}}$
    end
    $\tau \leftarrow \max\{\tau_1, \tau_2, \tau_3\}$
    if $\tau \leq \tau^*$ then
        $\tau^* \leftarrow \tau$
        $(\alpha^*, \beta^*) \leftarrow (\alpha, \beta)$
    end
end
Output: $(\tau^*, \alpha^*, \beta^*)$
```

Performance Evaluation In performance evaluation, we demonstrate the performance of cooperative networking for HetNets with mm-wave communications. For simplicity, deterministic settings are assumed for assessment, i.e., $C_{M,S_1} =$ 12 Gbps, $C_{M,S_2} = 8$ Gbps, $C_{S_1,S_2} = 7$ Gbps, $C_{S_1,U} = 0.8$ Gbps, and $C_{S_2,U} =$ 2 Gbps, and the size of file for the downlink transmission is $L = 2$ Mb.

The E2E latency with respect to different traffic allocations (α, β) is shown in Fig. 1.22, where SeNB 1 and SeNB 2 play as the coordinator in Fig. 1.22a and Fig. 1.22b, respectively. The dark blue regions emerge inside the square $[0, 1] \times [0, 1]$ consisting of all potential α and β, i.e., $\alpha \in \mathcal{A} \setminus \{0, 1\}$ and $\beta \in \mathcal{B} \setminus \{0, 1\}$. Thus, exploiting the mm-wave backhaul between SeNBs (if applicable) is always advantageous with proper traffic allocations, and the resulting E2E latency is significantly lower than those without traffic allocations, i.e., strategies with $\alpha \in \{0, 1\}$ or $\beta \in \{0, 1\}$. Comparing Fig. 1.22a and Fig. 1.22b, the minimum E2E latency is 0.396 ms when taking SeNB 1 the coordinator, while it is 0.426 ms when taking



(a) SeNB 1 as the coordinator

(b) SeNB 2 as the coordinator

Figure 1.22: End-to-end latency with traffic allocation coefficient pairs (α, β): (a) taking SeNB 1 as the coordinator; (b) taking SeNB 2 as the coordinator.

SeNB 2 as the coordinator. Hence, it can be concluded that taking SeNB 1 as the coordinator and performing proper traffic allocations at the MeNB and SeNB 2 enable the minimum latency for downlink transmission.

Concluding Remarks In this paper, we have considered one HetNet consisting of one MeNB, two SeNBs and one UE, and have investigated the low-latency strategy for the downlink transmission from the MeNB to the UE. Results have shown that:

- Cooperative networking is a promising technique to reduce the E2E latency in buffer-aided HetNets.
- It is critical to properly perform traffic (re)allocations with cooperative networking schemes.

J6: Traffic Allocation for Low-Latency Multi-Hop Millimeter-Wave Networks with Buffers [YHX18]

In this paper, we have considered two traffic allocation schemes for multi-hop mm-wave networks with buffers, and have investigated the E2E latency for delivering files. Besides, considering the Nakagami-m fading for mm-wave channel, we have derived lower bounds on the average E2E latency for the mentioned two traffic allocations.

Backgrounds Low latency in 5G mobile communications has attracted massive attention for research in the past few years. The problem of minimizing the average sum queue length under an HD constraint in a two-hop network was investigated in [CLY15]. Several QoS routing problems, i.e., loss rate, average delay, and delay distribution, were discussed in [LH08] for multi-hop networks. A cross-layer framework was established in [CLSC16] via a tuple-based multidimensional conflict graph model, to study the distributed scheduling and delay-aware routing in multi-hop multi-radio multi-channel networks. In [PPT17], a distributed flow allocation scheme was proposed for random access wireless multi-hop networks with multiple disjoint paths, with the objective of maximizing the average aggregate flow throughput and ensuring a bounded packet delay.

In spite of many efforts dedicated to multi-hop networks with buffers, e.g., [LH08, AZLB16, BMP15, JZSS15], the works regarding traffic allocations for low-latency communications with mm-wave are still rather few. In our research [YXP18], the low-latency potential of traffic dispersion and network densification were investigated via using network calculus and effective capacity. However, traffic allocations for reducing the latency were not investigated in [YXP18], and the performance gain via traffic allocation is not yet clear.

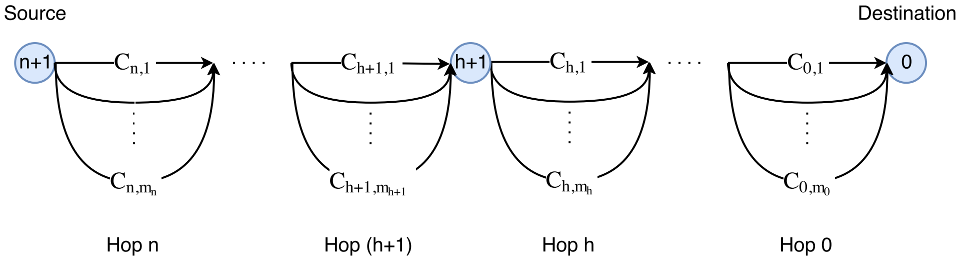


Figure 1.23: Illustration of a multi-hop system, consisting of multiple relay nodes and multiple channels in each hop.

Motivation & Contributions Traffic allocation plays a crucial role in reducing the latency [CY13], since the traffic congestion at the relay nodes may lead to larger queues. However, these graph-based approaches (refer to [Ber98] for example) do not apply to the scenarios with buffers for optimizing the traffic allocation. In addition, the potential of traffic allocation in the presence of buffers has not be investigated previously. It is these two facts that motivate our research.

We aim to develop traffic allocation schemes that enable low-latency mm-wave networks. Main contributions of our work are summarized as follows:

- We formulate two traffic allocation schemes, namely local allocation and global allocation, and discuss their performance with respect to the resulting E2E latency.
- For multi-hop buffered networks with multiple channels in each hop, we exploit the recursive nature of the global allocation scheme, which substantially simplifies the traffic allocation procedure. We also show the overall computational complexity for local and global allocations.
- We conduct the asymptotic analysis to explore the limiting capability for low latency. For mm-wave channels with Nakagami-m fading incorporated, lower bounds of the average E2E latency of two allocation schemes for a two-hop linear network are derived.

Model, Methodology & Analysis A multi-hop network with multiple parallel channels in each hop is considered, as shown in Fig. 1.23, where buffers that apply the FIFO rule are employed at relay nodes. For simplifying illustration, for a $(n+1)$-hop network, the destination and the source are labeled as node 0 and node $n+1$, respectively, and all relays are labeled in orders from 1 to n. The hop between node h and node $h+1$ is denoted by hop h, for all $h \in \{0\} \cup [n]$. Assuming m_h channels in hop h, we denote by $C_{h,k}$ the capacity of the k^{th} channel in hop h for all $k \in [m_h]$.

The traffic allocation fulfilled at the source and relays is illustrated as follows:

(i) The incoming traffic at one node is first partitioned into several smaller fractions according to the given allocation scheme.

(ii) Those fractions are subsequently pushed onto the channels and delivered to the next node, where each fraction can be partitioned again for further delivery.

We define $\underline{\alpha}_h \triangleq [\alpha_{h,1}, \ldots, \alpha_{h,m_h}] \in \mathbb{R}_+^{m_h}$ the traffic allocation with respect to channels in hop h for all $h \in \{0\} \cup [n]$, subject to $\|\underline{\alpha}_h\|_1 \triangleq \sum_{i=1}^{m_h} \alpha_{h,i} = 1$. The traffic allocation $\underline{\alpha}_h$ depends on the capacities of outgoing channels, and the traffic allocation $\underline{\alpha}_h$ is performed at node $h+1$. Relays are FD but only equipped with one single buffer, such that the data reception and transmission can be performed at the same time and the received fractions depart from the buffer one by one. At each relay node, one fraction is not served (chopped into smaller fractions and forwarded to the next node) until it is completely received, and any fraction is infinitely divisible.

In this work, we consider two traffic allocation schemes, namely, local traffic allocation $\mathcal{M}_{\text{local}}$ and global traffic allocation $\mathcal{M}_{\text{global}}$, which are described as below:

- $\mathcal{M}_{\text{local}}$: Node h for all $h \in [n+1]$ only has the capacity information of the channels in hop $h-1$. The traffic allocation performed at node h only optimizes the transmission over channels in hop $h-1$. This scheme ensures that the latency in the local hop is minimized, but is oblivious to the traffic allocations in other hops.

- $\mathcal{M}_{\text{global}}$: Node h for all $h \in [n+1]$ has the entire capacity information of all channels from hop 0 to hop $h-1$. The traffic allocation performed at node h not only relies on channels in hop $h-1$, but also relies on channels in the remaining hops, i.e., from hop 0 to hop $h-1$. This scheme minimizes the latency through h hops.

We denote by τ_n the E2E latency of the tandem queuing system with n relays for delivering one file of normalized size (without loss of generality), where the delivery is completed only when all fractions are received at the destination. For networks with buffers incorporated, it is worth noting that there is a common formulation of E2E latency for two schemes, while distinct traffic allocation schemes result in different latencies.

Based on mechanisms of $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ described above, we present the results regarding the minimum E2E latency in the following Theorem 1.2.13 and Theorem 1.2.14, respectively.

Theorem 1.2.13. *For the tandem network in Fig. 1.23 with n relay nodes and m_h channels in the h^{th} hop, the minimum E2E latency with $\mathcal{M}_{\text{local}}$ is*

$$\tau_n^* = \sum_{h=0}^{n} \left(\sum_{k=1}^{m_h} C_{h,k} \right)^{-1}, \tag{30}$$

achieved by $\alpha_{h,k} = C_{h,k}\left(\sum_{k=1}^{m_h} C_{h,k}\right)^{-1}$.

Theorem 1.2.14. *For the tandem network in Fig. 1.23 with* n *relay nodes and* m_h *channels in the* h^{th} *hop, the minimum E2E latency with* $\mathcal{M}_{\text{global}}$ *is*

$$\tau_n^* = \frac{\left(C_{n,m_n}^{-1} + \tau_{n-1}^*\right)\prod\limits_{k=1}^{m_n-1} C_{n,k+1}\left(C_{n,k}^{-1} + \tau_{n-1}^*\right)}{1 + \sum\limits_{i=2}^{m_n}\prod\limits_{k=1}^{i-1} C_{n,k+1}\left(C_{n,k}^{-1} + \tau_{n-1}^*\right)}, \tag{31}$$

with initial condition $\tau_0^* \triangleq \left(\sum_{i=1}^{m_0} C_{0,i}\right)^{-1}$, *achieved by*

$$\alpha_{h,k} = \begin{cases} C_{h,k}\left(\sum\limits_{k=1}^{m_h} C_{h,k}\right)^{-1}, & h = 0 \\ \dfrac{\prod\limits_{i=1}^{k-1} C_{h,i+1}\left(C_{h,i}^{-1} + \tau_{h-1}^*\right)}{1 + \sum\limits_{j=2}^{m_h}\prod\limits_{i=1}^{j-1} C_{h,i+1}\left(C_{h,i}^{-1} + \tau_{h-1}^*\right)}, & h \geq 1. \end{cases} \tag{32}$$

It is worth mentioning that $\mathcal{M}_{\text{global}}$ outperforms $\mathcal{M}_{\text{local}}$ at the expense of higher computational complexity. Specifically, given n and m, we denote by $f_{\text{local}}(m,n)$ and $f_{\text{local}}(m,n)$ the numbers of channels in total for conducting $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, respectively. We have

$$f_{\text{local}}(m,n) = m(n+1) \quad \text{and} \quad f_{\text{global}}(m,n) = \frac{m(n+1)(n+2)}{2}. \tag{33}$$

Comparing $\mathcal{M}_{\text{global}}$ to $\mathcal{M}_{\text{global}}$, we find

$$\frac{f_{\text{global}}(m,n)}{f_{\text{local}}(m,n)} = \frac{n+2}{2} \in \mathcal{O}(n), \tag{34}$$

which grows with the number of relay nodes, linearly, indicating an n times higher computational complexity.

Subsequently, we focus on the performance gain by adopting $\mathcal{M}_{\text{global}}$ relative to $\mathcal{M}_{\text{local}}$. For fairness and convenience, we assume $m_h = m$ and $C_{h,k} = C$ for all $h \in \{0\} \cup [n]$ and $k \in [m]$. Given n relay nodes and m channels per hop, we define the performance gain by $\mathcal{M}_{\text{global}}$ relative to $\mathcal{M}_{\text{global}}$ as

$$\rho(n,m) \triangleq \frac{\tau_n^*|_{\text{local}}}{\tau_n^*|_{\text{global}}}, \tag{35}$$

where $\tau_n^*|_{\text{global}}$ and $\tau_n^*|_{\text{local}}$ are the minimum E2E latencies from Theorem 1.2.14 and Theorem 1.2.13, respectively.

A recursive method for computing the relative performance gain $\rho(n,m)$ is given in the following Theorem 1.2.15.

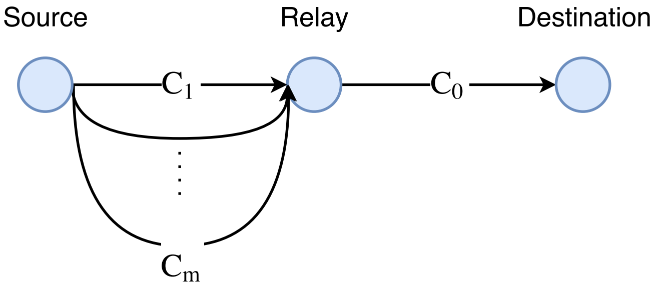


Figure 1.24: Illustration of a two-hop network, where multiple channels are between the source and the relay, and one channel between the relay to the destination.

Theorem 1.2.15. *Given n relay nodes and m channels in each hop, the relative performance gain $\rho(n,m)$ is*

$$\rho(n,m)=\frac{n+1}{m}\left(u_n^*\right)^{-1}, \tag{36}$$

where u_k^* for all $k\in[n]$ is given as $u_k^*=\left(1-\left(1+u_{k-1}^*\right)^{-m}\right)^{-1}u_{k-1}^*$, with initial condition $u_0^*=m^{-1}$.

In Theorem 1.2.16, an asymptotic analysis on the relative performance gain is conducted, i.e., infinite channels per hop. Specifically, with any given number of relay nodes n, the asymptotic relative performance gain $\bar{\rho}(n)$ is defined as

$$\bar{\rho}(n)\triangleq\lim_{m\to\infty}\rho(n,m). \tag{37}$$

Theorem 1.2.16. *Given n relay nodes, the asymptotic performance gain $\bar{\rho}(n)$ for any $n\geq 0$ is recursively given as*

$$\bar{\rho}(n)=\frac{n+1}{n}\left(1-\exp\left(-\frac{n}{\bar{\rho}(n-1)}\right)\right)\bar{\rho}(n-1), \tag{38}$$

with initial condition $\bar{\rho}(0)=1$.

In what follows, we investigate the latency performance of buffered networks with mm-wave. For mm-wave channels with Nakagami-m fading, the normalized capacity is given as $C=\log_2(1+g\cdot\xi)$, where ξ is the SNR and the random variable g represents the channel power gain following the Gamma distribution, i.e., $g\sim\Gamma\left(M,M^{-1}\right)$ with a positive Nakagami parameter M. For the sake of simplicity and tractability, we here particularly consider a two-hop network as shown in Fig. 1.24. Assuming that mm-wave channels C_i, $i\in\{0,1,2\}$, have the identical Nakagami parameter[4] M, we have

$$\begin{aligned}\mathbb{E}\left[C_i\right] =& \int_0^{\infty} \log_2\left(1 + x\xi_i\right) f\left(x; M\right) dx \\ =& \frac{M^M}{\xi_i^M \Gamma\left(M\right) \ln\left(2\right)} \cdot G_{2,3}^{3,1}\left(\begin{matrix} -M, 1-M \\ 0, -M, -M \end{matrix} \middle| \frac{M}{\xi_i} \right),\end{aligned} \tag{39}$$

where $G_{p,q}^{m,n}\left(\begin{smallmatrix} a_1, a_2, \ldots, a_p \\ b_1, b_2, \ldots, b_q \end{smallmatrix} \middle| z \right)$ denotes the Meijer G-function [Olv10] for $0 \leq m \leq q$ and $0 \leq n \leq p$, and parameters a_j, b_j and $z \in \mathbb{C}$.

In terms of $\mathbb{E}\left[C_i\right]$ given above, lower bounds on the average minimum E2E latency for $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ are given in the following Proposition 1.2.5 and Proposition 1.2.6, respectively.

Proposition 1.2.5. *A lower bound on the average minimum E2E latency for* $\mathcal{M}_{\text{local}}$ *is*

$$\mathbb{E}\left[\tau_2^*\right] \geq \left(\mathbb{E}\left[C_0\right]\right)^{-1} + \left(\mathbb{E}\left[C_1 + C_2\right]\right)^{-1}. \tag{40}$$

Proposition 1.2.6. *A lower bound on the average minimum E2E latency for* $\mathcal{M}_{\text{global}}$ *is*

$$\mathbb{E}\left[\tau_2^*\right] \geq \left(\mathbb{E}\left[C_0\right]\right)^{-1} + \left(\mathbb{E}\left[C_1 + C_2\right] + \epsilon_0 \mathbb{E}\left[C_1 C_2\right]\right)^{-1}, \tag{41}$$

where ϵ_0 *for* $M > 1$ *is defined as*

$$\epsilon_0 \triangleq \ln\left(2\right)\left(\frac{1}{2} + \xi_0^{-1}\left(1 + \left(M-1\right)^{-1}\right)\right). \tag{42}$$

Performance Evaluation The minimum E2E latency τ^* against the number of relay nodes n is illustrated in Fig. 1.25. τ^* can be significantly reduced when increasing m for both $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$. Compared to $\mathcal{M}_{\text{local}}$, the advantage of $\mathcal{M}_{\text{global}}$ is growingly significant as n increases. We also find that the curve with $m = 5$ for $\mathcal{M}_{\text{local}}$ coincides with the curve with $m = 10$ for $\mathcal{M}_{\text{global}}$, at $n = 6$, indicating that $\mathcal{M}_{\text{global}}$ with fewer channels is still competitive in outperforming $\mathcal{M}_{\text{local}}$ with more channels, as long as there are sufficiently many relay nodes.

Given n relays, the minimum E2E latency τ^* against the number of channels m is shown in Fig. 1.26. For both $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, the decaying rate of τ^* slows down when m becomes large, indicating that the merits of having multiple channels vanish as the number of channels increases.

The relative performance gain $\rho\left(n, m\right)$ with respect to the number of relay nodes n is shown in Fig. 1.27. $\mathcal{M}_{\text{local}}$ is equivalent to $\mathcal{M}_{\text{global}}$ for $m = 1$, i.e., $\rho\left(n, 1\right) = 1$ for all n, since only one single channel per hop exists. However, $\rho\left(n, m\right)$ substantially grows when increasing m or n.

[4] Normally, the randomness in mm-wave channels is relatively weak, such that the Nakagami parameter $M \geq 3$ as in [BH15, YZHL17].

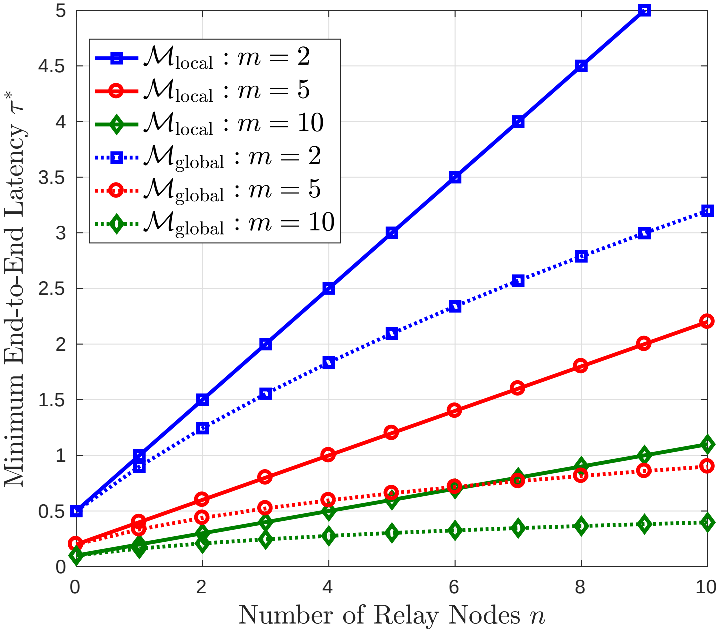


Figure 1.25: Minimum E2E latency τ^* vs. number of relay nodes n, where the number of channels per hop is $m = 2$, 5 or 10, the capacity of each channel is $C = 1$, and the size of transmitted file is 1.

Then we investigate the average E2E latency of a linear two-hop mm-wave network. We denote by r and L the (normalized) distances for source-relay and source-destination, respectively. Applying the path loss model for LoS mm-wave communications [RSM+13, XMH+17], the SNR ξ_i in $C_i = \log_2(1 + g_i\xi_i)$ can be written as

$$\xi_i = \begin{cases} \gamma (L - r)^{-\alpha}, & i = 0 \\ \gamma r^{-\alpha}, & i \in \{1, 2\}, \end{cases} \tag{43}$$

where the Gamma-distributed random variables g_i for $i \in \{0, 1, 2\}$ are independent and identically distributed with Nakagami parameter M. The average E2E latency and the lower bounds for mm-wave networks are illustrated in Fig. 1.28. The lower bounds given in Proposition 1.2.5 and Proposition 1.2.6 are tight. The minimum τ^* is achieved only if a proper r is used, thereby indicating the critical role of relay deployment in minimizing the E2E latency. We also find that the values of r to achieve the minimum τ^* are different between $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$. Thus, different relay deployments may be needed for different allocation schemes.

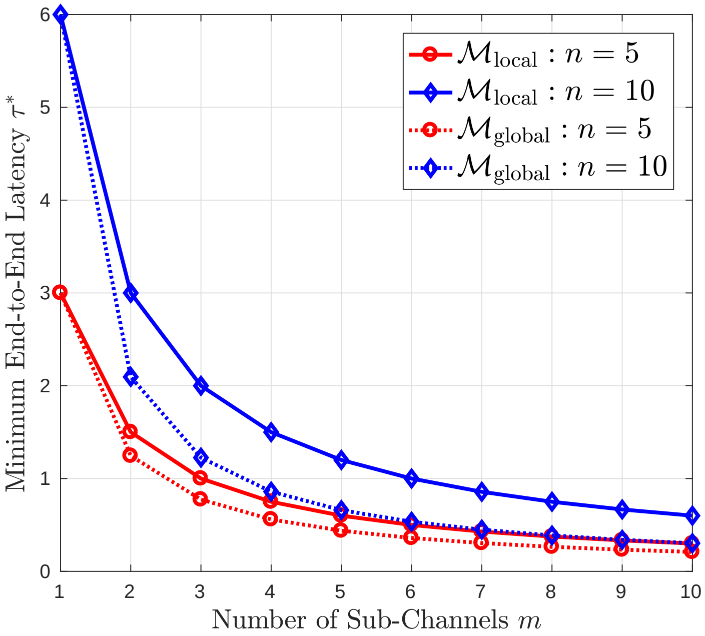


Figure 1.26: Minimum E2E latency τ^* vs. number of channels m, where the number of relay nodes is $n = 5$ or 10, the capacity of each channel is $C = 1$, and the size of transmitted file is 1.

Concluding Remarks In this paper, we have studied the E2E latency in multi-hop networks, where local allocation and global allocation are considered. We have presented the resulting E2E latency and computational complexity of different traffic allocation schemes, and also have derived lower bounds for a two-hop mm-wave network with Nakagami-m fading. Results have shown that:

- Compared to $\mathcal{M}_{\text{local}}$, $\mathcal{M}_{\text{global}}$ achieves a gain in latency reduction that grows as the number of relay nodes n increases, but pays for n times higher computational complexity.

- More parallel channels enable a lower latency, while the benefit diminishes as the number of channels increases. The asymptotic performance gain only depends on the number of relays.

- A proper deployment of relay nodes, which may be different for distinct traffic allocation schemes, plays a critical role in minimizing the E2E latency.

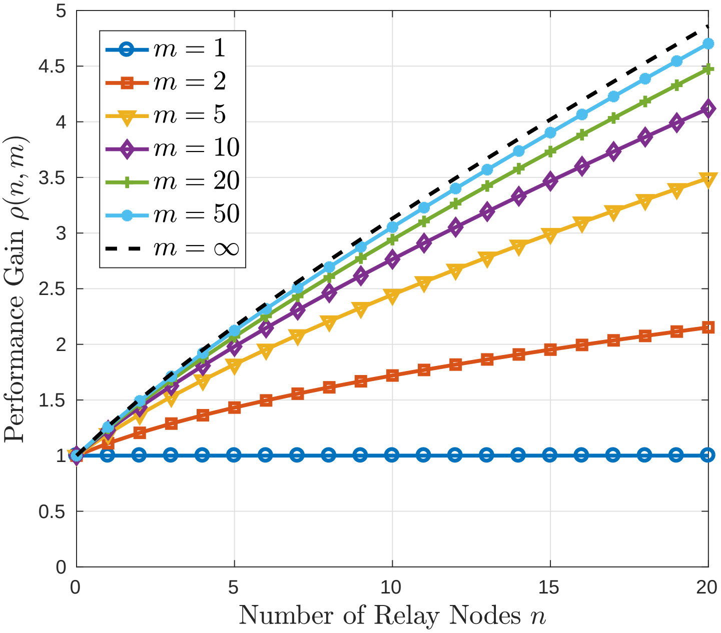


Figure 1.27: Relative performance gain $\rho(n, m)$ vs. number of relay nodes n, where the number of channels per hop is $m = 1$, 2, 5, 10, 20, 50 or ∞.

1.3 Summary

Regarding throughput and latency in future mobile communications, this dissertation aims to investigate the performance of mm-wave networks and to explore the key principles potentially for network designs in practice. Several general but key conclusions drawn from our research are summarized as below:

Throughput of mm-wave relay networks:

- Multiple relays associated with proper routing algorithms can effectively combat severe throughput degradation resulted by random blockages.
- For relaying with mm-wave, FD outperforms HD only if both the beamwidth and the self-interference coefficient are small.
- For scenarios with high sum-power budget or narrow beams, it is better to adopt the direct transmission for higher throughput.
- First-order reflections may constructively or destructively contribute to the throughput, and properly using reflections is viable for preserving the link connectivity.

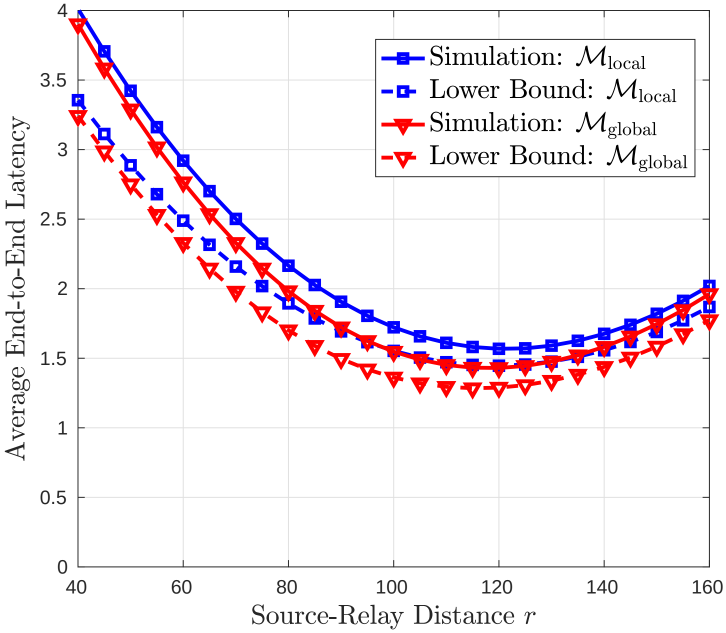


Figure 1.28: Average E2E latency $\mathbb{E}\left[\tau_2^*\right]$ and the lower bounds, against source-relay distance r, where (normalized) source-destination distance $L = 200$, transmit power $\gamma = 60$ dB, path loss exponent $\alpha = 3$, and Nakagami parameter $M = 5$.

Latency analysis via stochastic network calculus:

- The derived analytic bounds are asymptotically tight, indicating the feasibility of keeping the track of probabilistic performance guarantees for diverse mm-wave networks via stochastic network calculus.
- In the presence of sum-power budget, optimized power allocation and small self-interference are critical. Also, the relay density should be properly designed if self-interference is not negligible.
- Traffic dispersion, network densification, and the hybrid scheme are advantageous with respect to high, lower and medium sum-power budgets, respectively.
- Without self-interference, it is always beneficial to have more independent paths or higher relay density, while the arrival rate and the sum-power budget jointly decide the performance gain.

Traffic allocation for low-latency mm-wave systems:

- Cooperative networking can be adopted for reducing the E2E latency in buffer-aided HetNets, where proper traffic allocations are needed.
- For n-hop networks with multiple parallel channels per hop, global traffic allocation achieves a gain in latency reduction that grows as n compared to its local counterpart, but pays for n times higher complexity.
- When increasing parallel channels in each hop, the achieved benefit diminishes and converges to a gain that only relies on the number of relays.
- It is important to properly deploy relay nodes, and optimized relay deployments are different for distinct traffic allocation schemes.

Part II

Included Papers

Paper J1

Maximum Throughput Path Selection with Random Blockage for Indoor 60 GHz Relay Networks

Guang Yang, Jinfeng Du, and Ming Xiao

Refereed article published in
IEEE Transactions on Communications, vol. 63, no. 10, Oct. 2015.

Maximum Throughput Path Selection with Random Blockage for Indoor 60 GHz Relay Networks

Guang Yang, Jinfeng Du, and Ming Xiao

Abstract

Indoor communications in the 60 GHz band is capable to support multi-gigabit wireless access thanks to the abundant spectrum and the possibility of using dense antenna arrays. However, the high directivity and penetration loss make it vulnerable to blockage events which can be frequent in indoor environments. Given network topology information in sufficient precision, we investigate the average throughput and outage probability when the connection between any two nodes can be established either via the line-of-sight *(LOS) link, through a reflection link, or by a half-duplex relay node. We model the reflection link as an LOS with extra power loss and derive the closed-form expression for the relative reflection loss. For networks with a central coordinator and multiple relays, we also propose a generic algorithm,* maximum throughput path selection *(MTPS), to select the optimal path that maximizes the throughput. The complexity of the MTPS algorithm is $\mathcal{O}(n^2)$ for networks equipped with n relays, whereas a brute-forced algorithm has complexity of $\mathcal{O}(n \cdot n!)$. Numerical results show that increasing the number of relays can significantly increase the average throughput and decrease the outage probability, and resorting to reflection paths provides significant gains when the probability of link blockage is high.*

A Introduction

Wireless communications in the 60 GHz band can provide multi-gigabit short-range wireless access in indoor environments, such as the Wireless Gigabit Alliance (WiGig) technology [WiG10], thanks to the abundant spectrum and high power gain antenna arrays. The high directional antennas and high penetration loss can greatly reduce the power of interfering signals but at the same time make it vulnerable to blockage events. Connectivity and throughput can be seriously impaired by blockage owing to the extremely weak capability of penetration and diffraction for 60 GHz radio waves [MC04, Smu09, MC06, GKZV09, WAN97], which is due to the physical property of radios with very short wavelength. As such blockage events may happen frequently due to object mobility (e.g., the movement of a human body), it is challenging to provide multi-gigabit throughput while reducing the outage probability to a sufficiently low level.

Numerous efforts devoted to 60 GHz communications focus on various approaches to tackle the problem of blockage. We only provide a small sample of previous contributions here and readers are kindly referred to [DH07, PR07, GQMT07] for more comprehensive reviews on 60 GHz communications systems.

A large amount of experiments have been performed in [JPM$^+$11, GT12, CZZ04, JPK$^+$13, MEP$^+$10] to characterize the property of indoor blockage events and their effect on signal strength by obstacle physical properties, such as size, shape, position, placement and density. The link blockage caused by random human activities has been analysed in [DLZ12] for typical indoor environments and in [SZM$^+$09] in the context of 60 GHz wireless personal area networks (WPANs). One way to tackle the problem of link blockage and improve the robustness of indoor networks is to deploy relays. A pyramid relaying system proposed in [LLNS04] shows superior coverage and capacity under various human shadowing densities. Significant reduction of the path loss demonstrated in [GOON10] reveals the importance of relay placement in improving the connectivity of 60 GHz indoor communications. The outage performance of decode-and-forward (DF) and amplify-and-forward (AF) relaying in [LPCF12] is analyzed by taking co-channel interference into account. On the other hand, the feasibility of data transmission via the first-order reflected radio waves has been verified in [GKZV09, GRON10] where no *line-of-sight* (LOS) path exists and the path loss is high. Two beam switching strategies proposed in [ASP$^+$09] switch the beam path from LOS link to a *non-line-of-sight* (NLOS) link to resolve the link blockage. It is further demonstrated in [SRH$^+$14] that beamforming and beam combining techniques can be beneficial to significantly increase the SNR for the NLOS transmissions. Since reflection comes with almost no extra cost, it is valuable to evaluate and quantify its benefit as a complementary component of actively deployed relay nodes.

For multi-hop LOS transmissions, an optimal geographic routing protocol is proposed in [CHS$^+$09] and a link scheduling scheme for 60 GHz multi-channel wireless mesh networks is investigated in [SZ09]. A routing algorithm is developed in [LSW$^+$09] to find the optimal relay path with the least interference to maximize the throughput. With high directional antennas, a multi-hop LOS relaying protocol proposed in [SZM$^+$09] achieves high network utilization with low overhead despite high link blockage probability in a 60 GHz indoor WPAN. The robustness of routing schemes is discussed in [SLC11] and neighbor discovery protocols are proposed in [Fan08] to improve the signal quality and maintain the connectivity. A randomized exclusive region based scheduling scheme is proposed in [CCSM10] to explore the potentials of spatial reuse, and the benefit of multi-hop concurrent transmissions is investigated in [QCSM11] for networks with linearly deployed nodes and no link blockage.

In this paper, we consider an indoor 60 GHz relay network where the connection between any two user nodes can be established either via the LOS link, through a reflection link, or by a half-duplex relay node. We focus on scenarios where the network has a central coordinator and multiple half-duplex relay nodes deployed as a fixed infrastructure. We assume all devices (including relays and user nodes) are equipped with directional antennas and their topology can be measured with sufficient precision [PN02, HWGL09]. Given the topology information of distance and direction, we investigate how such topology information can be used to improve the system performance. Because the small-scale multipath effect in 60 GHz communications is negligible [WAN97, GKZV09], the data rate supported by a specific link is determined by the transmit power, antenna gains, pass loss exponent and the transmission distance. By modeling the reflection link as an LOS transmission with extra power loss, we can calculate the throughput of each of the three options, namely, the LOS link, the reflection link, and the relay link. We investigate the average throughput and the outage probability under two random blockage models where the probability of blockage is identical for all links (topology independent) or is proportional to the length of the link (topology dependent). We propose the *maximum throughput path selection* (MTPS) algorithm with low complexity for multi-relay scenarios to select the best path that consists of one or more hops to maximize the throughput when the knowledge of blockage events is available.

To highlight our objective, we do not consider the combining of signals from different propagation paths. Although the extension of our analysis to scenarios with multiple propagation paths can be straightforward via approaches such as beam combining [SRH$^+$14], the extension based on radio wave broadcast via a single beam is highly non-trivial due to the high directivity of 60 GHz antenna arrays. Besides, we do not take into account the interference from concurrent transmissions when multiple relays are selected to assist transmission for the following two reasons. Firstly, the communicating nodes (users and relays) in our network are deployed inside a hall in contrast to the linear deployment in [CCSM10, QCSM11], and therefore the probability that concurrent transmissions fall into the boresight scenario at the same receiving node at the same time (hence causing severe interference) is very small given the high directivity of antenna arrays and the random positions of the communicating nodes. Secondly, the effect of interference from concurrent transmissions can be partially modeled by link blockage events since severely degraded link quality can be treated as link blockage.

Our study differs from the existing results from the following aspects:

- The routing schemes in [CHS$^+$09, SZ09, LSW$^+$09, CCSM10, QCSM11] do not consider the link blockage, while the main objective of our study is to develop a routing algorithm for the network with link blockage;

- In [LLNS04, GOON10, LPCF12] the performance is studied for 60 GHz relaying channels. Yet, they only consider fixed relaying networks with 2 hops and do not optimize relaying paths. We focus on generic routing schemes including hop selection and reflection utilization such that the throughput can be maximized in multi-hop networks;

- Although the approaches on reflection signals in [DLZ12, SZM$^+$09, GRON10, ASP$^+$09, SLC11, Fan08] are viable to maintain the link connectivity when the LOS channel is blocked, they are based on measured results. There is no closed-form analysis on 60 GHz reflection signals yet;

- For the multi-hop concurrent transmission (MHCT) proposed in [CCSM10, QCSM11], the application scenarios are limited to nodes deployed in one line. In our framework nodes can be arbitrarily placed. Moreover, MHCT in [CCSM10, QCSM11] does not consider the link blockage problem, whereas our schemes are mainly proposed to address the problem of link blockage.

The rest of the paper is organized as follows. We present the system model in Sec. B along with preliminaries on the reflection loss and relaying strategy. We investigate the average throughput and outage probability based on two random blockage models in Sec. C, and present the MTPS algorithm in Sec. D. Our models and analysis are validated by numerical results in Sec. E and conclusions are given in Sec. F.

B System Model and Adaptive Relaying Scheme

B.1 Link Model and Network Architecture

Given the transmit power P_t, the transmitter antenna gain G_t, and the receiver antenna gain G_r, the power of the received signals can be determined as follows

$$P_\mathrm{r}(l) = P_\mathrm{t} G_\mathrm{t} G_\mathrm{r} \left(\frac{\lambda}{4\pi}\right)^2 \left(\frac{1}{l}\right)^n, \tag{1}$$

where λ is wavelength, l is the transmission distance, and n is the path loss exponent which ranges from 2 to 6 from 60 GHz measurement in [GKZV09]. Note that (1) is a modified version of the standard Friis free space transmission equation in which the path loss exponent is $n = 2$. However, in practical situations the path loss exponent n in the empirical and deterministic model (1) can be higher due to shadowing and oxygen absorption. Given system bandwidth W and one-side power spectral density of white Gaussian noise N_0, the achievable[1] rate is given by

$$R(l) = W \log_2(1 + \alpha l^{-n}), \tag{2}$$

where $\alpha = \frac{P_\mathrm{t} G_\mathrm{t} G_\mathrm{r} \lambda^2}{16\pi^2 N_0 W}$ represents the signal-to-noise ratio (SNR) measured at $l = 1$ meter distance. To simplify notations, we assume that all transmission links have the same α, and extensions to general setups are straightforward.

We consider a 60 GHz wireless relay network where a pair of communication nodes N_1 and N_2 are randomly placed within a circular space, as shown in Fig. 1.1, with the center C and radius R_0. One or more half-duplex relay nodes are deployed inside this area to assist transmission. Supposing N_1 and N_2 are uniformly

[1] Strictly speaking, the rate in (2) is actually the capacity, i.e., the theoretical upper bound on data rate that may be achieved asymptotically.

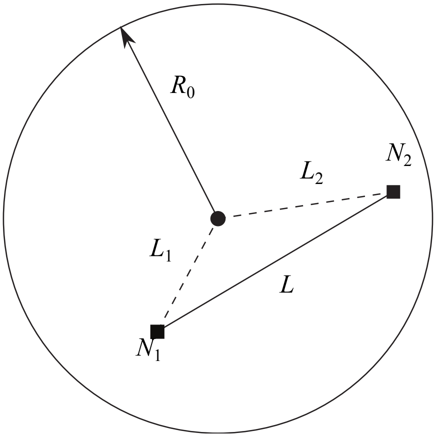


Figure 1.1: Randomly distributed nodes in a circular space, where the relay node is placed at the center and two communicating user nodes N_1 and N_2 are randomly located in the hall with radius R_0.

distributed in the circular area, and their distances to the center C, denoted by L_1 and L_2, respectively, are random variables with a probability density function (p.d.f.)

$$f_{L_i}(l_i) = \frac{2l_i}{R_0^2}, \quad \text{where } 0 < l_i < R_0 \text{ and } i = 1, 2. \tag{3}$$

The distance between N_1 and N_2, denoted as L, follows the p.d.f. [BDF63]:

$$f_L(l) = \frac{2l}{R_0^2}\left[1 - \frac{2}{\pi}\arcsin(\frac{l}{2R_0}) - \frac{l}{\pi R_0}\sqrt{1 - \frac{l^2}{4R_0^2}}\right], \tag{4}$$

where $0 < l < 2R_0$.

The LOS link between any two nodes may be blocked by an obstacle (e.g., human body) in indoor scenarios. We further assume that the reflection path between any two nodes is always available [MEP$^+$10], since such reflection may happen via the floor, the ceiling, or walls[2].

We further assume that the network has a central coordinator installed to take care of the synchronization and management of the network information. It collects and updates necessary information, manages the path selection process, and coordinates the communications according to an optimized scheduling. Such central coordinator can be realized, for example, in the form of a piconet coordinator for the high rate WPAN, or by the access point that connects the indoor network with other networks. We will discuss the operation of the central coordinator, its status update process, and the associated overhead in Sec. D.

[2] Given a single beam at the transmitter and the receiver, there is only one feasible reflection path due to the high directivity of 60 GHz antenna arrays.

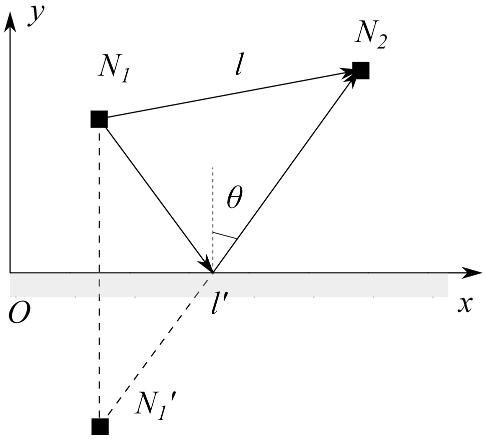


Figure 1.2: The LOS path between a transmitter N_1 at (x_1, y_1) and a receiver N_2 at (x_2, y_2), and a reflection path via the surface on the x-axis. $N_1^{'}$ at $(x_1, -y_1)$ is a mirror node of N_1 w.r.t. the reflection surface, θ is the incident angle, l and l' are the distances of the LOS link and reflection, respectively.

B.2 Relative Reflection Loss

In indoor scenarios, signals from the reflection path suffer from extra path loss compared to LOS signals due to the extended transmission distance and the power loss on the surface of reflective materials. The attenuation from reflection depends on the material thickness, permittivity, and the incident angle [DHHV08]. The power of the first-order reflected wave is (in total) about 15 dB lower than that of the LOS wave [GKZV09, GRON10]. Despite of the power loss, wave reflection from the ceiling has been proven to be a viable way of preserving connectivity and avoiding blockage [MEP$^+$10].

Definition B.1 (Relative Reflection Loss). The *relative reflection loss* χ is defined as the extra power attenuation experienced by the reflection path compared to the LOS path. That is, denoting P_{LOS} and $P_{\text{reflection}}$ (L_{LOS} and $L_{\text{reflection}}$) as the received power in decibel (path loss in dB) for the LOS link and reflection link, respectively, we have

$$\chi = P_{\text{LOS}} - P_{\text{reflection}} = L_{\text{reflection}} - L_{\text{LOS}} \text{ [dB]}.$$

Proposition B.1. *Let l and $l^{'}$ be the length of LOS path $N_1 \rightarrow N_2$ and reflection path $N_1^{'} \rightarrow N_2$, respectively, as shown in Fig. 1.2, the* relative reflection loss *is given by*

$$\chi(l, l^{'}, \theta) = 10n \log \frac{l^{'}}{l} - 20 \log |\eta(\theta)| \text{ [dB]}, \tag{5}$$

Table 1.1: The relative reflection loss: χ obtained from (5) versus χ^* calculated based on the measurement data in [GRON10].

Reflection	$\left(l, l^{'}, \theta\right)$	χ^*(dB)	χ(dB)
ceiling	$\left(2, 2\sqrt{5}, \arctan\left(\frac{1}{2}\right)\right)$	15.29	15.24
outer wall	$\left(2, 2\sqrt{10}, \arctan\left(\frac{1}{3}\right)\right)$	17.08	17.77
inner wall	$\left(2, 2\sqrt{10}, \arctan\left(\frac{1}{3}\right)\right)$	21.63	22.52
ground	$\left(2, 2\sqrt{2}, \frac{\pi}{4}\right)$	31.62	31.83

where n is the path loss exponent, and θ denotes the incident angle, and $\eta(\theta)$ is the reflection coefficient given by [Azz86]

$$\eta(\theta) = \frac{-\omega\cos\theta + \sqrt{\omega - \sin^2\theta}}{\omega\cos\theta + \sqrt{\omega - \sin^2\theta}}, \tag{6}$$

where ω is dielectric constant determined by the inherent physical property of reflective material.

Proof. See Appendix G.1 for the proof. □

To validate our model for the relative reflection loss, we present in Table 1.1 the theoretical values of χ given by (5) and the measurement data χ^* based on the experimental results from [GRON10], where the transmitter and the receiver are placed at the same horizontal plane (1 meter above the ground) in a 3-meter high empty room. The theoretical results match the measurement results to a good precision, which enables us to estimate the data rate via reflection paths based on the topology information. For the special case where the transmitter and the receiver have the same fixed distance to the reflection plane, $\chi\left(l, l^{'}, \theta\right)$ in (5) degenerates to $\chi(l)$ and the rate of the reflection path can be written as

$$R_{\text{refl}}(l) = W\log_2\left(1 + \frac{\alpha}{\chi(l)}l^{-n}\right). \tag{7}$$

B.3 Optimized Time Splitting of Half-duplex Relaying

Definition B.2 (Optimized Time Splitting)**.** Given a time slot, the *optimized time splitting* provides the time allocation between reception and transmission phases for the half-duplex DF relaying that maximizes the throughput.

Denoting $\beta \in (0,1)$ the normalized time splitting parameter for relaying, the maximum throughput of a two-hop relay path is therefore

$$R_{\text{relay}}(l_1, l_2) = \max_{\beta\in(0,1)} \min\left\{\beta R(l_1), (1-\beta) R(l_2)\right\}, \tag{8}$$

where l_1 and l_2 are the length of two hops respectively.

Proposition B.2. *Let non-negative $R_1{=}R(l_1)$ and $R_2{=}R(l_2)$ represent the rates in two hops with transmission distance l_1 and l_2, respectively, the maximum throughput of the two-hop relaying is given by*

$$R_{\text{relay}}(l_1, l_2) = \rho\left(R_1, R_2\right) \triangleq \begin{cases} 0, & \text{if } R_1R_2{=}0, \\ \frac{R_1R_2}{R_1+R_2}, & \text{otherwise.} \end{cases} \tag{9}$$

Proof. (8) is maximized when $\beta R_1 = (1{-}\beta)R_2$, i.e., when $\beta = R_2/\left(R_1 + R_2\right)$, which leads to (9). □

B.4 Relay-prioritized Region and Critical Distance

Definition B.3 (Relay-prioritized Region)**.** *Relay-prioritized region* is the region (of the circular space) in which relaying can provide higher transmission rate than the LOS path.

With half-duplex relaying, denoting l, l_1, l_2 the length of the LOS path and the two relaying paths, respectively, the relay-prioritized region exists if and only if

$$\rho\left(R\left(l_1\right), R\left(l_2\right)\right) > R\left(l\right), \tag{10}$$

for some $l_1, l_2 > 0$ and $l_1 + l_2 \geq l$. Note that the relay-prioritized region may not exist if the distance between the transmitter and the receiver is too close. To study the existence of the region, we have following definition.

Definition B.4 (Critical Distance)**.** *Critical distance* is the minimum distance between a pair of nodes for the existence of the relay-prioritized region.

Proposition B.3. *Given α and the path loss exponent n, the relay-prioritized region exists if and only if*

$$l \geq l^* \triangleq \sqrt[n]{\frac{\alpha}{2^n - 2}}, \tag{11}$$

where l is the communication distance and l^* is the critical distance.

Proof. Since $R(l)$ defined by (2) is a monotonically decreasing function of l, the left-hand side of (10) is maximized when $l_1 = l_2 = l/2$. Substituting this into (10) and combining it with (2), we obtain the critical distance shown in (11). □

The critical distance l^* decreases as n increases. As the path loss gets severer, it is more likely that a random deployed relay nodes can improve the throughput. Hence we can benefit more from utilizing relays in high path loss environments, compared to those with smaller n.

C Average Throughput and Outage Probability

In contrast to the broadcasting wave propagation in lower frequency radios with omnidirectional antennas, the high directivity of 60 GHz antenna arrays changes the wave propagation characteristic significantly. In the boresight scenario, the received signal strength can be several order of magnitude higher than that of off-boresight scenarios. As a consequence, the communication between two nodes in an indoor environment as considered in Fig. 1.2 can only be established either via the LOS link of length l, the reflection path (e.g., via the ceiling) of length l', or via a half-duplex DF relaying node.

To quantify the benefits of using relaying and/or reflection, we investigate the average throughput and outage performance for the following four scenarios:

- Case I (LOS): only LOS link is available;
- Case II (LOS, Relay): both LOS and relay paths are available;
- Case III (LOS, Reflection): both LOS and reflection paths are available;
- Case IV (LOS, Relay, Reflection): LOS, relay, and reflection are available.

We start the analysis with a single relay, and then extend the results to multiple-relay scenarios.

C.1 Random Blockage Models

When a transmission path is blocked, the strength of the received signal will be severely degraded. For simplicity we assume that the transmission rate drops to zero once the link is blocked, and we call such an event the *link blockage*. The probability of blockage for a specific link depends on many aspects, such as the area of indoor space, the beamwidth of antenna array, the size/shape of obstacles, and the link length. To model the link blockage events, we assume that there are N obstacles that behave independently and randomly. Each obstacle can block a link with probability p. That is, on average there are totally Np obstacles blocking some links. Here we consider two random blockage models, topology-independent and topology-dependent, that are described as follows. Given M links, labeled by $k = 1, \ldots, M$, if a link blockage event has occurred, the probability that link k is blocked in the topology-independent model is given by

$$\tau_k \triangleq \Pr(\text{link } k \text{ blocked} \mid \text{a link blockage}) = \frac{1}{M}, \tag{12}$$

and in the topology-dependent model, we instead have

$$\tau_k = \frac{l_k}{l_s}, \quad \text{where } l_s \triangleq \sum_{k=1}^{M} l_k, \tag{13}$$

and l_k is the length of link k. This is, in the topology-dependent model an active obstacle may block any link with probability proportional to the length of the link, which is motivated by the fact that the longer links are more likely to be blocked.

Therefore, the probability that link k is blocked is given by

$$p_k = \tau_k \Pr(\text{a link blockage occurs}) = \tau_k p, \tag{14}$$

where $p = \sum_k p_k$ is the *link blockage probability* (or blockage probability in short). Note that each active obstacle cannot cause more than one link blockage at a time, unless it rightly stands in the area of the intersection of multiple links, where the area of intersected links depends on the beam-width. Here, the case of a single obstacle blocking multiple links is not considered in our model, since the probability of such events, which can be approximated by the ratio of the intersected area to the whole area of consideration, is much smaller than the probability given by (14), especially for the scenarios with narrow beamwidth. Yet, we will consider the scenario in which multiple obstacles may obstruct the same link simultaneously.

Supposing the nodes are randomly placed within a circular hall, as illustrated in Fig. 1.1, we first analyze the blockage probability (14) of each possible link based on the associated distance vector $\mathbf{L}$. The performance in terms of average throughput and outage probability are then evaluated by taking average over all possible $\mathbf{L}$. If no relay is deployed, the distance vector $\mathbf{L} = \{l\}$ contains only one element l whose distribution is given by (4). For the cases with one relay, we have $\mathbf{L} = \{l, l_1, l_2\}$, where l_1 and l_2 follow the marginal distribution given by (3). Note that l, l_1 and l_2 are not independent.

C.2 Average Throughput with Random Blockage

For the single relay scenario with distance vector $\mathbf{L} = \{l, l_1, l_2\}$, we denote A_1 as the number of obstacles in the LOS link (length l) and A_2 the number of obstacles in the relaying path (consisting of two links with length l_1 and l_2). To simplify the notation, we denote B_1 the event that the LOS link is blocked and B_2 the event that the relaying path is blocked. Thus, $A_i = 0$ implies that the corresponding path is available (indicated by $\bar{B}_i$); otherwise it is blocked (indicated by B_i). Since the relay path consists of two hops and obstacles on either hop will block the relying path, the blockage event B_2 for the relay path can be decomposed into two sub-events $B_2 = B_2^{(1)} \cup B_2^{(2)}$.

Case I (LOS)

Since $\mathbf{L} = \{l\}$, we have $\tau_0 = 1$. Thus $p_0 = p$. The probability of no blockage is

$$\Pr\left(\bar{B}_1\right) = \Pr\left(A_1 = 0\right) = (1-p)^N. \tag{15}$$

Therefore we obtain the average throughput:

$$\bar{R}_{\mathrm{I}}(p) = \mathbb{E}\left[R(l)\right] \cdot \Pr\left(\bar{B}_1\right) = (1-p)^N \mathbb{E}\left[R(l)\right], \tag{16}$$

where $\mathbb{E}\left[R(l)\right]$ denotes the expectation over l.

Case II (LOS, Relay)

Since $\mathbf{L} = \{l, l_1, l_2\}$, we have $p_k = \tau_k p$, $k = 0, 1, 2$. Only if both LOS and relay path are blocked, the connection between the transmitter and the receiver is lost. Based on path availability, the throughput can be analyzed for four circumstances as follows:

(1) If both LOS and relay path are available, we have

$$\Pr\left(E_1 \middle| \mathbf{L}\right) \triangleq \Pr\left(\bar{B}_1 \cap \bar{B}_2\right) = (1-p)^N. \tag{17}$$

Since the transmitter-receiver pair can choose either path for communication, the maximum throughput is obtained as $R_{\max}^{(1)} = \max\{R(l), \rho\left(R(l_1), R(l_2)\right)\}$.

(2) If the only relaying path is not blocked, we have

$$\begin{aligned}\Pr\left(E_2 \middle| \mathbf{L}\right) \triangleq& \Pr\left(B_1 \cap \bar{B}_2\right)\\ =&\sum_{n=1}^{N} \Pr\left(\sum_{i=1}^{2} A_i = n\right) \cdot \Pr\left(A_2 = 0 \middle| \sum_{i=1}^{2} A_i = n\right)\\ =&\sum_{n=1}^{N} \binom{N}{n} p_0^n \left(1-p\right)^{N-n} = \left[1-(1-\tau_0)p\right]^N - \left(1-p\right)^N,\end{aligned} \tag{18}$$

and the corresponding rate is $R_{\max}^{(2)} = \rho\left(R(l_1), R(l_2)\right)$.

(3) If only LOS is not block, the probability is

$$\begin{aligned}\Pr\left(E_3 \middle| \mathbf{L}\right) \triangleq& \Pr\left(\bar{B}_1 \cap B_2\right)\\ =&\sum_{n=1}^{N} \Pr\left(\sum_{i=1}^{2} A_i = n\right) \cdot \Pr\left(A_1 = 0 \middle| \sum_{i=1}^{2} A_i = n\right)\\ =&\sum_{n=1}^{N} \binom{N}{n} (p_1 + p_2)^n \left(1-p\right)^{N-n} = \left(1-\tau_0 p\right)^N - \left(1-p\right)^N,\end{aligned} \tag{19}$$

and the corresponding rate is $R_{\max}^{(3)} = R(l)$.

(4) If both LOS and relay paths are blocked, we have $R_{\max}^{(4)} = 0$ with the probability

$$\begin{aligned}\Pr\left(E_4 \middle| \mathbf{L}\right) \triangleq& \Pr\left(B_1 \cap B_2\right)\\ =&1 - \Pr\left(\bar{B}_1 \cap \bar{B}_2\right) - \Pr\left(B_1 \cap \bar{B}_2\right) - \Pr\left(\bar{B}_1 \cap B_2\right).\end{aligned} \tag{20}$$

Thus, the throughput of Case II (LOS, Relay) for a given distance vector $\mathbf{L}$ is

$$\mathbb{E}\left[R_{\mathrm{II}}(p) \middle| \mathbf{L}\right] = \sum_{i=1}^{4} R_{\max}^{(i)} \Pr\left(E_i \middle| \mathbf{L}\right), \tag{21}$$

and the average throughput is obtained by averaging over all the possible realizations of $\mathbf{L} = \{l, l_1, l_2\}$, i.e.,

$$\bar{R}_{\mathrm{II}}(p) = \iiint\limits_{(l,l_1,l_2)} \mathbb{E}\left[R_{\mathrm{II}}(p) \middle| \mathbf{L}\right] f_{\mathbf{L}}\left(l, l_1, l_2\right) dl dl_1 dl_2, \tag{22}$$

where $f_{\mathbf{L}}\left(l, l_1, l_2\right)$ is the joint probability distribution function.

Case III (LOS, Reflection)

Due to the extra power loss by reflection, the reflection path will not be used unless the LOS is blocked. Thus the average throughput is

$$\begin{aligned} \bar{R}_{\mathrm{III}}(p) = & \mathbb{E}\left[R(l)\right] \Pr\left(\bar{B}_1\right) + \mathbb{E}\left[R_{\mathrm{refl}}(l)\right] \Pr\left(B_1\right) \\ = & (1-p)^N \mathbb{E}\left[R(l)\right] + \left[1-(1-p)^N\right] \mathbb{E}\left[R_{\mathrm{relf}}(l)\right], \end{aligned} \tag{23}$$

which indicates that, when $\bar{R}_{\mathrm{I}}(p)$ approaches zero, there is still extra $\mathbb{E}\left[R_{\mathrm{relf}}(l)\right]$ provided by transmission via reflection.

Case IV (LOS, Relay, Reflection)

The analysis is similar to in Case II (LOS, Relay), except that reflection is now taken into account. We can enumerate all the random blockage events and their corresponding throughput, and the results $(P\{E_i | \mathbf{L}\}, R_{\max}^{(i)})$ for $i \in \{1, 2, \ldots, 8\}$ are summarized as below:

- Blockage Event $\left\{E_1 \middle| \mathbf{L}\right\} \triangleq \bar{B}_1 \cap \bar{B}_2^{(1)} \cap \bar{B}_2{}^{(2)}$:

$$\begin{aligned} & R_{\max}^{(1)} = \max\left\{R(l),\ \rho\left(R(l_1), R(l_2)\right)\right\} \\ & \Pr\left(E_1 \middle| \mathbf{L}\right) = (1-p)^N \end{aligned}$$

- Blockage Event $\left\{E_2 \middle| \mathbf{L}\right\} \triangleq \bar{B}_1 \cap B_2^{(1)} \cap \bar{B}_2^{(2)}$:

$$\begin{aligned} & R_{\max}^{(2)} = \max\left\{R(l),\ \rho\left(R_{\mathrm{refl}}(l_1), R(l_2)\right)\right\} \\ & \Pr\left(E_2 \middle| \mathbf{L}\right) = \left(1-(1-\tau_1)p\right)^N - (1-p)^N \end{aligned}$$

- Blockage Event $\left\{E_3 \middle| \mathbf{L}\right\} \triangleq \bar{B}_1 \cap \bar{B}_2^{(1)} \cap B_2^{(2)}$:

$$\begin{aligned} & R_{\max}^{(3)} = \max\left\{R(l),\ \rho\left(R(l_1), R_{\mathrm{refl}}(l_2)\right)\right\} \\ & \Pr\left(E_3 \middle| \mathbf{L}\right) = \left(1-(1-\tau_2)p\right)^N - (1-p)^N \end{aligned}$$

- Blockage Event $\left\{E_4 \middle| \mathbf{L}\right\} \triangleq \bar{B}_1 \cap B_2^{(1)} \cap B_2^{(2)}$:

$$R_{\max}^{(4)} = \max\left\{R(l),\ \rho\left(R_{\mathrm{refl}}(l_1), R_{\mathrm{refl}}(l_2)\right)\right\}$$

$$\Pr\left(E_4 \middle| \mathbf{L}\right) = (1-\tau_0 p)^N + (1-p)^N - \left[(1-(1-\tau_1)p)^N + (1-(1-\tau_2)p)^N\right]$$

- Blockage Event $\left\{E_5 \middle| \mathbf{L}\right\} \triangleq B_1 \cap \bar{B}_2^{(1)} \cap \bar{B}_2^{(2)}$:

$$R_{\max}^{(5)} = \max\left\{R_{\text{refl}}(l),\ \rho\left(R(l_1), R(l_2)\right)\right\}$$
$$\Pr\left(E_5 \middle| \mathbf{L}\right) = (1-(1-\tau_0)p)^N - (1-p)^N$$

- Blockage Event $\left\{E_6 \middle| \mathbf{L}\right\} \triangleq B_1 \cap \bar{B}_2^{(1)} \cap B_2^{(2)}$:

$$R_{\max}^{(6)} = \max\left\{R_{\text{refl}}(l),\ \rho\left(R(l_1), R_{\text{refl}}(l_2)\right)\right\}$$
$$\Pr\left(E_6 \middle| \mathbf{L}\right) = (1-\tau_1 p)^N + (1-p)^N - \left[(1-(1-\tau_0)p)^N + (1-(1-\tau_2)p)^N\right]$$

- Blockage Event $\left\{E_7 \middle| \mathbf{L}\right\} \triangleq B_1 \cap B_2^{(1)} \cap \bar{B}_2^{(2)}$:

$$R_{\max}^{(7)} = \max\left\{R_{\text{refl}}(l),\ \rho\left(R_{\text{refl}}(l_1), R(l_2)\right)\right\}$$
$$\Pr\left(E_7 \middle| \mathbf{L}\right) = (1-\tau_2 p)^N + (1-p)^N - \left[(1-(1-\tau_0)p)^N + (1-(1-\tau_1)p)^N\right]$$

- Blockage Event $\left\{E_8 \middle| \mathbf{L}\right\} \triangleq B_1 \cap B_2^{(1)} \cap B_2^{(2)}$:

$$R_{\max}^{(8)} = \max\left\{R_{\text{refl}}(l),\ \rho\left(R_{\text{refl}}(l_1), R_{\text{refl}}(l_2)\right)\right\}$$
$$\Pr\left(E_8 \middle| \mathbf{L}\right) = 1 + (1-p)^N - \textstyle\sum_{i=0}^{2}\left[(1-\tau_i p)^N + (1-(1-\tau_i)p)^N\right]$$

Hence, the average throughput is

$$\bar{R}_{\text{IV}}(p) = \iiint\limits_{(l,l_1,l_2)} \mathbb{E}\left[R_{\text{IV}}(p) \middle| \mathbf{L}\right] f_{\mathbf{L}}\left(l, l_1, l_2\right) dl dl_1 dl_2, \tag{24}$$

where $\mathbb{E}\left[R_{\text{IV}}(p) \middle| \mathbf{L}\right]$ is given by

$$\mathbb{E}\left[R_{\text{IV}}(p) \middle| \mathbf{L}\right] = \sum_{i=1}^{8} R_{\max}^{(i)} \Pr\left(E_i \middle| \mathbf{L}\right). \tag{25}$$

The extension to the multiple-relay scenario is straightforward and similar to the single-relay case.

C.3 Outage Probability with Random Blockage

Given a threshold rate R_ϵ, we say an outage event happens if the rate R is lower than R_ϵ. Thus the outage probability is

$$\Pr_{\text{out}}\left(R_\epsilon\right) = \Pr\left(R < R_\epsilon\right). \tag{26}$$

Take Case II (LOS, Relay) for example. Given blockage probability p, the outage probability $\Pr\left(R_{\mathrm{II}}(p) < R_\epsilon\right)$ is a function of p that can be written as

$$\iiint\limits_{(l,l_1,l_2)} \Pr\left(R_{\mathrm{II}}(p) < R_\epsilon \middle| \mathbf{L}\right) f_{\mathbf{L}}\left(l,l_1,l_2\right) dl dl_1 dl_2, \tag{27}$$

where $\Pr\left(R_{\mathrm{II}} < R_\epsilon \middle| \mathbf{L}\right)$ is the outage probability for given $\mathbf{L}$ and averaged over all the four different scenarios, i.e.,

$$\Pr\left(R_{\mathrm{II}}(p) < R_\epsilon \middle| \mathbf{L}\right) = \sum_{i=1}^{4} \Pr\left(R_{\max}^{(i)} < R_\epsilon\right) \cdot \Pr\left(E_i \middle| \mathbf{L}\right), \tag{28}$$

where $\Pr\left(E_i \mid \mathbf{L}\right)$ is the probability of the corresponding event E_i given $\mathbf{L}$, and it is different for the topology-independent or topology-dependent models.

To gain more insights of the outage performance, we set the threshold R_ϵ to a sufficiently low value such that the transmission rates of LOS or by the relay path are still above the threshold even if the terminals are the farthest apart. Thus, $\Pr\left(R_{\max}^{(i)} < R_\epsilon\right) = 0$, for $i = 1,2,3$, and $\Pr\left(R_{\mathrm{II}}(p) < R_\epsilon \mid \mathbf{L}\right) = \Pr\left(E_4 \mid \mathbf{L}\right)$. The outage probability is

$$\begin{aligned} \Pr_{\text{out}}^{(1\_\text{relay})}(p) &= \iiint\limits_{(l,l_1,l_2)} \Pr\left(E_4 \mid \mathbf{L}\right) \cdot f_{\mathbf{L}}\left(l,l_1,l_2\right) dl dl_1 dl_2 \\ &= 1 - \iiint\limits_{(l,l_1,l_2)} \left[(1-\tau_0 p)^N + (1-(1-\tau_0)p)^N\right] \cdot f_{\mathbf{L}}\left(l,l_1,l_2\right) dl dl_1 dl_2 + (1-p)^N, \end{aligned} \tag{29}$$

where τ_0 is described in (13) and (12).

Particularly, for the topology-independent random blockage model, the outage probability in (29) is given by:

$$\Pr_{\text{out}}^{(1\_\text{relay})}(p) = g_{p,3}^N(1) + g_{p,3}^N(2) - g_{p,3}^N(3), \tag{30}$$

where the function $g_{p,M}^N(k)$ is defined as

$$g_{p,M}^N(k) = 1 - \left(1 - k\frac{p}{M}\right)^N. \tag{31}$$

As the number of relays increases to 2, there are totally $M = \binom{2+2}{2} = 6$ links in the relay network. The outage probability with given blockage probability p and obstacle number N is

$$\Pr_{\text{out}}^{(2\_\text{relays})}(p) = g_{p,6}^N(1) + 2g_{p,6}^N(2) - 7g_{p,6}^N(4) + 7g_{p,6}^N(5) - 2g_{p,6}^N(6), \tag{32}$$

and the derivation of (32) is in Appendix G.2.

In the region with low blockage probability (for example, $p \leq 0.1$), by using the binomial theorem over (31) and omitting the higher order terms, we can approximate (30) and (32) by, respectively,

$$\underset{\text{out}}{\overset{(1\_\text{relay})}{\Pr}}(p) \approx 4\binom{N}{2}\left(\frac{p}{3}\right)^2, \tag{33}$$

and

$$\underset{\text{out}}{\overset{(2\_\text{relay})}{\Pr}}(p) \approx 12\binom{N}{3}\left(\frac{p}{6}\right)^3. \tag{34}$$

Similarly, given $m = 3$ or $m = 4$ relays as well as N obstacles, we can also derive the approximation of outage probability by following the method shown in Appendix G.2. With a similar approach in $m = 1, 2, 3$ and 4, $\Pr_{\text{out}}^{(m-\text{relays})}(p)$ for small p and general m in the topology-independent model can be expressed as

$$\underset{\text{out}}{\overset{(\text{m\_relay})}{\Pr}}(p) \approx 2(m+1)!\binom{N}{m+1}\left(\frac{p}{\binom{m+2}{2}}\right)^{m+1}. \tag{35}$$

However, the analysis becomes much more involved if the reflection is also exploited, and we will resort to the numerical evaluations for such situations.

D Maximum Throughput Path Selection

In this section we will develop an algorithm termed as *maximum throughput path selection* (MTPS) to find the path that maximizes the throughput for each source-destination pair. We denote the topology of the network by a graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ is the set of all nodes in the network and $\mathcal{A}$ is the set of undirected connections between any pair of nodes. let $\mathcal{N}_r = \{1, 2, \ldots, |\mathcal{N}_r|\}$ be the set of relay nodes and let $\mathcal{N}_u$ be the set of user nodes, we have $\mathcal{N} = \mathcal{N}_r \cup \mathcal{N}_u$.

As discussed in Sec. B.1, a central coordinator is installed to manage and schedule the communications. It collects and updates user nodes' location information, monitors the status of blockage for user-relay links and relay-relay links, determines and executes path selections based on communications request, and coordinates concurrent communication tasks within the network. Since some parameters, such as the dimension of the indoor space, dielectric constants ω of the reflection materials (walls, ceiling, floor), the position of deployed relaying nodes, and the antenna gains G_t and G_r, can be obtained at the installment and updated when necessary, we will assume that they are known at no extra cost. The path loss exponent n, the transmit power budget P_t at each individual node, and the background noise N_0 are long-term parameters and therefore can be updated with negligible overhead.

However, short-term parameters such as user nodes' current positions, the availability of the LOS links for user-relay and relay-relay paths, and the status of direct communications between any two nodes have to be updated at the central coordinator periodically to enable the proposed scheme. To ensure that the control signaling between the central coordinator and all other nodes can be delivered successfully, we assume that the central coordinator works in the omni- or quasi-omni mode and the control signaling channels are orthogonal to data transmission channels. Such orthogonal control channels can be established by dedicated time slots or system bandwidth. For instance, a time-division-duplex (TDD) based control plane is adopted in the WiGig specifications [WiG10]. The allocation of dedicated control channels can be justified as follows. When the load of the system is not high, there are sufficient resources for both data transmission and control signaling; when the load is high, it is crucial to have orthogonal control signaling to minimize the possibility of collision, which could otherwise cause severe system performance degradation. The associated overhead will be analyzed in Sec. D.3.

The proposed path selection can be done in three steps:

1. The coordinator monitors the network and updates a modified adjacency matrix following Algorithm 2;

2. Upon the request from a source-destination pair, the coordinator selects the path according to Algorithm 3;

3. The coordinator sends the communication schedule to the source-destination pair and the selected relay nodes.

Note that although our Algorithms 2 and 3 only focus on a single source-destination pair, multiple source-destination pairs can be scheduled concurrently by using orthogonal channels (e.g., in different time/frequency/beam). We will not elaborate on this as it is not the focus of this work.

D.1 Algorithm Description

The modified adjacency matrix $\mathbf{D} = [d_{uv}]$ indicates the effective distance between all pairs of nodes by taking into account link blockage and reflection. We first use a distance matrix $\tilde{\mathbf{D}}$ to denote the physical distance between any pair of nodes. That is, the entry $\tilde{d}_{uv}$ of $\tilde{\mathbf{D}}$ means the distance between the nodes u and v. Based on the status of the link, the adjacency matrix $\mathbf{D} = [d_{uv}]$ is updated by taking $d_{uv} = \tilde{d}_{uv} \cdot b_{uv}$, where $b_{uv} = 1$ means the existence of the LOS link, and $b_{uv} = \infty$ for the scenario with link blockage but no reflection, and $b_{uv} = \sqrt[n]{\chi(d_{uv})}$ for the scenario where the reflection path is available in the event of link blockage. The link status matrix $\mathbf{B} = [b_{uv}]$ is determined and updated based on the feedback from the relay and user nodes. The process of calculating the effective distance is shown in Algorithm 2.

Note that although the size of the modified adjacency matrix $\mathbf{D}$ is $|\mathcal{N}| \times |\mathcal{N}|$, which can be large when there are many user nodes, only a few elements need to be updated regularly. First of all, $\mathbf{D}$ is symmetric, with a small sub-matrix of size

$|\mathcal{N}_r| \times |\mathcal{N}_r|$ that is crucial for all kinds of communication requests. Secondly, for each new user node added into the system, we only need to add a new vector of length $|\mathcal{N}_r|$ to monitor the links between the new user node and all the relay nodes. Last but not the least, the links among all user nodes are only updated based on their updated position information and the corresponding link status parameters b_{uv} are not updated unless $\{u, v\}$ form a source-destination pair.

Algorithm 2: Weighted Distance Calculation

```
Data: system location parameters, $\omega$, and $n$
Input: distance matrix $\hat{\mathbf{D}} = \left[\hat{d}_{uv}\right]$
if the reflection participates then
    ReflectionOption ← True;
else
    ReflectionOption ← False;
end
foreach $(u, v) \in \mathcal{A}$ do
    if $(u, v)$ is blocked then
        if ReflectionOption = True then
            $b_{uv} \leftarrow \sqrt[n]{\chi(d_{uv})}$;
        else
            $b_{uv} \leftarrow \infty$;
        end
    else
        $b_{uv} \leftarrow 1$;
    end
end
Hadamard product: $\mathbf{D} \leftarrow \tilde{\mathbf{D}} \circ \mathbf{B}$;
Output: modified adjacency matrix $\mathbf{D} = [d_{uv}]$
```

With the updated modified adjacency matrix $\mathbf{D}$, we find a path with the maximum throughput based on the MTPS procedure presented in Algorithm 3. We focus on a single source-destination pair $\{s, t\}$ and take $\mathcal{N} = \mathcal{N}_r \cup \{s, t\}$ to simplify the notation in the algorithm. In the algorithm, the sorted set $\mathcal{U}$ is used to denote the set of unselected relays, which is initialized by $\mathcal{N}_r$, and the entry $\mathcal{U}^{(i)}, (i \in \{1, 2, \ldots, |\mathcal{U}|\})$, represents the i^{th} entry of $\mathcal{U}$. We use an indicator matrix $\mathcal{F} = [f_{ij}]$ to indicate that whether the rate of a link can be improved by using a relaying node. We use $\mathcal{P}$ to denote the current optimal path. The main steps of Algorithm 3 can be outlined as follows:

1) Initialize $\mathcal{P} = \{(s, t)\}$, $\mathcal{U} = \mathcal{N}_r$, and $\mathbf{D} = \mathbf{0}$;

2) For each hop $i \to j$ of the current path $\mathcal{P}$, if it is not locked $(f_{ij} = 0)$ and at least one relay from $\mathcal{U}$ is in the relay-prioritized region, we find the relay node $r \in \mathcal{U}$ with the highest throughput, and remove it from $\mathcal{U}$, and split the original hop $i \to j$ two new hops $i \to r$ and $r \to j$.

3) Update the current optimal path;

4) If the updated path remains the same as the previous one, output the current path and exit; otherwise, go back to 2).

D.2 Maximum Throughput Calculation

Assuming that there are L hops for a path with half-duplex relays, and the rate over the i-th hop is denoted by R_i, where $i \in \{1, 2, \dots, L\}$, then the throughput of the L-hop relay path is given by Proposition D.1.

Proposition D.1. *For an L-hop channel, let R_i be the capacity of each individual channel i, $i = 1, 2, \dots, L$, the maximum achievable rate R^* over the cascaded channel under the half-duplex constraint is given by:*

$$R^* = \min \left\{ \frac{R_1 R_2}{R_1 + R_2}, \frac{R_2 R_3}{R_2 + R_3}, \cdots, \frac{R_{L-1} R_L}{R_{L-1} + R_L} \right\}. \tag{36}$$

Proof. For an L-hop network consisting of $L - 1$ half-duplex relaying nodes, we assume each relaying node k, $k \in \{1, 2, \dots, L-1\}$, connecting the k^{th} and $(k+1)^{\text{th}}$ hop, has a transmitting state, a receiving state and a silent state with corresponding durations $\xi_k^{\mathfrak{t}}$, $\xi_k^{\mathfrak{r}}$ and $\xi_k^{\mathfrak{s}}$, respectively. We have a duration constraint as follows:

$$\xi_k^{\mathfrak{t}} + \xi_k^{\mathfrak{r}} + \xi_k^{\mathfrak{s}} = 1, \quad k \in \{1, 2, \dots, L-1\}. \tag{37}$$

Clearly, the end-to-end throughput is the maximum achievable flow through all intermediate nodes. We assume that a packet of size R goes through each node within a unit duration. At relaying node k, the durations of receiving and transmitting states are respectively given by:

$$\xi_k^{\mathfrak{r}} = \frac{R}{R_k}, \quad \xi_k^{\mathfrak{t}} = \frac{R}{R_{k+1}}, \tag{38}$$

where R_{k-1} and R_k are the channel capacities of k^{th} and $(k+1)^{\text{th}}$ hop, respectively. Combining with the constraint in (37), we have the silence duration as the function of R:

$$\xi_k^{\mathfrak{s}}(R) = 1 - \frac{R_k + R_{k+1}}{R_k R_{k+1}} R = 1 - s_k R. \tag{39}$$

Thus, for all $k \in \{1, 2, \dots, L-1\}$, we can obtain a cluster of linear functions $\xi_k^{\mathfrak{s}}(R)$. We find that $\xi_k^{\mathfrak{s}}(R)$ decays by increasing R with the slope $-s_k$. In addition, by considering the fact that for each relaying node k, we always have $\xi_k^{\mathfrak{s}}(R) \geq 0$, the problem can be reformulated as:

$$R^* = \max_{\substack{\xi_k^{\mathfrak{s}}(R) \geq 0 \\ k \in \{1,2,\dots,L-1\}}} R = \min_{k \in \{1,2,\dots,L-1\}} \frac{1}{s_k} = \min_{k \in \{1,2,\dots,L-1\}} \frac{R_k R_{k+1}}{R_k + R_{k+1}}. \tag{40}$$

Algorithm 3: Maximum Throughput Path Selection

```
Data:
    • effective distance $\mathbf{D} = [d_{uv}]$;
    • aggregated parameter $\alpha$ and path loss exponent $n$;
Input:
    • initial path set $\mathcal{P} = \{(s,t)\}$;
    • unselected relays set $\mathcal{U} = \mathcal{N}_r$;
Initialize the link-lock matrix $\mathbf{F} = [f_{uv}] = \mathbf{0}_{|\mathcal{N}_r| \times |\mathcal{N}_r|}$;
Calculate critical distance $l^* \leftarrow \sqrt[n]{\frac{\alpha}{2^n - 2}}$;
Set termination condition TerminateFlag $\leftarrow$ False;
while TerminateFlag $=$ False do
    $\mathcal{P}' \leftarrow \emptyset$;
    foreach $(i,j) \in \mathcal{P}$ do
        SubstituteFlag $\leftarrow$ False;
        $\mathcal{P}^* \leftarrow \{(i,j)\}$;
        if $d_{ij} \geq l^*$ and $f_{ij} = 0$ then
            $R_{\max} \leftarrow R(d_{ij})$;
            for $k \leftarrow 1$ to $|\mathcal{U}|$ do
                $r \leftarrow \mathcal{U}^{(k)}$;
                $R^*_{\max} \leftarrow \rho(R(d_{ir}), R(d_{rj}))$;
                if $R^*_{\max} \geq R_{\max}$ then
                    $R_{\max} \leftarrow R^*_{\max}$;
                    $\mathcal{P}^* \leftarrow \{(i,r),(r,j)\}$;
                    $r^* \leftarrow r$;
                    SubstituteFlag $\leftarrow$ True;
                end
            end
            if SubstituteFlag $=$ True then
                $\mathcal{U} \leftarrow \mathcal{U} - \{r^*\}$;
            else
                $f_{ij} \leftarrow 1$;
            end
        else
            $f_{ij} \leftarrow 1$;
        end
        $\mathcal{P}' \leftarrow \mathcal{P}' \cup \mathcal{P}^*$;
    end
    if $\mathcal{P}' = \mathcal{P}$ then
        $\mathcal{P}_{opt} \leftarrow \mathcal{P}$;
        TerminateFlag $\leftarrow$ True;
    else
        $\mathcal{P} \leftarrow \mathcal{P}'$;
    end
end
Output: the optimal path set $\mathcal{P}_{opt}$
```

Therefore, within a time slot, a packet with the size at most R^* can be transmitted from the source to the sink. This concludes the maximum achievable rate shown in (36). □

Note that the maximum throughput provided by Proposition D.1 holds only if the amount of transmitted data goes to infinity. As a simple example, we assume there is one optimal routing path $\mathcal{P}$ between the source and destination, consisting of L hops with the maximum transmission rate R in each hop. By following (36), we obtain the theoretically maximum throughput $R^* = R/2$. Now we assume K packets, each size of R_p, are to be transmitted. Thus the effective rate over the path is given by:

$$R_{\text{eff}} = \frac{KR_p}{\frac{R_p}{R}L + \frac{R_p}{R}(2K-1)} = \frac{KR}{2K+L-1} \leq R^*. \tag{41}$$

It is clear that for finite K, the effective rate is strictly smaller than the theoretical optimal results given by Proposition D.1. Yet for a large K, R_{eff} shall approach R^*. Hence, here we assume K is large enough such that

$$\frac{L-1}{K} \to 0 \Longrightarrow R_{\text{eff}} \to R^*. \tag{42}$$

D.3 Complexity Analysis

The computational complexity of the MTPS algorithm consists of two parts: calculating the throughput of paths and comparing the throughput. From (40), the complexity of the throughput calculation can be measured by the number of the $\min(\cdot,\cdot)$ operation, which outputs the smaller value of two arguments. It is evident that if we want to find the minimum of q inputs, the $\min(\cdot,\cdot)$ operator will be recursively applied $q-1$ times. For comparing the throughput, the complexity for finding the optimal path is evaluated by the number of the $\max(\cdot,\cdot)$ operation, which outputs the larger input.

For the brute-force algorithm with exhaustive search, given n relays, there are $k!\binom{n}{k}$ distinct paths of $k+1, (k = 0, 1, \ldots, n)$ hops connecting the transmitter and the receiver. To calculate the throughput of an i-hop path, $i-1$ $\min(\cdot,\cdot)$ operations are needed. Thus the complexity is given by $\sum_{k=0}^{n} k \cdot k!\binom{n}{k}$. To find the paths with the maximum throughput, $\sum_{k=0}^{n} k!\binom{n}{k}$ comparisons are needed since there are $\sum_{k=0}^{n} k!\binom{n}{k}$ different paths. Let $T_1(n)$ denote the number of $\min(\cdot,\cdot)$ or $\max(\cdot,\cdot)$ operations for the network with n relays, we have

$$T_1(n) = \sum_{k=0}^{n} \frac{k \cdot n!}{(n-k)!} + \sum_{k=0}^{n} \frac{n!}{(n-k)!} - 1 \in \mathcal{O}(n \cdot n!), \tag{43}$$

since

$$\lim_{n\to\infty} \frac{T_1(n)}{n \cdot n!} = \lim_{n\to\infty} \left(\frac{1}{n} \sum_{k=0}^{n} \frac{k+1}{(n-k)!} - \frac{1}{n \cdot n!} \right) = e, \tag{44}$$

where $e = 2.718\cdots$ is the Euler's number.

Similarly, we define $T_2(n)$ the number of $\min(\cdot,\cdot)$ or $\max(\cdot,\cdot)$ operations for the MTPS algorithm. For the worst case, only one available relay is selected within each iteration to replace one hop, and other hops will be locked for the path. In such scenario, within each iteration, the number of $\min(\cdot,\cdot)$ equals to that of $\max(\cdot,\cdot)$. This is because each hop will be replaced by possible 2-hops, which need comparison operations subsequently. Thus, the complexity of the MTPS algorithm is given by:

$$T_2(n) = 2\left(n + 2\sum_{k=1}^{n-1}(n-k)\right) = 2n^2 \in \mathcal{O}(n^2). \tag{45}$$

Comparing (43) and (45), we see that the MTPS algorithm can significantly reduce the complexity compared to the brute-forced algorithm when the number of relays are large.

The signaling overhead of the protocol is small. On one hand, the amount of data that needs to periodically updated is small. For example, the availability of the LOS paths for user-relay or relay-relay links and the status of any direct user-user transmission only account for a few bits from each node (a few bits for link identity and another bit for LOS availability). The positioning information is at most 30 bytes [CLCS09, LDBL07]. Given the fact that the data transmission rate is in the order of 100 *Mbps*, the time needed for transmitting control signaling for each user node is around 10 microseconds[3]. On the other hand, the periodicity of the parameter update is relatively large since the mobility of users/obstacles is very slow in the indoor environment. For example, given a beamwidth of 12° and communication distance of 5 meters, it takes roughly 500 milliseconds for the node to move from the center of the beam to its edge and thus invalidate the previous position information, if the velocity is 1 meter per second. Combining the above two factors, and given the fact that there cannot be too many users in indoor environment (say maximum 1 user per square meter), the aggregate overhead of the control signaling is less than 1%.

E Performance Evaluation

In this section, we present simulation results on average throughput and outage performance for both topology-dependent and topology-independent blockage models with different numbers of relay nodes deployed inside a circular hall of radius $R_0 = 15$ meters, where $N = 20$ obstacles are randomly distributed with link blockage probability $p \in [0, 1]$. To simplify simulation, we assume that all the nodes are placed on the same horizontal plane above the ground, and the ceiling of the indoor environment is supposed to be always available for wave reflection. The distance between the transmitter-receiver plane and the reflection ceiling is defined as h, and

[3] Here we also take into account the rate degradation caused by using omni-directional antenna pattern at the central coordinator for control signaling.

Table 1.2: Simulation Parameters

Parameters	**Symbol**	**Value**
Bandwidth	W	1200 MHz
Transmitting Power	P_{t}	0.1 mW
Transmitter Antenna Gain	G_{t}	15 dB
Receiver Antenna Gain	G_{r}	15 dB
Path Loss Exponent	n	3
Background Noise	N_0	-114 dBm/MHz
Wavelength	λ	5×10^{-3} m
Dielectric Constant of Ceiling	ω	$6.14 - \jmath 0.3015$
Distance to Reflective Ceiling	h	3 m
Number of Obstacles	N	20
Radius of Relay Placement	r	3 m
Radius of Circular Hall	R_0	15 m

the common simulation parameters are summarized in Table 1.2, which corresponds to the SNR of 35 dB at 1-meter LOS transmission distance and 5 dB at 10-meter distance.

Furthermore, we assume that all the antennas are situated at the same horizontal plane as defined by the height of the nodes, which will allow us to characterize the radiation pattern in a 2-D fashion[4]. Given azimuthal angle coordinate ϕ and beamwidth Δ_b, the antenna gain can be determined by using the idealized *flat-top model* [WNE02] as follows,

$$G(\phi) = \begin{cases} \dfrac{2\pi}{\Delta_b}, & |\phi| \leq \dfrac{\Delta_b}{2} \\ 0, & \text{otherwise.} \end{cases} \tag{46}$$

With the antenna gain of 15 dB given by Table 1.2, we have the beamwidth $\Delta_b = 11.4°$.

E.1 Average Throughput

We first evaluate the average throughput for four communications scenarios, namely, LOS, (LOS, Relay), (LOS, Reflection), and (LOS, Relay, Reflection). We increase the link blockage probability from 0 to 1 to demonstrate the benefits of resorting to relaying and/or reflection, the dependence of the two random blockage models, and the influence of the number of deployed relaying nodes.

[4] Usually, directional antennas are characterized by the radiation pattern in the 3-D fashion, which relates to the elevation and azimuth beamwidth.

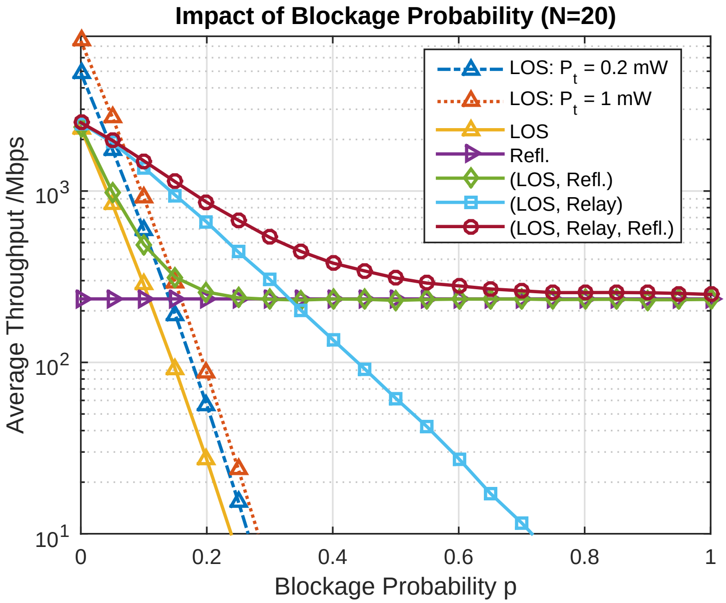


Figure 1.3: Average throughput [Mbps] versus link blockage probability p for topology-independent models, where $N = 20$ obstacles are randomly distributed within a circular hall of radius $R_0 = 15$ meters and the half-duplex relay node is deployed at the center of the hall. The default transmit power is $P_t = 0.1$mW with bandwidth $W = 1200$ MHz.

One Relay Scenario

In Fig. 1.3 we demonstrate the benefits of resorting to relaying and/or reflection under the topology-independent random blockage model. We can see that, when there are many obstacles (N=20) with link blockage probability $p \in (0.1, 0.2)$, the average throughput can be increased by approximately 10 times by introducing a half-duplex relay node. The benefit of resorting to reflection paths, which is not significant for small p (< 0.05), outweighs that of relaying for $p > 0.35$. Note that, without reflection, the average throughput for the LOS and (LOS, relay) scenarios both decreases dramatically, but adding a relay node can significantly slow down the performance degradation trend. When reflection is taken into account, average throughput of roughly 290 Mbps is still available even for very high link blockage probability, which is well in line with our analysis in (23). Therefore, resorting to reflection is capable to provide the minimum guarantee on average throughput when the other transmission paths are unavailable. We also plot two curves for LOS scenarios with increased transmit power $P_{\mathrm{t}} = 0.2$ mW (dashed line, 3 dB power gain) and $P_{\mathrm{t}} = 1$ mW (dotted line, 10 dB power gain), respectively. Although

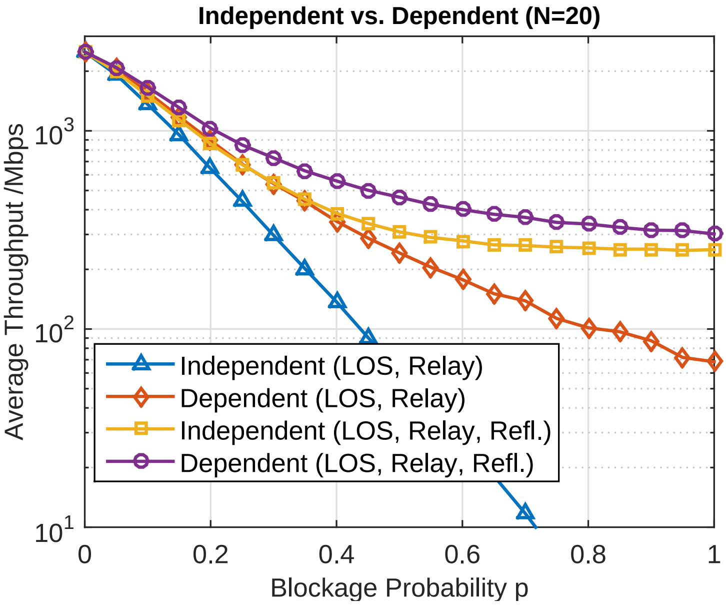


Figure 1.4: Impact of blockage probability p on average throughput [Mbps] in the topology-dependent or topology-independent model, where the (LOS, Relay) and (LOS, Relay, Reflection) scenarios are employed.

increasing the transmit power can improve the average throughput, it is not an effective way to combat link blockage.

In Fig. 1.4 we evaluate the influence of blockage modeling on the performance of average throughput. For the (LOS, Relay) scenario, the difference between the average throughput of two blockage models can be huge, especially when the blockage probability p is large. The gap is much smaller when reflection is also taken into consideration since the reflection paths are assumed to be always available (say via ceiling) and therefore not subject to link blockage.

Multi-Relay Scenarios

In Fig. 1.6 we evaluate the benefit of deploying multiple relaying nodes under the topology-independent model, where 2 to 4 relays are uniformly distributed on the circle with the radius $r = 3$ meters according to the placement patterns shown in Fig. 1.5. As the number of relaying nodes increases, the average throughput is improved significantly for almost all range of link blockage probability ($p > 0.05$). At $p = 0$, the improvement of increasing the number of relays is almost invisible since the probability of increasing the rate by using a half-duplex relaying node is small given the relatively small radius of the hall ($R_0 = 15$ meters). Another

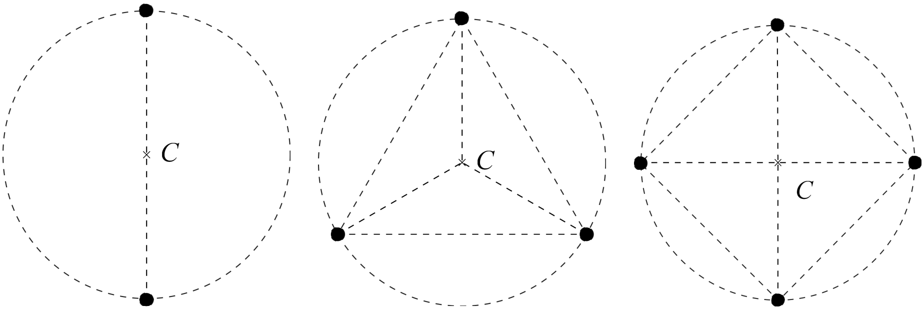


Figure 1.5: The placement patterns for multiple relays that are uniformly scattered on the circle with the radius $r = 3$ m in the room with radius $R_0 = 15$ m.

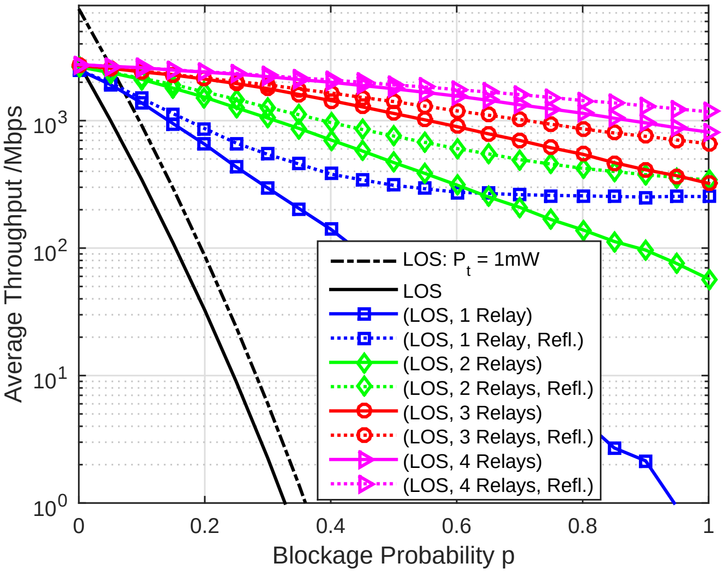


Figure 1.6: Average throughput performance for the multi-relay scenarios with blockage in topology-independent model, where the relays are following the deployment shown in Fig. 1.5.

interesting observation in Fig. 1.6 is that average throughput guarantee can also be realized by increasing the number of deployed relaying nodes, and the importance of reflection paths is no longer remarkable even when p is large.

On the other hand, Fig. 1.7 shows that the strong influence of blockage modeling still remains for multi-relay scenarios, very similar to what we have observed in Fig. 1.4. Besides, a more interesting phenomenon will happen if we have higher

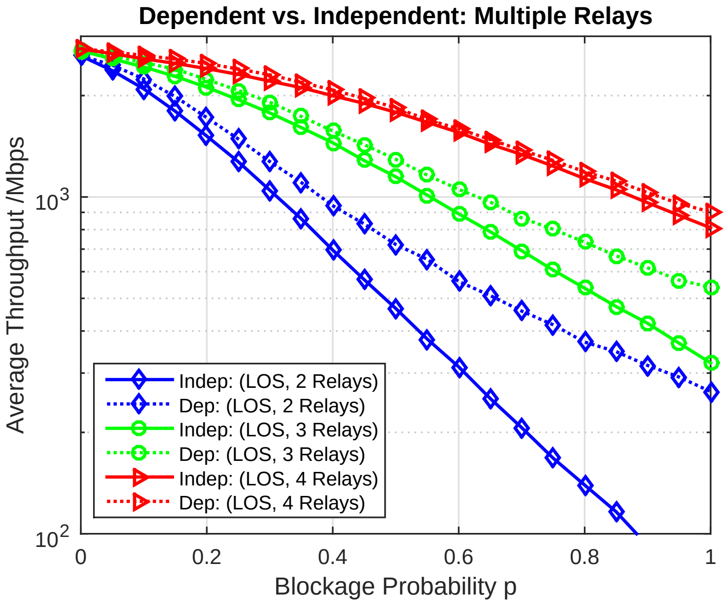


Figure 1.7: Average throughput [Mbps] for (LOS, relay) with multiple relaying nodes under topology-dependent or topology-independent models.

antenna gains, that is, the topology-dependent model will be surpassed by its topology-independent counterpart in terms of average throughput when the number of relays is increased to a certain value, which also implies the influence of blockage modeling. Here, we are not going to provide those simulations due to the space limitation.

E.2 Outage Probability

We first investigate the influence of blockage modelling on the availability of LOS and relay links. In Fig. 1.8 we plot the event probabilities $\Pr(E_i)$, $i \in \{1,2,3,4\}$, indicating the availability of the LOS and/or the relay paths for the (LOS, Relay) scenario, where

$$\Pr(E_i) = \iiint\limits_{(l,l_1,l_2)} \Pr\left(E_i \middle| \mathbf{L}\right) f_{\mathbf{L}}(l,l_1,l_2)\, dl dl_1 dl_2, \tag{47}$$

and $\Pr(E_i)$, $i = 1,2,3,4$, corresponding to the four blockage situations of (LOS, Relay) scenario, are presented in (17), (18), (19) and (20), respectively. We observe that, except for E_1, the probability for two blockage models differs significantly for some range of p. If we maintain connectivity by choosing a lower threshold,

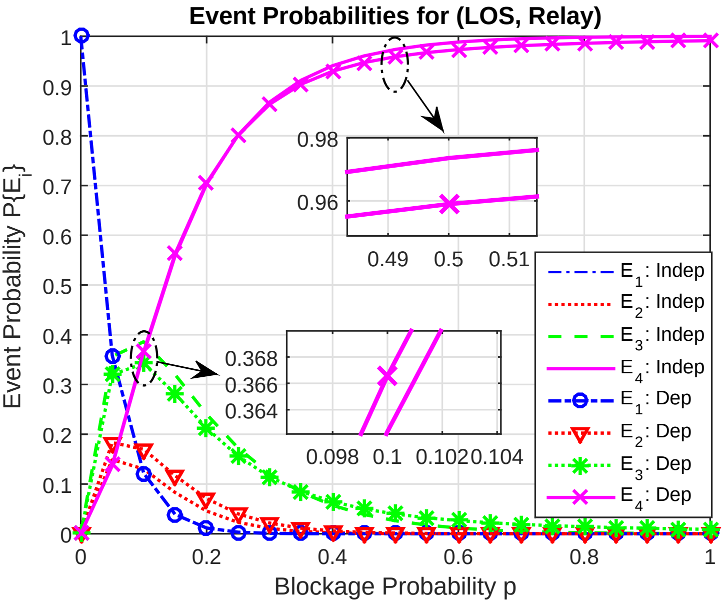


Figure 1.8: The event probabilities $\Pr(E_i)$, $i = 1, 2, 3, 4$, for the (LOS, Relay) scenario regarding the availability of the LOS and/or the relaying paths under topology-dependent or topology-independent models. $\Pr(E_i)$, $i = 1, 2, 3, 4$, correspond to four blockage situations presented in (17), (18), (19) and (20), respectively.

the outage probability $\Pr_{\text{out}}^{(1\_\text{relay})}(p) = \Pr(E_4)$ is shown in (29). From numerical evaluations we can see that the topology-dependent model has higher $\Pr(E_4)$ than that of the topology-independent model for $p \leq 0.23$ and smaller otherwise (as shown in the zoom-in details), which implies that the comparison of the two random blockage models should be related to the underlining link blockage probability, which coincides with what we conclude from Fig. 1.7.

In Fig. 1.9 we investigate the outage performance under the topology-independent random blockage model where the rate threshold is set to $R_\epsilon = R(2R_0) = 605$ Mbps. When p is small, the approximated outage probability given by (35) matches well with the simulation results. It also shows that the outage probability can be remarkably reduced by increasing the number of relays. Resorting to the reflection paths, however, can only provide marginal contribution for low blockage probability cases. This is in line with what we have observed in the simulation results for average throughput.

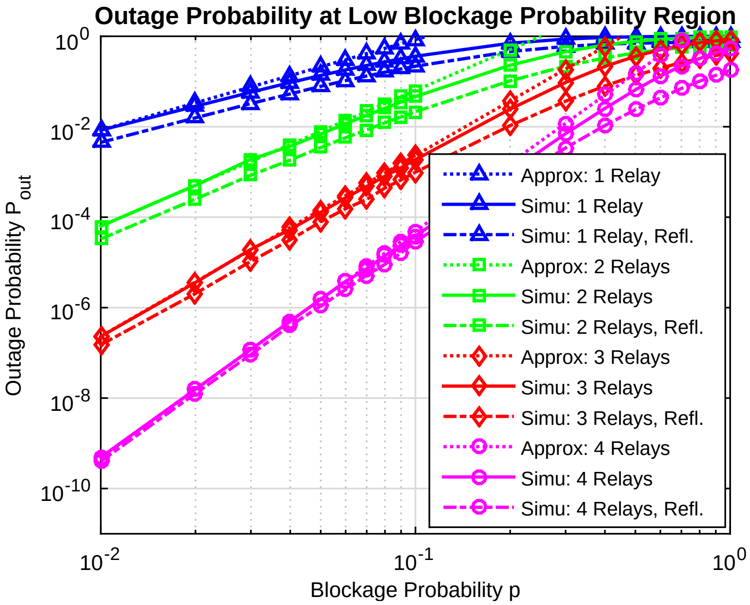


Figure 1.9: Outage probability for scenarios with different number of relays under the topology-independent model where the threshold rate is set to $R_\epsilon = 605$ Mbps. The approximated and simulated outage performance are also provided.

F Conclusions

We have investigated the throughput and outage performance of indoor 60 GHz communications over relay networks, where the communication between two nodes can be established either by the LOS path, via a half-duplex relaying node, or via the reflection path. A central coordinator is deployed to collect the topology information of nodes and to manage the path selection and scheduling. We consider both topology-independent and topology-dependent random blockage models and analyze the performance of average throughput and outage probability. To increase the throughput for networks with multiple relays, we propose the MTPS algorithm to select the optimal transmission path. Given n relay nodes, we show that the complexity of the proposed algorithm is $\mathcal{O}(n^2)$, in contrast to $\mathcal{O}(n \cdot n!)$ by the brute-force approach. Simulation results show that increasing the number of relays can substantially increase the average throughput and decrease the outage probability. Reflection path, on the other hand, is very useful to keep the connectivity and lower the outage probability when the link blockage probability is high. Furthermore, we have observed that different blockage models can significantly affect the performance of average throughput and outage probability. Therefore, it is crucial to determine the suitable types of random blockage models when analyzing and evaluating different communications schemes for a specific network configuration.

In future, we will explore the influence of interference from concurrent transmission on the performance of the MTPS algorithm. Besides, it would be interesting and crucial to determine which model is more suitable to model the random blockage events in different indoor situations, and what's the impact of the probabilistic path loss model [SSRRM15] on the design of optimal path selection algorithms.

G Appendices

G.1 Proof of Proposition B.1

We assume the propagation distance is much larger than the wavelength. We denote by E_{in} and E_{refl} the electric fields of incident and reflected waves at the reflection point, respectively. According to the *Fresnel Reflection Formula*, we have $E_{\text{refl}} = |\eta| E_{\text{in}}$, where η is given in (6). As shown in Fig. 1.2, the distance of the LOS path between two points $N_1(x_1, y_1)$ and $N_2(x_2, y_2)$ is $l = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ and the distance of the reflection path is $l^{'} = \sqrt{(x_1 - x_2)^2 + (y_1 + y_2)^2}$. The incident angle θ is determined as

$$\theta = \arctan \left| \frac{x_2 - x_1}{y_2 + y_1} \right| . \tag{48}$$

The path loss of the LOS path and the reflection path are therefore (measured in dB)

$$L_{N_1 N_2}(l) = 20 \log \left(\frac{4\pi}{\lambda} \right) + 10n \log(l), \tag{49}$$

$$L_{N_1^{'} N_2}(l^{'}, \theta) = 20 \log \left(\frac{4\pi}{\lambda} \right) + 10n \log(l^{'}) - 20 \log |\eta(\theta)|. \tag{50}$$

Subtracting (49) from (50) provides (5).

In a special case with $y_1 = y_2 = h$, we have $l = |x_1 - x_2|$ and $l^{'} = \sqrt{d^2 + 4h^2}$ and the incident angle θ can be obtained by $\theta = \arctan\left(\frac{d}{2h}\right)$. Once h is known and fixed, both $l^{'}$ and θ can be regarded as the functions of l and therefore $\chi(l, l^{'}, \theta)$ can be reduced to a function of only l for given h,

$$\chi(l) = 5n \log \left(1 + \frac{4h^2}{l^2} \right) - 20 \log \left| \frac{\sqrt{(\omega-1)l^2 + 4\omega h^2} - 2\omega h}{\sqrt{(\omega-1)l^2 + 4\omega h^2} + 2\omega h} \right| . \tag{51}$$

G.2 Derivation of Equation (32)

As depicted in Fig. 1.10, N_1 and N_2 are the user terminals, and R_1 and R_2 represent two relays, respectively. The label b_k, $k \in \{0, 1, \ldots, 5\}$ on each arrow of the graph denotes the blockage event over the corresponding link. Clearly, there are four link blockage patterns X_i, $i = 1, 2, 3, 4$, to disconnect the communication between N_1

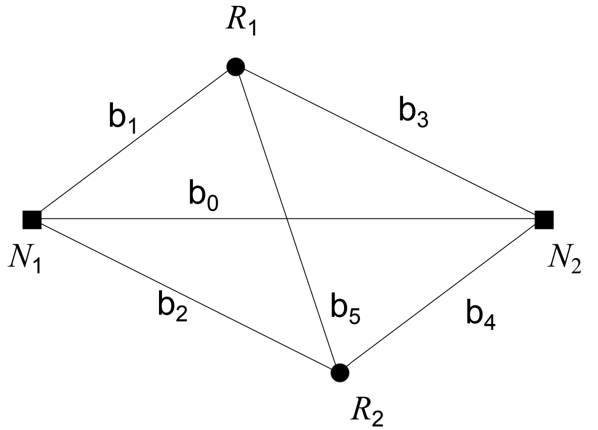


Figure 1.10: The scenario with two relays, namely R_1 and R_2, where N_1 and N_2 are two wireless devices.

and N_2, where $X_1 = b_0 b_1 b_2$, $X_2 = b_0 b_3 b_4$, $X_3 = b_0 b_1 b_5 b_4$ and $X_4 = b_0 b_2 b_5 b_3$. The outage happens if any of the four link blockage patterns occurs. Thus, the outage probability is given by:

$$\underset{\text{out}}{\overset{(2\_\text{relay})}{\Pr}} = \Pr\left(X_1 \cup X_2 \cup X_3 \cup X_4\right)$$
$$= \sum_{i=1}^{4} \Pr\left(X_i\right) - \sum_{i<j} \Pr\left(X_i X_j\right) + \sum_{i<j<k} \Pr\left(X_i X_j X_k\right) - \sum_{i<j<k<l} \Pr\left(X_i X_j X_k X_l\right). \tag{52}$$

By expressing X_i, $i = 1, 2, 3, 4$ in b_k, $k \in \{0, 1, \ldots, 5\}$ and using the inclusion-exclusion principle iteratively, the outage probability can be obtained.

Paper J2

Performance Analysis of Millimeter-Wave Relaying: Impacts of Beamwidth and Self-Interference

Guang Yang and Ming Xiao

Refereed article published in
IEEE Transactions on Communications, vol. 66, no. 2, Feb. 2018.

Performance Analysis of Millimeter-Wave Relaying: Impacts of Beamwidth and Self-Interference

Guang Yang and Ming Xiao

Abstract

We study the maximum achievable rate of a two-hop amplified-and-forward (AF) relaying millimeter-wave (mm-wave) system, where two AF relaying schemes, i.e., half-duplex (HD) and full-duplex (FD) are discussed. By considering the two-ray mm-wave channel and the Gaussian-type directional antenna, jointly, the impacts of the beamwidth and the self-interference coefficient on maximum achievable rates are investigated. Results show that, under a sum-power constraint, the rate of FD-AF mm-wave relaying outperforms its HD counterpart only when antennas with narrower beamwidth and smaller self-interference coefficient are applied. However, when the sum-power budget is sufficiently high or the beamwidth of directional antenna is sufficiently small, direct transmission becomes the best strategy, rather than the AF relaying schemes. For both relaying schemes, we show that the rates of both AF relaying schemes scale as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$ *with respect to beamwidth* θ_m*, and the rate of FD-AF relaying scales as* $\mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$ *with respect to self-interference coefficient* μ*. Besides, we show that, ground reflections may significantly affect the performance of mm-wave communications, constructively or destructively. Thus, the impact of ground reflections deserves careful considerations for analyzing or designing future mm-wave wireless networks.*

A Introduction

Thanks to abundant spectrum resources, wireless communications at millimeter-wave (mm-wave) bands (ranging from around 24 GHz to 300 GHz) is a key enabler for fulfilling the multi-Gbps rates in future mobile communications [XMH$^+$17, RSM$^+$13]. However, mm-wave communications still need to overcome many challenges. A typical challenge is the severe path loss. One common solution is to use highly directional antennas. Normally, a directional antenna with narrower beamwidth has a higher antenna gain at the main lobe, which can compensate for the path loss, and hence can significantly improve the transmission distance and reduce the outage probability [RRS$^+$05, YYMZ06, Epp06, WNLa$^+$14]. In previous works, the radiation pattern of directional antennas was usually modeled in an idealized fashion, i.e., a large constant antenna gain within the narrow main-lobe and zero elsewhere. This idealized radiation pattern (often referred to as the "flat-top" model) was widely used [WNE02, KPL03, SMM11] for performance analysis, due to its high tractability. Evidently, the side-lobe effect is neglected in the flat-top model, which however is not negligible when the beamwidth is less than 60°, as demonstrated in [Ram01]. In recent years, numerous efforts have been dedicated in directional antennas, and the effect of side-lobe leakages is incorporated. For instance, in [BB+10, AEAH12], a piece-wise model was applied. In [WKW10, WNLa+14, SGFF+15, BH15], a sectorized model that considers the side-lobe leakage is employed, which has both constant gains in the main and side lobes, respectively.

In addition to the severe path loss, mm-wave communications suffer high penetration loss and weak diffusion, diffraction, and high-order reflections [GKZV09]. Therefore, mm-wave radios are more suitable for short-range wireless systems. To achieve the higher coverage and robustness, relaying techniques may be used for mm-wave communications. According to the duplex mode, relaying techniques can be categorized as half-duplex (HD) and full-duplex (FD). In contrast to the HD relay, the FD relay can support simultaneous reception and transmission, while it suffers from the self-interference. Numerous efforts have been devoted to exploring potentials of performance enhancement via relays in mm-wave communications [GOON10, DSZC13, RST14, LJT14]. In [WZS+15], the energy efficiency was studied for 60 GHz indoor networks with amplify-and-forward (AF) relay. The comparison between two duplex modes was performed, and the impacts of imperfect self-interference suppression, drain efficiency and static circuit power were also studied. In recent works [WZS+16b, WZS+16a], the algorithm of energy-efficient cross-layer resource allocation with for the FD decode-and-forward (DF) relaying was developed.

However, for mm-wave relaying systems with directional antennas, there are two major limitations in the existing works:

(i) The sectorized model (e.g., [WNLa+14, SGFF+15, BH15]) is a prevailing option when modeling the radiation pattern of directional antennas, due to its high tractability for analysis. However, in the sectorized model, only two constant gains are used to characterize the main and side lobes, respectively, without any transition between them. An obvious drawback of this idealization is that, the crucial "roll-off" feature (a gradual decay from the main lobe to the side lobe) of the real-world radiation pattern for directional antennas is not reflected, and the resulting discontinuity may seriously affect the system performance evaluations.

(ii) It has been shown in [GRON10, YDX15, LvdBS+16, VH16] that, first-order reflections are not negligible in mm-wave communications. However, in most of preceding works (e.g., [GKZV09, SGFF+15, SKGA15, TBHJ16]), the impacts of ground reflections (first-order reflections) are rarely incorporated, since it is widely and deeply believed that the ground reflection is not a dominant factor that can dramatically affect performance evaluations. This conventional channel model for mm-wave radios (based on the LOS path only) may lead to an obvious overestimation (resp., underestimation) in performance evaluations, due to the omission of the non-negligibly constructive (resp., destructive) effects of reflections.

It is evident that, the aforementioned limitations may result in obvious inaccuracy when analyzing the performance of mm-wave wireless communications. To address above two problems, two heuristic models are considered in this paper. For the radiation pattern of directional antennas, the *Gaussian-type directional antenna model* [TSI06,Gag12,MiW14,TH15] is used, where the main-lobe gain attenuates to a non-zero side-lobe gain in the continuous manner, such that the "roll-off" feature of the real radiation pattern can be seized, while preserving the tractability. Furthermore, to incorporate ground reflections in mm-wave channels, the *two-ray channel model* [R$^+$96,Gol05,RJLMPG17] is employed. Due to the fact that ground normally acts as the commonest reflective surface in various scenarios[1], we consider ground reflections only in our study. The use of the two-ray model in mm-wave communications is fairly recent, which can be found in [JLGH09,WPK$^+$15,HSA$^+$13,HOI$^+$16].

The objective of our paper is to study the maximum achievable rates of a two-hop AF relaying system with mm-wave. More precisely, we first formulate the rates by HD-AF and FD-AF relaying schemes, respectively, where the optimized time-sharing scheme for the HD mode and optimal power allocations are studied. To the best of our knowledge, we are the first to consider a joint treatment of antenna model and channel model, which is important for analyzing the performance of mm-wave communications. Subsequently, we investigate the impacts of the beamwidth and the self-interference cancellation on mm-wave communications with AF relays. The main contributions of our work can be summarized from the following aspects:

- To overcome the limitations induced by the conventional oversimplification of the directional antenna and the propagation model for mm-wave communications, we consider the Gaussian antenna model and the two-ray channel model, *jointly*[2]. Note that, there exists a trade-off between the main-lobe and side-lobe gains in the Gaussian antenna model when varying the beamwidth. Paired with the two-ray channel model, evidently, the distribution of signal strength on direct and reflected propagation paths heavily depends on the beamwidth. Thus, a joint treatment of two heuristic models is essential, and it enables performance evaluations of mm-wave systems in a more accurate way.

- In terms of main-lobe beamwdith θ_m, we demonstrate that the maximum rates of AF relaying schemes scale as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$. Furthermore, let

[1] The channel model, of course, can be generalized to a "multi-ray" version, if multiple dominant reflective objects simultaneously exist.

[2] It is worth mentioning that, the Gaussian antenna model itself is not novel. However, the joint treatment of Gaussian antenna model and two-ray channel model has not been performed previously, which is one of our main contributions.

self-interference coefficient $\mu \in (0,1)$ characterize the self-interference cancellation. That is, a smaller μ indicates better self-interference cancellation, we find that the rate of FD-AF relaying scales as $\mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$. The convexity of the decay in the achievable rate with respect to μ reveals great benefits by strengthening the self-interference cancellation in FD-AF relay implementations. These results give a more convenient way to keep track of the performance, and provide insights for the future system design.

- In contrast to the conventional belief that the effect of reflections is negligible in mm-wave communications, our results show that, the constructive or destructive contribution of ground reflections is indeed nontrivial. Note that, the ground reflection heavily relies on the incident angle, the radiation pattern of directional antenna, and the transmission distance, the contribution of ground reflections may be fairly obvious. Thus, careful considerations regarding the impacts of ground reflections are valuable for analyzing or designing mm-wave networks.

The rest of the paper is organized as follows. In Sec. B, we give a system for two-hop AF relaying with mm-wave, where the two-ray channel and Gaussian antennas are jointly considered. In Sec. C, the maximum achievable rates of two-hop relaying mm-wave systems by two relaying schemes are presented, respectively, where time-sharing and power allocations are incorporated. In Sec. D, the impacts of the beamwidth and the self-interference cancellation on the rates of two relaying schemes are comprehensively studied. Numerical results are provided in Sec. E, which are followed by conclusions drawn in Sec. F.

In our paper, we use "big-$\mathcal{O}$" notation to denote the variation of rates with respect to the beamwidth or the self-interference coefficient. The big-$\mathcal{O}$ notation is defined as: let $u(x)$ and $f(x)$ be two functions defined on some subset of the real numbers, then we let $u(x) = \mathcal{O}(f(x))$ whenever $|u(x)/f(x)|$ is upper bounded for all feasible x.

B System Model

We consider a two-hop AF relaying system with mm-wave, which consists of a source node S, a destination node D, and a relay node R (in the HD or FD mode)[3], as illustrated in Fig. 1.1. All nodes are equipped with directional antennas. For notational simplicity, we denote by $h_i \in \mathbb{C}$, $i \in \{1,2\}$ the channel coefficients of links $S-R$ and $R-D$, respectively. We denote by H_S, H_R and H_D the deployment height of S, R and D, respectively. In addition, for $S-R$ and $R-D$ links, we let L_1 and L_2 be the respective horizontal separation distances.

In this section, we elaborately show the two-ray channel model and the Gaussian antenna model, respectively, which are used for our subsequent performance

[3] We assume that there exists no direct link between S and D in our scenario (e.g., due to extremely severe blockage or path loss).

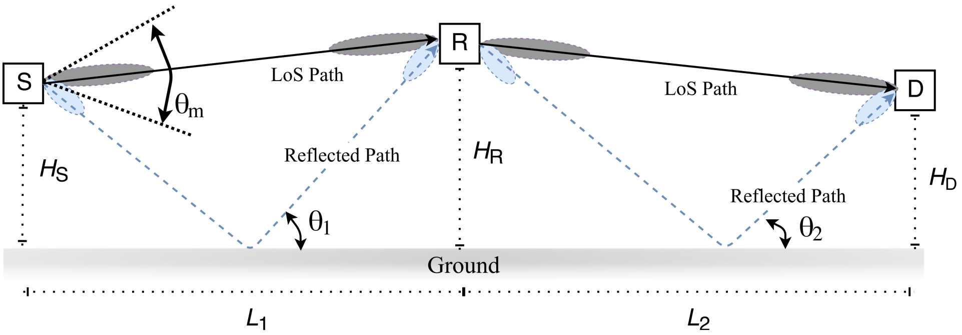


Figure 1.1: Two-hop AF relaying system: two-ray channel and directional antennas.

analysis for mm-wave communications. At mm-wave bands, the multi-path effect resulted by scattering is negligible, since electromagnetic waves with short wavelength have weak capability of diffraction or high-order reflections [GKZV09]. Also, the shadowing effect happens only when obstacles emerge in the propagation path of mm-wave radios, and severe shadowing may result in the link blockage. To mitigate the detrimental effect by shadowing, one can elevate the deployment heights of nodes, i.e., H_S, H_R, and H_D, such that the probability that obstacles appear in wave propagation path can be reduced. For focusing on impacts of the beamwidth and self-interference, both multi-path and shadowing effects are not considered for mm-wave channels in our study.

B.1 Two-Ray Channel

The two-ray model is a classic channel model that considers two major coexisting transmission paths, i.e., the light-of-sight (LOS) path and the reflection path, and the feasibility of modeling mm-wave channels has been investigated in [JLGH09, WPK+15, RJLMPG17]. The two-ray channel model is shown in Fig. 1.1.

By Friis transmission formula, the received signal power is written as $P_r = P_t \cdot |h_i|^2$, where P_t denotes the transmit power, and h_i for $i = 1, 2$ denotes the channel coefficient. In the presence of reflections from the ground surface, h_i is exactly written as

$$\begin{cases} h_1 = \dfrac{\lambda\left(G\left(0\right) + G\left(\theta_1\right)\Gamma\left(\theta_1\right)\cos\left(\theta_1\right)e^{-j\Delta\varphi_1}\right)}{4\pi\sqrt{\left(H_S - H_R\right)^2 + L_1^2}} \\ h_2 = \dfrac{\lambda\left(G\left(0\right) + G\left(\theta_2\right)\Gamma\left(\theta_2\right)\cos\left(\theta_2\right)e^{-j\Delta\varphi_2}\right)}{4\pi\sqrt{\left(H_R - H_D\right)^2 + L_2^2}} \end{cases}, \tag{1}$$

where λ is the wavelength of mm-wave radio, $G\left(*\right)$ denotes the radiation pattern of directional antenna (to be elaborated subsequently), and θ_i denotes the reflection angle relative to the ground plane, which is given as

$$\begin{cases} \theta_1 = \arctan\left(\dfrac{H_S + H_R}{L_1}\right) \\ \theta_2 = \arctan\left(\dfrac{H_R + H_D}{L_2}\right) \end{cases}.$$

Furthermore, the phase difference $\Delta\varphi_i$ for $i = 1, 2$, which is characterized by the length difference between the LOS path and the reflection path, is expressed as

$$\begin{cases} \Delta\varphi_1 = \dfrac{2\pi}{\lambda}\left(\sqrt{(H_S{+}H_R)^2{+}L_1^2} - \sqrt{(H_S{-}H_R)^2{+}L_1^2}\right) \\ \Delta\varphi_2 = \dfrac{2\pi}{\lambda}\left(\sqrt{(H_R{+}H_D)^2{+}L_2^2} - \sqrt{(H_R{-}H_D)^2{+}L_1^2}\right) \end{cases}.$$

In addition, reflection coefficient $\Gamma(\theta_i)$ associated with reflection angle θ_i [Gol05] is given as

$$\Gamma(\theta_i) = \frac{\sin\theta_i - Z(\theta_i)}{\sin\theta_i + Z(\theta_i)}.$$

Here, $Z(\theta_i)$ with respect to perpendicularly and horizontally polarized electromagnetic waves are respectively shown as

$$Z(\theta_i) = \begin{cases} \omega^{-1}\sqrt{\omega - \cos^2(\theta_i)}, & \text{perp. polarization} \\ \sqrt{\omega - \cos^2(\theta_i)}, & \text{horiz. polarization} \end{cases}, \tag{2}$$

where ω denotes the dielectric constant of ground.

B.2 Gaussian-Type Directional Antenna

For radiation pattern $G(*)$, we consider a Gaussian-type directional antenna model, which captures the "roll-off" feature of real radiation patterns [Gag12, MiW14, TH15], and matches well with measured results [TSI06]. For the validation of the Gaussian-type directional antenna model, interested readers are referred to related literature, e.g., [TSI06, Gag12, MiW14, TH15].

In the Gaussian-type model[4], let $\phi \in [-\pi, \pi)$ be the orientation angle relative to the boresight. With respect to the constraint for the total radiated power in all directions, i.e., $\int_{-\pi}^{\pi} G(\theta)d\theta = 2\pi$, the antenna gain along this orientation is given as

$$G(\phi) = \frac{2\pi}{V(\theta_m, \theta_h) + 2\pi - \theta_m} 10^{\frac{3}{10}\left[\frac{\theta_m^2 - 4\phi^2}{\theta_h^2}\right]_+}, \tag{3}$$

[4] Slightly different from the model adopted in [TSI06, Gag12, MiW14, TH15], we in this paper particular introduce the constraint of total radiated power in all directions. Details are provided in the subsequent section.

where $[*]_+ \triangleq \max\{*, 0\}$, θ_h is the half-power beamwidth, θ_m is the main-lobe beamwidth, and $V(\theta_m, \theta_h)$ is defined as

$$V(\theta_m, \theta_h) \triangleq \int_0^{\theta_m} 10^{\frac{3}{10}\left(\frac{\theta_m^2 - x^2}{\theta_h^2}\right)} dx.$$

According to evaluations performed in [TSI06, MiW14], we have the empirical expression $\theta_m = 2.6 \cdot \theta_h$ for $\frac{\pi}{12} \leq \theta_m \leq \frac{\pi}{3}$. Thus, (3) can be further reduced to

$$G(\phi) = \frac{2\pi}{2\pi + 42.64443\theta_m} \cdot 10^{2.028\left[1-\left(\frac{2\phi}{\theta_m}\right)^2\right]_+}, \tag{4}$$

which will be used throughout the paper.

C Rates of AF Relaying Schemes

In this section, we study the maximum achievable rates of two-hop AF relaying system with mm-wave, as illustrated in Fig. 1.1. The channel gain of a two-ray mm-wave channel on the i^{th} hop for $i = 1, 2$ is characterized by $g_i \triangleq |h_i|^2$, following (1), and the antenna gain of Gaussian-type directional antenna follows the expression in (4).

For notational simplicity, we define ξ_i as the transmit power on the i^{th} hop. Besides, let s be the signal-to-noise ratio (SNR) or the signal-to-interference-plus-noise ratio (SINR), then we use $C(s) \triangleq \log_2(1+s)$ to denote the achievable rate. In our study, sum-power constraint is considered for both AF relaying schemes[5], which will be elaborated subsequently.

C.1 Half-Duplex Mode (HD-AF)

Let x be the signal transmitted from S, where $x \sim \mathcal{CN}(0,1)$ is independent and identically distributed (i.i.d.). Then, the received signal at R is given as

$$y_R = \sqrt{\xi_1} h_1 \cdot x + n_1,$$

where the AWGN n_1 at R follows $\mathcal{N}(0,1)$. After amplifying y_R by scaling coefficient a, i.e.,

$$a = \sqrt{\frac{\xi_2}{g_1 \xi_1 + 1}}.$$

Then, the amplified signal is subsequently sent to D from R, and the signal that finally reaches D is obtained as

$$y_D = a\sqrt{\xi_1} h_1 h_2 \cdot x + a h_2 \cdot n_1 + n_2,$$

[5] Sum-power constraint can be realized by a centralized power controller, and it is already used in many works on optimizing multi-hop networks. e.g., [Saa14, WSW11]. The study also can be extended to the scenario with per-node power constraint, which however is not considered in our study, since our main focus is to investigate the impacts of beamwidth and self-interference.

where the AWGN n_2 at D follows $\mathcal{N}(0,1)$.

Given a sum-power constraint, power allocation is performed to maximize the rate of HD-AF relaying. We consider the optimized time-sharing strategy for HD-AF relaying. A normalized time-sharing parameter $0 \leq \beta \leq 1$ is used to characterize the duration for the transmission from S to R, and $(1-\beta)$ is left for the transmission from R to D. Thus, the sum-power constraint that incorporates the time-sharing scheme can be interpreted as

$$\beta \xi_1 + (1-\beta)\,\xi_2 = \xi, \tag{5}$$

where ξ represents the sum-power constraint in terms of ξ_1 and ξ_2. In the following proposition, we formulate the maximum achievable rate of HD-AF by applying time sharing and power allocation.

Proposition C.1. *For two-hop HD-AF relaying, given sum-power constraint $\beta\xi_1 + (1-\beta)\,\xi_2 = \xi$, the maximum achievable rate is formulated as*

$$\eta^*_{\mathrm{HD}} = \max_{\substack{0 \leq \beta \leq 1 \\ \xi_1, \xi_2 \geq 0 \\ \beta\xi_1 + (1-\beta)\xi_2 = \xi}} \min \left\{ \begin{array}{c} \beta C\left(g_1\xi_1\right), \\ (1-\beta)\,C\left(\dfrac{g_1 g_2 \xi_1 \xi_2}{1 + g_1\xi_1 + g_2\xi_2}\right) \end{array} \right\}.$$

Proof. From S to R, it is clear that, the throughput with respect to duration β is given as

$$T_{S-R} = \beta \log_2\left(1 + g_1\xi_1\right) \triangleq \beta C\left(g_1\xi_1\right).$$

From R to D, due to the AF policy, the transmitted signal from relay is a scaled version of its received signal, and hence the throughput is obtained as

$$T_{R-D} = (1-\beta)\,C\left(\frac{g_1 g_2 \xi_1 \xi_2}{1 + g_1\xi_1 + g_2\xi_2}\right).$$

Since the end-to-end throughput is restricted to the bottleneck hop, the overall throughput is $T = \min\left\{T_{S-R}, T_{R-D}\right\}$. With (5), the optimization problem can be formulated. □

However, it is worth noting that, it is a non-trivial task to solve the optimization problem formulated in Proposition C.1. The major difficulty lies in the constraint introduced by (5), where the selection of the optimal time-sharing parameter and power allocation are intertwined with each other. More precisely, the optimal β depends on ξ_1 and ξ_2, while ξ_1 and ξ_2 are constrained to ξ in terms of β via (5). In this case, we are unable to decouple the constraints straightforwardly, such that it is intractable to obtain a closed-form solution for η^*_{HD}. However, in what follows, we will study the maximum achievable rate in a special case, i.e., with fixed $\beta = \frac{1}{2}$, which corresponds to HD-AF relaying with equally assigned time slots for two transmission phases. Thus, with fixed $\beta = \frac{1}{2}$, the optimization problem in Proposition C.1 can be reduced to

$$\begin{aligned}\eta^*_{\mathrm{HD}}\Big|_{\beta=\frac{1}{2}} &= \max_{\substack{\xi_1+\xi_2=2\xi\\ \xi_1,\xi_2\geq 0}} \frac{1}{2}C\left(\frac{g_1g_2\xi_1\xi_2}{1+g_1\xi_1+g_2\xi_2}\right)\\ &\triangleq \max_{\substack{\xi_1+\xi_2=2\xi\\ \xi_1,\xi_2\geq 0}} \frac{1}{2}C\left(\gamma_{\mathrm{HD}}\left(\xi_1,\xi_2\right)\right)\\ &= \frac{1}{2}C\left(\max_{\substack{\xi_1+\xi_2=2\xi\\ \xi_1,\xi_2\geq 0}} \gamma_{\mathrm{HD}}\left(\xi_1,\xi_2\right)\right).\end{aligned}$$

For the maximum SNR, denoted by γ^*_{HD}, we have

$$\begin{aligned}\gamma^*_{\mathrm{HD}} \triangleq \max_{\substack{\xi_1+\xi_2=2\xi\\ \xi_1,\xi_2\geq 0}} \gamma_{\mathrm{HD}}\left(\xi_1,\xi_2\right) &= \max_{\substack{\xi_1+\xi_2=2\xi\\ \xi_1,\xi_2\geq 0}} \frac{g_1g_2\xi_1\xi_2}{1+g_1\xi_1+g_2\xi_2}\\ &= \left(\min_{0\leq\xi_1\leq 2\xi} \frac{1+\left(2g_2\xi\right)^{-1}}{g_1\xi_1} + \frac{1+\left(2g_1\xi\right)^{-1}}{g_2\left(2\xi-\xi_1\right)}\right)^{-1}\\ &= \frac{4g_1g_2\xi^2}{\left(\sqrt{1+2g_1\xi}+\sqrt{1+2g_2\xi}\right)^2},\end{aligned}$$

which concludes the maximum spectral efficiency, i.e.,

$$\eta^*_{\mathrm{HD}}\Big|_{\beta=\frac{1}{2}} = \frac{1}{2}C\left(\frac{4g_1g_2\xi^2}{\left(\sqrt{1+2g_1\xi}+\sqrt{1+2g_2\xi}\right)^2}\right), \tag{6}$$

whenever the power allocation below can be applied:

$$\begin{cases}\xi_1^* = \dfrac{2\xi\sqrt{1+2g_2\xi}}{\sqrt{1+2g_1\xi}+\sqrt{1+2g_2\xi}}\\ \xi_2^* = \dfrac{2\xi\sqrt{1+2g_1\xi}}{\sqrt{1+2g_1\xi}+\sqrt{1+2g_2\xi}}\end{cases}.$$

C.2 Full-Duplex Mode (FD-AF)

In FD-AF relaying, $\mu \in (0,1)$ is the self-interference coefficient, which characterizes the capability of self-interference techniques[6]. That is, a smaller μ indicates a more powerful self-interference suppression. Let $x\left[k\right]$ denote the transmitted signal from

[6] Commonly, the self-interference coefficient is determined by the back-end implementation, circuit design or signal processing techniques, which is independent of the beamwidth of directional antennas at the front-end.

S at time slot k, and we assume it is i.i.d. over time slots. At R, it is easy to obtain that

$$\begin{cases} y_R^{(r)}[k] = \sqrt{\xi_1} h_1 \cdot x[k] + \sqrt{\mu} y_R^{(t)}[k-1] + n_1[k] \\ y_R^{(t)}[k] = a \cdot y_R^{(r)}[k] \end{cases},$$

where $y_R^{(r)}[k]$ and $y_R^{(t)}[k]$ represent the received and transmitted signals at R in time slot k, respectively, the variance of AWGN $n_1[k]$ is 1, and the scaling coefficient a is given as

$$a = \sqrt{\frac{\xi_2}{\xi_1 g_1 + 1 + \mu \xi_2}}.$$

Finally, the received signal at D can be expressed as

$$y_D[k] = a h_2 \left(\sqrt{\xi_1} h_1 x[k] + \sqrt{\mu} y_R^{(t)}[k-1] + n_1[k] \right) + n_2[k],$$

where the variance of AWGN $n_2[k]$ is 1. Thus, the SINR can be easily obtained as

$$\gamma_{\mathrm{FD}}(\xi_1, \xi_2) = \frac{g_1 g_2 \xi_1 \xi_2}{(g_2 \xi_2 + 1)(\mu \xi_2 + 1) + g_1 \xi_1}, \tag{7}$$

and the resulting rate is given by $\eta_{\mathrm{FD}} = C(\gamma_{\mathrm{FD}}(\xi_1, \xi_2))$.

Given a sum-power constraint, in following Proposition C.2, the maximum achievable rate of FD-AF relaying by power allocation is presented, where the sum-power constraint is given as $\xi_1 + \xi_2 = \xi$.

Proposition C.2. *For two-hop FD-AF relaying, given the sum-power constraint $\xi_1 + \xi_2 = \xi$, the maximum achievable rate is*

$$\eta_{\mathrm{FD}}^* = C\left(\frac{g_1 g_2 \xi^2}{2 + (g_1 + g_2 + \mu)\xi + 2\sqrt{(1 + g_1 \xi)(1 + g_2 \xi)(1 + \mu \xi)}} \right),$$

where the optimal power allocation is given as

$$\begin{cases} \xi_1^* = \dfrac{\xi \sqrt{1 + (\mu + g_2)\xi + \mu g_2 \xi^2}}{\sqrt{1 + g_1 \xi} + \sqrt{1 + (\mu + g_2)\xi + \mu g_2 \xi^2}} \\ \xi_2^* = \dfrac{\xi \sqrt{1 + \xi g_1}}{\sqrt{1 + \xi g_1} + \sqrt{1 + (\mu + g_2)\xi + \mu g_2 \xi^2}} \end{cases}.$$

Proof. We know that $\eta_{\mathrm{FD}}^* = C(\gamma_{\mathrm{FD}}^*)$, where γ_{FD}^* denotes the maximized SINR at destination, i.e.,

$$\begin{aligned}\gamma_{\mathrm{FD}}^{*} &\triangleq \max_{\substack{\xi_1+\xi_2=\xi \\ \xi_1,\xi_2\geq 0}} \gamma_{\mathrm{FD}}\left(\xi_1,\xi_2\right) = \max_{\substack{\xi_1+\xi_2=\xi \\ \xi_1,\xi_2\geq 0}} \frac{g_1 g_2 \xi_1 \xi_2}{\left(g_2\xi_2+1\right)\left(1+\mu\xi_2\right)+g_1\xi_1} \\ &= \left(\min_{0\leq\xi_1\leq\xi} \frac{1+\mu\xi+\frac{\mu+\frac{1}{\xi}}{g_2}}{g_1\xi_1}+\frac{1+\frac{1}{g_1\mu}}{g_2\xi_2}-\frac{\mu}{g_1}\right)^{-1} \\ &= \frac{g_1 g_2 \xi^2}{\left(\sqrt{1+g_1\xi}+\sqrt{1+\left(\mu+g_2\right)\xi+\mu g_2\xi^2}\right)^2-\mu g_2\xi^2} \end{aligned} \tag{8}$$

whenever the power allocation below can be applied:

$$\begin{cases} \xi_1^{*} = \dfrac{\xi\sqrt{1+\left(\mu+g_2\right)\xi+\mu g_2\xi^2}}{\sqrt{1+g_1\xi}+\sqrt{1+\left(\mu+g_2\right)\xi+\mu g_2\xi^2}} \\ \xi_2^{*} = \dfrac{\xi\sqrt{1+\xi g_1}}{\sqrt{1+\xi g_1}+\sqrt{1+\left(\mu+g_2\right)\xi+\mu g_2\xi^2}} \end{cases}.$$

Thus, the proposition can be concluded. □

D Impacts of Beamwidth and Self-Interference Cancellation on Mm-wave AF Relaying

Based on the general results presented in Sec. C, in this section, we investigate the impacts of beamwidth and self-interference coefficient, respectively, on the maximum achievable rates of two-hop AF relaying systems with mm-wave.

D.1 Impact of Beamwidth

By Propositions C.1 and Propositions C.2, we know that both η_{HD}^{*} and η_{FD}^{*} monotonically increase in $g_i \triangleq |h_i|^2$. It is known from (1) that, h_i is related to the beamwidth of directional antennas, since $G(0)$ and $G(\theta_i)$ are determined by θ_m, as shown in (4). Thus, θ_m determines g_i, and hence affects the rates of relaying systems. It is easy to find that, (4) can be reformulated as

$$G\left(\theta_i\right) = G\left(0\right)\epsilon^{\min\left(1,\left(\frac{2\theta_i}{\theta_m}\right)^2\right)},$$

where $\epsilon = 0.0094$. Thus, g_i can be equivalently written as

$$g_i = \left(\frac{B_i}{1+\tau\theta_m}\right)^2\left|1+z_i\cdot\epsilon^{\min\left(1,\left(\frac{2\theta_i}{\theta_m}\right)^2\right)}\right|^2, \tag{9}$$

where, for notational simplicity, τ, B_i and z_i are respectively given as

$$\begin{cases} \tau \triangleq \dfrac{42.6443}{2\pi} \in \mathbb{R} \\ B_1 \triangleq \dfrac{\lambda}{4\pi\sqrt{\left(H_S - H_R\right)^2 + L_1^2}} \cdot 10^{2.028} \in \mathbb{R} \\ B_2 \triangleq \dfrac{\lambda}{4\pi\sqrt{\left(H_R - H_D\right)^2 + L_2^2}} \cdot 10^{2.028} \in \mathbb{R} \\ z_i \triangleq \Gamma\left(\theta_i\right) \cdot \cos\theta_i \cdot \exp\left(-j\Delta\varphi_i\right) \in \mathbb{C} \end{cases}.$$

In the following lemma, we demonstrate the decaying rule of g_i with respect to θ_m.

Lemma D.1. *With the beamwidth θ_m, channel gain g_i scales as $\mathcal{O}\left(\theta_m^{-2}\right)$.*

Proof. Let θ_i be the given reflection angle. According to (9), we have

$$\begin{aligned} g_i &< \left(\frac{B_i}{\tau}\right)^2 \cdot \left|1 + z_i \cdot \epsilon^{\min\left(1, \left(\frac{2\theta_i}{\theta_m}\right)^2\right)}\right|^2 \cdot \theta_m^{-2} \\ &\le \left(\frac{B_i}{\tau}\right)^2 \cdot \left|1 + |z_i| \cdot \epsilon^{\min\left(1, \left(\frac{2\theta_i}{\theta_m}\right)^2\right)}\right|^2 \cdot \theta_m^{-2} \\ &< \left(\frac{B_i \cdot \left(1 + |z_i|\right)}{\tau}\right)^2 \cdot \theta_m^{-2}, \end{aligned}$$

where the last line is obtained due to the fact that $\min\left(1, \frac{2\theta_i}{\theta_m}\right)$ is a non-increasing function of θ_m. Note that $\epsilon = 0.0094$ and $0 < |z_i| < 1$, it can be concluded that $g_i = \mathcal{O}\left(\theta_m^{-2}\right)$. □

Furthermore, we present series of inequalities in the following lemma.

Lemma D.2. *For $\forall x \ge 0$, we have*

$$\ln\left(1 + x\right) \le \frac{x}{\sqrt{1 + x}} \le \min\left\{\sqrt{x}, x\right\}.$$

Proof. For the first inequality, we define

$$f\left(x\right) \triangleq \ln\left(1 + x\right) - \frac{x}{\sqrt{1 + x}},$$

then the first derivative with respect to x is given as

$$\frac{df\left(x\right)}{dx} = \frac{2\sqrt{1 + x} - x - 2}{2\left(1 + x\right)^{\frac{3}{2}}} = -\frac{\left(\sqrt{1 + x} - 1\right)^2}{2\left(1 + x\right)^{\frac{3}{2}}}.$$

It is clear that, the first derivative of $f(x)$ over $x \geq 0$ is definitely non-positive, which indicates $f(x) \leq f(0) = 0$ for all $x \geq 0$. For the second inequality, it can be easily concluded by considering if $x \geq 1$ or $0 \leq x \leq 1$. □

In light of above two lemmas, the impact of the beamwidth on the maximum achievable rates of the AF relaying systems with mm-wave is revealed in the following theorem.

Theorem D.1. *With respect to beamwidth θ_m, the maximum achievable rate of two-hop AF relaying with mmWvae (given in Proposition C.1 or C.2) scales as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$.*

Proof. Please see Appendix G.1. □

The result in Theorem D.1 is unsurprising, and it coincides with the intuition that the directional antenna with a narrower beamwidth enables a higher achievable rate. The main contribution of Theorem D.1 is that, we characterize the rates of AF relaying schemes in the concise manner, with respect to the beamwidth of the specific Gaussian-type directional antenna.

Subsequently, we investigate the contribution of the reflection path relative to the LOS path. We know that, h_i consists of the LOS component and the reflection component, i.e.,

$$h_i = \underbrace{\frac{\lambda G(0)}{4\pi L_i}}_{\triangleq h_i^{\mathrm{LOS}} \in \mathbb{R}} + \underbrace{\frac{\lambda G(\theta_i)}{4\pi L_i} \cdot z_i}_{\triangleq h_i^{\mathrm{ref}} \in \mathbb{C}} \in \mathbb{C}.$$

We have $\left|h_i^{\mathrm{LOS}}\right| > \left|h_i^{\mathrm{ref}}\right|$, since the reflection component suffers the reflection loss (partial absorption by reflective material) and the additional path loss (due to extra transmission distance for the reflection path).

We define $g_i^{\mathrm{LOS}} \triangleq \left|h_i^{\mathrm{LOS}}\right|^2$ and $g_i^{\mathrm{ref}} \triangleq \left|h_i^{\mathrm{ref}}\right|^2$. To characterize the contribution of the reflection path relative to the LOS path, we define relative contribution ζ_i in terms of g_i^{LOS} and g_i^{ref} as

$$\zeta_i \triangleq \frac{g_i^{\mathrm{ref}}}{g_i^{\mathrm{LOS}}} \in (0, 1). \tag{10}$$

Note that $\left||h_i| - |h_i^{\mathrm{LOS}}|\right| \leq \left|h_i^{\mathrm{ref}}\right|$ always holds, the lower and upper bounds of g_i in terms of g_i^{LOS} and ζ_i are given as

$$\left(1 - \sqrt{\zeta_i}\right)^2 \cdot g_i^{\mathrm{dir}} \leq g_i \leq \left(1 + \sqrt{\zeta_i}\right)^2 \cdot g_i^{\mathrm{dir}}.$$

Given ϵ and z_i, we provide the lower and upper bounds for ζ_i in the proposition below. Moreover, the non-increasing monotonicity of ζ_i in θ_m is also revealed.

Proposition D.1. *Given reflection angle θ_i, relative contribution ζ_i defined in* (10) *is monotonically non-decreasing in θ_m, and we further have*

$$0 < \epsilon^2 |z_i|^2 \leq \zeta_i < |z_i|^2 < 1.$$

Proof. By the definition of ζ_i, we have

$$\zeta_i = \frac{G^2\left(\theta_i\right)|z_i|^2}{G^2(0)} = \epsilon^{2\min\left(1,\left(\frac{2\theta_i}{\theta_m}\right)^2\right)} \cdot |z_i|^2.$$

Due to the monotonicity $\min\left(1, \frac{2\theta_i}{\theta_m}\right)$ with respect to θ_m, it is straightforward to conclude the bounds for ζ_i, as well as the non-decreasing monotonicity. □

Proposition D.1 indicates that, ground reflections affects the rate of AF relaying systems more when the directional antenna has a broader beamwidth. It is worth mentioning that, the contribution may be exhibited either in the constructive or the destructive way, since ζ_i is only counted by the absolute value.

D.2 Impact of Self-Interference Cancellation

We know that, the SINR for FD-AF relaying, denoted by γ_{FD}^*, heavily relies on μ (see Proposition C.2). In this subsection, we investigate the impact of the self-interference coefficient on the maximum achievable rate of FD-AF mm-wave relaying system, i.e., the impact of μ on $\eta_{\mathrm{FD}}^* = C\left(\gamma_{\mathrm{FD}}^*\right)$.

For simplifying illustration, we start with γ_{FD}^*, which can be rewritten as

$$\gamma_{\mathrm{FD}}^* = \frac{A_0}{A_1 + \xi\mu + A_2\sqrt{1+\xi\mu}}, \tag{11}$$

where A_0, A_1 and A_2 respectively denote

$$\begin{cases} A_0 = g_1 g_2 \xi^2 \\ A_1 = 2 + \left(g_1 + g_2\right)\xi \\ A_2 = 2\sqrt{\left(1+g_1\xi\right)\left(1+g_2\xi\right)} \end{cases}.$$

In the following lemma, we demonstrate the convexity and monotonicity of γ_{FD}^*.

Lemma D.3. *SINR γ_{FD}^* is strictly convex and monotonically decreasing in self-interference coefficient $\mu \in (0,1)$.*

Proof. In terms of μ, we can verify that, γ_{FD}^* is continuous and twice differentiable over $\mu \in (0,1)$. The first-order and second-order partial derivatives of (11), with respect to μ, are respectively given as

$$\frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu} = -\frac{\xi A_0\left(1 + \frac{A_2}{2\sqrt{1+\xi\mu}}\right)}{\left(A_1 + \xi\mu + A_2\sqrt{1+\xi\mu}\right)^2}$$

and

$$\frac{\partial^2 \gamma_{\mathrm{FD}}^*}{\partial \mu^2} \triangleq \frac{\partial}{\partial \mu}\left(\frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu}\right) \frac{\xi^2 A_0 \left(3A_2^2\sqrt{1+\xi\mu}\right)+8\left(1+\xi\mu\right)^{\frac{3}{2}}+A_2\left(8+A_1+9\xi\mu\right)}{4\left(1+\xi\mu\right)^{\frac{3}{2}}\left(A_1+\xi\mu+A_2\sqrt{1+\xi\mu}\right)^3}.$$

Evidently, we have

$$\frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu} < 0 \text{ and } \frac{\partial^2 \gamma_{\mathrm{FD}}^*}{\partial \mu^2} > 0,$$

which concludes the decreasing monotonicity and convexity, respectively. □

The decreasing monotonicity of γ_{FD}^* in μ, revealed in Lemma D.3, indicates the an upper bound for SINR, i.e.,

$$\gamma_{\mathrm{FD}}^* < \overline{\gamma}_{\mathrm{FD}}^* \triangleq \lim_{\mu \to 0} \gamma_{\mathrm{FD}}^* = \frac{A_0}{A_1 + A_2}.$$

Given any self-interference coefficient $\mu > 0$, to characterize the difference between γ_{FD}^* and $\overline{\gamma}_{\mathrm{FD}}^*$, we define the ratio of γ_{FD}^* to $\overline{\gamma}_{\mathrm{FD}}^*$, i.e.,

$$\kappa \triangleq \frac{\overline{\gamma}_{\mathrm{FD}}^*}{\gamma_{\mathrm{FD}}^*} = \frac{A_1 + \xi\mu + A_2\sqrt{1+\xi\mu}}{A_1 + A_2}, \tag{12}$$

and the increasing monotonicity of κ with respect to sum-power constraint ξ is presented in Proposition D.2.

Proposition D.2. *Given channel gain g_i on the i^{th} hop and any $\mu > 0$, the difference metric κ defined in* (12) *is monotonically increasing with respect to sum-power constraint ξ.*

Proof. Please see Appendix G.2. □

Proposition D.2 indicates that, in the presence of imperfect self-interference cancellation, i.e., $\mu \neq 0$, the performance difference between γ_{FD}^* and $\overline{\gamma}_{\mathrm{FD}}^*$, characterized by κ in the format of ratio, gets enlarged when increasing the sum-power budget. Therefore, FD-AF relaying with a smaller μ needs to be devised especially when sum-power budget ξ is high, such that all the power can be fully exploited for further improving the achievable rate.

Based on Lemma D.3, the monotonicity and the convexity of η_{FD}^* with respect to μ are shown in the following theorem.

Theorem D.2. *Achievable rate η_{FD}^* by Proposition C.2 is strictly convex and monotonically decreasing in self-interference coefficient $\mu \in (0, 1)$.*

Proof. Please see Appendix G.3. □

The convexity and monotonicity revealed in Theorem D.2 indicate the great importance of reducing the self-interference coefficient for FD-AF mm-wave relaying. Besides, it is revealed that $\eta_{\text{FD}}^{*} = \mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$ in Theorem D.3.

Theorem D.3. *Achievable rate η_{FD}^{*} scales as $\mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$, with respect to $\mu \in (0,1)$.*

Proof. Based on (11), we have

$$\eta_{\text{FD}}^{*} = C\left(\frac{A_0}{A_1 + \xi\mu + A_2\sqrt{1+\xi\mu}}\right) < C\left(\frac{A_0}{A_1 + \xi\mu + A_2\sqrt{\xi\mu}}\right).$$

It is worth noting that, $\sqrt{\mu} > \mu$ can be achieved for $\mu \in (0,1)$. Applying Lemma D.2, we further obtain that

$$\eta_{\text{FD}}^{*} < C\left(\frac{A_0}{\xi + A_2\sqrt{\xi}} \cdot \mu^{-1}\right) \leq \frac{1}{\ln 2}\sqrt{\frac{A_0}{\xi + A_2\sqrt{\xi}}} \cdot \mu^{-\frac{1}{2}}.$$

Thus, we can conclude that $\eta_{\text{FD}}^{*} = \mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$. □

It can be seen from Theorem D.3 that, the maximum achievable rate of FD-AF mm-wave relaying acceleratingly increases as reducing the self-interference coefficient, which is in accordance with the result by Theorem D.2, and again demonstrates the huge benefit of suppressing the self-interference.

The following proposition gives the condition of μ for FD-AF to outperform direct transmission or HD-AF relaying.

Proposition D.3. *Let $\chi \geq 0$ be the given maximum achievable rate of direct transmission or HD-AF relaying. For $\eta_{\text{FD}}^{*} \leq \chi$, we have*

$$\mu \in \begin{cases} (0,1), & \Psi_\chi \leq 1 \\ \left[\frac{\Psi_\chi^2 - 1}{\xi}, 1\right), & 1 < \Psi_\chi < \sqrt{1+\xi}\,, \\ \emptyset, & \Psi_\chi \geq \sqrt{1+\xi} \end{cases}$$

where Ψ_χ is given as

$$\Psi_\chi \triangleq \xi\sqrt{g_1 g_2\left(1 - \frac{1}{2^\chi - 1}\right)} - \sqrt{(1+g_1\xi)(1+g_2\xi)}.$$

Proof. For $\eta_{\text{FD}}^{*} = C\left(\gamma_{\text{FD}}^{*}\right) \leq \chi$, we have

$$\gamma_{\text{FD}}^{*} = \frac{A_0}{A_1 + \xi\mu + A_2\sqrt{1+\xi\mu}} \leq 2^\chi - 1.$$

Applying the change of variables $v = \sqrt{1+\xi\mu}$, it is easy to obtain that

$$\begin{aligned} v \geq & \sqrt{1+\frac{A_0}{2^{\chi}-1}-A_1+\frac{A_2^2}{4}}-\frac{A_2}{2} \\ = & \xi\sqrt{g_1 g_2\left(1-\frac{1}{2^{\chi}-1}\right)}-\sqrt{\left(1+g_1\xi\right)\left(1+g_2\xi\right)} \triangleq \Psi_{\chi}. \end{aligned}$$

According to the possible values of Ψ_χ and the condition that $1 < v \leq \sqrt{1+\xi}$, we can solve μ via recovering μ in terms of v, i.e., $\mu = \xi^{-1}\left(v^2-1\right)$, which then completes the proof of the proposition. □

Briefly, given channel gains g_i $(i = 1, 2)$ and sum-power constraint ξ, Ψ_χ characterizes the resulting performance relative to threshold χ. Regarding the range of Ψ_χ, critical value 1 characterizes the minimum requirement on channel gains and sum power for Ψ_χ to achieve χ (given perfect self-interference cancellation, i.e., $\mu \to 0$), while critical value $\sqrt{1+\xi}$ characterizes the maximum requirement on channel gains and sum power for Ψ_χ to achieve χ (assuming no self-interference cancellation, i.e., $\mu \to 1$). Thus, by Proposition D.3, η^*_{FD} is definitely less than χ if channel gains and sum power for Ψ_χ cannot meet the minimum requirement, i.e., $\Psi_\chi \leq 1$, while η^*_{FD} definitely exceeds χ if channel gains and sum power for Ψ_χ can meet the maximum requirement, i.e., $\Psi_\chi \geq \sqrt{1+\xi}$. For Ψ_χ between 1 and $\sqrt{1+\xi}$, η^*_{FD} can achieve the given χ, conditioning on certain feasible set for μ, i.e., $\xi^{-1}\left(\Psi_\chi^2-1\right) \leq \mu < 1$.

E Numerical Results

In this section, the maximum achievable rates of two AF relaying schemes with mm-wave by Proposition C.1 and C.2, respectively, are demonstrated. Moreover, the impacts of the beamwidth and the self-interference coefficient are shown. We assume the used mm-wave radio is perpendicularly polarized (refer to (2)). In addition, we assume that nodes S, R and D are placed in a line, i.e., $L_1+L_2 = L$, and deployed at with the identical height, i.e., $H_S = H_R = H_D = H$. Several key system parameters and corresponding notations are summarized in Table 1.3.

E.1 Performance Comparison

With respect to sum-power constraint ξ, the maximum achievable rates of direct transmission, HD-AF and FD-AF mm-wave relaying systems are provided in Fig. 1.2. For HD-AF relaying, rate η^*_{HD} by Proposition C.1 and that with fixed $\beta = \frac{1}{2}$ (refer to (6)) are both provided. It is shown that, compared to the case of fixed $\beta = \frac{1}{2}$, a performance gain, i.e., ≈ 5 dB, can be achieved when applying the optimized time-sharing scheme. For FD-AF relaying, according to Proposition C.2, the achievable rates with different self-interference coefficients are shown. Evidently,

Table 1.3: System Settings

Parameter	Notation
Wavelength	$\lambda = 5 \times 10^{-3}$ m
Sum-Power Constraint	ξ
Main-lobe Beamwidth	θ_m
Self-interference Coefficient	μ
Deployment Height	$H = 5$ m
Source-Destination Distance	$L = 200$ m
Source-Relay Distance	L_1
Ground Dielectric Constant	$\omega \approx 15$ [Gol05]

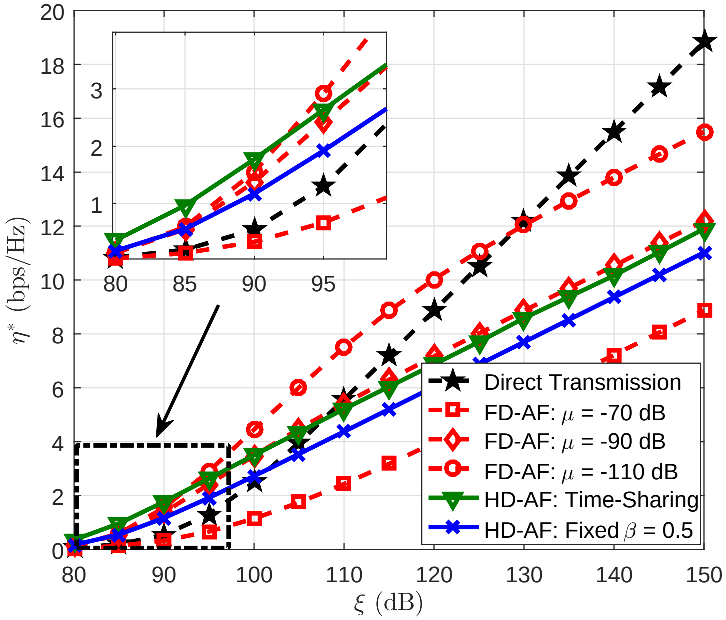


Figure 1.2: Maximum rates of direct transmission and both AF relaying schemes under different sum-power constraints, where $L_1 = 80$ m and $\theta_m = \frac{\pi}{6}$.

varying μ from -70 dB to -110 dB, we find that η_{FD}^{*} can be significantly improved when μ reduces. Comparing the cases with different values of μ, it can be seen that, the self-interference coefficient plays a crucial role in the rate of FD-AF relaying with mm-wave, particularly when a medium or high sum-power budget is given, i.e., $\xi \geq 100$ dB, while the impacts of the self-interference coefficient are relative smaller for $\xi \leq 90$ dB. To be more precise, taking the performance of HD-AF relaying as a benchmark, for $\xi \geq 100$ dB, we notice that η_{FD}^{*} is much less than η_{HD}^{*} when $\mu = -70$ dB, while it evidently outperforms the latter when $\mu = -90$ dB or smaller. However, within the low ξ region, i.e., $\xi \leq 90$ dB, it can be seen that, η_{FD}^{*} is still inferior to η_{HD}^{*} even though μ has been reduced to -110 dB. When ξ is lower, e.g., $\xi \leq 100$ dB, direct transmission is inferior to HD-AF but beyond FD-AF with $\mu = -70$ dB, while it surpasses the relaying counterparts if ξ is sufficiently high, e.g., $\xi \geq 130$ dB. The findings above indicate that, given high sum-power budget, direct transmission is the best choice. For relaying schemes, FD-AF relaying with a small self-interference coefficient should be adopted for scenarios with medium sum-power budget, while HD-AF relaying is more suitable for scenarios with low sum-power budget. Intuitively, HD-AF largely benefits from the absence of interference at the low sum-power region, and FD-AF with a small self-interference coefficient largely benefits from the full utilization of time slots for transmission at the intermediate sum-power region. However, once the sum-power is sufficiently high, the severe path loss from the source to the destination can be completely compensated, such that direct transmission in mm-wave communications becomes the most efficient option, instead of using any relaying technique.

E.2 Impacts of Beamwidth

The impacts of beamwidth on the rates of different transmission schemes are shown in Fig. 1.3, where the direct transmission scheme and both AF relaying schemes are compared. Clearly, all schemes suffer performance degradation as θ_m increases, while the direct transmission encounters the most severe deterioration, and FD-AF follows. It can be seen that, for both FD-AF and HD-AF relaying schemes, the rate degrades with θ_m as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$, as analyzed in Theorem D.1. In addition, compared to η_{HD}^{*}, we notices that η_{FD}^{*} rapidly decays when θ_m grows, which indicates the performance by FD-AF mm-wave relaying is more sensitive to the variation of beamwidth. The slow decay of η_{HD}^{*} with respect to θ_m mainly comes from the time fractions in fulfilling the two-phase transmission (in contrast to the full utilization of time duration in η_{FD}^{*}), which mitigate the impact of the variation of g_1 and/or g_2 on the achievable rate. Furthermore, higher achievable rates can be achieved by FD-AF mm-wave relaying, when a smaller self-interference coefficient (e.g., $\mu \leq -90$ dB) and a narrower beamwidth of directional antenna (e.g., $\theta_m \leq \frac{\pi}{6}$) are provided. Otherwise, HD-AF mm-wave relaying or direct transmission strategy should be adopted. We can also see that, given $\xi = 100$ dB, when applying very sharp beams, i.e., small θ_m, direct transmission outperforms the HD-AF relaying. Note that, narrowing the beamwidth increases the channel gain, which is equivalent

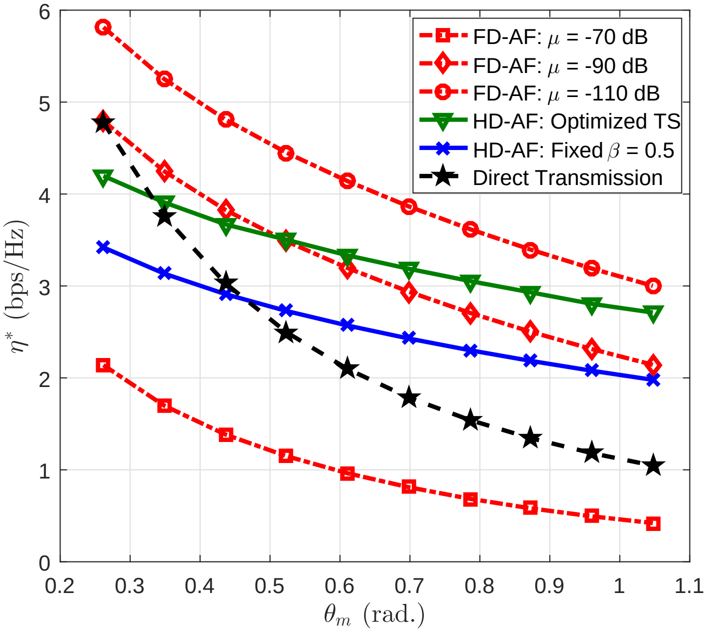


Figure 1.3: Maximum rates of AF relaying schemes (FD-AF, HD-AF with time-sharing, and HD-AF with $\beta = \frac{1}{2}$) with different beamwidth under a sum-power constraint $\xi = 100$ dB are applied, and $L_1 = 80$ m.

to elevating the sum-power budget while keeping the channel gain untouched. In this sense, the gain by sharpening beams is equivalent to the gain by increasing the sum-power budget with fixing the channel gains. Recalling that direct transmission outperforms HD-AF in the high sum-power region (as shown in Fig. 1.2), the observation that HD-AF is inferior to direct transmission when applying narrower beams then can be explained. Therefore, the findings above demonstrate that, to achieve higher rates of FD-AF mm-wave relaying, a careful joint treatment of the beamwidth and the self-interference coefficient during system designs and implementations is important. Besides, direct transmission is still a promising candidate when the antennas are highly directional.

Fig. 1.4 illustrates the contribution of ground reflections in two AF relaying schemes with mm-wave. For comparison, we take the curves labeled with “1-ray” as references, which correspond to the conventional modeling method that considers the LOS path only for mm-wave channels. Evidently, with the specific two-ray model (labeled with “2-ray”), the rates heavily rely on L_1 and L_2. More precisely, given θ_m, for $L_1 = 60$ m (simultaneously, $L_2 = L - L_1 = 140$ m), the achieved rate is beyond that under the “1-ray” model. However, the rate is remarkably inferior

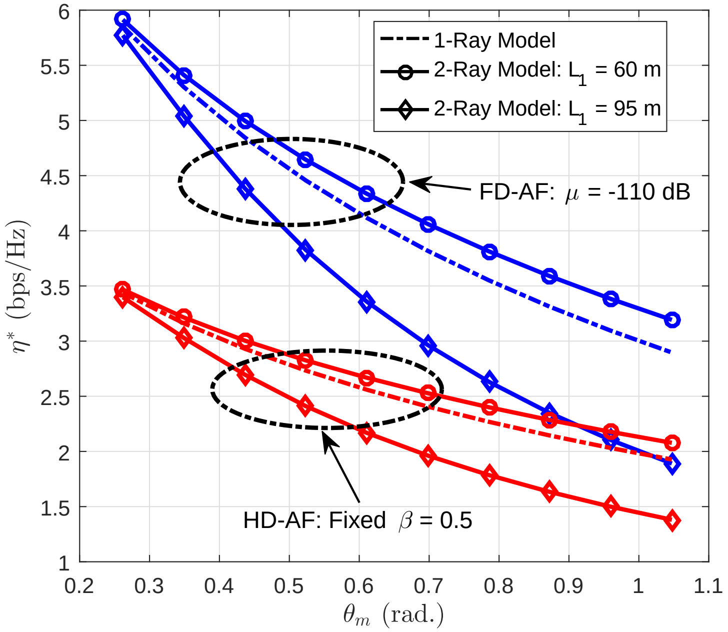


Figure 1.4: Contribution of ground reflections to rates with respect to different beamwidth, where $\xi = 100$ dB.

to that under the "1-ray" model when $L_1 = 95$ m (simultaneously, $L_2 = 105$ m). This surprisingly significant performance fluctuation stems from the contribution of ground reflections. It can be seen from (1) that, the reflection component of the channel coefficient depends on the incident angle of reflected signal, the radiation pattern of directional antenna, and the transmission distance. Thus, for $L_1 = 60$ m, the ground reflections contribute to the channel gain in the constructive manner, thereby improving the rates of the two-hop mm-wave relaying system, while the ground reflections for $L_1 = 95$ m adversely affect the rate. In light of above findings, an important insight for practical applications here is that, when the beamwidth is not very small, considerable improvement can be achieved via properly deploying the relay, such that the constructive ground reflections can be fully exploited. Furthermore, we can see that, when θ_m decreases, the rate with the two-ray model converges to that with the "1-ray" model. This observation indicates that, the impacts of ground reflections diminish when highly directional antennas are employed, and it agrees to Proposition D.1. Hence, from Fig. 1.4, it can be concluded that, the two-ray model provides a general framework for arbitrary beamwidth, while the "1-ray" model is only suitable for scenarios with very narrow beams.

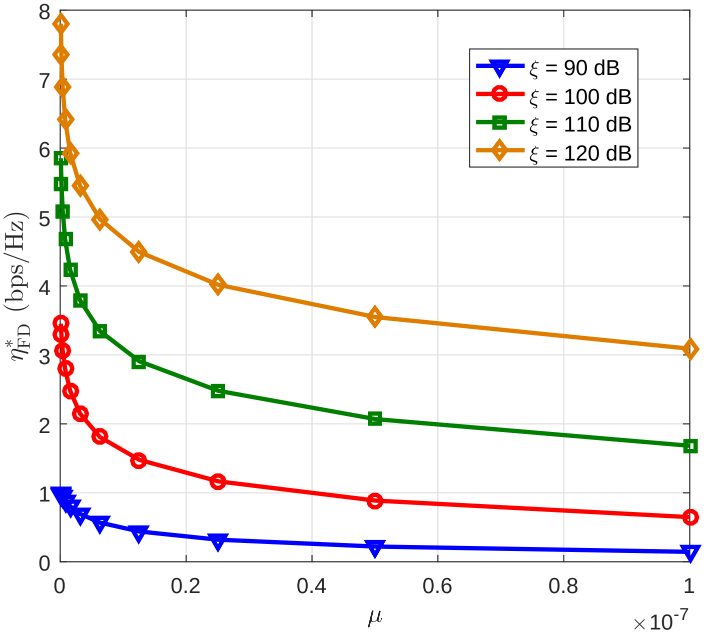


Figure 1.5: Impact of μ on η^*_{FD}, where $\theta_m = \frac{\pi}{4}$ and $L_1 = 100$ m.

E.3 Impacts of Self-Interference Cancellation

Fig. 1.5 depicts the impact of the self-interference coefficient on the rate of FD-AF mm-wave relaying. We find that, η^*_{FD} dramatically drops when increasing μ from -120 dB to -70 dB. As aforementioned in Theorem D.2, η^*_{FD} with respect to μ varies in the convex and decreasing manner. Thus, a slight reduction of μ is capable of providing remarkable performance improvement, and the improvement accelerates at lower μ and higher ξ, as shown in Fig. 1.5. This finding indicates the great importance of devising powerful self-interference cancellation techniques for FD-AF mm-wave relaying, especially when the two-hop AF mm-wave relaying system is assigned with a higher sum-power budget.

With respect to different sum-power constraints, in Fig. 1.6, we compare the performance of FD-AF mm-wave relaying in the presence of non-zero μ, and its upper limit is achieved by assuming $\mu = 0$. Given sum-power constraint ξ, when μ decreases from -70 dB to -130 dB, it is not surprising that η^*_{FD} gradually approaches its upper limit $\overline{\eta}^*_{\mathrm{FD}}$. We notice that, with any given μ, the rate difference between η^*_{FD} and $\overline{\eta}^*_{\mathrm{FD}}$ (vertical performance gap) gets enlarged as ξ grows, which is in accordance with Proposition D.2. Moreover, in the presence of larger self-interference coefficients, it is evident that only a smaller ξ is required for η^*_{FD} to

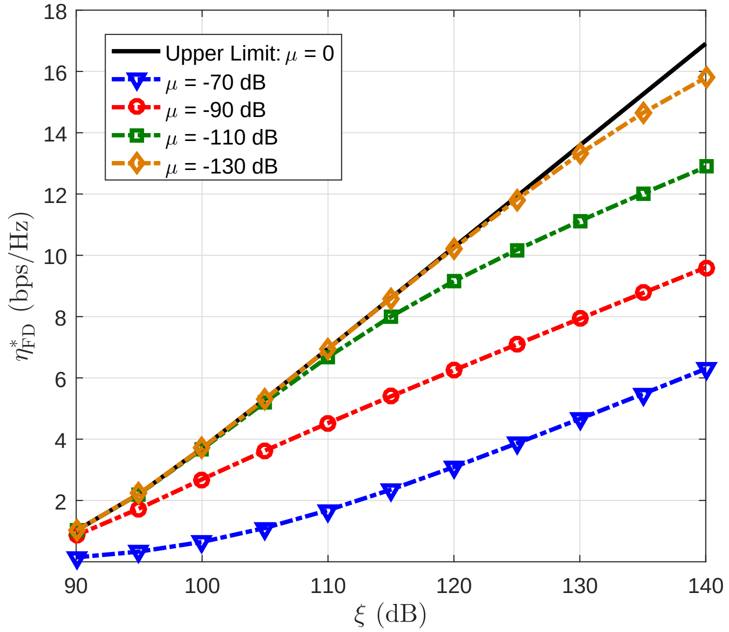


Figure 1.6: Performance gap between η^*_{FD} and its upper limit $\overline{\eta}^*_{\mathrm{FD}}\big|_{\mu=0}$, where $\theta_m = \frac{\pi}{4}$ and $L_1 = 100$ m.

approach its upper limit $\overline{\eta}^*_{\mathrm{FD}}$. This interesting observation can be explained by (11), since the SINR approaches its upper limit only if the product term $\mu\xi$ is sufficiently small. Therefore, a much smaller μ will be accordingly required for any given higher ξ, such that the condition $\mu\xi \to 0$ can be satisfied to make η^*_{FD} approach $\overline{\eta}^*_{\mathrm{FD}}$. In light of the relation between μ and ξ for guaranteeing $\mu\xi \to 0$, an important insight for mm-wave relay implementation is that, when the given sum-power budget is not high, it is not necessary to devise strong self-interference cancellation techniques, since the limiting performance can be easily approached even if μ is not sufficiently small. Thereby, the implementation cost and complexity of FD-AF relays can be largely reduced.

F Conclusions

We have studied the achievable rate of the two-hop AF relaying system with mm-wave, where two relaying schemes, i.e., HD-AF and FD-AF relaying are considered. With the joint treatment of the two-ray mm-wave channel and Gaussian-type directional antenna, we have investigated the impact of the beamwidth and the self-interference coefficient on the maximum achievable rate. Under a sum-power constraint, it has been demonstrated that, FD-AF relaying outperforms its HD counterpart, only when directional antennas with a smaller beamwidth and the AF relay with a smaller self-interference coefficient are applied. Thus, HD-AF relaying scheme is still a competitive candidate for mm-wave communications whenever above conditions cannot be satisfied. We also have found that, when the sum-power budget is sufficiently high or the beam of directional antenna is sufficiently sharp, the performance of direct transmission is beyond that of two relaying schemes, such that AF relaying is not necessary for in this case. In addition, it has been demonstrated that, for both mm-wave relaying schemes, the maximum achievable rate scales as $\mathcal{O}\left(\min\left\{\theta_m^{-1}, \theta_m^{-2}\right\}\right)$ with respect to beamwidth θ_m, and scales as $\mathcal{O}\left(\mu^{-\frac{1}{2}}\right)$ for FD-AF relaying with respect to self-interference coefficient μ. Furthermore, it has been revealed that, the ground reflection may significantly affect the performance of mm-wave communications, constructively or destructively. Therefore, it is crucial to incorporate the impacts of ground reflections for analyzing or designing mm-wave networks with directional antennas.

G Appendices

G.1 Proof of Theorem D.1

By Lemma D.1, it is known that, there exists some positive K_i, such that for all feasible θ_m we have

$$g_i \leq K_i \cdot \theta_m^{-2}.$$

For HD-AF relaying, we know that η_{HD}^* is monotonically increasing with g_i. By Proposition C.1, we obtain that

$$\begin{aligned}\eta_{\mathrm{HD}}^* &\overset{(a)}{<} \max_{\substack{0\leq\beta\leq 1\\ \xi_1,\xi_2\geq 0\\ \beta\xi_1+\bar{\beta}\xi_2=\xi}} \min\left\{\beta C\left(g_1\xi_1\right), \left(1-\beta\right) C\left(g_1\xi_1\right)\right\}\\ &= \max_{\substack{0\leq\beta\leq 1,\xi_1\geq 0\\ \beta\xi_1\leq\xi}} \min\left\{\beta C\left(g_1\xi_1\right), \left(1-\beta\right) C\left(g_1\xi_1\right)\right\} \overset{(b)}{\leq} \max_{\substack{0\leq\beta\leq 1,\xi_1\geq 0\\ \beta\xi_1\leq\xi}} \beta C\left(g_1\xi_1\right),\end{aligned}$$

where (a) applies the fact that

$$\frac{g_1 g_2 \xi_1 \xi_2}{1+g_1\xi_1+g_2\xi_2} = \frac{g_2\xi_2}{1+g_1\xi_1+g_2\xi_2} \cdot g_1\xi_1 < g_1\xi_1,$$

and (b) follows from the property that $\min\{a,b\} \leq a$ for any $a, b \in \mathbb{R}$. By Lemma D.2, it is easy to obtain that

$$C\left(g_1\xi_1\right) < \log_2\left(1+K_1\xi_1\theta_m^{-2}\right) \leq \frac{1}{\ln 2} \cdot \min\left\{K_1\xi_1\theta_m^{-2}, \sqrt{K_1\xi_1}\theta_m^{-1}\right\},$$

which subsequently yields

$$\eta_{\mathrm{HD}}^* < \max_{\substack{0\le\beta\le1,\xi_1\ge0\\ \beta\xi_1\le\xi}} \frac{K_1\beta\xi_1}{\ln 2}\theta_m^{-2} = \frac{K_1\xi}{\ln 2}\cdot\theta_m^{-2} = \mathcal{O}\left(\theta_m^{-2}\right).$$

or

$$\eta_{\mathrm{HD}}^* < \max_{\substack{0\le\beta\le1,\xi_1\ge0\\ \beta\xi_1\le\xi}} \frac{\beta\sqrt{K_1\xi_1}}{\ln 2}\theta_m^{-1} \le \frac{\sqrt{K_1\xi}}{\ln 2}\cdot\theta_m^{-1} = \mathcal{O}\left(\theta_m^{-1}\right).$$

Since $\mathcal{O}\left(\theta_m^{-2}\right) \subset \mathcal{O}\left(\theta_m^{-1}\right)$ as $\theta_m \ge 1$, we then have $\eta_{\mathrm{HD}}^* = \mathcal{O}\left(\theta_m^{-1}\right)$ for all $\theta_m > 0$, or $\eta_{\mathrm{HD}}^* = \mathcal{O}\left(\theta_m^{-2}\right)$ particularly when $\theta_m \ge 1$.

Likewise, for FD-AF relaying, we have

$$\begin{aligned}\eta_{\mathrm{FD}}^* <& C\left(\frac{\xi K_1K_2\cdot\theta_m^{-2}}{K_1+K_2+2\sqrt{K_1K_2\left(1+\mu\xi\right)}}\right)\\ \le&\frac{1}{\ln 2}\sqrt{\frac{\xi K_1K_2}{K_1+K_2+2\sqrt{K_1K_2\left(1+\mu\xi\right)}}}\cdot\theta_m^{-1} = \mathcal{O}\left(\theta_m^{-1}\right).\end{aligned}$$

Again, it is easy to see that, $\eta_{\mathrm{FD}}^* \in \mathcal{O}\left(\theta_m^{-2}\right)$ if $\theta_m \ge 1$.

G.2 Proof of Proposition D.2

Note that κ can be reformulated as

$$\kappa = 1 + \underbrace{\frac{\xi\mu}{A_1+A_2}}_{\triangleq\kappa_1} + \underbrace{\frac{A_2\left(\sqrt{1+\xi\mu}-1\right)}{A_1+A_2}}_{\triangleq\kappa_2},$$

where κ_1 and κ_2 are exactly shown as

$$\kappa_1 = \frac{\xi\mu}{\left(\sqrt{1+g_1\xi}+\sqrt{1+g_2\xi}\right)^2}$$

and

$$\kappa_2 = \frac{2\sqrt{\left(1+g_1\xi\right)\left(1+g_2\xi\right)}\cdot\left(\sqrt{1+\xi\mu}-1\right)}{\left(\sqrt{1+g_1\xi}+\sqrt{1+g_2\xi}\right)^2}.$$

For the first-order partial derivatives of κ_1 and κ_2 with respect to ξ, it is easy to verify that

$$\frac{\partial\kappa_1}{\partial\xi} > 0 \text{ and } \frac{\partial\kappa_2}{\partial\xi} > 0,$$

which naturally yields

$$\frac{\partial \kappa}{\partial \xi} = \frac{\partial \kappa_2}{\partial \xi} + \frac{\partial \kappa_2}{\partial \xi} > 0.$$

Therefore, κ is a monotonically increasing function of ξ, which concludes the proposition.

G.3 Proof of Theorem D.2

For η_{FD}^*, its first-order partial derivative is

$$\frac{\partial \eta_{\mathrm{FD}}^*}{\partial \mu} = \frac{\partial \eta_{\mathrm{FD}}^*}{\partial \gamma_{\mathrm{FD}}^*} \cdot \frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu} = \frac{1}{(1+\gamma_{\mathrm{FD}}^*)\ln 2} \cdot \frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu}.$$

Likewise, the second-order partial derivative is given by

$$\begin{aligned} \frac{\partial^2 \eta_{\mathrm{FD}}^*}{\partial \mu^2} =& \frac{\partial}{\partial \mu}\left(\frac{\partial \eta_{\mathrm{FD}}^*}{\partial \gamma_{\mathrm{FD}}^*} \cdot \frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu}\right) = \frac{\partial^2 \eta_{\mathrm{FD}}^*}{\partial {\gamma_{\mathrm{FD}}^*}^2} \cdot \left(\frac{\partial \gamma_{\mathrm{FD}}^*}{d\mu}\right)^2 + \frac{\partial \eta_{\mathrm{FD}}^*}{\partial \gamma_{\mathrm{FD}}^*} \cdot \frac{\partial^2 \gamma_{\mathrm{FD}}^*}{\partial \mu^2} \\ =& \frac{1}{(1+\gamma_{\mathrm{FD}}^*)\ln 2} \cdot \frac{\partial^2 \gamma_{\mathrm{FD}}^*}{\partial \mu^2} - \frac{1}{(1+\gamma_{\mathrm{FD}}^*)^2 \ln 2} \cdot \left(\frac{\partial \gamma_{\mathrm{FD}}^*}{\partial \mu}\right)^2, \end{aligned}$$

where the chain rule of derivatives for composite functions is applied. With the aid of derivations in Lemma D.3, it is not difficult to verify that,

$$\frac{\partial \eta_{\mathrm{FD}}^*}{\partial \mu} < 0 \text{ and } \frac{\partial^2 \eta_{\mathrm{FD}}^*}{\partial \mu^2} > 0.$$

Therefore, the decreasing monotonicity and the strict convexity are concluded, respectively.

Paper J3

Analysis of Millimeter-Wave Multi-Hop Networks With Full-Duplex Buffered Relays

Guang Yang, Ming Xiao, Hussein Al-Zubaidy, Yongming Huang, and James Gross

Refereed article published in
IEEE/ACM Transactions on Networking, vol. 26, no. 1, Feb. 2018.

Analysis of Millimeter-Wave Multi-Hop Networks With Full-Duplex Buffered Relays

Guang Yang, Ming Xiao, Hussein Al-Zubaidy, Yongming Huang, and James Gross

Abstract

The abundance of spectrum in the millimeter-wave (mm-wave) bands makes it an attractive alternative for future wireless communication systems. Such systems are expected to provide data transmission rates in the order of multi-gigabits per second in order to satisfy the ever-increasing demand for high rate data communication. Unfortunately, mm-wave radio is subject to severe path loss which limits its usability for long-range outdoor communication. In this work, we propose a multi-hop mm-wave wireless network for outdoor communication where multiple full-duplex buffered relays are used to extend the communication range, while providing end-to-end performance guarantees to the traffic traversing the network. We provide a cumulative service process characterization for the mm-wave propagation channel with self-interference in terms of the moment generating function (MGF) of its channel capacity. We then use this characterization to compute probabilistic upper bounds on the overall network performance, i.e., total backlog and end-to-end delay. Furthermore, we study the effect of self-interference on the network performance and propose an optimal power allocation scheme to mitigate its impact in order to enhance network performance. Finally, we investigate the relation between relay density and network performance under a sum power constraint. We show that increasing relay density may have adverse effects on network performance, unless the self-interference can be kept sufficiently small.

A Introduction

With rapidly increasing demands on network service, wireless communications in millimeter-wave (mm-wave) bands (ranging from 24.25 GHz to 300 GHz) becomes a promising technology to improve the network throughput for future communication systems [RSM$^+$13, XMH$^+$17]. Compared to conventional wireless communications in lower frequency bands, i.e., sub 6 GHz, mm-wave wireless communications have significant advantages, including considerably broader bandwidth, lower cost electronics, and higher gain directional antenna implementations [DH07]. These attributes make mm-wave a promising solution for the wireless backhaul [GQMT07], since the initial cost of fiber optic backhaul tends to be quite high and the conventional microwave based backhaul networks cannot support the throughput requirements of future networks.

The mm-wave technology can be utilized for both indoor and outdoor communications. A significant amount of experimental results for investigating the indoor mm-wave wireless personal area networks (WPAN) were reported in recent years [MC06, Smu09, GKZV09, RMIR^{+}14, YDX15]. For outdoor environments, in [VEDS88], antennas with narrow beamwidth were used to measure the path loss in urban street environments for the line-of-sight (LOS) and non-line-of-sight (NLOS) scenarios. In [BDRQL11], a channel sounder was deployed to estimate the outdoor 60 GHz channel using a 59 GHz horn antenna. Results show that the path loss exponent for the 60 GHz channel is between 2 and 2.5 for the outdoor environment, such as airport fields, urban streets or tunnels. In the recent work [ALS^{+}14], the spatial statistical models of mm-wave channels at 28 GHz and 73 GHz were established based on real-world urban measurements. Also, the small-scale fading effects were shown to be negligible in mm-wave bands due to the short wavelength [WAN97, GKZV09]. Hence, the channel fading is dominated by the shadowing effect, which is generally modeled as a log-normal random variable.

In light of the above, it is clear that, the use of mm-wave bands is limited to short-distance LOS communications, e.g., usually below 500 meters. To overcome larger distances or obstructed paths, especially in outdoor applications for high data rate transmissions, a strategically placed store-and-forward relay node may be used to form a multi-hop wireless network. Thus, it is our claim that multi-hop communications can be utilized to mitigate the effects of path loss over long distances and/or the effect of NLOS, while maintaining the traffic flows' quality of service (QoS) requirements. In this case, an understanding of corresponding network performance in terms of end-to-end delay and loss probability becomes the key to support real-time missions and critical applications, e.g., online banking, remote health, transportation systems operation and control, and electric power systems. Nevertheless, an analytical model for the multi-hop network in mm-wave bands does not exist, and its performance is not yet understood.

In order to further improve network throughput in the proposed network, we consider full-duplex relays to enable simultaneous transmission and reception. Although it suffers from self-interference, full-duplex relaying can still enhance the network throughput by applying interference cancellation techniques [DS10, JCK^{+}11, CJS^{+}10]. Numerous efforts dedicated to the study of full-duplex relaying for mm-wave applications can be found in [LJT14, MKJ^{+}16, WZS^{+}16a, AH16, DCK16, WZSH16]. Without loss of generality, we use a self-interference coefficient, which is a discounting parameter for the service offered by the channel to characterize the interference at each relaying transceiver.

In this work, we provide a probabilistic end-to-end delay and total backlog analysis of such networks in terms of the underlying channel parameters. This analysis can be used as a guideline for planning and operating QoS-driven multi-hop mm-wave network. The analysis of multi-hop wireless networks in mm-wave bands poses two main challenges: (i) the service process characterization for mm-wave fading channel, and (ii) multi-hop network performance analysis. The first challenge comes from the random nature of the mm-wave fading channel which results in time varying channel capacity, and the second challenge is a direct result of the limitations and strict assumptions of the traditional queuing theory, which is the main tool for network analysis, when applied to queuing networks. To address these two challenges, we adopt a moment generating function (MGF)-based stochastic network calculus approach [Fid10] for the analysis of networks of tandem queues. Then the service process, which is a function of the instantaneous channel capacity, is given in terms of the MGF of the fading channel distribution. This addresses the first challenge. Furthermore, we utilize network calculus to address the second challenge by using the service concatenation property.

A.1 Methodologies for Wireless Network Analysis

Network calculus is an effective methodology for network performance analysis. It was originally proposed by Cruz [Cru91a, Cru91b] in the early 90's for the worst-case analysis of deterministic networked systems. Since then, the methodology has been extended to probabilistic settings. Following the pioneering works in [Cha00, CCS01, LBL07] on MGF-based traffic and service characterization, in order to model traffic and service processes with independent increments and to utilize independence among multiplexed flows, the MGF-based network calculus was proposed [Fid10]. Typically, the MGF approach to network calculus employs a finite-state Markov channel abstraction for the analysis of wireless fading channels [Fid06a, MRJ11]. It is worth noting that MGF-based approach was used, outside the network calculus framework, for the analysis of various fading channels and relaying channels, e.g., [RLMAG15, TDHA14, AS00, YA12]. In contrast, the $(\min, \times)$ network calculus approach, proposed by [AZLB16], provides probabilistic performance bounds directly in terms of the fading channel parameters. It does that by transferring the problem from the "bit domain", where traffic and service quantities are measured in bits, to the "SNR domain", where these quantities are described by their SNR equivalence when measured at the channel capacity limit, using the exponential function. To apply the $(\min, \times)$ network calculus to non-identically distributed multi-hop wireless networks, a recursive formula for delay bound computation was developed in [PAZKG15].

To evaluate the performance of our specific mm-wave multi-hop wireless network, we model a multi-hop path in the network by a tandem of queues with service processes that represent the time-varying service offered by the underlying mm-wave channel. Then we follow an MGF-based network calculus approach to compute probabilistic bounds on end-to-end delay and total backlog, respectively, for that network.

A.2 Motivations and Contributions

Although network calculus has been around for some years, its application to wireless networks analysis is fairly recent. Furthermore, in the existing related work, the self-interference factor was not taken into account. In mm-wave communications, the large-scale fading, i.e., shadowing effect follows the log-normal distribution. Due to the unique nature of the mm-wave channel, a service process that characterizes this channel is important for any performance analysis. To our best knowledge, the performance guarantees of mm-wave multi-hop wireless networks considering heterogeneous self-interfered channels have not yet been investigated before. Coupled with the importance of mm-wave networks for the next generation mobile communications, it motivates us to investigate the backlog and delay performance, as well as the constrained sum power budget and QoS trade-off corresponding to the self-interfered channel.

Based on the motivation and specific challenges above, the main contributions of this paper are two-fold: (i) contribution to the theory of network calculus that is represented by a simplified closed-form expression for the network service curve applicable to both homogeneous and heterogeneous wireless networks, and (ii) contribution to the application by providing a service process characterization for mm-wave fading channels with self-interference, in terms of the MGF of the fading distribution. Note that, (i) is a general contribution to stochastic network calculus for multi-hop networks, while (ii) is a contribution specific to mm-wave communications. Additional contributions of this work include:

- An optimal power allocation scheme, based on the proposed methodology, more precisely, for mm-wave multi-hop networks with independent and identically distributed (i.i.d.) shadowing under end-to-end delay constraint.

- An insight into the impact of self-interference on mm-wave network performance. Under optimal power allocation, the end-to-end performance bounds exponentially degrade with the self-interference coefficient. This suggests that managing self-interference can be extremely rewarding.

This work builds on our own previous work [YXG$^+$16], where a service characterization for a single-hop 60 GHz system without self-interference was presented.

The remainder of the paper is organized as follows. In Sec. B, we provide the basics of MGF-based stochastic network calculus. We construct a model for the mm-wave multi-hop network and derive its probabilistic backlog and delay bounds in Sec. C. In Sec. D we propose an optimal power allocation strategy under a sum power constraint that results in better performance bounds. An asymptotic performance analysis of the network with self-interference is presented in Sec. E. Numerical results and simulations are presented in Sec. F, where we discuss the validity and the effectiveness of our analytic upper bounds, and investigate the impacts of self-interference coefficient and relay density on network performance. Conclusions are presented in Sec. G.

B Preliminaries

In this section we mainly provide a brief review of network calculus fundamental results and the MGF-based stochastic network calculus framework in particular.

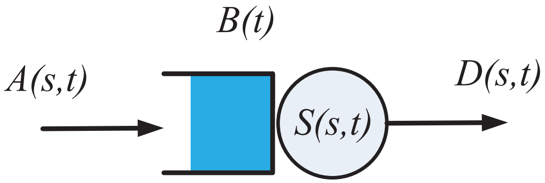


Figure 1.1: A queuing model for a store-and-forward node.

More details and the proofs for the presented fundamentals results can be found for example in [Cha94, LBT01, Cha00, Fid06a, JL08].

B.1 Model and Notation

Assuming a fluid-flow, discrete-time queuing system with a buffer of infinite size, and given a time interval $[s, t)$, $0 \leq s \leq t$, we define the non-decreasing (in t) bivariate processes $A(s, t)$, $D(s, t)$ and $S(s, t)$ as the cumulative arrival to, departure from and service offered by the system as shown in Fig. 1.1. We further assume that $A(s, t)$, $D(s, t)$ and $S(s, t)$ are stationary non-negative random processes with $A(t, t) = D(t, t) = S(t, t) = 0$ for all $t \geq 0$. The cumulative arrival and service processes are given in terms of their instantaneous values during the i^{th} time slot, a_i and s_i respectively, as follows:

$$A(s,t) = \sum_{i=s}^{t-1} a_i \quad \text{and} \quad S(s,t) = \sum_{i=s}^{t-1} s_i\,, \tag{1}$$

for all $0 \leq s \leq t$. We assume that time slots are normalized to 1 time unit. We denote by $B(t)$ the backlog (the amount of buffered data) at time t. Furthermore, $W(t)$ denotes the virtual delay of the system at time t.

Network calculus is based on $(\min, +)$-algebra, for which in particular the convolution and deconvolution are important to obtain bounds on the system performance. More precisely, given the non-decreasing and strictly positive bivariate processes $X(s, t)$ and $Y(s, t)$, the $(\min, +)$ convolution and deconvolution are respectively defined as

$$(X \otimes Y)(s,t) \triangleq \inf_{s \leq \tau \leq t} \{X(s,\tau) + Y(\tau,t)\}$$

and

$$(X \oslash Y)(s,t) \triangleq \sup_{0 \leq \tau \leq s} \{X(\tau,t) - Y(\tau,s)\}.$$

B.2 Network Calculus Basics

In network calculus, the queuing system in Fig. 1.1 is analyzed with the arrival process $A(s, t)$ as input and the departure process $D(s, t)$ as system output. Input and output are related to each through the $(\min, +)$ convolution of the input with the service process $S(s,t)$. In particular, we consider in the following time varying systems known as *dynamic servers*, for which for all $t \geq 0$ the network element offers a time varying service $S\,(s,t)$ that satisfies the following input-output inequality [Cha00] $D(0,t) \geq (A \otimes S)(0,t)$, which holds with strict equality when the system is linear [LBT01]. One typical example is a work-conserving link with a time-variant capacity, with the available service $S(s,t)$ during interval $[s,t)$.

Based on this server model, the total backlog and end-to-end delay, which are critical measures in system evaluation, can be studied via network calculus. On one hand, buffer dimensioning is a major factor to consider when designing and implementing broadband networks. This is true due to the space restriction and cost of storage in intermediate network devices, e.g., routers in high data rate networks. On the other hand, end-to-end delay is closely related to the quality of service (QoS) and user experience for many networked applications, e.g., voice and video services. For a given queuing system with cumulative arrival $A(0,t)$ and departure $D(0,t)$ and for $t \geq 0$, the backlog at time t, $B(t)$ is defined as the amount of traffic remaining in the system by time t. Therefore,

$$B(t) \triangleq A(0,t) - D(0,t)\,. \tag{2}$$

Likewise, the virtual delay $W(t)$ is defined as the time it takes the last bit received by time t to depart the system under a first-come-first-serve (FCFS) scheduling regime. Hence,

$$W(t) \triangleq \inf\{w \geq 0 : A(0,t) \leq D(0,t+w)\}. \tag{3}$$

Substituting $D(0,t) \geq (A \otimes S)(0,t)$ in the above expressions and after some manipulation and using definitions of $(\min, +)$ convolution and deconvolution, we can obtain the following bounds on $B(t)$ and $W(t)$ respectively, as

$$B(t) \leq (A \oslash S)(t,t) \tag{4}$$

and

$$W(t) \leq \inf\{w \geq 0 : (A \oslash S)(t+w,t) \leq 0\}. \tag{5}$$

A main attribute of network calculus is its ability to handle concatenated systems, e.g., multi-hop store-and-forward networks. This is mainly achieved using the server concatenation theory, which states that a network service process can be computed as the $(\min, +)$ convolution of the individual nodes' service processes [LBT01]. More precisely, given n tandem servers, the network service process $S_{\text{net}}(s,t)$ is given by

$$S_{\text{net}}(s,t) = \left(S_1 \otimes S_2 \otimes \cdots \otimes S_n\right)(s,t), \tag{6}$$

where $S_i(s,t)$, for any $1 \leq i \leq n$, represents the service process of the i^{th} server.

B.3 MGF-based Probabilistic Bounds

Deterministic network calculus [LBT01] can provide worst-case upper bounds on the backlog and the delay if traffic envelopes (an upper bound on the arrival process) as well as a service curve (a lower bound on the service process) are considered. However, when analyzing systems with random input and/or service (like wireless networks), due to a possibly non-trivial probability for the service increment or arrival increment to be zero, the worst-case analysis is no longer useful to describe the performance any more. In such cases, probabilistic performance bounds provide more useful and realistic description of the system performance than worst-case analysis. In the probabilistic setting (where the arrival process and/or the service process are stationary random processes), the backlog and delay bounds defined in (2) and (3) respectively are reformulated in a stochastic sense as:

$$\Pr\left(B(t) > b^{\varepsilon'}\right) \leq \varepsilon' \quad \text{and} \quad \Pr\left(W(t) > w^{\varepsilon''}\right) \leq \varepsilon'', \tag{7}$$

where $b^{\varepsilon'}$ and $w^{\varepsilon''}$ denote the target probabilistic backlog and delay associated with violation probabilities ε' and ε'' respectively. These performance bounds can be obtained by the distributions of the processes, i.e., in terms of the arrival and service processes MGFs [Fid10] or their Mellin transforms [AZLB16]. These approaches constitute what we refer to as stochastic network calculus and they are most suitable for the analysis of wireless networks.

In general, the MGF-based bounds are obtained by applying Chernoff's bound, that is, given a random variable X, we have

$$\Pr(X \geq x) \leq e^{-\theta x}\mathbb{E}\left[e^{\theta X}\right] = e^{-\theta x}\mathbb{M}_X(\theta),$$

whenever the expectation exists, where $\mathbb{E}[Y]$ and $\mathbb{M}_Y(\theta)$ denote the expectation and the moment generating function (or the Laplace transform) of Y, respectively, and θ is an arbitrary non-negative free parameter. Given the stochastic process $X(s,t), t \geq s$, we define the MGF of X for any $\theta \geq 0$ as [Fid06b]

$$\mathbb{M}_X(\theta, s, t) \triangleq \mathbb{E}\left[e^{\theta X(s,t)}\right].$$

Moreover, $\overline{\mathbb{M}}_X(\theta, s, t) \triangleq \mathbb{M}_X(-\theta, s, t) = \mathbb{E}\left[e^{-\theta X(s,t)}\right]$ is also defined in a similar way.

A number of properties of MGF-based network calculus are summarized in [Fid06b]. In this work, we consider a queuing system comprised of a set of tandem queues. Using (6), the MGF of the end-to-end service process, written as $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t)$, of N tandem queues with service processes $S_i, i = 1, \dots N$, is bounded by

$$\begin{aligned}\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t) \triangleq & \overline{\mathbb{M}}_{S_1 \otimes S_2 \otimes \cdots \otimes S_N}(\theta, s, t) \\ = & \mathbb{E}\left[\exp\left(-\theta \cdot \inf_{s \leq u_1 \leq \cdots \leq u_{N-1} \leq t}\left\{\sum_{i=1}^{N} S_i(u_{i-1}, u_i)\right\}\right)\right]\end{aligned}$$

$$\leq \sum_{s \leq u_1 \leq \cdots \leq u_{N-1} \leq t} \prod_{i=1}^{N} \overline{\mathbb{M}}_{S_i}(\theta, u_{i-1}, u_i), \tag{8}$$

where $u_0 = s$ and $u_N = t$, and (8) is obtained via applying the union bound and independence assumption [Fid06b].

Let $\mathbb{M}_A(\theta, s, t)$ be the MGF of the arrival process. Assume the arrival and service processes are independent, and define $\mathsf{M}(\theta, s, t)$ as

$$\mathsf{M}(\theta, s, t) \triangleq \sum_{u=0}^{\min(s,t)} \mathbb{M}_A(\theta, u, t) \cdot \overline{\mathbb{M}}_{S_{\text{net}}}(\theta, u, s). \tag{9}$$

Then the probabilistic backlog bound in (7) can be obtained using (4) as follows

$$\Pr\left(B(t) > b^{\varepsilon'}\right) \leq \Pr\left((A \oslash S_{\text{net}})(t,t) > b^{\varepsilon'}\right) \leq e^{-\theta b^{\varepsilon'}} \cdot \mathsf{M}(\theta, t, t) \triangleq \varepsilon',$$

where we applied the Chernoff's bound, the union bound and the independence assumption of A and S_{net}. Solving for $b^{\varepsilon'}$ yields [Fid06b]

$$b^{\varepsilon'} = \inf_{\theta > 0} \left\{ \frac{1}{\theta} \left(\log \mathsf{M}(\theta, t, t) - \log \varepsilon' \right) \right\}. \tag{10}$$

Similarly, for the probabilistic delay bound in (7), using (5), we have

$$\Pr\left(W(t) > w^{\varepsilon''}\right) \leq \Pr\left((A \oslash S_{\text{net}})(t + w^{\varepsilon''}, t) > 0\right) \leq \mathsf{M}\left(\theta, t + w^{\varepsilon''}, t\right) \triangleq \varepsilon'',$$

where, again we use the Chernoff's bound, the union bound and the independence assumption of A and S_{net}. Solving for $w^{\varepsilon''}$ we get [Fid06b, AZLB16]

$$w^{\varepsilon''} = \inf\left\{ w : \inf_{\theta > 0} \{\mathsf{M}(\theta, t + w, t)\} \leq \varepsilon'' \right\}. \tag{11}$$

C Performance Analysis of mm-wave Multi-Hop Wireless Network

C.1 System Model

We consider a multi-hop wireless network in mm-wave bands, as shown in Fig. 1.2, consisting of a source S, n ($n \geq 1$) full-duplex relays R_i, $i = 1, 2, \ldots, n$ and a destination D. For simplifying illustration, we assign the labels 0, 1, …, $n+1$ to the ordered nodes. That is, S and D correspond to nodes 0 and $(n+1)$, respectively. Furthermore, we label the channel between nodes $(i-1)$ and i as the i^{th} hop in the set of hops $\mathcal{I}_{\mathcal{H}}$, i.e., $i \in \mathcal{I}_{\mathcal{H}} = \{1, 2, \ldots, n+1\}$, and the distance between the two nodes by l_i (in meter). We denote the channel gain coefficient of the i^{th} hop

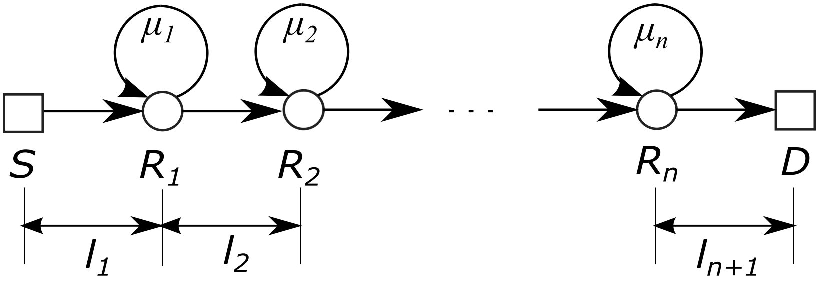


Figure 1.2: A multi-hop wireless network with n full-duplex relays.

by g_i. Given the separation distance l_i, a generalized model of g_i (in dB) for the mm-wave channel is given by [RRE14, RMSS15]

$$g_i[\text{dB}] = -\left(\alpha + 10\beta \log_{10}(l_i) + \xi_i\right), \tag{12}$$

where α and β are the least square fits of floating intercept and slope of the best fit, and $\xi_i \sim \mathcal{N}(0, v_i^2)$ corresponds to the log-normal shadowing effect with variance v_i^2. The values of the parameters α and β greatly depend on the environment configurations. Our service characterization and performance analysis are carried out in terms of these two parameters in order to incorporate all such configurations.

In addition to the (large scale) fading effect, without loss of generality, the proposed model also considers the self-interference at each full-duplex relay node. Compared to self-interference, the interference impact from other neighboring nodes is small, due to the rapid attenuation of the millimeter waves, and thus it is ignored. A common approach to model the self-interference is to use a coefficient $0 \leq \mu \leq 1$ that characterizes the coupling between the transmitter and the receiver of a full-duplex device. It has been shown that the self-interference is linearly related the transmission power [DS10]. In the presence of self-interference, the signal to interference plus noise ratio (SINR) in the i^{th} hop, denoted by γ_i, for the described channel is expressed as

$$\begin{aligned} \gamma_i =& \kappa \cdot \omega_i \cdot g_i \,, \\ s.t., \quad \omega_i =& \begin{cases} \dfrac{\lambda_{i-1}}{1+\mu_i \lambda_i}, & i \in \{1, 2, \ldots, n\} \\ \lambda_{i-1}, & i = n+1 \end{cases}, \end{aligned} \tag{13}$$

where κ is a scalar depending on system configuration, i.e., the antenna gains of the communication pair, g_i is the channel gain coefficient given by (12), $\lambda_i \triangleq \frac{P_i}{N_0}$ denotes the transmitted signal-to-noise ratio (SNR) at node i corresponding to transmit power P_i and background noise power N_0.

For the multi-hop scenario, we assume the stochastic process of each hop to be stationary and independent in time. That is, we can use a series of independent random variables γ_i to characterize the multi-hop channels, namely, $\gamma_i^{(k)} \overset{\ell}{=} \gamma_i$ in all time slot k, where $\overset{\ell}{=}$ denotes equality in law (i.e., in distribution). The shadowing effect, which is due to objects obstructing the propagation of mm-wave radios, is considered in the channel gain coefficient model given by (12). Regarding the stochastic behavior of different links, generally, shadowing is not spatially independent. Considering the fact that highly directional antennas are commonly used for mm-wave communications, it is safe to assume that the obstructions of radio propagation behave independently, which justifies the independence assumption across hops.

In general, the fading distributions of the subsequent channels in multi-hop wireless network may not be identical. Nevertheless, it is worthwhile to decompose the set of hops into subsets of hops with identically distributed channel gains. More precisely, we decompose the set of hops, $\mathcal{I}_{\mathcal{H}}$, into m subsets, $\mathcal{X}_k, k \in \mathcal{I}_{\mathcal{M}} = \{1, \ldots, m\}$, where, $\mathcal{I}_{\mathcal{H}} = \bigcup_{k=1}^{m} \mathcal{X}_k$, with $\mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset$ for all $i, j \in \mathcal{I}_{\mathcal{M}}$ such that $i \neq j$, where $\mathcal{X}_k$ is defined as the set of indices such that

$$\mathcal{X}_k = \{j \in \mathcal{I}_{\mathcal{H}}, k \in \mathcal{I}_{\mathcal{M}} : F_{\gamma_j}(x) = F^{\langle k \rangle}(x)\}, \tag{14}$$

where $F_X(x)$ is the probability distribution function of the random variable X, $F^{\langle k \rangle}(x)$ represents a unique distribution function corresponding to the subset of i.i.d. hops denoted by the index $k \in \mathcal{I}_{\mathcal{M}}$. We emphasize that $|\mathcal{I}_{\mathcal{H}}| \geq |\mathcal{I}_{\mathcal{M}}|$, where $|\mathcal{Y}|$ represents the cardinality of the set $\mathcal{Y}$, and the equality is attained when the multi-hop network is fully heterogeneous. In such extreme case, each subset $\mathcal{X}_k$ contains only one element, i.e., the channel at each hop has a distinct fading distribution.

To address the transmit power allocation problem for the multi-hop network, we consider a system with sum power P_{tot} constraint, i.e., $\sum_{i=0}^{n} P_i = P_{\text{tot}}$. Equivalently, given a constant background noise power N_0 for all hops, the sum power constraint can be reformulated as

$$\sum_{i=0}^{n} \lambda_i = \lambda_{\text{tot}} \triangleq \frac{P_{\text{tot}}}{N_0}. \tag{15}$$

C.2 MGF Bound for the Cumulative Service Process

The performance bounds given by (10) and (11) require the computation of MGFs of arrival and service processes. For the arrival process, we consider in this work the $(\sigma(\theta), \rho(\theta))$ traffic characterization introduced by [Cha94, Cha00], where the MGF of arrival processes, $\mathbb{M}_A(\theta, s, t)$, is upper bounded by

$$\mathbb{M}_A(\theta, s, t) \leq e^{\theta(\rho(\theta)(t-s)+\sigma(\theta))} \triangleq e^{\theta\sigma(\theta)} \left(p_a(\theta)\right)^{t-s}, \tag{16}$$

for any $\theta > 0$, where $p_a(\theta) = e^{\theta\rho(\theta)}$. This class of arrival processes includes a variety of traffic models, e.g., the exponentially bounded burstiness (EBB), effective bandwidth arrival characterization and deterministically bounded arrivals.

Regarding the service process, we present in the following a series of results that apply to a Shannon-capacity type service process. Given a channel SINR γ, in this model the service process (in bit per second) of the channel is given by

$$C(\gamma) = \eta \ln(1+\gamma),$$

where $\eta = \frac{W}{\ln 2}$ with channel bandwidth W. By (1), the cumulative service process for hop i, $S_i(s,t)$, is given by

$$S_i(s,t) = \sum_{k=s}^{t-1} C(\gamma_i(k)) = \eta \sum_{k=s}^{t-1} \ln\left(1+\gamma_i(k)\right), \tag{17}$$

where $\gamma_i(k)$ is the instantaneous SINR in the k^{th} time slot for the i^{th} hop given in terms of g_i in (12) and ω_i in (13).

However, associated with the specific fading characteristic of mm-wave channels, an exact expression for the MGF of this Shannon-type cumulative service process in (17) is intractable, since the *shifted log-normal random variable*, i.e., $(1+\gamma_i(k))$, has no closed-form inverse moment. Instead, in the following lemma, we present an upper bound on the MGF of such Shannon-type service processes $S_i(s,t)$.

Lemma C.1. *Let $F_X(x)$ denote the cumulative distribution function (c.d.f.) of a non-negative random variable X, for any $\delta > 0$ and $\theta \geq 0$ we have*

$$\mathbb{E}\left[(1+X)^{-\theta}\right] \leq \mathcal{U}_{\delta,X}(\theta),$$

where

$$\mathcal{U}_{\delta,X}(\theta) = \min_{u\geq 0}\left\{\left(1+\delta N_\delta(u)\right)^{-\theta} + \sum_{i=1}^{N_\delta(u)} a_{\theta,\delta}(i) F_X(i\delta)\right\},$$

where $N_\delta(u)$ and $a_{\theta,\delta}(i)$ are respectively given by $N_\delta(u) = \lfloor \frac{u}{\delta} \rfloor$ and $a_{\theta,\delta}(i) = (1+(i-1)\delta)^{-\theta} - (1+i\delta)^{-\theta}$.

For the proof of Lemma C.1, please refer to [YXG+16]. Note that the tightness of the bound obtained in Lemma C.1 depends on the parameter δ, the discretization step size. Technically, a smaller step size yields a tighter upper bound while leading to higher computational costs.

Based on Lemma C.1, a bound on the MGF of the service process for any single-server wireless system with service process increments given by the Shannon capacity is given:

Theorem C.1. *Given $S(s,t) = \eta \sum_{k=s}^{t-1} \ln(1+\gamma(k))$ with independent positive $\gamma(k)$, an upper bound on $\overline{\mathbb{M}}_S(\theta,s,t)$ is given by*

$$\overline{\mathbb{M}}_S(\theta,s,t) \leq \prod_{k=s}^{t-1} q(-\theta,k),$$

where $q(-\theta,k) \triangleq \mathcal{U}_{\delta,\gamma(k)}(\eta\theta)$*. Furthermore, if* $\gamma(k) \overset{\ell}{=} \gamma$ *holds for all* k*, then* $q(-\theta,k) = q(-\theta)$ *and the above expression reduces to* $\overline{\mathbb{M}}_S(\theta,s,t) \leq (q(-\theta))^{t-s}$*.*

Proof. Starting from the definition of $\overline{\mathbb{M}}_S(\theta,s,t)$ and using the independence assumption of $\gamma(k)$ in k, we have

$$\begin{aligned}\overline{\mathbb{M}}_S(\theta,s,t) =& \mathbb{E}\left[\exp\left(-\theta \cdot S(s,t)\right)\right]\\ =& \mathbb{E}\left[\prod_{k=s}^{t-1} \exp\left(-\theta\eta \ln(1+\gamma(k))\right)\right] = \prod_{k=s}^{t-1} \mathbb{E}\left[(1+\gamma(k))^{-\theta\eta}\right].\end{aligned}$$

Applying Lemma C.1 to the right hand side of the expression above, Theorem C.1 immediately follows. □

Next, we provide a MGF bound on the network service for multi-hop wireless networks with heterogeneous, independent Shannon-type service processes per link. The channel categorization, shown in (14), is used for expression simplifications.

Theorem C.2. *The network service process* $S_{\text{net}}(s,t)$ *of a multi-hop wireless network consisting of* n *relays and characterized by the decomposable set of hops* $\mathcal{I}_{\mathcal{H}} = \bigcup_{i=1}^{m} \mathcal{X}_i$ *following* (14)*, where the subset* $\mathcal{X}_i$ *is associated with the randomly varying SINR* $\hat{\gamma}_i$ *and has a Shannon-type service process increment* $\ln(1+\gamma)$*, has the following MGF bound*

$$\overline{\mathbb{M}}_{S_{\text{net}}}(\theta,s,t) \leq \sum_{\sum\limits_{i=1}^{m} \pi_i = t-s} \prod_{i=1}^{m} \binom{\pi_i + |\mathcal{X}_i| - 1}{|\mathcal{X}_i| - 1} \hat{q}_i^{\pi_i}(-\theta), \tag{18}$$

where $\hat{q}_i(-\theta) \triangleq \mathcal{U}_{\delta,\hat{\gamma}_i}(\eta\theta)$ *for all* $i \in \mathcal{I}_{\mathcal{M}}$*.*

Proof. Using equation (6) and (8), we can bound the MGF for the $n+1$ hops network service process by

$$\begin{aligned}\overline{\mathbb{M}}_{S_{\text{net}}}(\theta,s,t) \leq& \sum_{s=\tau_1 \leq \cdots \leq \tau_{n+1}=t} \prod_{i=1}^{n+1} \overline{\mathbb{M}}_{S_i}(\theta,\tau_{i-1},\tau_i)\\ \leq& \sum_{\sum_{i=1}^{n+1} \tau_i = t-s} q_1^{\tau_1}(-\theta) q_2^{\tau_2}(-\theta) \cdots q_{n+1}^{\tau_{n+1}}(-\theta)\\ =& \sum_{\sum_{i=1}^{m} \pi_i = t-s} \prod_{i=1}^{m} \hat{q}_i^{\pi_i}(-\theta) \sum_{\sum_{k\in\mathcal{X}_i} \tau_k = \pi_i} 1\end{aligned}$$

where the first inequality is obtained by using (8), and the second inequality is obtained by using the change of variables $\tau_i = \tau_i - \tau_{i-1}$ and the stationarity of the service processes, i.e., $\overline{\mathbb{M}}_{S_i}(\theta, \tau_{i-1}, \tau_i) = \overline{\mathbb{M}}_{S_i}(\theta, 0, \tau_i - \tau_{i-1})$, then applying Theorem C.1. The equality in the last line is obtained by aggregating similar terms, i.e., $\pi_i = \sum_{k \in \mathcal{X}_i} \tau_k$. Applying the combinations of multi-sets theory [Bru92], it is known that

$$\sum_{\sum_{k \in \mathcal{X}_i} \tau_k = \pi_i} 1 = \binom{\pi_i + |\mathcal{X}_i| - 1}{|\mathcal{X}_i| - 1},$$

where $|\mathcal{X}_i|$ denotes the cardinality of $\mathcal{X}_i$, and then the theorem is concluded. □

We emphasize that, all results presented so far apply to general cases of distributions of link SINR γ. Thus, the results have wide applicability to wireless (and wired) network analysis, as long as Shannon-type service processes are assumed.

C.3 Probabilistic Performance Bounds

The general probabilistic total backlog and end-to-end delay bounds for a multi-hop wireless network are given by (10) and (11), respectively. Both bounds are given in terms of $\mathsf{M}(\theta, s, t)$ in (9). Theorem C.3 provides an upper bound on the function $\mathsf{M}(\theta, s, t)$, and hence, probabilistic performance bounds for multi-hop wireless networks, when the arrival is characterized by (16) and the network service is provided by Theorem C.2.

Let us first define $\mathcal{K}_{\tau,n,m}(x)$ as

$$\mathcal{K}_{\tau,n,m}(x) \triangleq x^{\tau} \cdot \binom{n+1-m+\tau}{n+1-m} \cdot {}_2F_1(1, n+2-m+\tau; \tau+1; x),$$

where the *Generalized Hypergeometric Function* ${}_pF_q(\underline{a}; \underline{b}; x)$, with vectors $\underline{a} = [a_1, \ldots, a_p]$ and $\underline{b} = [b_1, \ldots, b_q]$, is given as

$${}_pF_q(\underline{a}; \underline{b}; x) \triangleq \sum_{k=0}^{\infty} \left(\frac{\prod_{i=1}^{p} (a_i)_k}{\prod_{j=1}^{q} (b_j)_k} \right) \cdot \frac{x^k}{k!},$$

and $(a_i)_k$ and $(b_i)_k$ are *Pochhammer symbols.*

Theorem C.3. *Let m be the number of subsets of identically distributed channels, $\tau \triangleq \max(s-t, 0)$, and $V_i(\theta) \triangleq p_a(\theta)\hat{q}_i(-\theta)$, a upper bound for $\mathsf{M}(\theta, s, t)$, $\theta > 0$, for the $(n+1)$-hop wireless network is given by*

$$\mathsf{M}(\theta, s, t) \leq \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{i=1}^{m} \psi_i(\theta) V_i^{m-1}(\theta) \mathcal{K}_{\tau,n,m}(V_i(\theta)),$$

whenever the stability condition, $\max_{i \in \mathcal{I}_{\mathcal{M}}} \{V_i(\theta)\} < 1$, is satisfied. Here, $\psi_i(\theta)$ for all $i \in \{1, \ldots, m\}$ is defined as

$$\psi_i(\theta) \triangleq \begin{cases} \prod_{j \neq i} (V_i(\theta) - V_j(\theta))^{-1}, & m \geq 2 \\ 1, & m = 1 \end{cases}. \tag{19}$$

Proof. Please see Appendix H.1. □

The stability condition in Theorem C.3 can be reasoned as follows. If we take the log of both sides of the expression, the condition can be stated as, "the difference between log MGF of the arrival and that of the service is less than 0," i.e., for a given QoS measure θ, the effective capacity must exceed the effective bandwidth, for the same θ, for the system to be stable. This intuitive result was hinted in [Cha00, Ch. 7].

Clearly, the expression provided by Theorem C.3 depends on the generalized hypergeometric function, which is computational costly. In what follows, we provide a looser but more simplified upper bound on $\mathsf{M}(\theta, s, t)$ in the homogeneous case specifically, i.e., $m = 1$. For that, we first need the following lemma regarding the upper bound on $\mathcal{K}_{\tau,n,1}(x)$.

Lemma C.2. *For non-negative integers n and τ, the inequality*

$$\mathcal{K}_{\tau,n,1}(x) \le \mathcal{G}_{\tau,n}(x) \triangleq \min\{\mathcal{G}_1(x), \mathcal{G}_2(x)\}$$

holds for $0 \le x < 1$, where $\mathcal{G}_1(x)$ and $\mathcal{G}_2(x)$ are respectively given by

$$\mathcal{G}_1(x) = \frac{\min\left(1, x^{\tau}\binom{n+\tau}{n}\right)}{(1-x)^{n+1}}$$

and

$$\mathcal{G}_2(x) = \frac{1}{(1-x)^{n+1}} - \binom{n+\tau}{n+1} x^{\tau-1}.$$

Proof. Please see Appendix H.2. □

The lemma above allows us to use a simpler expression, which is shown in the following theorem, to describe the upper bound (compared to Theorem C.3) on $\mathsf{M}(\theta, s, t)$ for homogeneous networks in particular.

Theorem C.4. *For homogeneous $(n+1)$-hop wireless networks characterized by the MGF service bound $\hat{q}(\theta)$, and for any $\theta > 0$, given $p_a(\theta)$ we have*

$$\mathsf{M}(\theta, s, t) \le \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \cdot \mathcal{G}_{\tau,n}\left(p_a(\theta)\hat{q}(-\theta)\right),$$

where $\tau \triangleq \max(s-t, 0)$, whenever the stability condition, $p_a(\theta)\hat{q}(-\theta) < 1$, holds.

Proof. Applying Lemma C.2 in (34) whenever $0 \le p_a(\theta)\hat{q}(-\theta) < 1$, then Theorem C.4 follows. □

Inserting the results from Theorem C.3 (or Theorem C.4 for merely homogeneous cases), in (10) and (11), we obtain the desired probabilistic bounds on backlog b^{ε} and delay w^{ε} in the network, respectively.

D Power Allocation for Multi-Hop Networks

In this section, we study the optimal power allocation under a sum power constraint for a mm-wave multi-hop network[1], applying the results presented in the previous section regarding the MGF-based network calculus. Here, more precisely, we are interested in finding the optimal transmit power allocation of multi-hop network with independent and identically distributed shadowing per hop, ξ_i, $\forall i \in \mathcal{I}_{\mathcal{H}}$, assuming that ξ_i is stationary. In particular, we limit our study to the case of homogeneous log-normally distributed shadowing over all links, i.e. we set $\xi_i \overset{\ell}{=} \xi \sim \mathcal{N}(0, v^2)$ for all hops $i \in \mathcal{I}_{\mathcal{H}}$. The case with non-identical shadowing is more involved and is left for future work.

From (10) and (11), it is clearly shown that, probabilistic bounds on the total network backlog and end-to-end delay performance are determined in terms of the function $\mathsf{M}(\theta, s, t)$ defined in (9). This implies that the performance optimization, e.g., with respect to transmit power allocation, of a multi-hop network is equivalent to optimizing the function $\mathsf{M}(\theta, s, t)$, which, with a given arrival process $\mathbb{M}_A$, is equivalent to optimizing $\overline{\mathbb{M}}_{S_{\text{net}}}$. Therefore, in what follows, the optimization subject is the MGF for the network service process, i.e., $\overline{\mathbb{M}}_{S_{\text{net}}}$. Since an exact expression for the MGF of network service process is not attainable, we opt for optimizing a bound on $\overline{\mathbb{M}}_{S_{\text{net}}}$, given by (8), instead. A power allocation that elevates the lower bound on network service process corresponds to lower $\overline{\mathbb{M}}_{S_{\text{net}}}$ and thus reduces the probabilistic bounds on network performance.

Theorem C.2 provides an upper bound on $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t)$ and hence a lower bound on the network service process. Therefore, in order to maximize the lower bound on the service process, we must minimize the upper bound on $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t)$. The function $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t)$ is related to the power allocation vector $\mathbf{P} \in \mathbb{R}_+^{n+1}$. From (8) we have

$$\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t) \leq \sum_{\sum_{i=1}^{n+1} \pi_i = t-s} \prod_{i=1}^{n+1} \left(\mathbb{E}\left[(1+\gamma_i)^{-\theta\eta} \right] \right)^{\pi_i}, \tag{20}$$

where the per node SINR, γ_i, $\forall i \in \mathcal{I}_{\mathcal{H}}$, is given by (17) and is in turn related to the allocated transmit power for the corresponding node. Let $\Xi_n \subset \mathbb{R}_+^{n+1}$ be the set of feasible power allocation schemes with respect to n intermediate nodes. Furthermore, the sum of all allocated power should be constrained by the total power budget, i.e.,

$$P_\Sigma \triangleq \sum_{i \in \mathcal{I}_{\mathcal{H}}} P_i \leq P_{\text{tot}},$$

where P_{tot} is the total power budget. For any power allocation $\mathbf{P} \triangleq \{P_i\}_{i \in \mathcal{I}_{\mathcal{H}}} \in \Xi_n$, we define $\mathbf{P}$ as a *feasible* power allocation scheme if the power constraint above is

[1] Our power allocation scheme applies to, for example, a multi-hop network that is managed by a centric controller for global power assignment, while having no (or very weak) power constraints on all transmitting nodes. More examples on power allocations for multi-hop wireless networks under sum power constraint can be found in [Ron11, HDL11].

satisfied.

To determine the optimal power allocation strategy, we need the following two lemmas first. To be more precise, in Lemma D.1 below, we show that, given a feasible power allocation, the maximal network service process (equivalently, the minimal $\overline{\mathbb{M}}_{S_{\text{net}}}$) is achieved only when the SINRs for all hops are identically distributed. Next, in Lemma D.2 we show the existence and uniqueness of the optimal power allocation vector $\mathbf{P}^*$ when $P_\Sigma = P_{\text{tot}}$.

Lemma D.1 (Sufficiency). *Given the sum transmit power budget P_{tot} for a $(n+1)$-hop wireless network, a feasible power allocation $\mathbf{P}^*$, where $P_\Sigma = P_{\text{tot}}$, that results in identically distributed SINR over all hops maximizes the lower bound on network service process whenever such $\mathbf{P}^*$ exists.*

Proof. Please see Appendix H.3. □

Lemma D.1 shows that the key enabler of optimal multi-hop network operation is the maintenance of identically distributed channels' SINR across all hops. Note that this result applies to arbitrary networks where the service process is characterized by a Shannon-type link model, as long as the individual links are not coupled in terms of interference (only self-interference is considered). In contrast, in what follows we present an approach for allocating transmit power[2] to achieve identically distributed SINR in a mm-wave multi-hop network. Given a background noise power of N_0, we define $\lambda_i = \frac{P_i}{N_0}$. Due to the one-to-one correspondence of λ_i and P_i and for convenience we opt to work with λ_i for the derivations that follows.

For γ_i, $i \in \mathcal{I}_{\mathcal{H}}$, we use

$$\gamma_i = \kappa \cdot 10^{-0.1\alpha} \omega_i l_i^{-\beta} \cdot 10^{-0.1\xi_i}$$

to rewrite (13). Note that the shadowing effects are assumed to be homogeneous and log-normally distributed over all hops, i.e., $\xi_i \sim \mathcal{N}(0, v^2)$, where the log-normal shadowing variance v^2 is independent of the transmit power, since $\omega_i l_i^{-\beta}$ depends on the power allocated to transmitter i, then a power allocation scheme that enables the equality $\omega_i l_i^{-\beta} = \omega_j l_j^{-\beta}$ for any $i, j \in \mathcal{I}_{\mathcal{H}}$ is sufficient to ensure identically distributed SINR $\gamma_i, \forall i$.

We follow an iterative approach starting from the last hop of the network, i.e., the $(n+1)^{\text{th}}$ hop, and moving backwards. Assume that $\omega_i l_i^{-\beta} = c$ holds for any $i \in \mathcal{I}_{\mathcal{H}}$, where c is a constant. Let $i = n+1$, then according to (13), $\omega_{n+1} = \lambda_n$ and therefore, $\lambda_n = c \cdot l_{n+1}^{\beta}$. Similarly, for $i = n$, we obtain

$$\lambda_{n-1} = (1 + \mu_n \lambda_n) \cdot c \cdot l_n^{\beta} = c \cdot l_n^{\beta} + c^2 \mu_n \left(l_n l_{n+1}\right)^{\beta}.$$

[2] The power is assumed to be infinitely divisible during the power allocation.

Then by recursively applying $\lambda_{i-1} = (1 + \mu_i \lambda_i) \cdot c \cdot l_i^{\beta}$, for all $i \in \mathcal{I}_{\mathcal{H}}$, and after some manipulation we obtain

$$\lambda_i = \sum_{k=1}^{n-i+1} c^k \left(\mu_{i+k}^{-1} \prod_{u=1}^{k} \mu_{i+u} \cdot l_{i+u}^{\beta} \right) \triangleq \sum_{k=1}^{n-i+1} \nu_{i,k} \cdot c^k \tag{21}$$

for all $0 \leq i \leq n$.

With the constraint $P_{\Sigma} \leq P_{\mathrm{tot}}$, which corresponds to

$$\lambda_{\Sigma} \triangleq \frac{P_{\Sigma}}{N_0} \leq \frac{P_{\mathrm{tot}}}{N_0} \triangleq \lambda_{\mathrm{tot}},$$

we can obtain the value of c using Lemma D.1 and by solving the equation $\lambda_{\Sigma} = \lambda_{\mathrm{tot}}$ subject to

$$\lambda_{\Sigma} = \sum_{i=0}^{n} \left(\sum_{k=1}^{n-i+1} \nu_{i,k} c^k \right) = \sum_{k=1}^{n+1} \left(\sum_{i=0}^{n+1-k} \nu_{i,k} \right) \cdot c^k, \tag{22}$$

where the last equality is obtained by collecting terms containing c^k for $1 \leq k \leq n+1$. Then $\hat{\gamma}$ can be expressed as

$$\hat{\gamma} = \kappa \cdot 10^{-0.1(\alpha+\xi)} \cdot c. \tag{23}$$

Lemma D.2 (Existence). *Given a $(n+1)$-hop wireless network operating under transmit power budget P_{tot}, there always exists a unique optimal power allocation $\mathbf{P}^*$ such that $P_{\Sigma} = P_{\mathrm{tot}}$, among all feasible $\mathbf{P} \in \Xi_n$ in terms of maximizing a lower bound on network service process.*

Proof. Please see Appendix H.4. □

Lemma D.1 shows that, allocating power in such a way that it results in i.i.d. SINRs across the hops is optimal, while Lemma D.2 states that utilizing all the available power for transmission is optimal since it provides the best network performance. The intuition behind these results is that, avoiding bottleneck is the best strategy, while utilizing more power for transmission enhances network performance. Furthermore, Lemma D.1 and Lemma D.2 provide the minimization on the MGF bound of the service process rather than minimizing the actual process. Nevertheless, such minimization results in maximizing the lower bound on network service which in turn enables the computation of better network backlog and end-to-end delay bounds and more efficient resource allocation and network dimensioning when based on the computed results. In light of the above, we have the following theorem to show the exact power allocations.

Theorem D.1. *Given the total power budget P_{tot}, i.e., $P_{\Sigma} \leq P_{\mathrm{tot}}$, and the background noise power N_0, for the mm-wave channel described in (12), and let x^* denote the positive solution for the algebraic equation*

$$\sum_{k=1}^{n+1} \left(\sum_{i=0}^{n+1-k} \nu_{i,k} \right) x^k = \frac{P_{\mathrm{tot}}}{N_0},$$

with $\nu_{i,k}$ given by

$$\nu_{i,k} = \mu_{i+k}^{-1} \cdot \prod_{u=1}^{k} \mu_{i+u} \cdot l_{i+u}^{\beta},$$

where μ_i and l_i are the model parameters defined in Sec. C, then, there exists a unique optimal power allocation strategy $\mathbf{P}^* \in \Xi_n$, such that

$$P_i^* = N_0 \sum_{k=1}^{n-i+1} \nu_{i,k} \cdot (x^*)^k\,, \quad \textit{for } i \in \mathcal{I}_{\mathcal{H}}\,.$$

Proof. Using Lemmas D.1 and D.2, the theorem immediately follows by applying the mapping between the transmit power and the SINR, i.e., $P_{\text{tot}} = \lambda_{\text{tot}} N_0$ and $P_i = \lambda_i N_0$. □

E Self-interference Impact in mm-wave Networks

To investigate the impact of self-interference on network performance, we consider a particular case, where the separation distances between adjacent nodes are assumed to be equal to l, e.g., $l_i = l$ for $\forall i \in \mathcal{I}_{\mathcal{H}}$, and all relays have an identical self-interference coefficient μ. Closed-form expressions for the network performance can be obtained under these assumptions which provide more insights to the network operation.

In the following analysis, we assume that the optimal power allocation scheme proposed in Theorem D.1 is used. Under this power allocation, we have $c = \omega_i l_i^{-\beta}$ for $\forall i \in \mathcal{I}_{\mathcal{H}}$. Note that c in this case is a measure of the SINR of channels, and hence it directly influences the network performance. (13) shows that, the parameter ω_i, and therefore the function c, are functions of μ. Therefore, in this section we represent this measure by the function $c(\mu) = \omega_i(\mu) l_i^{-\beta}$. Applying (22) under the proposed power allocation scheme as well as the conditions $P_\Sigma = P_{\text{tot}}$ and $\mu_i = \mu$, the optimal λ_i in (21), denoted by λ_i^*, reduces to

$$\begin{aligned}\lambda_i^* &= \mu^{-1} \sum_{k=1}^{n-i+1} \left(c(\mu)\,\mu l^{\beta}\right)^k \\ &= \begin{cases} c(\mu)\, l^{\beta} \cdot \dfrac{\left(c(\mu)\,\mu l^{\beta}\right)^{n-i+1} - 1}{c(\mu)\,\mu l^{\beta} - 1}, & \text{if } c(\mu)\,\mu l^{\beta} \neq 1 \\ \dfrac{2\lambda_{\text{tot}}(n-i+1)}{(n+1)(n+2)}, & \text{otherwise.} \end{cases}\end{aligned} \tag{24}$$

In the above, we used the geometric sum to obtain the first case. It is easy to find that, the case $c(\mu)\,\mu l^{\beta} = 1$ corresponds to a particular situation for the self-interference coefficient μ, i.e., $\mu = (2\lambda_{\text{tot}})^{-1}(n+1)(n+2)$, which immediately gives the optimal $c(\mu)$ as $c(\mu) = 2\lambda_{\text{tot}}\left[(n+1)(n+2)\,l^{\beta}\right]$ by applying $c(\mu) =$

$\left(\mu l^{\beta}\right)^{-1}$. It is worth noting that, $c\left(\mu\right)\mu l^{\beta}=1$ here corresponds to a very special case that rarely occurs in realistic scenarios, since μ, which is only related to the relay implementation, is independent from λ_{tot} and n. In this sense, for most cases $\mu=\left(2\lambda_{\text{tot}}\right)^{-1}\left(n+1\right)\left(n+2\right)$ is not satisfied. Therefore, from the perspective of generality, our interest mainly focuses on the case $c\left(\mu\right)\mu l^{\beta}\neq 1$.

According to the assumption of optimal power allocation that gives $P_{\Sigma}=P_{\text{tot}}$ according to Lemma D.2, we have

$$\begin{aligned}\lambda_{\text{tot}}=&\sum_{i=0}^{n}\lambda_i^*=\frac{c(\mu)l^{\beta}}{c(\mu)\mu l^{\beta}-1}\sum_{i=0}^{n}\left(\left(c\left(\mu\right)\mu l^{\beta}\right)^{n-i+1}-1\right)\\=&\frac{c(\mu)l^{\beta}}{c(\mu)\mu l^{\beta}-1}\left(c\left(\mu\right)\mu l^{\beta}\frac{\left(c\left(\mu\right)\mu l^{\beta}\right)^{n+1}-1}{c\left(\mu\right)\mu l^{\beta}-1}-n-1\right),\end{aligned}\tag{25}$$

where we used the change of variables $j=n-i+1$ and then applied the geometric sum formula in the last step.

For notational simplicity, we set $t=c(\mu)\mu l^{\beta}$ with $t\neq 1$, then (25) can be written as

$$\mu\lambda_{\text{tot}}=\frac{t}{t-1}\left(t\cdot\frac{t^{n+1}-1}{t-1}-(n+1)\right),$$

which subsequently yields

$$t^{n+3}-\left(n+2+\mu\lambda_{\text{tot}}\right)t^2+\left(n+1+2\mu\lambda_{\text{tot}}\right)t=\mu\lambda_{\text{tot}}.\tag{26}$$

Recovering t to $c(\mu)\mu l^{\beta}$ and dividing both sides of (26) by μ, in terms of $c(\mu)$, we immediately obtain

$$a_1(\mu)c^{n+3}\left(\mu\right)+a_2(\mu)c^2\left(\mu\right)+a_3(\mu)c\left(\mu\right)-\lambda_{\text{tot}}=0,\tag{27}$$

where $a_i(\mu)$, $i=1,2,3$ are respectively given by

$$\begin{cases}a_1(\mu)=\mu^{n+2}l^{(n+3)\beta},\\a_2(\mu)=-(n+2+\mu\lambda_{\text{tot}})\mu l^{2\beta},\\a_3(\mu)=(n+1+2\mu\lambda_{\text{tot}})l^{\beta}.\end{cases}$$

To this end, the *Descartes' Sign Rule* [AJS98] can be applied to determine the number of positive roots of (27). The rule states that, when the terms in a polynomial are ordered according to their variable exponent, then the number of positive real roots of that polynomial is either the number of sign changes, say n, between consecutive non-zero coefficients, or is less than that by an even number, i.e., $n,n-2,n-4,\ldots$. Clearly, (27) has one or three real positive root(s), if there exist real positive solutions. In addition, we can see that, there are two repeated positive roots $c_1\left(\mu\right)=c_2\left(\mu\right)=\left(\mu l^{\beta}\right)^{-1}$. By excluding two repeated roots that violate the condition $c\left(\mu\right)\mu l^{\beta}\neq 1$, we are left with one unique positive solution

$c(\mu) \neq \left(\mu l^{\beta}\right)^{-1}$ for (27), under optimal power allocation, which coincides with the claim in Lemma D.2.

For the arbitrary positive integer n, there is no explicit generic analytical solution of (27) for $c(\mu)$. Hence, in order to keep track of $c(\mu)$ with respect to μ, we compute the first derivative over μ on both sides of (27), then we have

$$\begin{aligned} 0 = & a_1'(\mu)c^{n+3}(\mu) + (n+3)\, a_1(\mu)c^{n+2}(\mu)c'(\mu) \\ & + a_2'(\mu)c^2(\mu) + 2a_2(\mu)c(\mu)c'(\mu) \\ & + a_3'(\mu)c(\mu) + a_3(\mu)c'(\mu), \end{aligned}$$

which gives

$$c'(\mu) = -\frac{a_1'(\mu)c^{n+3}(\mu) + a_2'(\mu)c^2(\mu) + a_3'(\mu)c(\mu)}{(n+3)a_1(\mu)c^{n+2}(\mu) + 2a_2(\mu)c(\mu) + a_3(\mu)}. \tag{28}$$

The asymptotic behaviors of $c(\mu)$ in the two cases, i.e., when $\mu \to 0$ and $n \to \infty$, respectively, are of particular interest, which will be investigated in what follows.

1) For $\mu \to 0$: Note that $c(\mu)$ is bounded by $\frac{\lambda_{\text{tot}}}{(n+1)l^{\beta}}$, since $\sum_1^{n+1} c(\mu)l^{\beta} = \sum_1^{n+1} \omega_i \leq \sum_{i=0}^{n} \lambda_i = \lambda_{\text{tot}}$. Then, (28) can be expressed by

$$\lim_{\mu \to 0} c'(\mu) = \frac{(n+2)l^{2\beta}c^2(\mu) - 2\lambda_{\text{tot}}l^{\beta}c(\mu)}{(n+1)l^{\beta}} \leq 0.$$

In the sense of asymptote, we obtain

$$c(\mu) = \frac{2\lambda_{\text{tot}}l^{-\beta}}{n\left(1 + \exp\left(\frac{2\lambda_{\text{tot}}\mu}{n+1}\right)\right) + 2}, \tag{29}$$

by applying the initial condition $c(0) = \frac{\lambda_{\text{tot}}}{(n+1)l^{\beta}}$.

2) For $n \to \infty$: Defining $D \triangleq \mu\lambda_{\text{tot}}$ and $t(\mu) \triangleq c(\mu)\,\mu l^{\beta}$, the derivative (28) can be written as

$$c'(\mu) = -\frac{c(\mu)}{\mu} \frac{\left(n^2 + (D+3)n + 2\right)t(\mu) - (n+2)D}{(n^2 + (D+3)n + 2 + D)\,t(\mu) - (n+3)D}. \tag{30}$$

Evidently, when λ_{tot} is fixed, $n \to \infty$ yields $n^{-1}D \to 0$. Dividing the numerator and denominator of (30) by n^2, we then have

$$\lim_{n \to \infty} c'(\mu) = -\frac{c(\mu)}{\mu},$$

which immediately yields

$$c(\mu) = K\mu^{-1}, \tag{31}$$

where $K > 0$ is a constant scalar.

Following the notation in (23), we denote by $\hat{\gamma}^*(\mu)$ the SINR with respect to the optimal power allocation proposed by Theorem D.1 as a function of self-interference coefficient μ. Regarding the two asymptotic cases above, we have the following theorem and corollaries to address the relationship between network service process under optimal power allocation and the self-interference coefficient.

Theorem E.1. *Assuming the optimal power allocation in Theorem D.1, in the two extreme cases discussed above, the maximized lower bound of network service process decreases as μ grows.*

Proof. Recalling that the lower bound of network service processes is characterized by $\mathbb{E}\left[(1+\hat{\gamma}^*(\mu))^{-\theta\eta}\right]$, we can see that

$$\begin{aligned}\frac{d}{d\mu}\mathbb{E}\left[(1+\hat{\gamma}^*(\mu))^{-\theta\eta}\right] =& \frac{d}{dc(\mu)}\mathbb{E}\left[(1+\hat{\gamma}^*(\mu))^{-\theta\eta}\right]c'(\mu)\\ =& -\theta\eta\cdot\mathbb{E}\left[\frac{\kappa 10^{-0.1(\alpha+\xi)}}{\left(1+\kappa c(\mu)10^{-0.1(\alpha+\xi)}\right)^{1+\theta\eta}}\right]\cdot c'(\mu)\\ =& -\theta\eta\cdot\mathbb{E}\left[\frac{\hat{\gamma}^*(\mu)}{\left(1+\hat{\gamma}^*\left(\mu\right)\right)^{1+\theta\eta}}\right]\cdot\frac{c'(\mu)}{c(\mu)}>0,\end{aligned}$$

since it has been shown that $c'(\mu)<0$ for asymptotic situations, e.g., when $\mu\to 0$, which corresponds to case that μ is extremely small, or $n\to\infty$, which corresponds to the case that n is sufficiently large. This indicates $\mathbb{E}\left[(1+\hat{\gamma}^*(\mu))^{-\theta\eta}\right]$ monotonically increases with μ, which equivalently shows the degradation of the maximized lower bound of network service process. Thus, the theorem is concluded. □

F Numerical Results and Discussion

In this section, we provide numerical results for the probabilistic bounds on total backlog and end-to-end delay performance for mm-wave multi-hop wireless network discussed in Sec. C, and validate them by simulations. Moreover, in the presence of self-interference, the performance of optimal power allocation presented in Sec. D and E is demonstrated and further discussed.

Here, in particular, the 60 GHz band (ranging from 57 GHz to 64 GHz) is selected for our mm-wave multi-hop network, and the common system parameters, including the parameters specifically associated with 60 GHz channels, are summarized in Table 1.4. In addition, we assume the following regarding the network configuration:

- We assume deterministically bounded arrival process[3] with constant parameters $\sigma(\theta)=0$ and $\rho(\theta)=\rho_a$;

[3] We opt to use deterministic arrival bound to highlight the effect of the mm-wave channel variability on the queuing performance of the network.

Table 1.4: System Parameters

Parameters	**Symbol**	**Value**
Bandwidth	W	500 MHz
Antenna Gain Scalar	κ	70 dBi
Power Budget	P_{tot}	50 Watt
Noise Power Density	N_0/W	−114 dBm/MHz
Hop Length	l	0.5 km
Path Loss Intercept	α	70
Path Loss Slope	β	2.45
STD of Shadowing	v	8 dB

- All relays have identical μ, and are uniformly deployed along the path from source to destination;

- Sufficiently large (or infinite) buffer size at each relay, i.e., overflow effects are neglected;

- A time-slotted system with time intervals of 1 second are assumed.

Under this scenario, we investigate the following:

1. We first validate the derived upper bounds for probabilistic end-to-end delay and total backlog from Theorem C.3.

2. Secondly, based on the validated bounds, we investigate the impact of optimal power allocation, self-interference and relay density on the probabilistic performance guarantees of mm-wave multi-hop networks.

For the sake of simplicity, in what follows the analytic bounds for homogeneous scenarios are all illustrated by Theorem C.4, while heterogeneous counterparts are provided by applying Theorem C.3.

F.1 Bound Validation

We start with considering a tandem 60 GHz network consisting of $n = 10$ relays that have identical self-interference coefficient $\mu = -80$ dB. From Table 1.4 and the power constraint formulated in the form of (15), we determine $\lambda_{\text{tot}} = 134$ dB. Figs. 1.3 and 1.4 show the total backlog and the end-to-end delay bounds respectively, compared to their corresponding simulated values. Recall that the SINR distributions are identical per hop due to applying the optimal power allocation policy from Theorem D.1, resulting in $m = 1$ of Theorem C.3.

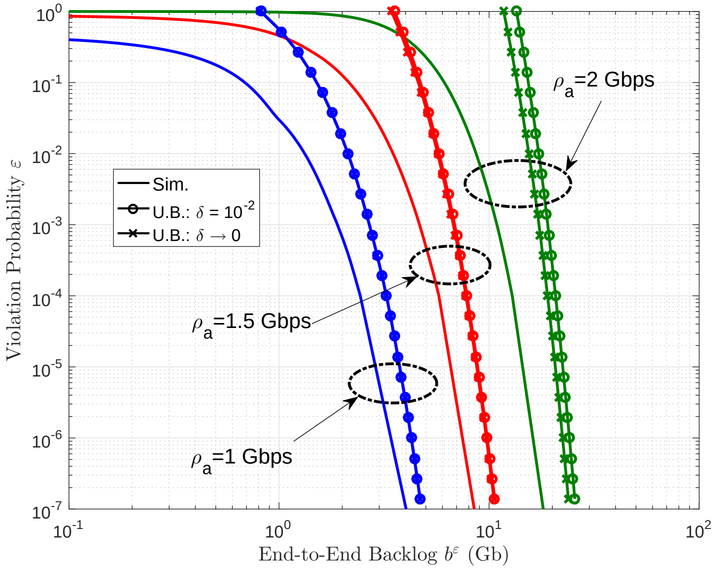


Figure 1.3: Violation probability ε vs. targeted theoretical backlog bounds b^ε, compared to simulations for different $\rho_a = 1$, 1.5 and 2 Gbps, with $n = 10$, and $\delta = 10^{-2}$ and $\delta \to 0$, respectively.

In Fig. 1.3, we validate the backlog bound using simulation for different arrival rates, 1 Gbps, 1.5 Gbps and 2 Gbps. In all cases we notice that the bound accurately predicts the slope of the simulation curve. Furthermore, the gap between the bound and the simulated backlog diminishes asymptotically. The plot furthermore contains the information on the accuracy of bound provided by Lemma C.1, as we compare in the plot the corresponding analytical bounds for a step size of $\delta = 10^{-2}$ as well as for letting $\delta \to 0$. We observe that the two curves are quite close together, which confirms also our findings in [YXG$^+$16], that the MGF bound provided by Lemma C.1 is close to the true MGF of the service process for reasonably small δ. As the system utilization becomes higher (i.e., when the arrival rate $\rho_a = 2$ Gbps), Fig. 1.3 shows that smaller step size δ is required.

Fig. 1.4 depicts the delay violation probability ε (bound and simulated) versus the end-to-end delay w^ε. We notice trends similar to those observed in Fig. 1.3. We also noticed that the bound becomes less tight as the system utilization increases, e.g., the case when $\rho_a = 2$ Gbps.

From the two figures above, we note that the simulated system performance and our computed performance bounds have the same slopes asymptotically, thereby concluding our validations for the derived bounds.

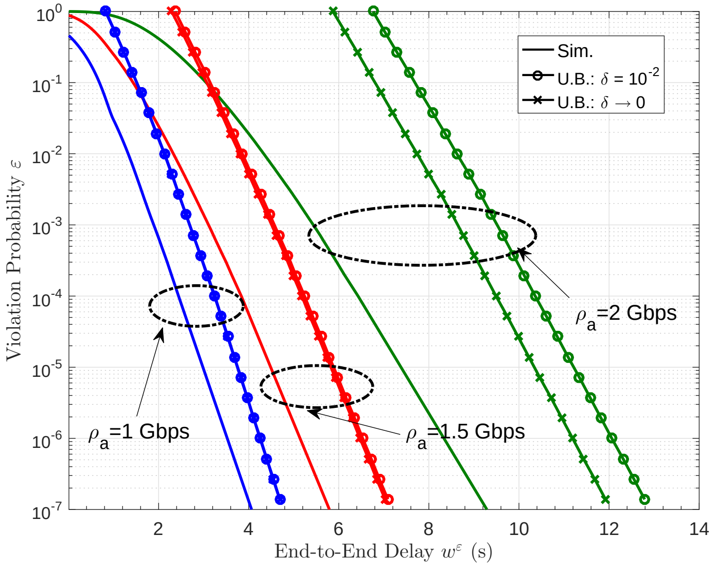


Figure 1.4: Violation probability ε vs. targeted theoretical delay bounds w^{ε}, compared to simulations for different $\rho_a = 1$, 1.5 and 2 Gbps, with $n = 10$, and $\delta = 10^{-2}$ and $\delta \to 0$, respectively.

F.2 Impact of Optimal Power Allocation

Fig. 1.5 demonstrates the merit of adopting the optimal power allocation, with respect to its impact on the end-to-end delay. We consider a network that consists of $n = 10$ relays, the self-interference coefficient is $\mu = -80$ dB, and the arrival rate is fixed to $\rho_a = 1$ Gbps. We use a uniformly allocated powers $\{P_i\}_{i=0}^{n} = \frac{50}{11}$W as a baseline for comparison. From the figure, it is evident that, the upper bound associated with the optimal power allocation (referring to Theorem C.4 for the homogeneous case) is asymptotically tight, while its counterpart (here applying the result for $m \geq 2$ in Theorem C.3, since the non-optimal power allocation yields the heterogeneity) is not. The slackness of bounds for the heterogeneous scenario comes from producing the binomial coefficient of (33) in Theorem C.3, where the upper bound is generalized in a simplified and unified manner. In other words, compared to the recursive approach by [PAZKG15], the tightness of our proposed method is sacrificed for gaining a lower computational complexity for the heterogeneous cases. Fortunately, the asymptotic tightness can be guaranteed for both homogeneous and heterogeneous scenarios, and this allows us to keep track of realistic performance behaviors. Furthermore, under a sum power constraint, we can see that, the network performance without the optimal power allocation suffers severe degradation, in

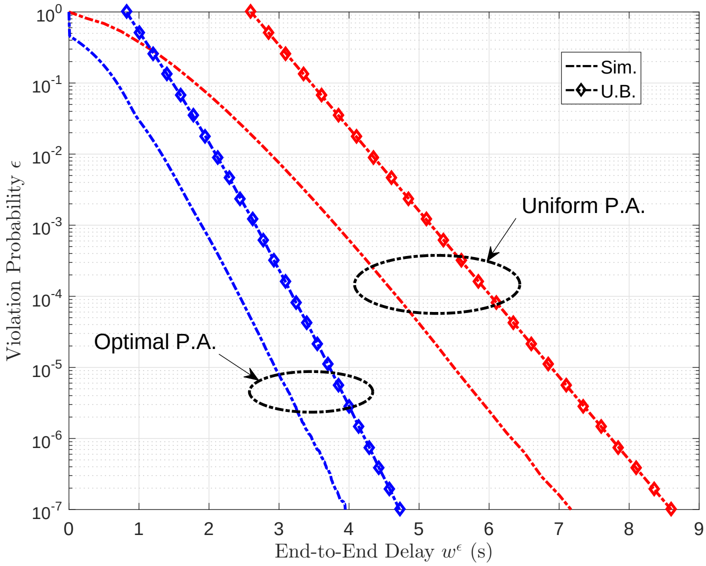


Figure 1.5: Violation probability ε vs. targeted theoretical delay bounds w^{ε}, compared to simulations for two power allocation strategies, with $\mu = -80$ dB, $n = 10$, $\rho_a = 1$ Gbps and $\delta = 10^{-2}$.

terms of the end-to-end delay, and the deterioration similarly happens to the derived bound.

Given different self-interference coefficients, i.e., $\mu = -100$ dB, -90 dB and -80 dB, in Fig. 1.6, we furthermore investigate the impact of self-interference on the network performance bound through the optimal power allocation. It is evident that, the improvement of bound performance by applying the optimal power allocation is significantly exacerbated as μ grows, which can be found through comparing, for example, the gap between optimal and uniform power allocations schemes with $\mu = -80$ dB and that with $\mu = -100$ dB, for any given violation probability ε. Despite the slackness of upper bound for the heterogeneous case (see curves with the uniform power allocation in Fig. 1.5), we are still able to conclude that the optimal power allocation is more important in case of high interference coupling, or low SINR scenarios, in general.

F.3 Impacts of Self-Interference and Relay Density

In the following, we further investigate the performance of 60 GHz networks operated by optima power allocation while varying the self-interference coupling coefficient μ. Here, the separation distance between source and destination is fixed

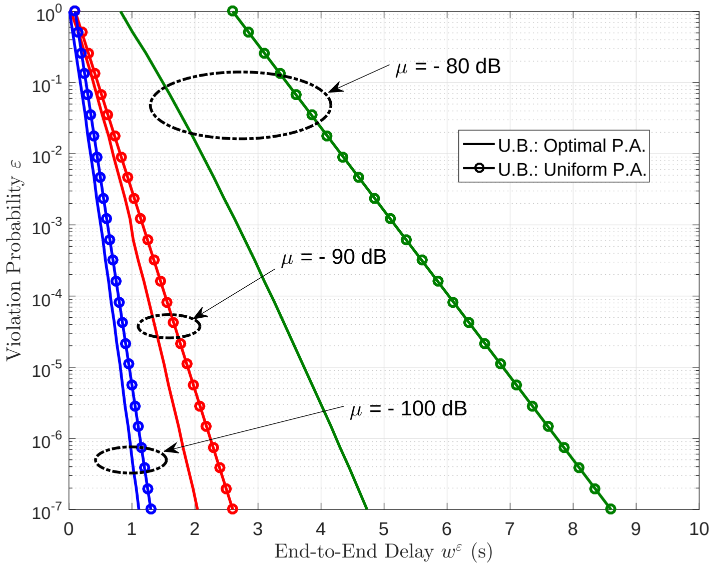


Figure 1.6: Violation probability ε vs. targeted theoretical delay bounds w^{ε}, for two power allocation strategies, with $\mu = -80$ dB, -90 dB and -100 dB, respectively, where $n = 10$, $\rho_a = 1$ Gbps, and $\delta = 10^{-2}$.

to $L = 5$ km, and an arbitrary number of relays with a sum power constraint is uniformly placed between the source and destination nodes. As we deploy more relays, the separation distance between adjacent nodes is shortened as $l = \frac{L}{n+1}$.

From Lemma D.1, we know that the SINR per link, determined by $c\left(\mu\right) = \omega^{*}\left(\mu\right) l^{-\beta}$, yields the service performance, where $\omega^{*}\left(\mu\right)$ denotes the optimal $\omega_i\left(\mu\right)$ for $\forall i \in \mathcal{I}_{\mathcal{H}}$ obtained by applying the optimal power allocation. We study hence the impact of μ on $c(\mu)$ by this relation, rather than straightforwardly to aim at the probabilistic backlog and delay violation probability bounds. The behavior of $c(\mu)$ with respect to the varying self-interference coefficient μ for different relay densities is shown in Fig. 1.7. For all curves, there exists a "waterfall" tendency with regard to $c(\mu)$ as μ increases, i.e., dividing the curve into a "flat" and "falling" stages, respectively. We find that, the point from which on this transition happens depends on the node density n. Taking the scenario $n = 50$ as an example, more precisely, when μ is below $\mu_c \approx -120$ dB, $c\left(\mu\right)$ remains flat by increasing μ. Keeping on elevating μ, however, $c\left(\mu\right)$ will encounter a dramatical decay once μ exceeds μ_c. This behavior relates to the fact that, the system switches from a noise-limited one to an interference-limited one as μ grows, such that the self-interference becomes the dominant factor restricting the level of $c\left(\mu\right)$. We also observe that, generally speaking, a higher relay density will result in a higher $c\left(\mu\right)$. However, when the

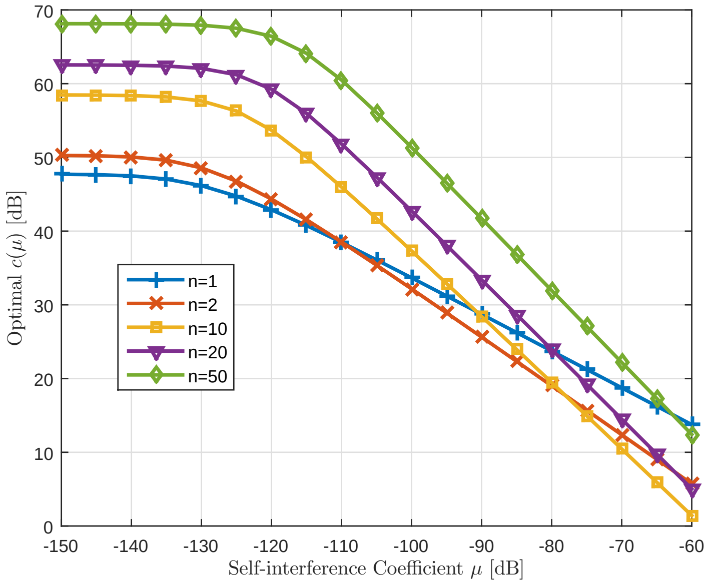


Figure 1.7: $c(\mu)$ vs. μ for different relay densities characterized by $n = 1$, 2, 10, 20 and 50, respectively, with $L = 5$ km, $P_{\text{tot}} = 50$ W.

self-interference coefficient is significant and the system being limited by interference, i.e., $\mu \geq -90$ dB, a sparser relay deployment is surprisingly able to provide a higher $c\,(\mu)$. From Fig. 1.7, we summarize that despite improving $c\,(\mu)$ by means of increasing n, the overall performance improves only if μ relates not to a strongly interference-limited system. Hence, optimizing the network performance by changing the node density must take the self-interference coefficient into account, since higher relay densities do not always imply higher performance.

Given $\varepsilon = 10^{-6}$, in Fig. 1.8, the probabilistic end-to-end delay bound w^{ε} with respect to a varying self-interference coefficient μ and a varying number of relays n is further demonstrated. For more noise-limited systems, the delay bound is barely sensitive to different number of relays (in fact a higher number of relays has a beneficial impact on the delay bound, which is not significantly visible in this plot), while for strongly interference-limited systems, e.g., when $\mu = -70$ dB, either a low or a higher number of relays outperforms relay densities in between significantly. Recall that as the node density increases, the link distances get shorter while the transmit power per relay also decreases. Still, as the results for noise-limited systems demonstrate, the resulting SINR improves as the relay density increases if the self-interference coefficient is small. When the self-interference coefficient increases, the performance degrades as long as the emitted transmit power per relay

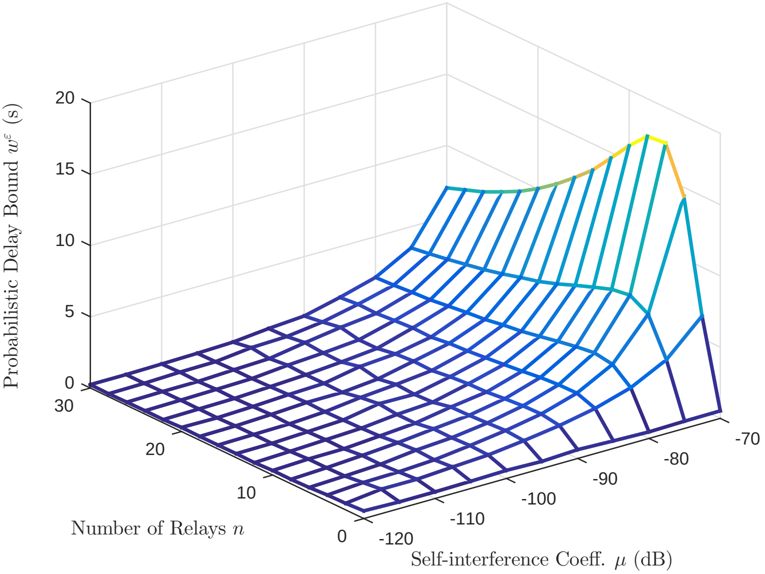


Figure 1.8: Probabilistic delay bound w^{ε} vs. n and μ jointly, with $\varepsilon = 10^{-6}$.

creates a *significant* self-interference with the own receiver. This happens precisely for medium number of relays, while for a larger relay density, the resulting self-interference per node drops below the noise level (asymptotically), thereby leading to a better system performance.

G Conclusions

We investigate stochastic performance guarantees, i.e., the probabilistic end-to-end backlog and delay, for mm-wave multi-hop wireless networks with full-duplex buffered relays, by means of MGF-based stochastic network calculus. According to specific propagation features of mm-wave radios, a cumulative service process characterization with self-interference is proposed, in terms of the MGF of its channel capacity. Based on this characterization, probabilistic upper bounds associated with overall network performance are developed. In addition, we propose an optimal power allocation scheme in the presence of self-interference, aiming at enhancing the network performance. The analytic framework of this paper supports a broad class of multi-hop networks, in terms of homogeneous and heterogeneous, where the asymptotic tightness of computed upper bounds has been validated. Results reveal that, the self-interference coefficient plays a crucial role in improving network performance. Another interesting and important finding is that, given the sum power constraint, increasing the relay density does not always improve network performance unless the self-interference coefficient is sufficiently small. We believe that approaches developed in this paper will have a variety of applications in designing and optimizing networks for next generation wireless communications, in terms of performance guarantees and enhancements.

H Appendices

H.1 Proof of Theorem C.3

By the definition of $\mathsf{M}\left(\theta, s, t\right)$ in (9), we start with the substitution of (16) and (18) in (9), then we have

$$\mathsf{M}(\theta,s,t) \leq e^{\theta\sigma(\theta)} \sum_{u=0}^{\min(s,t)} p_a^{t-u}(\theta) \cdot \sum_{\sum_{i=1}^{m} \pi_i = s-u} \prod_{i=1}^{m} \binom{\pi_i + |\mathcal{X}_i| - 1}{|\mathcal{X}_i| - 1} \hat{q}_i^{\pi_i}(-\theta). \tag{32}$$

Then using the change of variable $u = s - u$ and rearranging terms, we equivalently have

$$\mathsf{M}(\theta,s,t) \leq \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{u=\tau}^{s} p_a^{u}(\theta) \cdot \sum_{\sum_{i=1}^{m} \pi_i = u} \prod_{i=1}^{m} \binom{\pi_i + |\mathcal{X}_i| - 1}{|\mathcal{X}_i| - 1} \hat{q}_i^{\pi_i}(-\theta),$$

Based on above, by splitting the power u for $p_a^u\left(\theta\right)$ into components in terms of π_i, we immediately have

$$\begin{aligned}
\mathsf{M}(\theta,s,t) \leq& \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{u=\tau}^{s} \sum_{\sum_{i=1}^{m} \pi_i = u} \prod_{i=1}^{m} \binom{\pi_i + |\mathcal{X}_i| - 1}{|\mathcal{X}_i| - 1} V_i^{\pi_i}\left(\theta\right) \\
\leq& \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{u=\tau}^{\infty} \sum_{\sum_{i=1}^{m} \pi_i = u} \prod_{i=1}^{m} \binom{\pi_i + |\mathcal{X}_i| - 1}{|\mathcal{X}_i| - 1} V_i^{\pi_i}\left(\theta\right),
\end{aligned}$$

where $V_i\left(\theta\right) \triangleq p_a(\theta)\hat{q}_i(-\theta)$ and $\tau \triangleq \max\left(s-t, 0\right)$ are respectively defined for notional simplicity, and the last inequality is obtained by pushing s to infinity. Then, we further have

$$\mathsf{M}\left(\theta,s,t\right) \leq \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{u=\tau}^{\infty} \binom{u+n+1-m}{n+1-m} \sum_{\sum_{i=1}^{m} \pi_i = u} \prod_{i=1}^{m} V_i^{\pi_i}\left(\theta\right), \tag{33}$$

where the following inequality for combinatorics is used, i.e.,

$$\binom{n_1}{k_1}\binom{n_2}{k_2}\cdots\binom{n_M}{k_M} \leq \binom{n_1+n_2+\cdots+n_M}{k_1+k_2+\cdots+k_M}$$

for all $n_i \geq k_i \geq 0$, $i \in \{1, 2, \ldots, M\}$.

From (33) on, we need to consider two situations: (i) $m = 1$ for the homogeneous network, and (ii) $m \geq 2$, which represents the heterogeneous scenario.

For $m = 1$, the expression in (33) directly reduces to

$$\mathsf{M}\left(\theta, s, t\right) \leq \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{u=\tau}^{\infty} \binom{u+n}{n} V^u\left(\theta\right), \tag{34}$$

where $V\left(\theta\right) = p_a\left(\theta\right)\hat{q}\left(-\theta\right)$, and $\hat{q}\left(-\theta\right)$ characterizes the homogeneous channel gain.

Regarding $m \geq 2$, the upper bound of $\mathsf{M}\left(\theta, s, t\right)$ can be formulated as

$$\begin{aligned}\mathsf{M}\left(\theta, s, t\right) \leq & \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{u=\tau}^{\infty} \binom{u+n+1-m}{n+1-m} \sum_{i=1}^{m} \varphi_i\left(\theta\right) V_i^{u+m-1}\left(\theta\right) \\ = & \frac{e^{\theta\sigma(\theta)}}{p_a^{s-t}(\theta)} \sum_{i=1}^{m} \varphi_i\left(\theta\right) V_i^{m-1}\left(\theta\right) \sum_{u=\tau}^{\infty} \binom{u+n+1-m}{n+1-m} V_i^u\left(\theta\right),\end{aligned} \tag{35}$$

where $\varphi_i\left(\theta\right) = \prod_{j\neq i}\left(V_i(\theta) - V_j(\theta)\right)^{-1}$. Here, the inequality comes from the application of following homogeneous polynomials identity [BB06], that is,

$$\sum_{k_1+\cdots+k_M=N} x_1^{k_1} x_2^{k_2} \cdots x_M^{k_M} = \sum_{i=1}^{M} \frac{x_i^{N+M-1}}{\prod_{j\neq i}(x_i - x_j)}$$

for any $M \geq 2$ distinct variables, $x_1, x_2, \ldots, x_M$.

Furthermore, we have a closed-form expression in terms of generalized hypergeometric function for the infinite sum [Wol18]

$$\sum_{k=u}^{\infty} \binom{n+k}{k} x^k = x^u \binom{n+u}{u} {}_2F_1\left(1, n+u+1; u+1; x\right),$$

whenever $|x| < 1$, which subsequently produces the newly defined $\mathcal{K}_{\tau,n,m}\left(x\right)$. It is worth noting that, the condition $\max\limits_{i\in\mathcal{I}_{\mathcal{M}}}\{V_i(\theta)\} < 1$ should be satisfied, for the sake of stability. Combining (34) and (35) and using (19), then the theorem can be concluded.

H.2 Proof of Lemma C.2

Note that $\mathcal{K}_{\tau,n,1}\left(x\right)$ is explicitly formulated as

$$\mathcal{K}_{\tau,n,1}\left(x\right) = \sum_{k=\tau}^{\infty} \binom{n+k}{k} x^k,$$

then the derivation can be performed from the the following two cases:

(i) It is easy to know that, $i+k \leq (i+k-\tau)(1+\frac{\tau}{i})$ holds for all $i \geq 1$ and all $k \geq \tau$, then we have

$$\binom{n+k}{k} \leq \binom{n+k-\tau}{k-\tau}\binom{n+\tau}{n}.$$

Then we have

$$\mathcal{K}_{\tau,n,1}(x) \leq x^{\tau}\binom{n+\tau}{n}\sum_{k=0}^{\infty}\binom{n+k}{k}x^{k}, \tag{36}$$

where we used the change of variables, i.e., $k=k-\tau$. It immediately gives $\mathcal{G}_1(x)$ by applying *Newton's Generalized Binomial Theorem* and by taking 1 as the upper limit into account as well.

(ii) On the other hand, it is easy to know that

$$\begin{aligned}\sum_{k=\tau}^{\infty}\binom{n+k}{k}x^{k} =& \frac{1}{(1-x)^{n+1}} - \sum_{k=0}^{\tau-1}\binom{n+k}{k}x^{k}\\ \leq& \frac{1}{(1-x)^{n+1}} - x^{\tau-1}\sum_{k=0}^{\tau-1}\binom{n+k}{k}\\ =& \frac{1}{(1-x)^{n+1}} - \binom{n+\tau}{n+1}x^{\tau-1},\end{aligned}$$

where inequality is true since $0 \leq x < 1$, and the last equality is obtained by applying a combinatorial property [Bru92]. Then $\mathcal{G}_2(x)$ is obtained.

H.3 Proof of lemma D.1

Note that the network service process is characterized by its MGF bound, $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta,s,t)$, $s>0$. Furthermore, a lower bound on the service process corresponds to an upper bound on $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta,s,t)$, which is given by (20). Since $P_{\Sigma}=P_{\text{tot}}$ by assumption, the optimization problem can be formulated as

$$\mathbf{P}^{*} = \underset{\mathbf{P}:P_{\Sigma}=P_{\text{tot}}}{\operatorname{argmin}} \sum_{\sum_{i=1}^{n+1}\pi_i=t-s}\prod_{i=1}^{n+1}\left(\mathbb{E}\left[(1+\gamma_i)^{-\theta\eta}\right]\right)^{\pi_i}. \tag{37}$$

To prove the lemma, we use the *generalized Haber inequality* [BA08]. That is, for any m non-negative real numbers $x_1, x_2, \ldots, x_m$, and for all $n \geq 0$, we have

$$\sum_{i_1+\cdots+i_m=n}\prod_{k=1}^{m}x_i^{i_k} \geq \binom{n+m-1}{m-1}\left(\frac{1}{m}\sum_{k=1}^{m}x_k\right)^{n}.$$

The above holds with equality *if and only if* $x_i = x_j$ for any $1 \leq i, j \leq m$. Using this result, the minimum in (37) can be achieved by choosing a power allocation vector $\mathbf{P}^*$ such that

$$\mathbb{E}\left[(1+\gamma_i)^{-\theta\eta}\right] = \mathbb{E}\left[(1+\gamma_j)^{-\theta\eta}\right]$$

holds for all $i, j \in \mathcal{I}_{\mathcal{M}}$, which equivalently indicates that the SINRs are identically distributed across all $n+1$ hops, i.e., $\gamma_j \overset{d}{=} \gamma_i$, for all $i \in \mathcal{I}_{\mathcal{H}}$. The resulting MGF service bound under $\mathbf{P}^*$ power allocation is then given by

$$\begin{aligned}\overline{\mathbb{M}}^*_{S_{\mathrm{net}}}(\theta, s, t) \leq & \binom{t-s+n}{n} \left(\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E}\left[(1+\gamma_i)^{-\theta\eta}\right] \right)^{t-s} \\ = & \binom{t-s+n}{n} \left(\mathbb{E}\left[(1+\hat{\gamma})^{-\theta\eta}\right] \right)^{t-s} . \end{aligned} \tag{38}$$

The first step is obtained by applying the generalized Haber inequality, with *strict equality* under optimal power allocation, to (20). The last step follows since $\gamma_i, \forall i \in \mathcal{I}_{\mathcal{H}}$ are i.i.d. under optimal power allocation, and the fact that a random variable is uniquely determined by its MGF, and by letting $\hat{\gamma} \overset{d}{=} \gamma_i$.

H.4 Proof of lemma D.2

The proof consists of two parts: (i) the existence and uniqueness of the power allocation scheme given $P_\Sigma = P_{\mathrm{tot}}$, and (ii) the optimality. The notations in derivations below follow the fashion we used previously.

(i) To show that there is exactly one real root c that meets the constraint on λ_{tot}, we first construct $f(x)$ as follows

$$f(x) \triangleq \sum_{k=1}^{n+1} \left(\sum_{i=0}^{n+1-k} \nu_{i,k} \right) x^k - \lambda_{\mathrm{tot}}. \tag{39}$$

Due to the facts that $f(0) = -\lambda_{\mathrm{tot}} < 0$ and $f(+\infty) = +\infty$, since $f(x)$ is continuous, by the *Intermediate Value Theorem*, there is a positive c such that $f(c) = 0$, which means that equation $f(x) = 0$ has a root. Furthermore, to prove the uniqueness of the root, we assume equation $f(x) = 0$ has at least two positive roots x_1 and x_2. Suppose that $x_1 < x_2$ such that $f(x_1) = f(x_2) = 0$. Note that $f(x)$ is continuous on the closed interval $[x_1, x_2]$ and differentiable on the open interval (x_1, x_2), it is easy to find that the three hypotheses of *Rolle's Theorem* are simultaneously satisfied. Thus, there should be a number $x^* \in (x_1, x_2)$ such that $f'(x^*) = 0$. However,

$$f'(x) = \sum_{k=1}^{n+1} \left(\sum_{i=0}^{n+1-k} \nu_{i,k} \right) kx^{k-1} > 0$$

holds for any point $x \in (0, +\infty)$, which contradicts the initial assumption and hence we conclude that $x_1 = x_2$ which proves the uniqueness of the root for the equation $f(x) = 0$ and hence the uniqueness of the optimal power control vector.

(ii) To prove the optimality of $\mathbf{P}^*$, we assume that the optimal solution is obtained with the sum power $P'_{\sum}$ that relates to the power allocation vector $\mathbf{P}'$, where $P'_{\sum} < P^*_{\sum} = P_{\text{tot}}$. In other words, we can obtain $\hat{\gamma}'$ associated with $\mathbf{P}'$, such that

$$\mathbb{E}\left[(1+\hat{\gamma}')^{-\theta\eta}\right] \leq \mathbb{E}\left[(1+\hat{\gamma}^*)^{-\theta\eta}\right], \tag{40}$$

where $\hat{\gamma}^*$ similarly relates to $\mathbf{P}^*$. For notational simplicity, by using (23), we define

$$g(c) \triangleq \mathbb{E}\left[(1+\hat{\gamma})^{-\theta\eta}\right] = \mathbb{E}\left[(1+\kappa 10^{-0.1(\alpha+\xi)}c)^{-\theta\eta}\right].$$

Let us consider all possible power allocation schemes $\mathbf{P}$ such that $P_\Sigma \leq P_{\text{tot}}$. It is evident that, the derivative of $g(c)$ over P_Σ gives

$$\begin{aligned}\frac{dg(c)}{dP_\Sigma} =& g'(c) \cdot \frac{dc}{dP_\Sigma} = g'(c) \cdot \left(\frac{dP_\Sigma}{dc}\right)^{-1} \\ =& -\theta\eta \cdot \mathbb{E}\left[\frac{\kappa 10^{-0.1(\alpha+\xi)}}{\left(1+\kappa 10^{-0.1(\alpha+\xi)}c\right)^{1+\theta\eta}}\right] \cdot \frac{N_0}{f'(c)} < 0.\end{aligned} \tag{41}$$

which indicates that $g(c)$ monotonically decreases in P_Σ. Since c is uniquely determined by P_Σ as shown above and because of (41), we conclude that

$$\mathbb{E}\left[(1+\hat{\gamma}')^{-\theta\eta}\right] > \mathbb{E}\left[(1+\hat{\gamma}^*)^{-\theta\eta}\right],$$

which contradicts the initial assumption in (40). Thus, the optimal solution is achieved only when $P_\Sigma = P_{\text{tot}}$ is satisfied, which completes the proof.

Paper J4

Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification?

Guang Yang, Ming Xiao, and H. Vincent Poor

Refereed article to appear in
IEEE Transactions on Communications, 2018.

Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification?

Guang Yang, Ming Xiao, and H. Vincent Poor

Abstract

Low latency is critical for many applications in wireless communications, e.g., vehicle-to-vehicle (V2V), multimedia, and industrial control networks. Meanwhile, for the capability of providing multi-gigabits per second (Gbps) rates, millimeter-wave (mm-wave) communication has attracted substantial research interest recently. This paper investigates two strategies to reduce the communication delay in future wireless networks: traffic dispersion and network densification. A hybrid scheme that combines these two strategies is also considered. The probabilistic delay and effective capacity are used to evaluate performance. For probabilistic delay, the violation probability of delay, i.e., the probability that the delay exceeds a given tolerance level, is characterized in terms of upper bounds, which are derived by applying stochastic network calculus theory. In addition, to characterize the maximum affordable arrival traffic for mm-wave systems, the effective capacity, i.e., the service capability with a given quality-of-service (QoS) requirement, is studied. The derived bounds on the probabilistic delay and effective capacity are validated through simulations. These numerical results show that, for a given sum power budget, traffic dispersion, network densification, and the hybrid scheme exhibit different potentials to reduce the end-to-end communication delay. For instance, traffic dispersion outperforms network densification when high sum power budget and arrival rate are given, while it could be the worst option, otherwise. Furthermore, it is revealed that, increasing the number of independent paths and/or relay density is always beneficial, while the performance gain is related to the arrival rate and sum power, jointly. Therefore, a proper transmission scheme should be selected to optimize the delay performance, according to the given conditions on arrival traffic and system service capability.

A Introduction

A.1 Background

Wireless communications in mm-wave bands (from around 24 GHz to 300 GHz) is a key enabler for multi-gigabits per second (Gbps) transmission [RSM$^+$13, RRE14, XMH$^+$17]. In contrast to conventional wireless communications in sub-6 GHz bands, many appealing properties, including the abundant spectral resources, lower component costs, and highly directional antennas, make mm-wave communications attractive for future mobile communications standards, e.g., ECMA, IEEE 802.15.3

Task Group 3c (TG3c), IEEE 802.11ad standardization task group, and Wireess Gigabit Alliance (WiGig).

As an important metric for evaluating the quality of service (QoS), low latency plays a crucial role in the forthcoming fifth generation (5G) mobile communications [FZM+17, SMS+17, OBB+14], especially for various delay-sensitive applications, e.g., high-definition television (HDTV), intelligent transport system, vehicle-to-everything (V2X), machine-to-machine (M2M) communication, and real-time remote control. The overall delay in wireless communications consists of four components as follows [BGH87, KR09]: propagation delay (time for sending the one bit to its designated end via the physical medium), transmission delay (time for pushing the packet into the communication medium in use), processing delay (time for analyzing a packet header and making a routing decision), and queuing delay (the time that a packet spends in the buffer or queue, i.e., waiting for transmission). Normally, the overall delay for queuing system is dominantly determined by the queuing delay, while the contributions by the other types of delay are nearly negligible. Thus, for low-latency buffer-aided systems, the major task is to largely decrease the queuing delay.

In recent years, many efforts from different aspects have been devoted to low-latency mm-wave communications. In [FZM+17], several critical challenges and possible solutions for delivering end-to-end low-latency services in mm-wave cellular systems were comprehensively reviewed, from the perspectives of protocols at the medium access control (MAC) layer, congestion control, and core network architecture. By applying the Lyapunov technique for the utility-delay control, the problem of ultra-reliable and low-latency in mm-wave-enabled massive multiple-input multiple-output (MIMO) networks was studied in [VLB+17]. Regarding hybrid beamforming in mm-wave MIMO systems, a novel algorithm for achieving the ultra-low latency of mm-wave communications was proposed in [CLW16], where the training time can be significantly reduced by progressive channel estimations. Furthermore, for systems with buffers at transceivers, the probabilistic delay for point-to-point mm-wave communications is analyzed in [YXG+16], where the delay bound is derived based on network calculus theory.

A.2 Motivation

Due to unprecedented data volumes in mm-wave communications, the transceivers for many applications are commonly equipped with large-size buffers, such that the data arrivals that cannot be processed in time will be temporarily queued up in the buffer until corresponding service is provided. Hence, the low-latency problem for mm-wave communications with buffers can be interpreted as a delay problem in queuing systems, equivalently. By queuing theory, it is known that the key idea for effectively reducing the queuing delay is to keep lower service utilization. That is, the average arrival rate of data traffic should be less than the service rate of server as much as possible.

Commonly, low service utilization can be fulfilled mainly through two distinct methods: offloading arrival traffic and improving the service capability. In a wireless network, offloading arrival traffic can be realized by adopting the traffic dispersion scheme, and service enhancement can be realized by adopting the network densification scheme. Traffic dispersion stems from the application of distributed antenna systems (DASs) or distributed remote radio heads (RRHs) in mm-wave communications, and network densification is motivated from the trend of dense deployment for mm-wave networks. Roughly, the traffic dispersion scheme applies the “divide-and-conquer” principle, which enables parallel transmissions to fully exploit the spatial diversity, such that a large single queue (or large delay, equivalently) can be avoided. On the other hand, the network densification scheme departs from reducing the path loss, via shortening the separation distance between adjacent nodes, such that the end-to-end service capability can be improved.

Clearly, both the traffic dispersion and network densification schemes are promising and competitive candidates for low-latency mm-wave communications. Though there are many research contributions in low-latency communications based on above two principles, the existing literature focus on either the dispersion scheme (e.g., [MPH06,PFLZ12,RFC12,HCCP16,FJ16,RPC16]) or multi-hop relaying scheme (e.g., [CBL06, LH08, BA09, AZLB16]). It is not clear yet which scheme can provide better delay performance. For designing or implementing mm-wave networks, it is essential to explore the respective strengths of traffic dispersion and network densification, and know the their applicability and capability for realizing low-latency mm-wave networks. Moreover, a combination of traffic dispersion and network densification, termed as “hybrid scheme”, is worthy of study. Intuitively, the hybrid strategy takes advantages of both traffic dispersion and network densification, and potentially can improve the delay performance under certain constraints.

A.3 Objectives and Contributions

The main objective of our work is to investigate the potential of delay performance of three transmission schemes for mm-wave communications, i.e., traffic dispersion, network densification, and the hybrid scheme. Considering buffer-aided mm-wave systems, the delay performance is studied in terms of probabilistic delay and effective capacity, respectively. More precisely, to characterize the violation probability of delay, we derive corresponding probabilistic delay bounds by applying *stochastic network calculus* based on moment generating function (MGF). Furthermore, *effective capacity* is investigated to characterize the maximum asymptotic service capability. The main contributions can be summarized as follows:

- To the best of our knowledge, the comparison between traffic dispersion and network densification has not been performed previously, and thus the advantage of each scheme for delay performance is not yet clear. In this paper, we comprehensively investigate the respective strengths of above two strategies. Furthermore, we propose a generic hybrid scheme, i.e., a flexible combination of traffic dispersion and network densification, and study its potentials in reducing the communication delay. Thus, our work not only exhibits the benefits of two basic transmission strategies in different scenarios, but also explores the capability of the novel hybrid scheme in reducing the delay for certain scenarios.

- Using MGF-based stochastic network calculus, probabilistic delay bounds for traffic dispersion, network densification, and the hybrid scheme are derived, respectively. Compared to most of existing results that only consider homogeneous networks, we study the probabilistic delay heterogeneous settings for the sake of generality. Our work contributes to stochastic network calculus theory by extending the application of stochastic network calculus from homogeneous cases to heterogeneous scenarios. Though the extension to heterogeneous scenarios is not complicated, we still provide detailed derivations regarding probabilistic delay bounds via the MGF-based method, resulting in a self-contained paper for better illustration.

- Another contribution relative to MGF-based stochastic network calculus is its application in mm-wave networks. Actually, the research regarding delay analysis in wireless communications using stochastic network calculus is rather limited, although the theory has developed for decades. A remarkable achievement is the development of $(\min, \times)$-algebra [AZLB13, PAZKG15, AZLB16], which was proposed to bridge the conventional stochastic network calculus and its applications in wireless scenarios. However, only Rayleigh fading channel is considered for discussion in related literature, while the investigation with respect to mm-wave fading characteristics, e.g., Nakagami-m fading, still remains blank. In our research, we not only provide closed-form expressions for the MGF of the service process, specifically for Nakagami-m fading channels, but also show the feasibility of using $(\min, +)$-algebra to address problems in wireless communications. In this sense, we explored a novel alternative parallel to the $(\min, \times)$-based methodology.

- It is known that, effective capacity can be used to analyze the maximum service capability in the asymptotic sense. Despite that significant progresses have been achieved in recent years, there however still remain several open issues for effective capacity, e.g., generic formulations for considered transmission schemes and analysis for networks with an arbitrary number of tandem servers. In our work, we show the maximum effective capacity in traffic dispersion with given sum power constraint, and identify the condition for achieving the optimum. Furthermore, although closed-form expressions of effective capacity for network densification and the hybrid scheme cannot be obtained (due to the convolution in $(\min, +)$-algebra), we derive lower and upper bounds to characterize their actual effective capacity. Numerical results demonstrate that the analytical lower and upper bounds are quite close to each other. Thus, the feasibility of using our derived bounds to capture the effective capacity of networks (partially or fully) with tandem architectures is validated.

- It is demonstrated that, traffic dispersion, network densification and the hybrid scheme have respective advantages, and resulting end-to-end delay performance depends on the sum power budget and the density of data traffic (e.g., average arrival rate). For instance, when the given sum power is large, traffic dispersion, the hybrid scheme, and network densification are suitable for the scenarios with heavy, medium, and light arrival traffic, respectively. However, when the given sum power is small, the corresponding strengths of above three schemes significantly change with respect to arrival traffic. These observations provide interesting insights for mm-wave network designs and implementations. That is, the transmission scheme for low-latency performance should be properly selected according to the density of arrival traffic and/or the feasible system gain.

The remainder of this paper is outlined as follows. In Sec. B, preliminaries for MGF-based stochastic network calculus are provided. In Sec. C, we give system models for traffic dispersion, network densification, and the hybrid scheme, respectively, and present MGF-based characterizations for traffic and service processes in mm-wave systems. For three low-latency schemes, corresponding probabilistic delay bounds are derived in Sec. D, by applying MGF-based stochastic network calculus. In Sec. E, we give a closed-form expression for the effective capacity for traffic dispersion, and derive lower and upper bounds to characterize the effective capacity for both network densification and the hybrid scheme. Performance evaluations are presented in Sec. F, where the derived results are validated, and the delay performance of three schemes is discussed. Conclusions are finally drawn in Sec. G.

B Preliminaries for Network Calculus

In this section, for illustration purpose, we will review network calculus theory briefly and present preliminaries for MGF-based stochastic network calculus. More details for the presented fundamental results can be found in [Cha94,LBT01,Cha00, JL08,FR15,Fid10].

B.1 Traffic and Service Process

Considering a fluid-flow, discrete-time queuing system with a buffer of infinite size, within time interval $[s,t)$, $0 \leq s \leq t$, the non-decreasing bivariate processes $A(s,t)$, $D(s,t)$ and $S(s,t)$ are defined as the cumulative arrival traffic to, departure traffic from, and service offered by server, respectively. We assume $A\,(s,t)$, $D\,(s,t)$ and $S\,(s,t)$ are stationary non-negative random processes, and their values are zeros whenever $s \geq t$. Furthermore, the cumulative arrival and service processes are the sum of instantaneous realizations of each time slot within the given time interval. More exactly, let a_i and s_i represent the instantaneous values of arrival and service in the i^{th} time slot, respectively, then $A(s,t)$ and $S(s,t)$ are given as

$$A(s,t) = \sum_{i=s}^{t-1} a_i \quad \text{and} \quad S(s,t) = \sum_{i=s}^{t-1} s_i\,, \tag{1}$$

for all $0 \le s \le t$, where time slots are normalized to 1 time unit.

For network calculus, the convolution and deconvolution in $(\min,+)$-algebra, denoted by $\otimes$ and $\oslash$, respectively, are two critical operators for deriving performance bounds of queuing systems. Their definitions with respect to the non-decreasing and strictly positive bivariate processes $X(s,t)$ and $Y(s,t)$ are respectively given as

$$(X \otimes Y)(s,t) \triangleq \inf_{s \le \tau \le t} \{X(s,\tau) + Y(\tau,t)\}$$

and

$$(X \oslash Y)(s,t) \triangleq \sup_{0 \le \tau \le s} \{X(\tau,t) - Y(\tau,s)\}.$$

With respect to cumulative arrival $A(s,t)$ and cumulative service process $S\left(s,t\right)$, according to network calculus, cumulative departure $D(s,t)$ is characterized as [Cha00]

$$D(s,t) \ge (A \otimes S)(s,t), \tag{2}$$

where the equality can be achieved if the system is linear [LBT01]. Then, in terms of $A\left(s,t\right)$ and $D\left(s,t\right)$, the virtual delay at time t is defined as $W(t) \triangleq \inf\{w \ge 0 : A(0,t) \le D(0,t+w)\}$, which is further upper bounded by

$$W(t) \le \inf\{w \ge 0 : (A \oslash S)(t+w,t) \le 0\}. \tag{3}$$

Moreover, an appealing and important property of network calculus theory is the capability of dealing with tandem systems, where the equivalent end-to-end network service process can be computed as the $(\min,+)$ convolution of the individual service processes. More exactly, given n concatenated servers, the end-to-end network service process $S_{\text{net}}\left(s,t\right)$ is given by

$$S_{\text{net}}\left(s,t\right) = \left(S_1 \otimes S_2 \otimes \cdots \otimes S_n\right)\left(s,t\right), \tag{4}$$

where $S_i\left(s,t\right)$ for any $1 \le i \le n$ denotes the service process on the i^{th} hop.

B.2 MGF-based Probabilistic Bounds

For queuing systems with stochastic traffic and/or service processes, the MGF-based stochastic network calculus [Fid10] is used to effectively characterize the probabilistic delay. Among various MGF-based approaches for stochastic processes, the Chernoff bound is widely used in stochastic network calculus for deriving probabilistic bounds. More exactly, for random variable X and given x, the Chernoff bound on $\Pr(X \geq x)$ is given as

$$\Pr(X \geq x) \leq e^{-\theta x}\mathbb{E}\left[e^{\theta X}\right] = e^{-\theta x}\mathbb{M}_X(\theta),$$

where $\mathbb{E}[Y]$ and $\mathbb{M}_Y(\theta)$ are the mean value and the MGF (or the Laplace transform) with respect to Y, respectively, and θ is a positive free parameter. For any stochastic process $X(s,t), t \geq s$, the MGF of X for any $\theta \geq 0$ is defined as [Fid06b]

$$\mathbb{M}_X(\theta, s, t) \triangleq \mathbb{E}\left[e^{\theta X(s,t)}\right].$$

Moreover, $\overline{\mathbb{M}}_X(\theta, s, t) \triangleq \mathbb{M}_X(-\theta, s, t) = \mathbb{E}\left[e^{-\theta X(s,t)}\right]$ is also defined, likewise.

The following two inequalities regarding the MGF, associated with convolution $\otimes$ and deconvolution $\oslash$, respectively, are extensively applied in MGF-based network calculus [Fid06b]. More exactly, let $X(s,t)$ and $Y(s,t)$ be independent random processes, we have

$$\overline{\mathbb{M}}_{X\otimes Y}(\theta, s, t) \leq \sum_{u=s}^{t} \overline{\mathbb{M}}_X(\theta, s, u) \cdot \overline{\mathbb{M}}_Y(\theta, u, t) \tag{5}$$

and

$$\mathbb{M}_{X\oslash Y}(\theta, s, t) \leq \sum_{u=0}^{s} \mathbb{M}_X(\theta, u, t) \cdot \overline{\mathbb{M}}_Y(\theta, u, s). \tag{6}$$

Based on (5) and (6), many properties for MGF-based stochastic network calculus are summarized in [Fid06b].

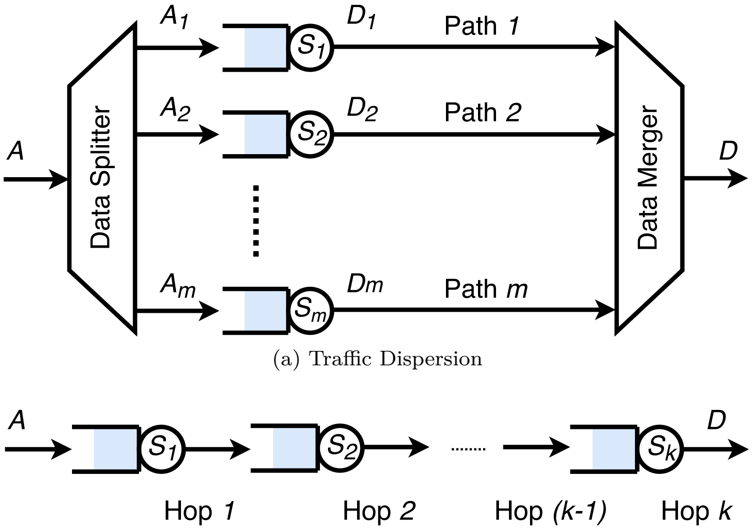
For the tandem network shown in (4), the corresponding MGF is written as $\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t)$ for the system with n tandem servers, which is bounded by

$$\overline{\mathbb{M}}_{S_{\text{net}}}(\theta, s, t) \triangleq \overline{\mathbb{M}}_{S_1\otimes S_2\otimes\cdots\otimes S_n}(\theta, s, t) \leq \sum_{s\leq u_1\leq\cdots\leq u_{n-1}\leq t} \prod_{i=1}^{n} \overline{\mathbb{M}}_{S_i}(\theta, u_{i-1}, u_i), \tag{7}$$

where $u_0 = s$ and $u_N = t$, and $S_i, i = 1, \ldots N$ denotes the service process on each hop. (7) is obtained via applying the union bound and independence assumption [Fid06b].

Assuming independent arrival traffic and service process, in terms of MGF-based characterizations for cumulative arrival traffic and cumulative service process, i.e., $\mathbb{M}_A(\theta, s, t)$ and $\overline{\mathbb{M}}_S(\theta, s, t)$, respectively, we define $\mathsf{M}_{A,S}(\theta, s, t)$ as

$$\mathsf{M}_{A,S}(\theta, s, t) \triangleq \sum_{u=0}^{\min(s,t)} \mathbb{M}_A(\theta, u, t) \cdot \overline{\mathbb{M}}_S(\theta, u, s). \tag{8}$$



(a) Traffic Dispersion

(b) Network Densification

Figure 1.1: Illustrations of two schemes for reducing latency for mm-wave communications: (a) traffic dispersion, and (b) network densification.

Then, based on (8), the violation probability is bounded as

$$\begin{aligned}\Pr\left(W(t) \geq w\right) &\leq \Pr\left((A \oslash S)(t+w,t) \geq 0\right)\\ &\leq \inf_{\theta>0} \mathbb{M}_{A\oslash S}\left(\theta, t+w, t\right) \leq \inf_{\theta>0} \mathsf{M}_{A,S}\left(\theta, t+w, t\right) \triangleq \epsilon^{w},\end{aligned} \tag{9}$$

where the Chernoff bound and the inequality in (6) are used, and ϵ^w denotes the violation probability of delay. The last line in (9) is obtained by applying the inequality for $\mathbb{M}_{A\oslash S}$ (refer to (6)) and the definition in (8). Solving w, we can obtain [Fid06b, AZLB16]

$$w = \inf\left\{\tilde{w} \geq 0 : \inf_{\theta>0}\left\{\mathsf{M}_{A,S}(\theta, t+\tilde{w}, t)\right\} \leq \epsilon^{w}\right\}. \tag{10}$$

C System Descriptions and MGF-based Traffic/Service Characterizations

C.1 System Model

Throughout this paper, we assume a constant arrival rate ρ for the incoming data traffic[1], i.e., $A\left(s,t\right) = \rho \cdot (t-s)$ for any $0 \leq s \leq t$. Two schemes, i.e., traffic disper-

[1] Constant-rate arrival traffic is mainly considered throughout this work for simplifying analyses, while discussions on the performance associated with stochastic arrival are also attached.

sion and network densification, are illustrated in Fig. 1.1. They work as follows:

- For traffic dispersion (as shown in Fig. 1.1a), the original arrival traffic is firstly partitioned into multiple sub-streams by the data splitter. More precisely, given a set of deterministic splitting coefficients $(z_1, z_2, \cdots, z_n)$, where $\sum_{i=1}^{n} z_i = 1$ and $z_i \in (0,1)$ for any $1 \leq i \leq n$, the i^{th} sub-stream $A_i(s,t)$ is obtained as $A_i(s,t) = z_i \cdot A(s,t)$. Then, each sub-stream gets served and delivered towards the receiver through the given path, independently. Finally, the receiver combines all sub-streams through the data merger from different paths, thereby forming the output traffic. Thanks to narrow beams (or highly directional antennas) in mm-wave communications, the inter-channel interference is negligible, and hence it is reasonable to assume independence for multiple propagation paths[2]. The principle of traffic dispersion is to decompose the original heavy arrival traffic into multiple lighter ones, thereby to avoid a large queue in the buffer.

- For network densification (as shown in Fig. 1.1b), multiple relay nodes[3] as servers are deployed along the source-destination transmission path. Due to the concatenation of relying nodes, the output traffic from one relay can be treated as the input traffic for the subsequent connected relay. The application of multi-hop relaying follows trend of ultra-dense mm-wave networks. Likewise, it is feasible to assume independent channel conditions on multiple hops, thanks to high directivity and propagation properties for mm-wave communications. In contrast to the principle of traffic dispersion, the mechanism of network densification is deploying a large number of relay nodes between the given source and destination, which can reduce the path loss on each hop via shortening the separation distance between adjacent nodes, thereby increase the end-to-end service capability.

Besides, combining the benefits of traffic dispersion and network densification, we consider hybrid scheme as shown in Fig. 1.2. The original arrival traffic is firstly divided into multiple sub-streams by data splitter. Subsequently, these sub-streams are allocated with independent paths for data transmission, and each path consists of multiple relay nodes. It is evident that, this combination takes advantages of traffic dispersion and network densification, i.e., offloading the arrival traffic and enhancing the service capability.

[2] In DAS or RRH, a large number of antenna elements can be divided into clusters and physically isolated. We also note that for the short wavelength mm-wave, lots of antenna can be co-located in a small area [XMH+17]. With the aid of proper beamforming techniques (precoding at the transmitter and signal shaping at the receiver), the orthogonality among distinct beams with high directivity is enabled, thereby providing multiple interference-free parallel paths from different clusters in mm-wave bands, where channel fading characteristics are independent due to the separation among distinct paths [SMGL12, AEALH14, GKBS16].

[3] Full-duplex relay nodes are used throughout this paper, where the self interference is ignored for simplifying analysis.

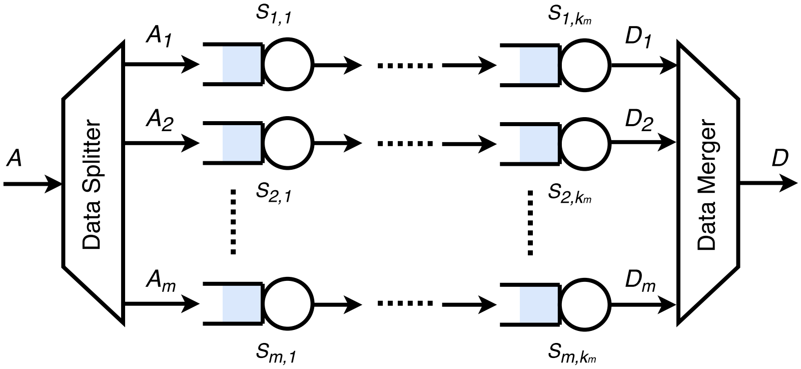


Figure 1.2: Hybrid scheme for low-latency mm-wave communications.

For the propagation characteristic in mm-wave bands, it is known that the small-scale fading in mm-wave channels is very weak [RSM+13], in contrast to that in sub-6 GHz bands. For the sake of tractability, the amplitude of the channel coefficient in mm-wave bands is commonly modeled as a Nakagami-m random variable, as in [BH15, YZHL17]. For simplicity, we assume the small-scale fading channel with Nakagami-m distribution is independent and identically distributed (i.i.d.) over time in terms of blocks, i.e., i.i.d. block fading. Given separation distance l and path loss exponent α, the capacity of the mm-wave channel is given as

$$C = B \log_2 \left(1 + \xi \gamma l^{-\alpha}\right), \tag{11}$$

where γ denotes the transmit power normalized by the background noise, B is the bandwidth, and the random variable ξ represents the channel gain, which follows the gamma distribution, i.e., $\xi \sim \Gamma\left(M, M^{-1}\right)$, with respect to Nakagami parameter M. The p.d.f. of the gamma-distributed ξ is given as

$$f\left(x; M, M^{-1}\right) \triangleq \frac{x^{M-1} \exp\left(-Mx\right)}{M^{-M} \Gamma\left(M\right)},$$

where $\Gamma\left(z\right) \triangleq \int_0^\infty z^{t-1} \exp\left(-t\right) dt$ denotes the gamma function for $\Re\left(z\right) > 0$.

C.2 MGF Bounds for Service Processes

According to (9), it can be found that, $\mathbb{M}_A\left(\theta, s, t\right)$ and $\overline{\mathbb{M}}_S\left(\theta, s, t\right)$ are required to compute the probabilistic delay bound. Regarding $\mathbb{M}_A\left(\theta, s, t\right)$, for simplifying illustration, deterministic arrivals with constant rate $\rho > 0$ are assumed in this paper, such that $\mathbb{M}_A\left(\theta, s, t\right)$ with respect to free parameter $\theta > 0$ can be written as

$$\mathbb{M}_A\left(\theta, s, t\right) = \exp\left(\theta \cdot \rho \cdot (t - s)\right) \triangleq \mu^{t-s}\left(\theta\right), \tag{12}$$

where $\mu(\theta) \triangleq \exp(\theta\rho)$. Moreover, for cumulative service process $S(s,t)$, in terms of channel capacity by (11), we have

$$S(s,t) = \sum_{q=s}^{t-1} C^{(q)} = \eta \sum_{q=s}^{t-1} \ln\left(1 + \xi^{(q)} \gamma l^{-\alpha}\right),$$

where $\eta \triangleq B \log_2 e$, and $C^{(q)}$ denotes the instantaneous channel capacity with respect to gamma-distributed $\xi^{(q)}$ at time q. Then, the corresponding MGF can be written as

$$\overline{\mathbb{M}}_S(\theta, s, t) = \left(\mathbb{E}\left[\left(1 + \xi\gamma l^{-\alpha}\right)^{-\eta\theta}\right]\right)^{t-s} \triangleq \mathcal{U}_C^{t-s}(\eta\theta), \tag{13}$$

where i.i.d. Nakagami-m fading across the time dimension is assumed. Here, $\mathcal{U}_C(x)$ for $x > 0$ is defined as

$$\mathcal{U}_C(x) \triangleq \left(\frac{Ml^\alpha}{\gamma}\right)^M U\left(M, 1 + M - x, \frac{Ml^\alpha}{\gamma}\right),$$

where

$$U(a,b,z) \triangleq \Gamma(a) \int_0^\infty \exp(-zt)\, t^{a-1} (1+t)^{b-a-1}\, dt$$

denotes the *confluent hypergeometric Kummer U-function.*

Note that, with stochastic service processes, it is intractable to obtain the closed-form probability distribution function (p.d.f.) for delay. Besides, the MGF-based approach gives bounds, instead of the actual delay. Therefore, for schemes to be investigated subsequently, it is infeasible and meaningless to formulate optimization problems with certain given constraints to optimize the actual delay performance. In this sense, our work mainly aims to characterize the delay performance via bounds, rather than to optimize the schemes.

D Probabilistic Delay Bounds for Low-Latency Transmission Schemes

In this section, we mainly focus on the performance analysis of traffic dispersion and network densification, and derive upper bounds for probabilistic delay. Subsequently, based on the derived results for above two basic schemes, the delay bound for the hybrid scheme is also presented.

D.1 Delay Bound for Traffic Dispersion

As shown in Fig. 1.1a, assuming $m \geq 1$ independent paths for the traffic dispersion scheme, associated with a set of deterministic splitting coefficients $(z_1, z_2, \cdots, z_m)$, where $\sum_{i=1}^{m} z_i = 1$ and $z_i \in (0,1)$ for any $1 \leq z_i \leq m$, the original cumulative arrival $A(s,t) = \rho(t-s)$ is decomposed into several sub-streams $A_i(s,t)$ for $1 \leq i \leq m$, i.e.,

$$A_i(s,t) = z_i A(s,t) = (z_i \cdot \rho) \cdot (t-s) \triangleq \rho_i (t-s).$$

In this sense, we have $\rho = \sum_{i=1}^{m} \rho_i$. Then, for each sub arrival traffic $A_i(s,t)$, we similarly have

$$\mathbb{M}_{A_i}(s,t) = \exp(\theta \cdot \rho_i \cdot (t-s)) \triangleq \mu_i^{t-s}(\theta). \tag{14}$$

Moreover, for each transmission path, the channel capacity of the i^{th} path at time q is given as

$$C_i'^{(q)} = B \log_2 \left(1 + \xi_i^{(q)} \gamma_i l^{-\alpha}\right),$$

where separation distance l is assumed for each path, and $\xi_i^{(q)}$ and γ_i denote the instantaneous gamma-distributed channel gain and normalized transmit power on the i^{th} path, respectively. Following (13), the MGF of $S_i'(s,t) \triangleq \sum_{q=s}^{t-1} C_i'^{(q)}$ can be written as

$$\overline{\mathbb{M}}_{S_i'}(\theta,s,t) = \mathcal{U}_{C_i'}^{t-s}(\theta) \triangleq \psi_i^{t-s}(\theta). \tag{15}$$

Without loss of generality, considering heterogeneous traffic dispersion, i.e., $\mu_i(\theta) \neq \mu_j(\theta)$ or $\psi_i(\theta) \neq \psi_j(\theta)$ may hold for $1 \leq i \neq j \leq m$, an upper bound on the violation probability for traffic dispersion is given in the following theorem.

Theorem D.1. *Let $W(t) \triangleq \max\{W_1(t), W_2(t), \cdots, W_m(t)\}$ be the delay for the traffic dispersion scheme with m independent paths, where $W_i(t)$ denotes the delay on the i^{th} path. Then, for any $w \geq 0$, the probabilistic delay is bounded as follows:*

$$\Pr(W(t) \geq w) \leq 1 - \prod_{i=1}^{m} \left[1 - \inf_{\theta_i > 0} \left\{ \frac{\psi_i^w(\theta_i)}{1 - \mu_i(\theta_i)\psi_i(\theta_i)} \right\}\right]^+,$$

whenever the stability condition $\mu_i(\theta_i)\psi_i(\theta_i) < 1$ holds for some $\theta_i > 0$ and all $i \in [m]$, where $[x]^+ \triangleq \max\{x, 0\}$ for $x \in \mathbb{R}$.

Proof. Please refer to Appendix H.1. □

We notice from Theorem D.1 that the definition of $W(t)$ indicates the synchronization constraint. The traffic dispersion scheme discussed here actually acts as a special variant of general *fork-join* systems [FJ16, RPC16], since all tasks of a job start execution at the same time, and the job is not completed until the final task leaves the system.

In Theorem D.1, the stability condition, i.e., $\mu_i(\theta_i)\psi_i(\theta_i) < 1$ for all $1 \leq i \leq m$, should be satisfied to obtain the above probabilistic delay bound. This stability condition stems from the fact that, to avoid infinite delay, the arrival rate of each sub-stream cannot exceed the service capability provided on the corresponding path. Furthermore, Theorem D.1 tells that, the path with higher service utilization, i.e., higher $\mu_i(\theta_i)\,\psi_i(\theta_i)$, is the main contributor to large delay.

With homogeneous settings, i.e., $\mu_i(\theta) = \mu(\theta)$ and $\psi_i(\theta) = \psi(\theta)$ for all $1 \leq i \leq m$, we have the following corollary.

Corollary D.1.1. *For homogeneous traffic dispersion, given any $w \geq 0$, the probabilistic delay bound of Theorem D.1 is given as*

$$\Pr(W(t) \geq w) \leq 1 - \left[\left(1 - \inf_{\theta>0}\left\{\frac{\psi^w(\theta)}{1 - \mu(\theta)\,\psi(\theta)}\right\}\right)^m\right]^+,$$

whenever $\mu(\theta)\,\psi(\theta) < 1$ holds for some $\theta > 0$.

Corollary D.1.1 demonstrates that, with fixed $\mu(\theta)$ and $\psi(\theta)$, the upper bound for the violation probability $\Pr(W(t) \geq w)$ grows with increasing m. The observation coincides with the result mentioned in [FJ16] and [RPC16] that the delay roughly scales up linearly with the number of independent paths, especially when the end-to-end delay on each path is small.

It is worth mentioning that Theorem D.1 and Corollary D.2.1 are built on the assumption that deterministic arrival is applied, such that the independence among different $W_i(t)$ is preserved. However, when considering stochastic arrival traffic, the independence of m flows after splitting does not hold, and thus $\mathbb{P}(W(t) \geq w) \neq 1 - \prod_{i=1}^{m}(1 - \mathbb{P}(W_i(t) \geq w))$. To address the difficulty induced by the dependence among stochastic sub-streams, we resort to a union bound for $\mathbb{P}(W(t) \geq w)$, i.e.,

$$\mathbb{P}(W(t) \geq w) \leq \min\left\{1, \sum_{i=1}^{m}\mathbb{P}(W_i(t) \geq w)\right\}.$$

Therefore, when $\mu_i(\theta)$ for all $1 \leq i \leq m$ are not independent, the results in Theorem D.1 and Corollary D.2.1 should be changed respectively to

$$\mathbb{P}(W(t) \geq w) \leq \min\left\{1, \sum_{i=1}^{m}\inf_{\theta_i>0}\left\{\frac{\psi_i^w(\theta_i)}{1 - \mu_i(\theta_i)\,\psi_i(\theta_i)}\right\}\right\}$$

and

$$\mathbb{P}(W(t) \geq w) \leq \min\left\{1, m\inf_{\theta>0}\left\{\frac{\psi^w(\theta)}{1 - \mu(\theta)\,\psi(\theta_i)}\right\}\right\},$$

where the asymptotic tightness of union bounds is roughly identical to that in Theorem D.1 and Corollary D.2.1, since

$$1 - \prod_{i=1}^{m}(1 - \mathbb{P}(W_i(t) \geq w)) \approx \sum_{i=1}^{m}\mathbb{P}(W_i(t) \geq w) \tag{16}$$

when $\mathbb{P}(W_i(t) \geq w)$ is sufficiently small.

D.2 Delay Bound for Network Densification

In the network densification scheme for multi-hop mm-wave networks, we again assume that $A(s,t) = \rho \cdot (t-s)$, such that

$$\mathbb{M}_A(s,t) = \mu^{t-s}(\theta).$$

Moreover, for the k-hop relaying channel, we denote by

$$C_i''^{(q)} = B\log_2\left(1 + \xi_i^{(q)}\gamma_i l_i^{-\alpha}\right)$$

the instantaneous channel capacity of the i^{th} hop at time q, where $\xi_i^{(q)}$, γ_i and l_i denote the instantaneous gamma-distributed channel gain, normalized transmit power, and transmission distance on the i^{th} hop, respectively. Based on (13), the MGF of $S_i''(s,t) \triangleq \sum_{q=s}^{t-1} C_i''^{(q)}$ can be written as

$$\overline{\mathbb{M}}_{S_i''}(\theta,s,t) = \mathcal{U}_{C_i''}^{t-s}(\theta) \triangleq \phi_i^{t-s}(\theta). \tag{17}$$

Considering heterogeneous multi-hop relaying, i.e., $\phi_i(\theta) \neq \phi_j(\theta)$ may hold for $1 \leq i \neq j \leq k$, we have the following theorem regarding the probabilistic end-to-end delay bound.

Theorem D.2. *Let $W(t)$ be the end-to-end delay of a k-hop network. Then, for any $w \geq 0$, the probabilistic delay is bounded as follows:*

$$\Pr(W(t) \geq w) \leq \inf_{\theta>0}\left\{\mu^{-w}(\theta)\sum_{v=w}^{\infty}\sum_{\sum_{i=1}^{k}\pi_i=v}\prod_{i=1}^{k}\left(\mu(\theta)\phi_i(\theta)\right)^{\pi_i}\right\},$$

whenever the stability condition $\mu(\theta)\phi_i(\theta) < 1$ holds for some $\theta > 0$ and all $i \in [k]$.

Proof. Please refer to Appendix H.2. □

Similarly, the stability condition indicates that, to avoid infinite delay, the arrival rate cannot exceed the service capability provided by each hop. Besides, by Theorem D.2, the hop with higher service utilization, i.e., higher $\mu(\theta)\phi_i(\theta)$, is the main contributor to a large delay.

In the following corollary, the probabilistic delay for multi-hop relaying with homogeneous settings is studied, where $\phi_i(\theta) = \phi(\theta)$ for all $1 \leq i \leq k$ are assumed.

Corollary D.2.1. *For homogeneous network densification, given any $w \geq 0$, the probabilistic delay bound of Theorem D.2 is given as*

$$\Pr(W(t) \geq w) \leq \binom{k-1+w}{w}\inf_{\theta>0}\left\{\phi^{w}(\theta)_2F_1\left(1,k+w;1+w;\mu(\theta)\phi(\theta)\right)\right\}$$

whenever the stability condition $\mu(\theta)\,\phi(\theta) < 1$ *holds for some* $\theta > 0$. *Here* ${}_2F_1(a,b;c;z)$ *is a* hypergeometric function, *defined as*

$$
{}_2F_1(a,b;c;z) \triangleq \sum_{n=0}^{\infty} \frac{(a)_n\,(b)_n}{(c)_n} \cdot \frac{z^n}{n!},
$$

where $(x)_n$ *denotes the* rising Pochhammer symbol, *given as*

$$
(x)_n \triangleq \begin{cases} 1, & n = 0 \\ \dfrac{(x+n-1)!}{(x-1)!}, & n > 0 \end{cases}.
$$

Proof. With the homogeneous setting, we obtain that

$$
\mathsf{M}_{A,S''}(\theta,s,t) \le \mu^{t-s}(\theta) \sum_{v=\tau}^{\infty} \binom{k-1+v}{v} (\mu(\theta)\phi(\theta))^{v}. \tag{18}
$$

Regarding the infinite sum in (18), we have

$$
\sum_{v=\tau}^{\infty} \binom{k-1+v}{v} (\mu(\theta)\,\phi(\theta))^{v} = (\mu(\theta)\phi(\theta))^{\tau} \binom{k-1+\tau}{\tau} {}_2F_1(1, k+\tau; 1+\tau; \mu(\theta)\phi(\theta)),
$$

where $\tau \triangleq \max\{s-t, 0\}$ and ${}_2F_1(a,b;c;z)$ is a hypergeometric function. Paired with (9) and the upper bound for $\mathsf{M}_{A,S''}(\theta,s,t)$, the theorem can be immediately concluded by letting $s = t + w$. □

Corollary D.2.1 demonstrates that, given fixed $\mu(\theta)$ and $\phi(\theta)$, the upper bound for the violation probability $\Pr(W(t) \ge w)$ grows when the number of hops k increases, i.e., scaling as $\mathcal{O}(k^w)$. The $\mathcal{O}(\cdot)$ is defined as a set of functions $u(x)$, i.e., $\mathcal{O}(f(x)) \triangleq \{u(x) \in \mathbb{R} : \sup|u(x)/f(x)| < \infty\}$, where $f(x) \in \mathbb{R}$.

D.3 Delay Bound for the Hybrid Scheme

In the hybrid scheme with m ($m \ge 1$) independent transmission paths, for the arrival traffic, we assume that the sub-stream $A_i(s,t)$ given in (14). Furthermore, we denoted by

$$
C_{i,j}^{(q)} = B \log_2 \left(1 + \xi_{i,j}^{(q)} \gamma_{i,j} l_{i,j}^{-\alpha}\right),
$$

the instantaneous channel capacity of the j^{th} hop on the i^{th} transmission path at time q, where $\xi_{i,j}^{(q)}$, $\gamma_{i,j}$ and $l_{i,j}$ represent the instantaneous gamma-distributed channel gain, normalized transmit power and propagation distance, respectively. Similarly, according to (13), the MGF of $S_{i,j}(s,t) \triangleq \sum_{q=s}^{t-1} C_{i,j}^{(q)}$ can be written as

$$
\overline{\mathbb{M}}_{S_{i,j}}(\theta,s,t) = \mathcal{U}_{C_{i,j}}^{t-s}(\theta) \triangleq \varphi_{i,j}^{t-s}(\theta). \tag{19}
$$

In light of above, the probabilistic delay bound for the hybrid scheme is presented in the following theorem.

Theorem D.3. *Assuming m ($m \geq 1$) independent paths for the hybrid scheme system, with k_i hops on the i^{th} path for $1 \leq i \leq m$, we define $\hat{p}_i$ for any given $w \geq 0$ as*

$$\hat{p}_i \triangleq \inf_{\theta_i > 0} \left\{ \mu_i^{-w}\left(\theta_i\right) \sum_{v=w}^{\infty} \sum_{\sum_{j=1}^{k_i} \pi_j = v} \prod_{j=1}^{k_i} \left(\mu_i\left(\theta_i\right) \varphi_{i,j}\left(\theta_i\right)\right)^{\pi_j} \right\}.$$

Then, the end-to-end probabilistic delay is upper bounded as

$$\Pr\left(W\left(t\right) \geq w\right) \leq 1 - \prod_{i=1}^{m} \left[1 - \hat{p}_i\right]^{+},$$

whenever the stability condition $\mu_i\left(\theta_i\right) \varphi_{i,j}\left(\theta_i\right) < 1$ holds for some $\theta_i > 0$ and all $1 \leq i \leq m$ and $1 \leq j \leq k_i$.

Proof. Applying Theorem D.2 in Theorem D.1 for each independent path, it is then straightforward to conclude the probabilistic delay bound for the hybrid scheme. □

Particularly, with homogeneous settings in the hybrid scheme, i.e., $\mu_i\left(\theta\right) = \mu\left(\theta\right)$, $\varphi_{i,j}\left(\theta\right) = \varphi\left(\theta\right)$, and $k_i = k$ for all $1 \leq i \leq m$ and $1 \leq j \leq k_i$, the result by Theorem D.3 can be further reduced, which can be presented via combining Corollary D.1.1 and Corollary D.2.1. For brevity, results related to the homogeneous scenario are omitted.

Again, if stochastic arrival traffic is applied, the independence among substreams does not hold. Then upper bound in Theorem D.3 should be changed to

$$\Pr\left(W\left(t\right) \geq w\right) \leq \min\left\{1, \sum_{i=1}^{m} \hat{p}_i\right\},$$

which is obtained via applying union bound (similar method in (16)).

E Effective Capacity Analysis with Given Average System Gain

From the analysis of Sec. D that, it is important to ensure that the arrival rate is below the service capability (refer to the stability conditions for three schemes). That is, the service capability characterizes the limiting potentials to deal with data traffic without causing infinite delays. Thus, in what follows, we use effective capacity [WN03], which is another important metric related to the delay performance departing from asymptotic service capability, to analyze three schemes.

E.1 Basics for Effective Capacity

In light of the MGF of the service process by (13), the effective capacity is defined as

$$\mathcal{C}\left(-\theta\right) \triangleq \lim_{t\to\infty} \frac{\log \overline{\mathbb{M}}_S\left(\theta, 0, t\right)}{-\theta t}, \tag{20}$$

where $\theta > 0$ represents the QoS exponent, which indicates a more stringent QoS requirement for a higher θ.

With certain positive θ that enables

$$\lim_{x\to\infty} \frac{\log\left(\zeta^{-1}\Pr\left(W\left(t\right) \geq x\right)\right)}{x} = -\theta\mathcal{C}\left(-\theta\right), \tag{21}$$

where ζ is the probability that the queue is not empty, the violation probability of delay, denoted by $\Pr\left(W\left(t\right) \geq w_{\max}\right)$, can be approximated as [TZ08,HZ12,KCH15, LGAZ$^+$16]

$$\Pr\left(W\left(t\right) \geq w_{\max}\right) \approx \zeta \exp\left(-\theta\mathcal{C}\left(-\theta\right) w_{\max}\right), \tag{22}$$

where $w_{\max}$ represents the maximum tolerance of delay.

In addition, for the stability consideration of first-in-first-out (FIFO) queuing systems in the asymptotic sense, according to the *Gärtner-Ellis Theorem* [Buc13], the arrival and service process should satisfy the following condition with given θ, i.e.,

$$\mathcal{R}\left(\theta\right) \triangleq \lim_{t\to\infty} \frac{\log \mathbb{M}_A\left(\theta, 0, t\right)}{\theta t} \leq \mathcal{R}^*\left(\theta\right) \triangleq \mathcal{C}\left(-\theta\right), \tag{23}$$

where $\mathcal{R}\left(\theta\right)$ in terms of QoS exponent θ is commonly termed as the "effective bandwidth" [Cha00], and the maximum effective bandwidth, denoted by $\mathcal{R}^*\left(\theta\right)$ is characterized by the effective capacity $\mathcal{C}\left(-\theta\right)$. We notice that, the relation between effective bandwidth and effective capacity, i.e., $\mathcal{R}\left(\theta\right) \leq \mathcal{C}\left(-\theta\right)$, follows the intuition of asymptotic stability (or stability in the long-term sense), and its principle resembles the stability condition stated in Theorem D.1, Theorem D.2, or Theorem D.3. It is worth mentioning that effective capacity $\mathcal{C}\left(-\theta\right)$ depicts the utmost service capability provided by the system, which is independent of the density of arrival traffic (refer to (20)).

According to the properties above, it is clear that, for any given QoS exponent $\theta > 0$, a larger effective capacity $\mathcal{C}\left(-\theta\right)$ not only indicates a stronger capability for serving heavier arrival traffic (see (23)), but also leads to faster decay in the probability delay (refer to (22)). Therefore, in the following analysis, we focus on the effective capacity (or the maximum effective bandwidth, equivalently) of traffic dispersion, network densification and the hybrid scheme, respectively.

In what follows, we investigate the effective capacity for traffic dispersion, network densification and the hybrid scheme. We denote by L the end-to-end distance between the source and the destination (in network densification and the hybrid scheme, distance L is assumed to each independent path). Besides, we assume that all transmitter are subject to a sum-power constraint γ. It is worth mentioning that the expressions of the effective capacity for networks with heterogeneous settings can be obtained by using $\overline{\mathbb{M}}_S(\theta,0,t)$ (refer to (15) and (17) for traffic dispersion and network densification, respectively), and hence they are omitted in this section to avoid redundancy. However, we in the following analyses mainly consider homogeneous settings for three schemes, aiming at obtaining closed-form expressions for fair comparisons without loss of tractability.

E.2 Effective Capacity of Traffic Dispersion

For traffic dispersion, the effective capacity of each independent path can be obtain as

$$\mathcal{C}_{S_i'}(-\theta) = -\frac{1}{\theta}\log\mathbb{E}\left[\exp\left(-\theta C_i'\right)\right],$$

where the independence of channel condition across the time dimension is applied. Then, considering n parallel independent paths, the effective capacity of traffic dispersion is given as

$$\mathcal{C}_{S'}(-\theta) \triangleq \mathcal{C}_{\sum_{i=1}^{n} S_i'}(-\theta) = \sum_{i=1}^{m}\mathcal{C}_{S_i'}(-\theta) = -\lim_{t\to\infty}\sum_{i=1}^{m}\frac{\log\overline{\mathbb{M}}_{S_i'}(\theta,0,t)}{\theta t}. \tag{24}$$

Then, the maximum effective bandwidth can be obtained as

$$\begin{aligned}\mathcal{R}^*_{S'}(\theta) &= -\frac{1}{\theta}\cdot\sum_{i=1}^{m}\log\mathbb{E}\left[\left(1+\xi_i\gamma_i L^{-\alpha}\right)^{-\eta\theta}\right]\\ &= -\frac{1}{\theta}\cdot\sum_{i=1}^{m}\log\left(\left(\frac{ML^{\alpha}}{\gamma_i}\right)^{M} U\left(M, 1+M-\eta\theta, \frac{ML^{\alpha}}{\gamma_i}\right)\right).\end{aligned}$$

It is worth mentioning that the QoS exponent θ here works for the entire system. That is, θ should make component probabilistic delays in n paths all satisfy the asymptotic condition in (21). In this sense, the worst component has been considered, such that we can disregard the arrival order of data and have the overall service of traffic dispersion as a sum of individual services.

With the Nakagami-m fading characteristic of the mm-wave channel, the maximum effective bandwidth $\mathcal{R}^*_{S'}(\theta)$ is given in Theorem E.1.

Theorem E.1. *Given sum power constraint $\sum_{i=1}^{m}\gamma_i = \gamma$, the maximum effective bandwidth $\mathcal{R}^*_{S'}(\theta)$ is upper bounded as*

$$\mathcal{R}^*_{S'}(\theta) \le \frac{m\log\left(\left(\frac{MmL^{\alpha}}{\gamma}\right)^{M} U\left(M, 1+M-\eta\theta, \frac{MmL^{\alpha}}{\gamma}\right)\right)}{-\theta},$$

where the equality holds if $\gamma_i = m^{-1}\gamma$ for all $1 \leq i \leq m$.

Proof. Please refer to Appendix H.3. □

Theorem E.1 shows that, for the traffic dispersion scheme, the maximum effective bandwidth can be achieved when applying homogeneous settings on n independent paths, i.e., $\gamma_i = m^{-1}\gamma$ for all $1 \leq i \leq m$.

E.3 Effective Capacity of Network Densification

For network densification, by (4), the network service process for the multi-hop relying scheme is characterized by

$$S''(0,t) \triangleq (S_1'' \otimes S_2'' \otimes \cdots \otimes S_k'')(0,t).$$

According the definition of effective capacity, we have

$$\mathcal{C}_{S''}(-\theta) \triangleq -\lim_{t\to\infty} \frac{\log \overline{\mathbb{M}}_{S_1''\otimes S_2''\otimes\cdots\otimes S_k''}(\theta,0,t)}{\theta t}. \tag{25}$$

It is worth mentioning that, since $(\min,+)$ convolution is involved for characterizing the concatenated service in network densification, it is intractable to derive the closed-form expression for the effective capacity. Thus, we will use upper and lower bounds to characterize the effective capacity of network densification. For the sake of tractability, we in what follows consider homogeneous settings for network densification, i.e., $\gamma_i = k^{-1}\gamma$ and $l_i = k^{-1}L$ for all $1 \leq i \leq k$, and we derive the closed-form upper and lower bounds on the effective capacity.

Based on the Nakagami-m fading characteristic of the mm-wave channel, the lower bound and upper bound on $\mathcal{R}^*_{S''}(\theta)$ are given in Theorem E.2.

Theorem E.2. *For homogeneous network densification with k independent hops, given $\theta > 0$, the maximum effective bandwidth is upper bounded as*

$$\mathcal{R}^*_{S''}(\theta) \leq -\frac{k}{\theta}\log\left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M, 1+M-\frac{\eta\theta}{k}, \frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right),$$

and it is lower bounded as

$$\mathcal{R}^*_{S''}(\theta) \geq -\frac{1}{\theta}\log\left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M, 1+M-\eta\theta, \frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right).$$

Proof. Please refer to Appendix H.4. □

As we can see in Theorem E.2, the upper bound meets the lower bound when $k = 1$, and the resulting maximum effective bandwidth is reduced to the closed-form expression for single-hop mm-wave networks.

E.4 Effective Capacity of the Hybrid Scheme

The effective capacity of the hybrid scheme is obtained as

$$\mathcal{C}_S\left(-\theta\right) = -\lim_{t\to\infty}\sum_{i=1}^{m}\frac{\log\overline{\mathbb{M}}_{S_{i,1}\otimes\cdots\otimes S_{i,k_i}}\left(\theta,0,t\right)}{\theta t}, \tag{26}$$

where we assume m independent paths and k_i relay nodes on the i^{th} path ($1 \leq i \leq m$).

Again, for tractability, we consider homogeneous settings for the hybrid scheme. That is, given $m \geq 1$ independent paths and $k \geq 1$ relay nodes per path, such that $m \cdot k = n$ (m or k is a divisor of n, equivalently), we assume that $\gamma_{i,j} = n^{-1}\gamma$ and $l_{i,j} = k^{-1}L$ for all $1 \leq i \leq m$ and $1 \leq j \leq k$. Then, in the following analysis, we derive the upper and lower bounds on the effective capacity.

Similarly, based on the Nakagami-m fading characteristic of the mm-wave channel, the lower bound for $\mathcal{R}_S^*\left(\theta\right)$ is presented in the following theorem.

Theorem E.3. *For the homogeneous hybrid scheme with m independent paths and $k = n/m$ relay nodes per path, given $\theta > 0$, the maximum effective bandwidth is upper bounded as*

$$\mathcal{R}_S^*\left(\theta\right) \leq -\frac{n}{\theta}\cdot\log\left(\left(\frac{M\left(mL\right)^{\alpha}}{\gamma n^{\alpha-1}}\right)^{M} U\left(M, 1+M-\frac{\eta\theta}{k}, \frac{M\left(mL\right)^{\alpha}}{\gamma n^{\alpha-1}}\right)\right),$$

and it is lower bounded as

$$\mathcal{R}_S^*\left(\theta\right) \geq -\frac{m}{\theta}\cdot\log\left(\left(\frac{M\left(mL\right)^{\alpha}}{\gamma n^{\alpha-1}}\right)^{M} U\left(M, 1+M-\eta\theta, \frac{M\left(mL\right)^{\alpha}}{\gamma n^{\alpha-1}}\right)\right).$$

Proof. The theorem is immediately concluded by straightforwardly applying the variable substitution for Theorem E.1 and Theorem E.2, and the details are omitted for brevity. □

From Theorem E.3, we can find that traffic dispersion and network densification can be treated as two extreme cases of the hybrid scheme, i.e., corresponding to $m = n$ and $m = 1$, respectively. When $m = n$, Theorem E.3 reduces to Theorem E.1.

F Performance Evaluation

In this section, we provide numerical results for the probabilistic end-to-end delay bound and effective capacity discussed in Sec. C and D, respectively. Through simulations, we firstly validate the derived bounds for probabilistic delay and effective capacity, where the respective advantages of traffic dispersion and network densification are evaluated and discussed[4]. Subsequently, for the hybrid scheme (including traffic dispersing and network densification), the factors and conditions for achieving low-latency mm-wave communications are extensively studied.

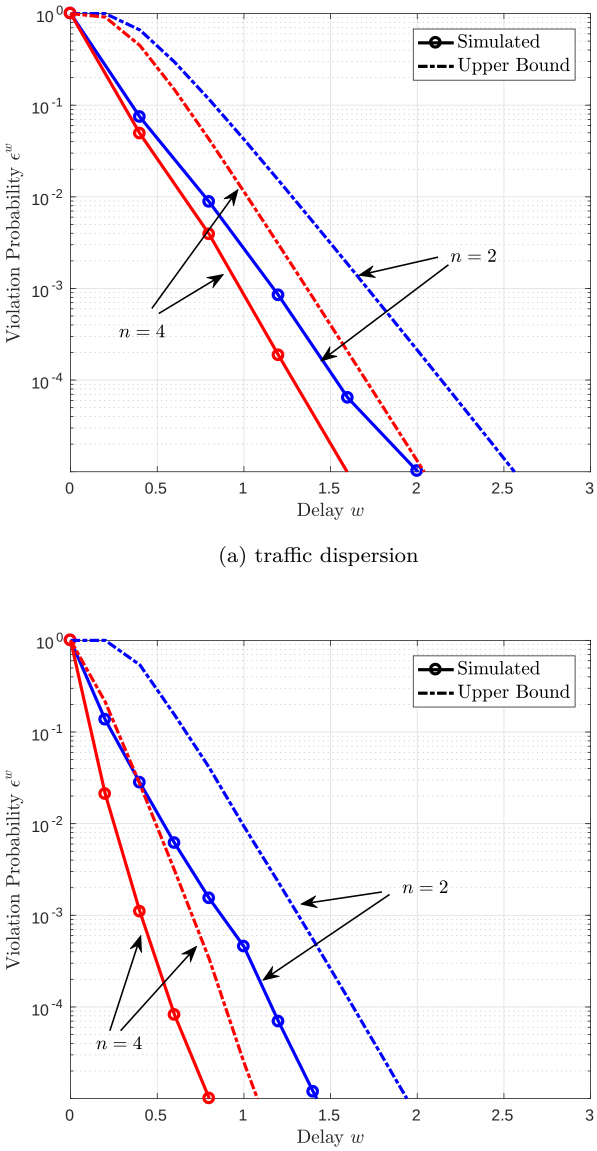
For fairness considerations, the homogeneous settings presumed in Sec. D are applied. The general system configurations are summarized as follows: the bandwidth is allocated with $B = 500$ MHz, the path loss exponent $\alpha = 2.45$, Nakagami-m parameter $M = 3$, and the source-destination distance $L = 1$ km.

F.1 Bound Validation

The violation probabilities of delay for traffic dispersion and network densification are illustrated in Fig. 1.3, where the sum transmit power is $\gamma = 85$ dB and the arrival rate is $\rho = 2$ Gbps. In Fig. 1.3a, for both cases $n = 2$ and $n = 4$, the probabilistic delay bound derived in Corollary D.1.1 accurately characterizes the slope of the simulated result. We notice that, although the violation probability of delay decreases as the number of independent paths grows, i.e., n increasing from 2 to 4, the resulting improvement is not remarkable under the given sum powerγ and arrival rate ρ. Likewise, in Fig. 1.3b, the bound by Corollary D.2.1 is also able to well predict the decaying rate of violation probability. However, in contrast to the traffic dispersion scheme, increasing the relay density (equivalently, increasing the number of relays) can significantly decrease the probabilistic delay.

Fig. 1.4 illustrates the effective capacity for traffic dispersion and network densification, with respect to the sum power γ varying from 50 dB to 100 dB and a given QoS exponent $\theta = 2$. Clearly, the derived lower and upper bounds of effective capacity in Theorem E.2 for network densification are quite close. Thus, those bounds are capable of capturing the actual effective capacity by network densification well. We find that traffic dispersion exhibits remarkable advantages when the sum power is high (the resulting effective capacity dramatically increases with γ), while the network densification scheme adversely outperforms its counterpart when having lower sum power, e.g., $\gamma \leq 80$ dB. Furthermore, for traffic dispersion, the gain achieved by elevating n becomes increasingly significant only when γ is high. However, for network densification the gain achieved by increasing n is relative steady. The findings above indicate that the benefit of traffic dispersion diminishes when the sum transmit power or the number of independent paths decreases. In this sense, the network densification is a better option for the scenarios with sparser relay deployment and lower sum power budget, and this insight is also in line with the results by comparing Fig. 1.3a and Fig. 1.3b.

[4] Two extreme cases, i.e., traffic dispersion and network densification, are considered only for simplifying comparison, and comprehensive results regarding the generic hybrid scheme are presented afterwards.



(a) traffic dispersion

(b) network densification

Figure 1.3: Violation probability ϵ^w vs. targeted delay bound w for two different schemes, where $\rho = 2$ Gbps and $\gamma = 85$ dB.

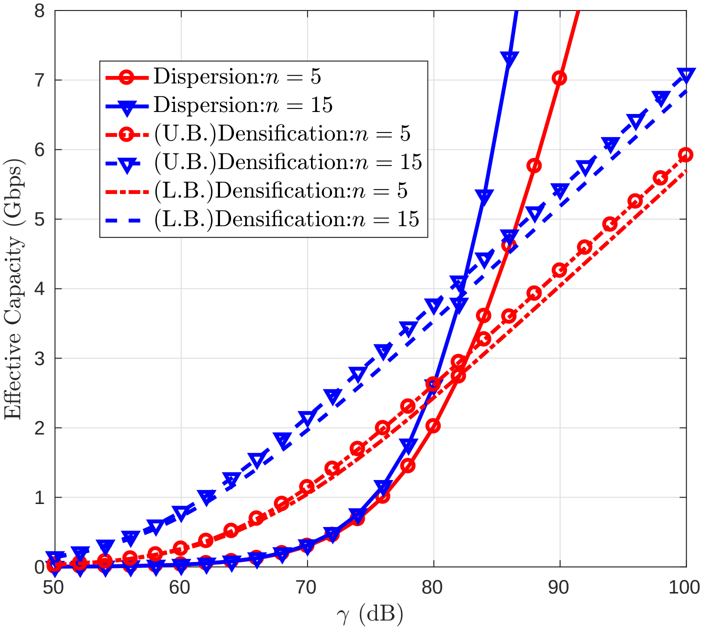


Figure 1.4: Effective Capacity $\mathcal{C}(-\theta)$ vs. sum transmit power γ for two transmission schemes, where QoS exponent $\theta = 2$. Here, U.B. and L.B. stand for "upper bound" and "lower bound", respectively.

F.2 Simulation and Discussion

Including two extreme cases, i.e., traffic dispersion ($m = n$) and network densification ($m = 1$), the effective capacity of the hybrid scheme with $1 \leq m \leq n$ for given $n = 12$ is illustrated in Fig. 1.5, where the QoS exponent is $\theta = 2$, and the sum transmit power γ varies from 70 to 90 dB. Due to the fact that the closed expression of the effective capacity for tandem networks cannot be obtained, we here again use lower and upper bounds in Theorem E.3 to characterize the effective capacity of the hybrid scheme. It can be seen that, with a lower sum power, e.g., $\gamma = 70$ dB, the effective capacity decays when m grows, and this observation indicates the advantage of network densification for the scenarios with a lower sum power budget. However, when γ becomes higher, we can see that the effective capacity increases first and decreases subsequently. For instance, the maximum effective capacity is achieved at $m = 2$ in the presence of $\gamma = 75$ dB. In this case, the best solution to minimize the end-to-end delay is to assign two independent transmission paths for traffic dispersion and five relay nodes per path. It is easy to see that traffic dispersion becomes the dominant contributor to the effective capacity when γ increases, and this ten-

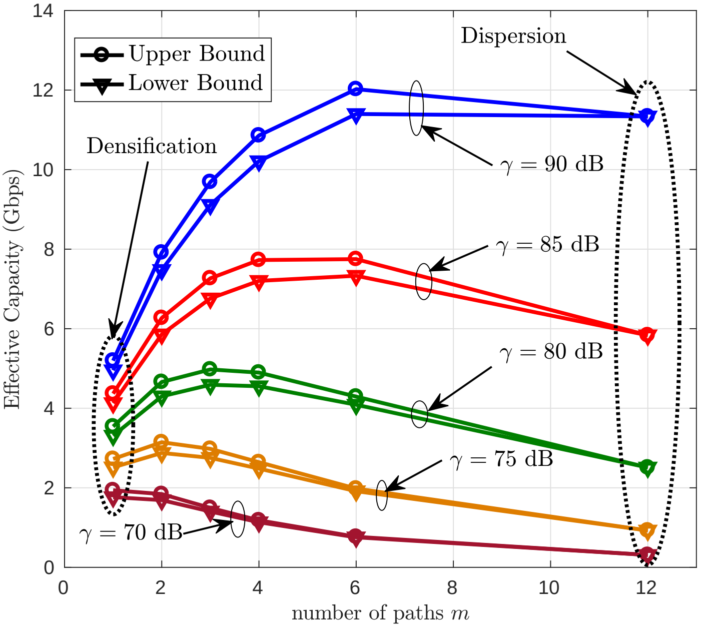


Figure 1.5: Effective capacity $\mathcal{C}(-\theta)$ vs. number of independent paths m for the hybrid scheme with respect to $\theta = 2$ and $n = 12$, where the number of independent paths is $m = 1$, 2, 3, 4, 6 or 12.

dency can be observed from the increase of the optimal number of paths. In light of the above findings, the hybrid scheme with proper configurations, i.e., the proper numbers of independent paths and hops per path, respectively, should be carefully considered to maximize the effective capacity. Also, the respective strengths of traffic dispersion and network densification revealed by Fig. 1.5 coincide with that from Fig. 1.4.

Coming back from effective capacity to probabilistic delay, the targeted delay tolerances versus different arrival rates for three transmission schemes i.e., traffic dispersion ($m = 12$), network densification ($m = 1$) and the hybrid scheme ($m = 3$), are provided in Fig. 1.6, where tolerance is given as $\epsilon^w = 10^{-3}$. For both groups in terms of different γ, clearly, the targeted delay exponentially increases when the arrival rate increases. These drastic growths result from the higher service utilization. That is, the arrival rate approaches the limiting service capability. Besides, comparing the three transmission schemes, the respective advantages demonstrated here coincide with the conclusions drawn from Fig. 1.5. From the perspective of probabilistic delay, we can conclude that it is critical to consider the arrival rate and the proper transmission scheme jointly to reduce the delay.

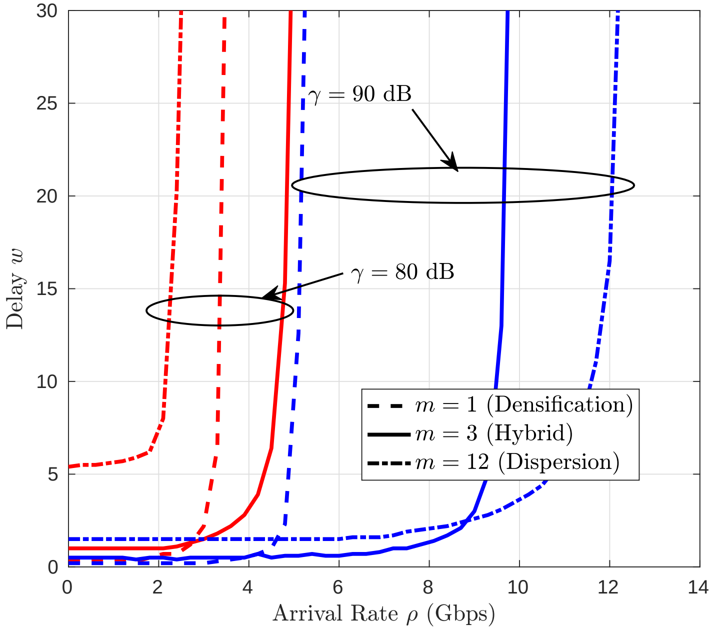


Figure 1.6: Probabilistic delay w vs. arrival rate ρ for traffic dispersion, network densification and the hybrid scheme, respectively, with respect to violation probability $\epsilon^w = 10^{-3}$, where $n = 12$.

G Conclusions

We have considered traffic dispersion and network densification for low-latency mm-wave communications, and have investigated their end-to-end delay performance. We have also proposed a hybrid scheme to further reduce latency in certain scenarios. Based on MGF-based stochastic network calculus and effective capacity theory, respectively, we have derived performance bounds for probabilistic delay and effective capacity for the three schemes, which have been validated through simulations. These results have demonstrated that, given the sum power budget, traffic dispersion, network densification, and the hybrid scheme show different potential in different scenarios for low-latency mm-wave communications. In addition, increasing the number of independent paths or the number of relays for network densification is always advantageous for reducing the end-to-end communication delay, while the performance gain heavily relies on the density of arrival traffic and the sum power budget, jointly. Thus, it is crucial to select the proper scheme according to the given arrival traffic and service capability.

H Appendices

H.1 Proof of Theorem D.1

We start with deriving the upper bound for M_{A_i,S'_i}:

$$\begin{aligned}\mathsf{M}_{A_i,S'_i}(\theta,s,t) &\overset{(a)}{\leq} \sum_{u=0}^{\min(s,t)} \mu_i^{t-u}(\theta)\,\psi_i^{s-u}(\theta) \overset{(b)}{=} \mu_i^{t-s}(\theta)\sum_{v=\tau}^{s}\left(\mu_i(\theta)\,\psi_i(\theta)\right)^v \\ &\overset{(c)}{\leq} \mu_i^{t-s}(\theta)\sum_{v=\tau}^{\infty}\left(\mu_i(\theta)\,\psi_i(\theta)\right)^v = \frac{\mu_i^{t-s}(\theta)\left(\mu_i(\theta)\,\psi_i(\theta)\right)^{\tau}}{1-\mu_i(\theta)\,\psi_i(\theta)}, \qquad (27)\end{aligned}$$

where $\tau \triangleq \max\{s-t,0\}$. Here, inequality (a) is obtained by plugging (14) and (15) into (8). By performing the change of variable, i.e., $v = s-u$, equality (b) is achieved. In (c), we let s go to infinity. In the final step, the geometric sum converges only when $\mu_i(\theta)\,\psi_i(\theta) < 1$ holds for certain θ.

By the definition of $W(t)$, it is easy to obtain that

$$\begin{aligned}\Pr(W(t)\geq w) &\triangleq \Pr\left(\max_{1\leq i\leq m}\{W_i(t)\}\geq w\right) \overset{(d)}{=} 1-\prod_{i=1}^{m}\left(1-\Pr(W_i(t)\geq w)\right) \\ &\overset{(e)}{\leq} 1-\prod_{i=1}^{m}\left(1-\inf_{\theta_i>0}\mathsf{M}_{A_i,S'_i}(\theta_i,t+w,t)\right) \\ &\overset{(f)}{\leq} 1-\prod_{i=1}^{m}\left(1-\inf_{\theta_i>0}\left\{\frac{\psi_i^w(\theta_i)}{1-\mu_i(\theta_i)\,\psi_i(\theta_i)}\right\}\right),\end{aligned}$$

where the independence assumption among distinct paths is used to derive (d), and inequality (e) applies the (9), and (f) follows from (27).

In addition, we notice that $\Pr(W(t)\geq w)\leq 1$ holds for any $w\geq 0$, then the theorem is concluded.

H.2 Proof of Theorem D.2

Applying (17) in (7), the MGF of k-hop network service process can be characterized as

$$\overline{\mathbb{M}}_{S''}(\theta,s,t) \triangleq \overline{\mathbb{M}}_{S''_1\otimes\cdots\otimes S''_k}(\theta,s,t) \leq \sum_{\sum_{i=1}^{k}\pi_i=t-s}\ \prod_{i=1}^{k}\phi_i^{\pi_i}(\theta).$$

Then, it is easy to obtain that

$$\mathsf{M}_{A,S''}(\theta,s,t) \leq \sum_{u=0}^{\min(s,t)} \mu^{t-s}(\theta) \sum_{\sum\limits_{i=1}^{k}\pi_i=s-u}\ \prod_{i=1}^{k}\phi_i^{\pi_i}(\theta)$$

$$
\begin{aligned}
&= \mu^{t-s}(\theta) \sum_{v=\tau}^{s} \sum_{\sum_{i=1}^{k} \pi_i = v} \prod_{i=1}^{k} \left(\mu(\theta)\,\phi_i(\theta)\right)^{\pi_i} \\
&\le \mu^{t-s}(\theta) \sum_{v=\tau}^{\infty} \sum_{\sum_{i=1}^{k} \pi_i = v} \prod_{i=1}^{k} \left(\mu(\theta)\,\phi_i(\theta)\right)^{\pi_i}, \qquad (28)
\end{aligned}
$$

where $\tau \triangleq \max\{s-t, 0\}$.

Similar to Theorem D.1, we notice that the stability condition, i.e., $\mu(\theta)\,\phi_i(\theta) < 1$ for all $1 \le i \le k$, needs to be satisfied to guarantee the convergence of (28). Regarding the convergence, it is easy to show that, once the stability condition is met, we have $\mu(\theta)\,\hat{\phi}(\theta) < 1$ equivalently, where $\hat{\phi}(\theta) \triangleq \max_{1\le i\le k}\{\phi_i(\theta)\}$. In this case, we obtain that

$$
\sum_{\sum_{i=1}^{k} \pi_i = v} \prod_{i=1}^{k} \left(\mu(\theta)\,\phi_i(\theta)\right)^{\pi_i} \le \sum_{\sum_{i=1}^{k} \pi_i = v} \left(\mu(\theta)\,\hat{\phi}(\theta)\right)^{\sum_{i=1}^{k} \pi_i}.
$$

According to combinatorics properties, we notice that

$$
\begin{aligned}
\sum_{v=\tau}^{\infty} \sum_{\sum_{i=1}^{k} \pi_i = v} \left(\mu(\theta)\,\hat{\phi}(\theta)\right)^{\sum_{i=1}^{k} \pi_i} &= \sum_{v=\tau}^{\infty} \left(\mu(\theta)\,\hat{\phi}(\theta)\right)^{v} \sum_{\sum_{i=1}^{k} \pi_i = v} 1 \\
&= \sum_{v=\tau}^{\infty} \binom{k-1+v}{v} \left(\mu(\theta)\,\hat{\phi}(\theta)\right)^{v} \\
&\le \sum_{v=0}^{\infty} \binom{k-1+v}{v} \left(\mu(\theta)\,\hat{\phi}(\theta)\right)^{v} = \left(1-\mu(\theta)\,\hat{\phi}(\theta)\right)^{-k}.
\end{aligned}
$$

Therefore, we can demonstrate that, $\mathsf{M}_{A,S''}(\theta, s, t)$ is convergent if the stability condition holds, since

$$
\mathsf{M}_{A,S''}(\theta, s, t) \le e^{\theta\sigma(\theta)} \mu^{t-s}(\theta) \left(1-\mu(\theta)\,\hat{\phi}(\theta)\right)^{-k} < \infty.
$$

Finally, with respect to (9), the theorem then can be concluded by letting $s = t+w$.

H.3 Proof of Theorem E.1

We define the function $y(r) \triangleq \log \mathbb{E}\left[(1+rX)^{-z}\right]$ for $r > 0$, where X represents a positive random variable and z is any positive number. Then, the first derivative of $y(r)$ with respect to r, written by $y'(r)$, is obtained as

$$y'(r) = -\frac{z\mathbb{E}\left[(1+rX)^{-(z+1)}X\right]}{\mathbb{E}\left[(1+rX)^{-z}\right]} < 0.$$

Besides, the second derivative of $y(r)$ with respect to r, written by $y''(r)$, is obtained as

$$\begin{aligned}
y''(r) =& \frac{1}{\mathbb{E}^2\left[(1+rX)^{-z}\right]} \cdot \left(z\mathbb{E}\left[\frac{X^2}{(1+rX)^{z+2}}\right]\mathbb{E}\left[\frac{1}{(1+rX)^{z}}\right] \right. \\
&\left. + z^2\left(\mathbb{E}\left[\frac{X^2}{(1+rX)^{z+2}}\right]\mathbb{E}\left[\frac{1}{(1+rX)^{z}}\right] - \mathbb{E}^2\left[\frac{X}{(1+rX)^{z+1}}\right]\right)\right) \\
=& \frac{1}{\mathbb{E}^2\left[(1+rX)^{-z}\right]} \cdot \left(z\mathbb{E}\left[\frac{X^2}{(1+rX)^{z+2}}\right]\mathbb{E}\left[\frac{1}{(1+rX)^{z}}\right] \right. \\
&\left. + z^2\left(\mathbb{E}\left[\frac{X^2}{(1+rX)^{z+2}}\right]\mathbb{E}\left[\frac{1}{(1+rX)^{z}}\right] - \mathbb{E}^2\left[\frac{X}{(1+rX)^{\frac{z}{2}+1}} \cdot \frac{1}{(1+rX)^{\frac{z}{2}}}\right]\right)\right) \\
\geq& \frac{z\mathbb{E}\left[(1+rX)^{-(z+2)}X^2\right]\mathbb{E}\left[(1+rX)^{-z}\right]}{\mathbb{E}^2\left[(1+rX)^{-z}\right]} > 0,
\end{aligned}$$

where the last line is obtained by applying the *Cauchy–Schwarz inequality*, i.e., $\mathbb{E}^2[AB] \leq \mathbb{E}\left[A^2\right]\mathbb{E}\left[B^2\right]$ for random variables A and B. Since $y'(r) < 0$ and $y''(r) > 0$, it is shown that $y(r)$ is a monotonically decreasing and strictly convex function with respect to $r > 0$.

Applying Jensen's inequality, we immediately have

$$\sum_{i=1}^{m} \log \mathbb{E}\left[\left(1 + \frac{\gamma_i \xi}{L^\alpha}\right)^{-\eta\theta}\right] \geq m \log \mathbb{E}\left[\left(1 + \sum_{i=1}^{m} \frac{\gamma_i \xi}{L^\alpha}\right)^{-\eta\theta}\right],$$

where the equality is achieved if and only if $\gamma_i = \gamma_j \triangleq m^{-1}\gamma$ holds for all $1 \leq i, j \leq m$. Thus, we can easily obtain, following the same lines as above, that

$$\mathcal{R}^*_{S'}(\theta) \leq -\frac{m}{\theta} \log \mathbb{E}\left[\left(1 + \gamma\xi L^{-\alpha}\right)^{-\eta\theta}\right],$$

and the proof is completed by applying (13).

H.4 Proof of Theorem E.2

For the upper bound, by the definition of $S''(0,t)$, we have

$$S''(0,t) \leq \min_{1\leq i\leq k}\left\{S_i''(0,t)\right\} \leq \frac{1}{k}\sum_{i=1}^{k} S_i''(0,t).$$

Then, for $\overline{\mathbb{M}}_{S''}(\theta,0,t)$, we have

$$\begin{aligned}\overline{\mathbb{M}}_{S''}(\theta,0,t) \geq & \mathbb{E}\left[\exp\left(-\frac{\theta}{k}\sum_{i=1}^{k} S_i''(0,t)\right)\right] = \mathbb{E}\left[\prod_{i=1}^{k}\exp\left(-\frac{\theta}{k}S_i''(0,t)\right)\right] \\ = & \left(\mathbb{E}\left[\exp\left(-\frac{\theta}{k}\sum_{k=0}^{t-1} C_i''^{(k)}\right)\right]\right)^k = \left(\mathbb{E}\left[\left(1+\xi_i\left(\frac{\gamma}{k}\right)\left(\frac{L}{k}\right)^{-\alpha}\right)^{-\frac{\eta\theta}{k}}\right]\right)^{kt} \\ = & \left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M, 1+M-\frac{\eta\theta}{k}, \frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right)^{kt},\end{aligned}$$

Finally, the upper bound on $\mathcal{R}_{S''}^*(\theta)$ can be obtained as

$$\mathcal{R}_{S''}^*(\theta) \leq -\frac{k}{\theta}\log\left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M, 1+M-\frac{\eta\theta}{k}, \frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right).$$

For the lower bound, based on (25) and the homogeneous settings, we have the upper bound on $\overline{\mathbb{M}}_{S''}(\theta,0,t)$ as

$$\begin{aligned}\overline{\mathbb{M}}_{S''}(\theta,0,t) \leq & \sum_{\sum_{i=1}^{k}\pi_i=t}\prod_{i=1}^{k}\left(\mathbb{E}\left[\left(1+\xi_i\gamma_i l_i^{-\alpha}\right)^{-\eta\theta}\right]\right)^{\pi_i} \\ = & \left(\mathbb{E}\left[\left(1+\xi k^{\alpha-1}\gamma L^{-\alpha}\right)^{-\eta\theta}\right]\right)^t \sum_{\sum_{i=1}^{k}\pi_i=t} 1 \\ = & \binom{t+k-1}{k-1}\left(\mathbb{E}\left[\left(1+\xi k^{\alpha-1}\gamma L^{-\alpha}\right)^{-\eta\theta}\right]\right)^t,\end{aligned}$$

where the second line is achieved due to the uniformly allocated transmit power (normalized) and the identical length of each hop. Therefore, the lower bound for $\mathcal{R}_{S''}^*(\theta)$ can be obtained as

$$\begin{aligned}\mathcal{R}_{S''}^*(\theta) \geq & \lim_{t\to\infty}\frac{\log\left(\binom{t+k-1}{k-1}\left(\mathbb{E}\left[\left(1+\xi k^{\alpha-1}\gamma L^{-\alpha}\right)^{-\eta\theta}\right]\right)^t\right)}{-\theta t} \\ \geq & \frac{\log\mathbb{E}\left[\left(1+\xi n^{\alpha-1}\gamma L^{-\alpha}\right)^{-\eta\theta}\right]}{-\theta} - \lim_{t\to\infty}\frac{\log\frac{(t+k-1)^{n-1}}{(k-1)!}}{\theta t}\end{aligned}$$

$$= \frac{\log \mathbb{E}\left[\left(1+\xi k^{\alpha-1}\gamma L^{-\alpha}\right)^{-\eta\theta}\right]}{-\theta},$$

where the inequality in the second line uses the property that

$$\binom{N}{k} \leq \frac{N^k}{k!},$$

for any integers $N \geq k \geq 0$.

Therefore, the maximum effective bandwidth for the network densification scheme is lower bounded by

$$\begin{aligned}\mathcal{R}^*_{S''}(\theta) &\geq -\frac{1}{\theta}\log \mathbb{E}\left[\left(1+\xi k^{\alpha-1}\gamma L^{-\alpha}\right)^{-\eta\theta}\right]\\ &= -\frac{1}{\theta}\log\left(\left(\frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)^M U\left(M, 1+M-\eta\theta, \frac{ML^\alpha}{\gamma k^{\alpha-1}}\right)\right).\end{aligned}$$

Paper J5

Low-Latency Heterogeneous Networks with Millimeter-Wave Communications

Guang Yang, Ming Xiao, Muhammad Alam, and Yongming Huang

Refereed article to appear in
IEEE Communications Magazine, 2018.

Low-Latency Heterogeneous Networks with Millimeter-Wave Communications

Guang Yang, Ming Xiao, Muhammad Alam, and Yongming Huang

Abstract

Heterogeneous network (HetNet) is a key enabler to largely boost network coverage and capacity in the forthcoming fifth-generation (5G) and beyond. To support the explosively growing mobile data volumes, wireless communications with millimeter-wave (mm-wave) radios have attracted massive attention, which is widely considered as a promising candidate in 5G HetNets. In this article, we give an overview on the end-to-end latency of HetNets with mm-wave communications. In general, it is rather challenging for formulating and optimizing the delay problem with buffers in mm-wave communications, since conventional graph-based network optimization techniques are not applicable when queues are considered. Toward this end, we develop an adaptive low-latency strategy, which uses cooperative networking to reduce the end-to-end latency. Then, we evaluate the performance of the introduced strategy. Results reveal the importance of proper cooperative networking in reducing the end-to-end latency. In addition, we have identified several challenges in future research for low-latency mm-wave HetNets.

A Introduction

A.1 Background and Motivation

To significantly improve the spectral efficiency and throughput, future wireless networks, e.g., the fifth-generation (5G) mobile network and beyond, are expected to be largely implemented in the heterogeneous manner, i.e., heterogeneous networks (HetNets) [ZZM+15]. In HetNets, diverse wireless applications, facility configurations, radio access techniques (RAT), and quality-of-service (QoS) requirements are supported.

With the proliferation of electronic devices and the rapid development of computer science, the traffic load of wireless communications increases continuously and tremendously. To meet the ever-increasing requirements in capacity, one of the most important technologies is millimeter wave (mm-wave), which enables multi-gigabits per second (Gbps) transmission rates, thanks to the abundant spectral resources [RSM+13]. Different from conventional mobile communications in sub-6 GHz bands, due to the short wavelength of mm-wave radio, it is easy to integrate tens-to-hundreds of antenna elements onto a small-size chip with lower costs. The resulting high directivity not only provides a higher antenna gain for combating the severe path loss in mm-wave bands, but also increases the spatial reuse [XMH+17].

Besides, due to the fast attenuation of mm-wave signals, the communication distance is commonly limited to short ranges, e.g., 150 to 200 meters. Thus, the inter-cell interference between neighboring small cells is commonly negligible.

In light of above, we notice that mm-wave can be flexibly utilized for diverse devices and network architectures to boost the the coverage and the spectral efficiency. Thus, mm-wave communications has been extensively considered as a promising candidate in HetNets in 5G mobile networks and beyond, and research from various aspects has been massively conducted, e.g., [MJA+17, MHR+17]. It is known that, in future mobile communications, latency plays a critical role in the QoS. However, low latency becomes a rather challenging task in 5G mm-wave HetNets, due to the following two facts:

- Buffers will be used in 5G to handle the unprecedentedly heavy traffic, while the queuing delay may seriously deteriorate the QoS in 5G.
- Diverse RATs and/or architectures of HetNets make it rather difficult to perform networking optimizations for lower latency.

Therefore, it is those open challenges that motivate us to investigate the low.latency mm-wave HetNets with buffers in this paper.

A.2 Low-Latency Communications

As aforementioned, to support massive and various delay-sensitive applications, low latency as an important QoS feature needs to be satisfied in future wireless communications [SMS+17]. In 5G networks, the end-to-end latency requirement will be on the order of 1 to 5 milliseconds (ms) [ABC+14], which is more stringent than that in 3G and 4G LTE systems. Thus, it is rather challenging for fulfilling ultra-low latency in future mobile communications.

In the past few years, many efforts have been devoted to low-latency communications. In [TS15], for ultra-low latency inter-BS communications, the technical challenge and possible solution of point-to-multipoint in-band mm-wave backhaul for 5G networks were studied. In [FZM+17], focusing on three critical higher-layer aspects, i.e., core network architecture, protocols at the medium access control (MAC) layer, and congestion control policy, the main challenges and potential solutions for ultra-low latency 5G cellular networks were comprehensively surveyed and discussed. For mm-wave MIMO systems, from the perspective of training time in hybrid beamforming, a novel algorithm based on progressive channel estimation was developed in [CLW16]. In [YXG+16], the upper bound on the probabilistic delay was proposed to keep the track of the latency of point-to-point buffer-aided systems with mm-wave.

Considering the unprecedented data volumes in 5G networks, large buffers are usually applied at the transceivers. It is known that, for wireless systems with buffers, the queuing delay dominantly affects the overall system latency [KR09].

Therefore, for buffer-aided HetNets, it is crucial to realize the ultra-low latency by largely reducing the queuing delay.

A.3 Low-Latency HetNets with Buffer

Although many remarkable progresses have been achieved, it is still an open and challenging topic on reducing the end-to-end latency in buffer-aided HetNets. The major difficulty lies in the incorporation of buffers, which makes the problem differ a lot from the conventional latency minimization problems. More exactly, in the presence of buffers, the end-to-end latency relies not only on the capacity of each link, but also on the arriving sequence and queuing state at the buffer. In this sense, the end-to-end latency for buffer-aided networks cannot be simply formulated as a conventional graph-based network optimization problem [YDX15, Ber98], e.g., shortest path problem, max-flow problem or min-cost flow problem. To the best of our knowledge, the latency minimization problem of the buffer-aided HetNet has not been studied previously.

In this article, we introduce an adaptive strategy for HetNets with buffers, where the cooperative networking is applied. Specifically, we restrict ourselves to a HetNet that consists of one micro cell, two small cells and one user. Results show that the proper cooperative networking plays a critical role in minimizing the end-to-end latency, and our work provides an insight for optimizing future HetNets.

The remainder of this article is organized as follows. In Sec. B, we present the system architecture for 5G HetNets with mm-wave communications, and elaborately discuss several potential scenarios in downlink communications. In Sec. C, we develop an adaptive strategy for minimizing the latency of downlink transmission, based on cooperative networking. In Sec. D, we evaluate the performance of the introduced adaptive low-latency strategy, which indicates the importance of proper cooperative networking. In Sec. E we identify several technical challenges in future research, and we summarize our work in Sec. F.

B HetNets with mm-wave Communications

A HetNet commonly consists of a macro-cell evolved NodeB (MeNB) and multiple small-cell evolved NodeBs (SeNBs). The MeNB is deployed to guarantee wide-range and seamless coverage, while the SeNBs, e.g., pico, femto, and relay eNBs, are deployed to increase the overall system throughput. In the HetNet with mm-wave communications, as illustrated in Fig. 1.1, the SeNB is connected to other SeNBs or the MeNB via mm-wave backhaul. The user equipment (UE) gets service from the SeNB via the mm-wave access if it is located in any small cell, and it communicates with the MeNB using microwave radios, otherwise. Thus, for link robustness considerations, dual bands, i.e., mm-wave and microwave bands, are supported at both the MeNB and the UE. It is also possible to have communications working in mm-wave and microwave bands simultaneously, where eleven distinct scenarios need to be considered (with or without the MeNB-UE connection). For

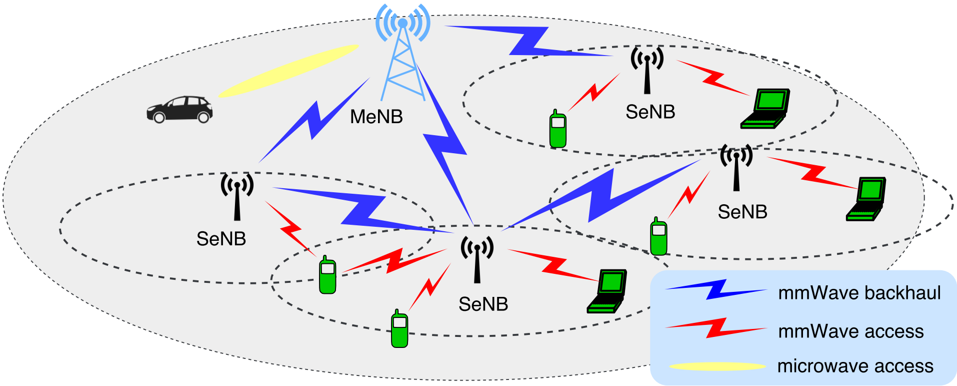


Figure 1.1: Illustration of heterogeneous networks (HetNets) with millimeter-wave (mm-wave) and microwave communications.

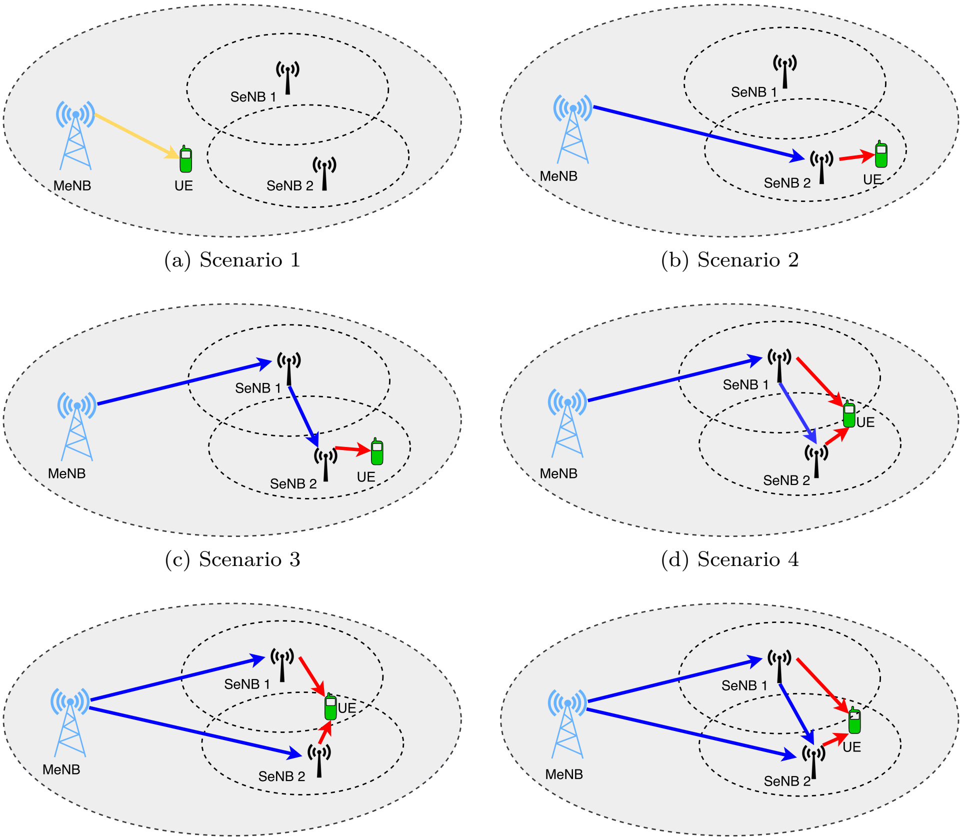
analytical simplicity, in this paper, we assume that the UE can only work in either mm-wave bands or microwave bands. In other words, the UE cannot simultaneously connect to the MeNB via the mm-wave link and to the SeNBs via the microwave link.

In what follows, for simplifying illustration, we specifically consider a HetNet that consists of one MeNB, two SeNBs and one UE. In such a network, there are several potential scenarios for downlink transmission from the MeNB to the UE, as illustrated in Fig. 1.2. These scenarios are elaborated on and discussed as follows:

Scenario 1: The UE belongs to neither of the small cells, and it is served by the MeNB via microwave radios, as shown in Fig. 1.2a. In this scenario, thanks to the direct connection between the MeNB and the UE, there is no extra queuing delay caused by any intermediate node. However, due to the limited bandwidth in microwave bands, the smaller channel capacity (compared to that in mm-wave bands) may produce a larger latency.

Scenario 2: As shown in Fig. 1.2b, the UE communicates with SeNB 2 via the mm-wave access, and SeNB 2 directly connects to the MeNB via the mm-wave backhaul. Thus, a two-hop network is formed. Unlike **Scenario 1**, in spite of two hops, the end-to-end latency can be largely reduced mainly thanks to mm-wave links.

Scenario 3: Slightly different from **Scenario 2**, in Fig. 1.2c, mm-wave backhauls MeNB–SeNB 1 and SeNB 1–SeNB 2 are available, while there is no direct connection between the MeNB and SeNB 2. Thus, the downlink communication is fulfilled through a three-hop network, following the routing MeNB–SeNB 1–SeNB 2–UE.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

Figure 1.2: Potential scenarios for a HetNet with mm-wave communications, where one MeNB, two SeNBs and one UE are considered.

Scenario 4: As shown in Fig. 1.2d, mm-wave backhaul transmissions are available only for MeNB–SeNB 1 and SeNB 1–SeNB 2. Besides, the UE emerges in the overlapped region of two neighboring small cells, i.e., edge UE, such that it can get served by both SeNB 1 and SeNB 2 via respective mm-wave access, simultaneously. Compared to **Scenario 3**, the difference lies in the UE's association to SeNB 1. In this scenario, the original data traverses from MeNB to SeNB 1. Then, the received traffic at SeNB 1 is divided into two parts: one fraction is directly delivered to the UE from SeNB 1, and the other fraction is delivered to the UE via SeNB 2. Thus, SeNB 2 actually works as a cooperative node, namely, coordinator, which helps SeNB 1 to offload and forward the traffic.

Scenario 5: For the edge UE emerging in the overlapped region of two neighboring small cells, if mm-wave backhauls between the MeNB to both SeNBs are available, while SeNBs cannot communicate with each other, then **Scenario 4** becomes **Scenario 5**. In this scenario, the original data traffic is partitioned into two parts at MeNB, which reach the UE via two SeNBs, respectively.

Scenario 6: For the edge UE emerging in the overlapped region of two neighboring small cells, if all mm-wave backhauls, i.e., MeNB–SeNB 1, MeNB–SeNB 2, and SeNB 1–SeNB 2, are available, then **Scenario 4** or **Scenario 5** becomes **Scenario 6**. In this scenario, the data traffic may be partitioned and reallocated at the MeNB and SeNB 1, and SeNB 2 as the coordinator is only responsible for merging and forwarding the potential incoming traffic.

For **Scenarios 4** to **Scenario 6**, we notice that there exists the process data splitting and merging. These processes correspond to the *fork-join* system [Che11] in practice, where the data can be correctly recovered at the UE with synchronization constrains for file transfer.

C Adaptive Low-Latency Strategy Based on Cooperative Networking

C.1 Adaptive Low-Latency Strategy

Normally, the MeNB plays the role of "decision maker" or "controller" on the control plane in the HetNet, which determines the networking scheme according to the collected information from SeNBs and the UE. Subsequently, the MeNB, SeNBs and the UE follow the decision distributed from the MeNB, such that the corresponding networking is performed afterwards on the data plane.

We in this section develop an adaptive low-latency transmission strategy in a HetNet with mm-wave communications, which gives the optimal networking scheme according to the acquired channel information. The adaptive strategy is illustrated in Fig. 1.3, where first-in-first-out (FIFO) queues are used for the buffer-aided HetNet. We can see that there are two planes in the HetNet, namely, the control

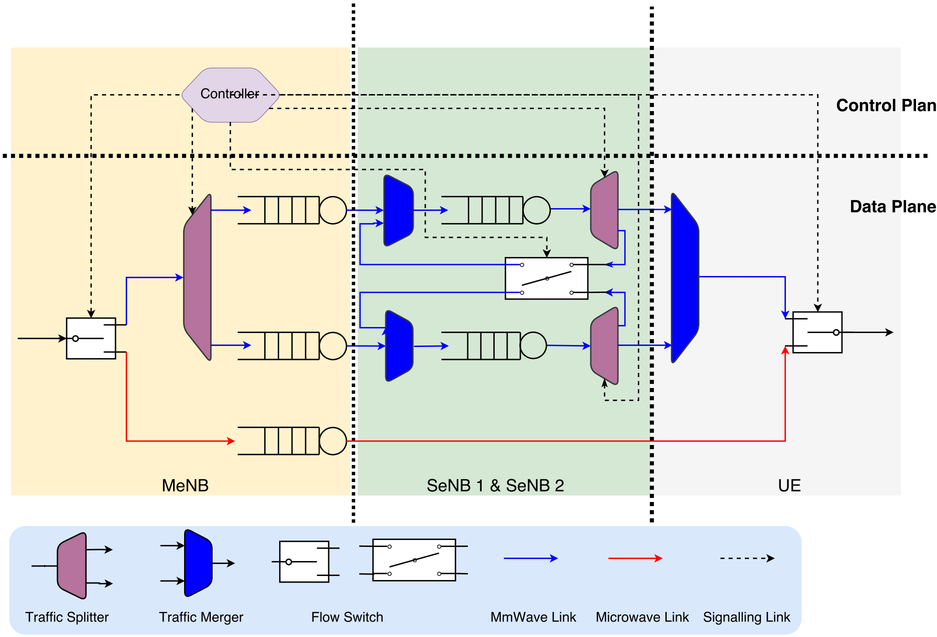


Figure 1.3: Diagram of adaptive low-latency strategy, where first-in-first-out (FIFO) queues are used for the buffer-aided HetNet.

plane and the data plane. The control plane is responsible for collecting the channel information, directing the networking scheme, and arranging the traffic allocation, where only control signals are operated in this plane. The data plane is only used for data transmission, where all operations are performed under the received control signals. Based on the UE's information, the MeNB will judge if the UE belongs to either small cell. If the UE is outside both small cells, then the MeNB decides to fulfill the downlink transmission in microwave bands. Otherwise, the MeNB needs to design a networking scheme, where the SeNB(s) will potentially participate in the networking. As shown in Fig. 1.3, data traffic can be partitioned through the traffic splitter, or combined through the traffic merger, under the guidance of the controller. Note that, if the mm-wave backhaul between two SeNBs is available, the controller will consider not only the proper traffic load allocated onto this mm-wave backhaul, but also the proper flow direction, i.e., from SeNB 1 to SeNB 2, or the opposite direction. Thus, both flow directions need to be considered, and the optimal networking is finally made by comparing the potential resulting end-to-end latency.

For $i \in \{1, 2\}$, we denote the channel capacity of mm-wave backhaul between the MeNB and SeNB i by C_{M,S_i}, the channel capacity of mm-wave access between SeNB i and the UE by $C_{S_i,U}$, and the channel capacity of the mm-wave backhaul between two SeNBs by C_{S_1,S_2}. Moreover, $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$ represent the traffic allocation coefficients at the MeNB and the non-cooperative SeNB, respectively, where $\mathcal{A} \subset [0,1]$ and $\mathcal{B} \subset [0,1]$ are corresponding feasible sets of traffic allocation coefficients. That is, if α (resp. β) fraction of the file is allocated onto one path, then the left $\bar{\alpha} \triangleq 1 - \alpha$ (resp. $\bar{\beta} \triangleq 1 - \beta$) fraction will be allocated onto the other path. The main idea of the algorithm is that, selecting SeNB 1 and SeNB 2 alternatively as the potential coordinator (the coordinator receives the traffic from both the MeNB and the non-cooperative SeNB, and then forwards the data to the UE), the algorithm traverses each feasible (α, β) in the spanned space $\mathcal{A} \times \mathcal{B}$, and computes all potential resulting end-to-end latency. Finally, the proper coordinator, i.e., SeNB ξ with $\xi = 1$ or 2, can be identified, and the optimal allocation pair (α, β) can be obtained.

The decision-making procedure at the MeNB is summarized as follows:

1. According to the information of the UE (channel information and location information), the MeNB first judges if Scenario 1 describes the current situation. If yes, a direct transmission in microwave bands will be performed, i.e., $\xi \leftarrow \emptyset$ and $(\alpha, \beta) \leftarrow (\emptyset, \emptyset)$. Otherwise, the downlink transmission requires the participation of SeNB(s), and the MeNB performs the following steps to make the networking decision.

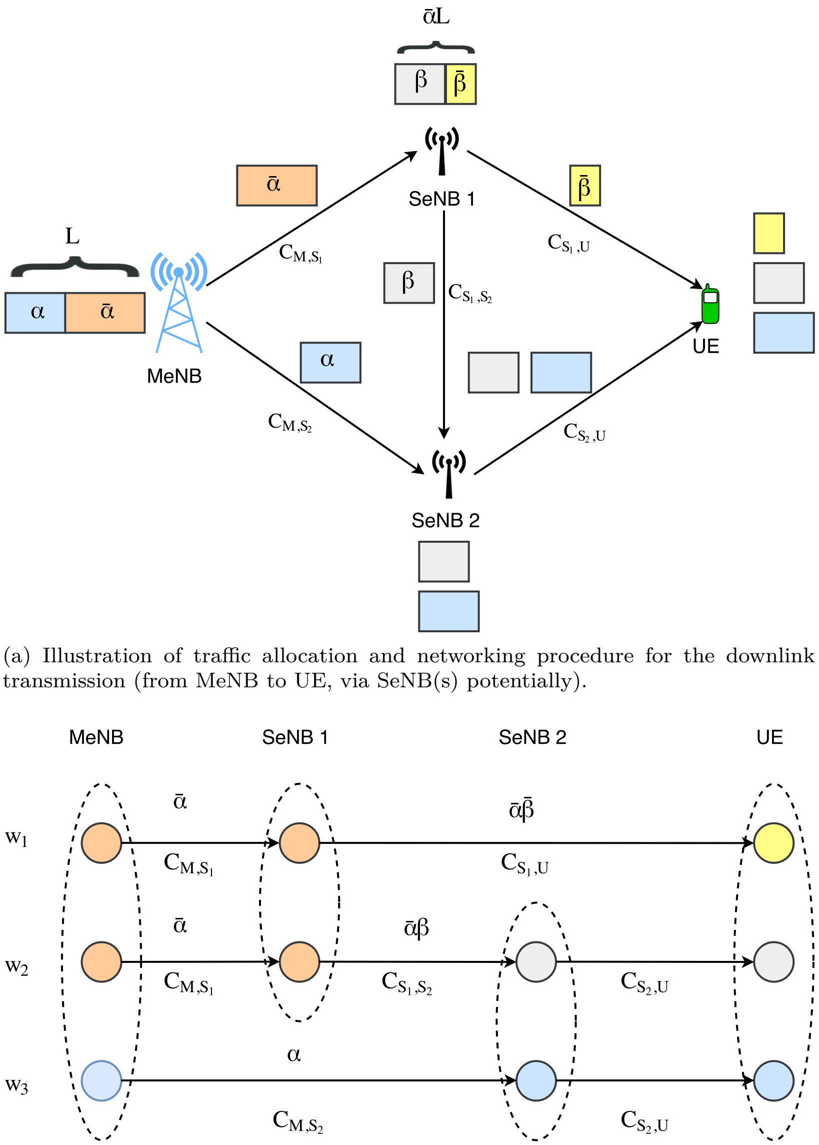
2. Treating SeNB 1 as the coordinator, the MeNB computes the minimum end-to-end latency τ_1^* and the associated optimal allocations (α_1^*, β_1^*), in the presence of known C_{M,S_1}, C_{M,S_2}, C_{S_1,S_2}, $C_{S_1,U}$, $C_{S_2,U}$, $\mathcal{A}$ and $\mathcal{B}$. Meanwhile, treating SeNB 2 as the coordinator, alternatively, the MeNB computes the minimum end-to-end latency τ_2^* and the associated optimal allocations (α_2^*, β_2^*), likewise.

3. Comparing τ_1^* and τ_2^*, the MeNB selects SeNB ξ with $\xi \leftarrow \arg_{i \in \{1,2\}} \min \tau_i^*$ as the coordinating SeNB, and the corresponding $(\alpha^*, \beta^*) \leftarrow \left(\alpha_\xi^*, \beta_\xi^*\right)$ will be adopted for networking as the optimal traffic allocations.

Note that the end-to-end delay only depends on the routing decision and the link capacities. Thus, the strategy proposed above can work for both multi-tier and multi-RAT HetNets.

C.2 Traffic Allocation for Cooperative Networking

The strategy of traffic allocation for cooperative networking is shown in Fig. 1.4. α and β denote the fractions of traffic allocated onto MeNB–SeNB 2 and SeNB 1–SeNB 2 mm-wave backhauls, respectively (SeNB 2 is taken as the coordinator in Fig. 1.4). Let the size of file for the downlink transmission be L units. As shown in Fig. 1.4a, the first allocation happens at the MeNB, where αL and $\bar{\alpha} L$ units of the file are pushed onto the mm-wave backhauls MeNB–SeNB 2 and MeNB–SeNB 1, respectively. The second allocation happens at SeNB 1, where the received $\bar{\alpha} L$



(a) Illustration of traffic allocation and networking procedure for the downlink transmission (from MeNB to UE, via SeNB(s) potentially).

(b) Abstraction of traffic allocation and networking.

Figure 1.4: Traffic allocation and networking procedure, and the corresponding abstraction.

units are divided into two parts, i.e., $\bar{\alpha}\beta L$ units and $\bar{\alpha}\bar{\beta}L$ units, for transmissions on the mm-wave backhaul SeNB 1–SeNB 2 and the mm-wave access SeNB 1–UE, respectively. SeNB 2 receives αL and $\bar{\alpha}\beta L$ units from both MeNB and SeNB 1, and buffers them in the queue. The downlink transmission is not completed until all units reach the UE, i.e., $\bar{\alpha}\bar{\beta}L$, $\bar{\alpha}\beta L$, and αL. An abstraction for this procedure is illustrated in Fig. 1.4b, where w_1, w_2 and w_3 denote component delays on three traversing paths, respectively, and the largest one among w_1, w_2 and w_3 defines the end-to-end latency.

We assume that the feasible sets of traffic allocations are given as $\mathcal{A} = \{\alpha : 0 \leq \alpha \leq 1\}$ and $\mathcal{B} = \{\beta : 0 \leq \beta \leq 1\}$, respectively. Taking the potential scenarios listed in Fig. 1.2 for example, the values for α and β are correspondingly given as follows:

- $\alpha \leftarrow \emptyset$ and $\beta \leftarrow \emptyset$ for **Scenario 1**
- $\alpha \leftarrow 1$ and $\beta \leftarrow \emptyset$ for **Scenario 2**
- $\alpha \leftarrow 0$ and $\beta \leftarrow 0$ for **Scenario 3**
- $\alpha \leftarrow 0$ and $\beta \leftarrow \beta^* \in (0, 1)$ for **Scenario 4**
- $\alpha \leftarrow \alpha^* \in (0, 1)$ and $\beta \leftarrow 0$ for **Scenario 5**
- $\alpha \leftarrow \alpha^* \in (0, 1)$ and $\beta \leftarrow \beta^* \in (0, 1)$ for **Scenario 6**

For minimizing the end-to-end latency, it is crucial to optimize α and β. Recalling the abstraction in Fig. 1.4b, for the first routing, i.e., MeNB–SeNB 1–UE, the component delay w_1 can be easily formulated. However, for the second routing and the third routing, i.e., MeNB–SeNB 1–SeNB 2–UE and MeNB–SeNB 2–UE, it is necessary to consider the order of arrivals of two distinct file fractions at SeNB 2. To be more precise, with FIFO queuing, the earlier arrival will be pushed onto mm-wave access SeNB 2–UE first, and the later one may have to wait in the queue until the earlier comer completely departs from the buffer. Comparing the arriving order of different file fractions, we then are able to formulate component delays w_2 and w_3. Finally, to achieve $\tau_1^* \leftarrow \min\max\{w_1, w_2, w_3\}$, the MeNB traverses all feasible $(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$ to identify the optimal traffic allocations, i.e., (α_1^*, β_1^*), which enables the minimal end-to-end latency when treating SeNB 2 as the coordinator.

It is worth mentioning that, the adaptive low-latency strategy based on cooperative networking can be extended to general scenarios with more than two SeNBs, as long as the channel information of all potential links is available for performing the global optimization, since a high-dimensional traffic allocation vector can always be generated for optimization with the global information. However, the major challenge with expanding size of HetNets is the computational complexity at the MeNB, since the cost for processing the channel information of all links and performing global optimization over a full-connected network (counting all potential links) dramatically increases. In this sense, it is necessary to consider the trade-off between the achieved latency and the computational complexity in practice. This issue is identified as a future challenge, stated in Sec. E.

D Performance Evaluation

In this section, we evaluate the performance of cooperative networking for HetNets with mm-wave communications. To investigate the impacts of traffic allocation pair (α, β), we assume deterministic settings for HetNets with mm-wave communications. That is, the channel capacities of mm-wave backhauls and accesses are $C_{M,S_1} = 12$ Gbps, $C_{M,S_2} = 8$ Gbps, $C_{S_1,S_2} = 7$ Gbps, $C_{S_1,U} = 0.8$ Gbps, and $C_{S_2,U} = 2$ Gbps, and the size of file for the downlink transmission is $L = 2$ Mb.

With different traffic allocation (α, β), the end-to-end latency is shown in Fig. 1.5, where SeNB 1 and SeNB 2 are selected as the coordinator in Fig. 1.5a and Fig. 1.5b, respectively. The region in dark blue represents the desired sets of feasible (α, β), which can provide a lower end-to-end latency for downlink transmission. We can see that, in both sub-figures, the dark blue regions emerge inside the square $[0, 1] \times [0, 1]$ for all potential α and β, i.e., $\alpha \in \mathcal{A}\backslash\{0, 1\}$ and $\beta \in \mathcal{B}\backslash\{0, 1\}$. This is resulted by the availability of mm-wave backhaul SeNB 1–SeNB 2, i.e., $C_{S_1,S_2} = 7 > 0$ Gbps. This observation indicates that, if the mm-wave backhaul between SeNBs is available, it is always beneficial to take advantages of this backhaul by properly performing traffic allocations, and the resulting end-to-end latency can be much less than those without traffic allocations, i.e., strategies with $\alpha \in \{0, 1\}$ or $\beta \in \{0, 1\}$. Furthermore, comparing Fig. 1.5a and Fig. 1.5b, we find that the minimum end-to-end latency is 0.396 ms when SeNB 1 is treated as the coordinator, while the minimum end-to-end latency is 0.426 ms when SeNB 2 is treated as the coordinator. Therefore, it is a better choice to take SeNB 1 as the coordinator, and perform the optimal traffic allocations at the MeNB and SeNB 2, respectively. The decision above is finally made and will be distributed from the MeNB for the subsequent low-latency cooperative networking.

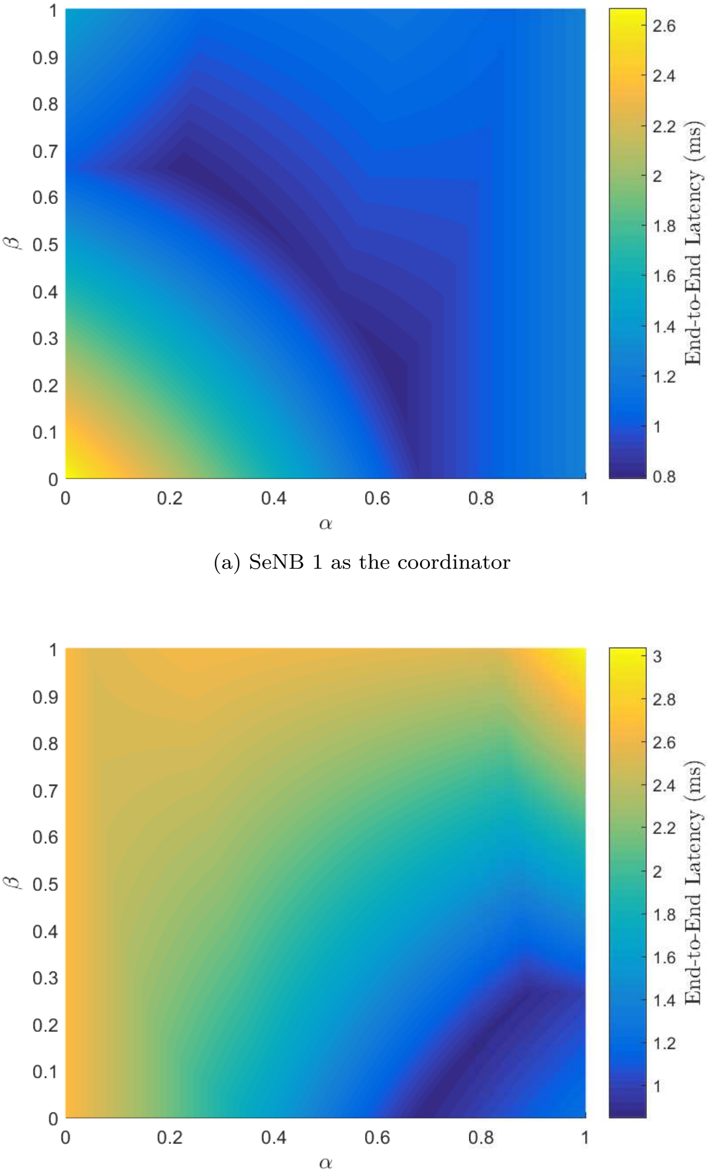
E Challenges in Future Research

Due to the densification tendency for small cells in future HetNets, the adaptive low-latency strategy developed in this article may face a few new technical challenges as follows:

- We only consider two small cells in our research. However, the overhead for collecting channel information and control signaling becomes tremendously heavy when more small cells are incorporated.

- The performance of the proposed approach depends on the channel state information. However, at mm-wave frequencies, it might be more difficult to obtain the channel state information due to the severe Doppler effect when mobility is involved. Thus, for mobile scenarios, it is a challenging task to overcome the degradation cased by Doppler effect.

- Since the capacities of mm-wave backhauls or accesses is not infinite, it is critical to properly schedule transmissions and manage the traffic in the presence



(a) SeNB 1 as the coordinator

(b) SeNB 2 as the coordinator

Figure 1.5: End-to-end latency with traffic allocation coefficient pairs (α, β): (a) taking SeNB 1 as the coordinator; (b) taking SeNB 2 as the coordinator.

of multiple UEs in the HetNet, which however is a non-trivial optimization problem.

Thus, our future work will focus on developing a low-complexity and scalable algorithm for low-latency wireless communications in HetNets with buffers.

F Conclusions

HetNets with mm-wave communications can significantly improve the network coverage and capacity, to satisfy ever-increasing requirements in data rates and latency. We have considered a HetNet consisting of one MeNB, two SeNBs and one UE, and investigated the low-latency strategy for the downlink transmission from the MeNB to the UE. For the HetNets with buffers, we have introduced an adaptive strategy based on cooperative networking, which largely minimizes the latency through optimizing traffic allocations. Results have demonstrated that, a proper cooperative networking is critical in reducing the end-to-end latency, thereby providing an insight on traffic management and network optimization for future HetNets. Besides, we have identified several challenges regarding cooperative communications in low-latency HetNets to be addressed in future research.

Paper J6

Traffic Allocation for Low-Latency Multi-Hop Networks with Buffers

Guang Yang, Martin Haenggi, and Ming Xiao

Refereed article submitted to
IEEE Transactions on Communications, 2018.

Traffic Allocation for Low-Latency Multi-Hop Networks with Buffers

Guang Yang, Martin Haenggi, and Ming Xiao

Abstract

For buffer-aided tandem networks consisting of relay nodes and multiple channels per hop, we consider two traffic allocation schemes, namely local allocation and global allocation, and investigate the end-to-end latency of a file transfer. We formulate the problem for generic multi-hop queuing systems and subsequently derive closed-form expressions of the end-to-end latency. We quantify the advantages of the global allocation scheme relative to its local allocation counterpart, and we conduct an asymptotic analysis on the performance gain when the number of channels in each hop increases to infinity. The traffic allocations and the analytical delay performance are validated through simulations. Furthermore, taking a specific two-hop network with millimeter-wave (mm-wave) as an example, we derive lower bounds on the average end-to-end latency, where Nakagami-m fading is considered. Numerical results demonstrate that, compared to the local allocation scheme, the advantage of global allocation grows as the number of relay nodes increases, at the expense of higher complexity that linearly increases with the number of relay nodes. It is also demonstrated that a proper deployment of relay nodes in a linear network plays an important role in reducing the average end-to-end latency, and the average latency decays as the channels become more deterministic. These findings provide insights for designing multi-hop networks with low end-to-end latency.

A Introduction

A.1 Background and Motivation

In many future applications, low latency is a crucial quality-of-service (QoS) constraint [SMS+17]. For instance, vehicle-to-everything, remote surgery, and industrial control need the support of low-latency communications. Accordingly, the requirement on end-to-end latency in the fifth-generation (5G) mobile network and beyond, on the order of 1 to 5 ms, is much more stringent than that in 3G and 4G systems [ABC+14, OBB+14]. Besides, to handle the unprecedented data volumes and heavy traffic load in future wireless communications, large buffers are usually used at transceivers. With these buffers, the data can be temporarily stored in a queue, until the corresponding service is available for its delivery. The queuing delay is defined as the waiting time of a packet in the buffer or queue before being transmitted [BGH87, KR09]. For future wireless systems with buffers, the queuing delay becomes a key contributor to the overall latency, since the heavy network traffic may produce significant data backlog in buffers. Thus, one of the most effective ways for achieving lower latency is to reduce the queuing delay.

It is worth mentioning that, as one key enabler of high data-rate transmissions, millimeter-wave (mm-wave) technologies have raised extensive research interest and have been regarded as promising candidates in future mobile networks [RSM+13, RRE14, XMH+17]. Motivated by the huge potential of using mm-wave in various scenarios, in this work, we restrict ourselves to mm-wave bands to investigate the traffic allocations over multi-hop networks. In this work, we consider a linear multi-hop network that consists of a source node, a destination node, and multiple buffer-aided relay nodes, where parallel channels in each hop are also assumed. This network architecture is promising for future mobile networks and motivated by the following two facts:

(i) Unlike wireless communications in sub-6 GHz bands, mm-wave radios used in future mobile systems encounter much more severe path loss, which may restrict the range of wireless communications. One solution to enlarge the range of mm-wave communications is to use relay nodes. With the multi-hop architecture, the distance between adjacent nodes is shortened, thereby mitigating the serious path loss in mm-wave bands.

(ii) The consideration of several parallel channels in each hop mainly stems from the application of distributed antenna systems (DAS) or remote radio heads (RRH). Note that sharp beams are generated by the dense antenna elements in mm-wave bands. With DAS or RRH, multiple channels can be established between communication nodes with negligible inter-channel interference. Multiple channels in good conditions can be selected via proper channel estimation and tracking techniques, thereby enabling higher performance for mm-wave communications.

It is important to investigate the end-to-end latency for networks with the aforementioned multi-hop multi-channel architecture. However, this system model is rarely studied, especially when buffers are incorporated at relay nodes.

A.2 Related Works

In the past few years, numerous efforts have been devoted to the research on latency in multi-hop networks with buffers, and remarkable progress has been reported. In [BA09], the average end-to-end delay in random access multi-hop wireless ad hoc networks was studied, and the analytical results were discussed and compared with the well established information-theoretic results on scaling laws in ad hoc networks. For an opportunistic multi-hop cognitive radio network, the average end-to-end latency in the secondary network was studied in [DA15] by applying queuing theoretic techniques and a diffusion approximation. In [BMP15], the queuing delay and medium access distribution over multi-hop personal area networks was investigated.

To reduce the end-to-end latency in multi-hop queuing systems, many works have focused on various aspects such as the routing, scheduling, and traffic control. Using back-pressure methods, algorithms and analysis were widely investigated in [JJS13, JWL15, SPB16, MMT16] for low-latency multi-hop wireless networks. For a two-hop half-duplex network with infinite buffers at both the source and the relay node, the problem of minimizing the average sum queue length under a half-duplex constraint was investigated in [CLY15]. Some efforts for systems with interference incorporated have also been made in the past decade. In [LH08], several QoS routing problems, i.e., end-to-end loss rate, end-to-end average delay, and end-to-end delay distribution, in multi-hop wireless network were considered, where an exact tandem queuing model was established, and a decomposition approach for QoS routing was presented. Using a tuple-based multidimensional conflict graph model, a cross-layer framework was established in [CLSC16], in order to investigate the distributed scheduling and delay-aware routing in multi-hop multi-radio multichannel networks. Considering a multi-hop system that consists of one source, one destination, and multiple relays, the end-to-end delay performance under a TDMA-ALOHA medium access control protocol, with interferers forming a Poisson point process, was investigated in [SH14], and insights regarding delay-minimizing joint medium access control/routing algorithms were provided for networks with randomly located nodes. In the recent work [PPT17], a distributed flow allocation scheme was proposed for random access wireless multi-hop networks with multiple disjoint paths, aiming to maximize the average aggregate flow throughput and guarantee a bounded packet delay.

In spite of many significant achievements in multi-hop networks with buffers, e.g., [LH08, AZLB16, BMP15, JZSS15], the research on the traffic allocation to achieve low-latency transmission is rather limited. In our recent work [YXP18], two low-latency schemes for mm-wave communications, namely traffic dispersion and network densification, were investigated in the framework of network calculus and effective capacity. The analysis in [YXP18] was performed for a given network setting, i.e., fixed transmission scheme, sum power budget, and arrival rate for the whole network, and bounding techniques were used to explore the potential for low-latency communications. However, traffic allocations for reducing the latency were not investigated, and the potential performance gain by optimized traffic allocations was not quantified.

A.3 Objective and Contributions

For multi-hop networks with multiple channels in each hop, an optimized traffic allocation scheme plays a crucial role when incorporating queues [CY13], since the traffic congestion at the relay nodes due to non-optimized allocation may produce long queues, resulting in larger end-to-end latency [KR09]. Conventionally, the method for studying low-latency networks is to transfer the objectives into network optimization problems, i.e., problems in [Ber98]. However, these graph-based approaches do not apply to the scenarios with buffers. Therefore, it is necessary to investigate the latency optimization problem for buffered networks in a different way.

The main objective of our work is to develop an efficient traffic allocation scheme for mm-wave networks that minimize the end-to-end latency. Specifically, we consider a linear multi-hop buffer-aided networks with multiple channels in each hop. The main contributions of our work are summarized as follows:

- Focusing on two traffic allocation schemes, namely, local allocation and global allocation, we calculate the end-to-end latency for delivering a fixed-length message from the source to the destination. Furthermore, we analytically compare these two allocation schemes through the relative performance gain and investigate the benefits of global allocation.

- For multi-hop buffered networks with multiple channels in each hop, we exploit the recursive nature of the global allocation scheme. Thus, there is no need to search for the optimal solution in an exhaustive manner. The recursive method introduced in this paper significantly simplifies the global minimization of the end-to-end latency and provides insights for analyzing tandem queuing systems. Besides, we give the overall computational complexity for performing local or global allocations.

- Following the recursive characterization for global allocation, we present the asymptotic relative performance gain when the number of per-hop channels goes to infinity. Furthermore, based on traffic allocation schemes that can be applied to generic multi-hop networks, we specifically consider a two-hop linear mm-wave network and investigate the average end-to-end latency, with Nakagami-m fading incorporated. We derive lower bounds of the average end-to-end latency for two allocation schemes.

B System Model and Problem Formulation

B.1 System Model

We consider a multi-hop system, which consists of multiple buffer-aided relay nodes and multiple channels in each hop. The first-in-first-out (FIFO) rule applies to the queues at the buffer-aided relay nodes. As illustrated in Fig. 1.1, given n tandem relay nodes, we label all nodes in reverse order, i.e., from the destination to the source, to simplify the notation in the following analysis. That is, the destination and the source are labeled as node 0 and node $n+1$, respectively. The hop between node h and node $h+1$ is denoted by hop h, for all $h \in \{0\} \cup [n]$. In addition, we assume there are m_h channels in hop h, and we denote by $C_{h,k}$ the capacity of the k^{th} channel in hop h for all $k \in [m_h]$. We define $[N] \triangleq \{1, 2, \ldots, N\}$ for any $N \in \mathbb{N}$.

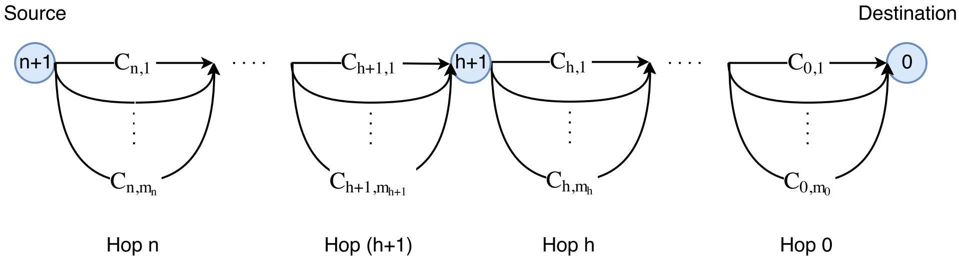


Figure 1.1: Illustration of a multi-hop system, consisting of multiple relay nodes are multiple channels in each hop.

Regarding the channels in each hop, as aforementioned, due to the negligible multi-path effect and the high directivity of mm-wave beams, the small-scale fading is very weak. We denote by $C_{h,k}$ the channel capacity. First we assume no fading for the investigation in Sec. C and Sec. D, and in Sec. E, Nakagami-m fading is added to the model. This modeling preserves the actual main characteristics of mm-wave channels in practice, and also provides high tractability for the following analysis.

In this work, we assume that the traffic allocation that allocates the traffic to the individual channels is performed at the source and relay nodes. That is, the traffic arriving at one node is decomposed into several fractions according to the given allocation scheme, and those fractions are subsequently pushed onto the channels and delivered to the next node, where each fraction is partitioned again. For the traffic allocation with respect to channels in hop h, we define $\underline{\alpha}_h \triangleq [\alpha_{h,1}, \alpha_{h,2}, \ldots, \alpha_{h,m_h}] \in \mathbb{R}_+^{m_h}$ for all $h \in \{0\} \cup [n]$, which is subject to the following constraint

$$\|\underline{\alpha}_h\|_1 \triangleq \sum_{i=1}^{m_h} \alpha_{h,i} = 1, \tag{1}$$

where $\|\cdot\|_1$ represents the 1-norm for vectors. The traffic allocation $\underline{\alpha}_h$ is determined by the capacities of outgoing channels, such that incoming fraction can be accordingly chopped down and reallocated onto the respective outgoing channels. It is worth noting that the traffic allocation $\underline{\alpha}_h$ is performed at node $h+1$.

We assume multiple parallel servers at the source node and the relay nodes (the number of servers equals the number of outgoing channels), such that fractions can be transmitted over the different channels at the same time. This model follows a special variant of general *fork-join* systems [FJ16, RPC16] (with a synchronization constraint), where all tasks of a job start execution simultaneously, and the job is completed when the final task leaves the system. Relay nodes are full-duplex but only equipped with a single buffer such that the data reception and transmission can be performed at the same time and the received fractions leave the buffer one by one. At each relay node, one fraction is not served (chopped into smaller fractions and forwarded to the next node) until it is completely received. In future mobile networks, the capacity of mm-wave channels can reach multi-gigabits per second and the packet size (referred to as the file size in this study) is around several kilobits or megabits at most [3GP17], and hence the latency is of the order of milliseconds or even smaller. In this sense, compared to the scheme that allows to receive and transmit a fraction simultaneously, the setting that a fraction has to be received entirely before it can be forwarded over the next hop definitely produces higher latency (but not significantly), while avoiding the potential interference induced by simultaneous transmission and reception. Furthermore, we assume that the file is infinitely divisible, i.e., any fraction can be divided into smaller pieces with arbitrarily small size. Aiming to investigate the limits of low latency, we assume infinite divisibility throughout our work for theoretical purposes. However, for communication systems in practice, there always exists an atomic unit for constituting packets, such that a file cannot be infinitely divisible. Hence, the delay performance obtained in this work is degraded if the practical constraint of finite divisibility is incorporated. It is worth mentioning that the assumption of infinite divisibility does not imply a fluid-flow model (although it is one of the typical features), since there are two phases, i.e., fraction reception and fraction transmission, at any intermediate node, rather than following the continuous manner in fluid-flow models.

B.2 Problem Formulation

For the transmission from the source to the destination, we consider two traffic allocation schemes, namely, local traffic allocation and global traffic allocation. To simplify the exposition, $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ are used to denote the local and global allocation scheme, respectively, which are described as follows.

- $\mathcal{M}_{\text{local}}$: Node h for all $h \in [n+1]$ only has the capacity information of the channels in hop $h-1$. The traffic allocation performed at node h only optimizes the transmission over channels in hop $h-1$. This scheme ensures that the latency in the local hop is minimized, but is oblivious to the traffic allocations in other hops.
- $\mathcal{M}_{\text{global}}$: Node h for all $h \in [n+1]$ has the entire capacity information of all channels from hop 0 to hop $h-1$. The traffic allocation performed at node h not only relies on channels in hop $h-1$, but also relies on channels in the remaining hops, i.e., from hop 0 to hop $h-1$. This scheme minimizes the latency through h hops.

The definition of the end-to-end latency in our study is given as follows:

Definition B.1 (End-to-End Latency)**.** Given traffic allocations $\{\underline{\alpha}_h\}$, $h \in \{0\} \cup [n]$, for all hops, the end-to-end latency $\tau_n\left(\underline{\alpha}_n, \underline{\alpha}_{n-1}, \cdots, \underline{\alpha}_0\right)$ (also written as τ_n in the sequel for notional simplicity) of the tandem queuing system with n relay nodes is defined as the time to deliver one fixed-length file of size 1 (without loss of generality) from the source to the destination[1], describing the time span from the moment the source starts transmission to the moment all fractions are received at the destination.

The definition above indicates that the latency takes into account both the time of traversing the wireless channels (service time) and the time of queuing in buffers (waiting time) at the relays. For exposition, we consider a simple network as an example to illustrate the end-to-end latency and the difference between $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, where $n = 1$, $m_1 = 2$, and $m_0 = 1$. In this specific network, traffic allocation $\underline{\alpha}_1 \triangleq [\alpha_{1,1}, \alpha_{1,2}]$ is performed over hop 1, while there is no traffic allocation over hop 0, i.e., $\underline{\alpha}_0 = 1$. The service time for fraction $\alpha_{i,j}$ over the channel with capacity $C_{i,j}$ is characterized by their quotient, i.e., $\alpha_{i,j} C_{i,j}^{-1}$. It is worth noting that, due to the adoption of buffer at node 1, the arrival order of different fractions should also be taken into account. Hence, in addition to the service time, the potential waiting time for the latter incoming fraction also contributes to the end-to-end latency τ_2. Then τ_2 is obtained as

$$\tau_2\left(\underline{\alpha}_1, \underline{\alpha}_0\right) = \begin{cases} \max\left\{C_{0,1}^{-1} + \alpha_{1,1}C_{1,1}^{-1}, \alpha_{1,2}\left(C_{0,1}^{-1} + C_{1,2}^{-1}\right)\right\}, & \alpha_{1,1}C_{1,1}^{-1} \leq \alpha_{1,2}C_{1,2}^{-1} \\ \max\left\{C_{0,1}^{-1} + \alpha_{1,2}C_{1,2}^{-1}, \alpha_{1,1}\left(C_{0,1}^{-1} + C_{1,1}^{-1}\right)\right\}, & \text{otherwise.} \end{cases} \tag{2}$$

From (2), we see that the end-to-end latency considered in this work is different from those that consider either the service time or the waiting time only. Thus, the conventional techniques for graph-based network optimization or queuing systems are not applicable.

Since the traffic allocation only occurs at hop 1, the resulting minimization problem regarding τ_2 can be formulated as

$$\tau_2^* = \min_{\|\underline{\alpha}_1\|=1} \tau_2\left(\underline{\alpha}_1, \underline{\alpha}_0\right) = \tau_2\left(\underline{\alpha}_1^*, \underline{\alpha}_0\right), \tag{3}$$

where $\underline{\alpha}_1^*$ representing the optimum traffic allocation differs for $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$. Following the distinct mechanisms of $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, we have:

- for $\mathcal{M}_{\text{local}}$, since the channel information at the local hop is adopted for optimization, $\underline{\alpha}_1^*$ is obtained as

$$\underline{\alpha}_1^* = \operatorname*{argmin}_{\|\underline{\alpha}_1\|=1} \max\left\{\alpha_{1,1}C_{1,1}^{-1}, \alpha_{1,2}C_{1,2}^{-1}\right\}. \tag{4}$$

[1] For generality and notational simplicity, we do not assign units to the file size and channel capacities, i.e., they are all normalized. For a concrete network, the most suitable units can be chosen, e.g., the file size (and, in turn, the fractions) could be measured in MB, and the capacities in Mb/s.

- for $\mathcal{M}_{\text{global}}$, since the channel information over all hops is adopted for optimization, $\underline{\alpha}_1^*$ is obtained as

$$\underline{\alpha}_1^* = \begin{cases} \underset{\|\underline{\alpha}_1\|=1}{\operatorname{argmin}} \max\left\{C_{0,1}^{-1} + \alpha_{1,1}C_{1,1}^{-1}, \alpha_{1,2}\left(C_{0,1}^{-1} + C_{1,2}^{-1}\right)\right\}, & \alpha_{1,1}C_{1,1}^{-1} \leq \alpha_{1,2}C_{1,2}^{-1} \\ \underset{\|\underline{\alpha}_1\|=1}{\operatorname{argmin}} \max\left\{C_{0,1}^{-1} + \alpha_{1,2}C_{1,2}^{-1}, \alpha_{1,1}\left(C_{0,1}^{-1} + C_{1,1}^{-1}\right)\right\}, & \text{otherwise.} \end{cases} \tag{5}$$

Evidently, $\mathcal{M}_{\text{global}}$ targets to the objective function τ_2 straightforwardly for optimization, while $\mathcal{M}_{\text{local}}$ can only give a sub-optimal solution via meeting the local optimization constraint.

It is worth noting that, when more hops and/or parallel channels on each hop are incorporated, it is not possible to provide a closed-form expression for the end-to-end latency (as given in (2)), since the orders of arrival of the fractions at different nodes become rather complicated.

C Traffic Allocation for Multi-Hop Networks

In this section, we will investigate the optimized end-to-end latency. In the analysis, the exact traffic allocations at the source and all relay nodes are presented, and the overall computational complexity for different traffic allocation schemes are briefly discussed subsequently. For generality, the capacities of parallel channels on each hop are distinguished by distinct notations. For specific scenarios considering the sum-capacity constraint, the bandwidth can be partitioned in any manner to implement parallel channels with arbitrarily distributed capacities.

We use “big-$\mathcal{O}$” notation to characterize the overall computational complexity for $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$. The definition of the “big-$\mathcal{O}$” notation is given as follows: assuming $u(x)$ and $f(x)$ are functions defined on some subset $\mathcal{X} \subset \mathbb{R}$, $\mathcal{O}(f(x))$ denotes the set of all functions $u(x)$ such that $|u(x)/f(x)|$ stays bounded, i.e.,

$$\mathcal{O}(f(x)) \triangleq \left\{u(x) : \sup_{x\in\mathcal{X}} |u(x)/f(x)| < \infty\right\}. \tag{6}$$

Clearly, we have $\mathcal{O}\left(f_1(x)\right) \subset \mathcal{O}\left(f_2(x)\right)$ if we have $|f_1(x)| \leq |f_2(x)|$ over all $x \in \mathcal{X}$.

C.1 Latency for Networks Using $\mathcal{M}_{\text{local}}$

Before deriving the minimum latency with $\mathcal{M}_{\text{local}}$, we start from a single-hop system, as shown in Fig. 1.2. To simplify the notation, we assume that there are m channels between the source and the destination, where C_i for $i \in [m]$ denotes the capacity of the i^{th} channel. The traffic allocation $\underline{\alpha} \triangleq [\alpha_1, \alpha_2, \ldots, \alpha_m]$ is performed at the source node.

In the following Lemma C.1, the optimal traffic allocation at the source node is presented, and the resulting minimum end-to-end latency for the system shown in Fig. 1.2 is also derived.

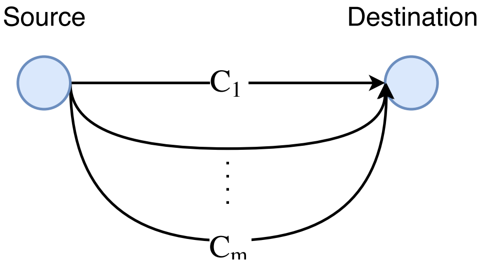


Figure 1.2: Illustration of a single-hop system with multiple channels between the source and the destination.

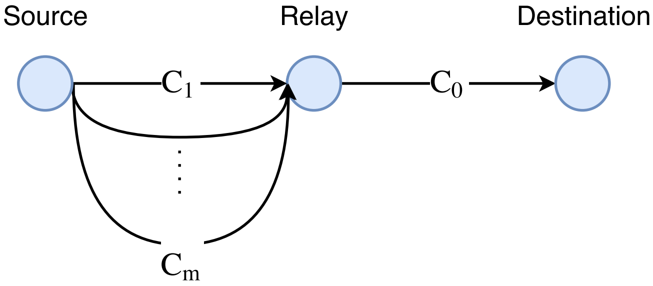


Figure 1.3: Illustration of a two-hop network, where multiple channels are between the source and the relay, and one channel between the relay to the destination.

Lemma C.1. *Given m channels with capacity C_i for $i \in [m]$ between the source and the destination, letting $\alpha_i \in (0,1)$ denote the fraction of the traffic allocated to the i^{th} channel with capacity C_i, the minimum end-to-end latency is*

$$\tau^* = \left(\sum_{i=1}^{m} C_i\right)^{-1}, \tag{7}$$

achieved by $\alpha_i = C_i \left(\sum_{i=j}^{m} C_j\right)^{-1}$ for all $i \in [m]$.

Proof. According to the mechanism of $\mathcal{M}_{\text{local}}$, we know that the minimum latency comes from applying the optimal traffic allocation $\underline{\alpha}^*$, i.e.,

$$\underline{\alpha}^* = \operatorname*{argmin}_{\|\underline{\alpha}\|=1} \max_{1\leq i\leq m} \left\{\alpha_i C_i^{-1}\right\}, \tag{8}$$

which is solved as $\alpha_i = C_i \left(\sum_{j=1}^{m} C_j\right)^{-1}$ for all $i \in [m]$. Then the minimum end-to-end latency τ^* can be obtained by applying $\underline{\alpha}^*$. □

From Lemma C.1, we notice that $\mathcal{M}_{\text{local}} = \mathcal{M}_{\text{global}}$ optimizes the traffic allocations such that all fractions arrive at the destination at the same time. Besides, from the perspective of the end-to-end delay, it is equivalent to having a single channel with the sum capacity when applying $\mathcal{M}_{\text{local}}$ for multiple channels.

Based on Lemma C.1, the optimal local traffic allocation and the resulting minimum latency for the multi-hop system in Fig. 1.1 are obtained in Theorem C.1.

Theorem C.1. *For the tandem network in Fig. 1.1 with n relay nodes and m_h channels in the h^{th} hop, the minimum end-to-end latency with $\mathcal{M}_{\text{local}}$ is*

$$\tau_n^* = \sum_{h=0}^{n} \left(\sum_{k=1}^{m_h} C_{h,k} \right)^{-1}, \tag{9}$$

achieved by $\alpha_{h,k} = C_{h,k} \left(\sum_{k=1}^{m_h} C_{h,k}\right)^{-1}$.

Proof. With $\mathcal{M}_{\text{local}}$, the objective is to ensure that all fractions can reach the next adjacent node simultaneously. According to Lemma C.1, in hop h the minimum latency is $w_h^* \triangleq \left(\sum_{k=1}^{m_h} C_{h,k}\right)^{-1}$.

Note that the traffic allocation is performed independently and sequentially from hop n to hop 0. In this case, the minimum end-to-end latency can be obtained by summing up w_h^* over all $h \in \{0\} \cup [n]$, i.e., $\tau_n^* \triangleq \sum_{h=0}^{n} w_h^* = \sum_{h=0}^{n} \left(\sum_{k=1}^{m_h} C_{h,k}\right)^{-1}$, when the proper local allocation scheme is applied. □

According to Theorem C.1, it is not difficult to find that the end-to-end transmission with $\mathcal{M}_{\text{local}}$ is equivalent to transmitting the entire file hop by hop. Paired with Lemma C.1, we know that all fractions are delivered from one node to the next simultaneously, which is equivalent to combining all sub-channels in each hop as a single channel with the sum capacity and moving the entire file through the network without partitioning. In this case, one relay node will buffer the whole file, while the buffers at other relay nodes remain empty. In terms of efficiency, the utilization of buffers for the transmission is relatively low, due to the uneven distribution of file fractions in the network. Furthermore, for networks with buffers, it is worth noting that the end-to-end latency (see the example with respect to (2)) differs from the non-buffered system, while the specific local scheme makes the resulting delay the same with the non-buffered system since arrivals on each hop can reach the node at the same time when $\mathcal{M}_{\text{local}}$ is applied.

C.2 Latency for Networks Using $\mathcal{M}_{\text{global}}$

Prior to investigating the optimized end-to-end latency with $\mathcal{M}_{\text{global}}$ for the multi-hop network shown in Fig. 1.1, we consider a two-hop system shown in Fig. 1.3, where a buffer-aided relay node is deployed between the source and the destination. We assume m channels between the source and the relay node, and we denote by C_i for $i \in [m]$ the capacity of the i^{th} channel. From the relay node to the destination, we assume that there is only one channel with capacity C_0. The traffic allocation $\underline{\alpha} \triangleq [\alpha_1, \ldots, \alpha_m]$ is performed at the source node, while no allocation is performed at the relay node, due to the single channel between the relay node and the destination.

In the following Lemma C.2, the optimal traffic allocation at the source node and the resulting minimum end-to-end latency for the system shown in Fig. 1.3 are derived.

Lemma C.2. *For the two-hop system shown in Fig. 1.3, the minimum end-to-end latency is*

$$\tau^* = \frac{\left(C_m^{-1} + C_0^{-1}\right) \prod_{k=1}^{m-1} C_{k+1} \left(C_k^{-1} + C_0^{-1}\right)}{1 + \sum_{i=2}^{m} \prod_{k=1}^{i-1} C_{k+1} \left(C_k^{-1} + C_0^{-1}\right)}, \tag{10}$$

achieved by

$$\alpha_i = \frac{\prod_{k=1}^{i-1} C_{k+1} \left(C_k^{-1} + C_0^{-1}\right)}{1 + \sum_{j=2}^{m} \prod_{k=1}^{j-1} C_{k+1} \left(C_k^{-1} + C_0^{-1}\right)}. \tag{11}$$

Proof. Please see Appendix H.1. □

Clearly, when applying the allocation in Lemma C.2, one fraction reaches the relay node always at the time when the previous fraction has completely left the buffer. In contrast to the scheme in Lemma C.1, all fractions arrive at the relay node sequentially, rather than simultaneously. Thus, the length of the resulting queue is the length of file fraction, which is usually much smaller than that of the whole file. It is worth mentioning that the optimum traffic allocation given in Lemma C.2 is not unique, i.e., there exist several other solutions that achieve the same minimum end-to-end latency.

$\mathcal{M}_{\text{global}}$ for a two-hop network

Assuming the sum capacity for channels between the source and the relay is fixed, we next investigate the impact of increasing the number of channels between the source and the relay on end-to-end latency for the two-hop network in Fig. 1.3.

Given a fixed sum capacity, it is evident that τ^* depends on the individual capacities C_i for $i \in [m]$. We will study the best channel capacity allocation for minimizing τ^* in the following corollary. Note that the channel capacity allocation determines the channel capacities potentially via frequency division techniques, while the traffic allocation partitions the data traffic according to the given channel capacities. We assume that the sum capacity for channels between the source and the relay is 1 without loss of generality.

Corollary C.1.1. *For the two-hop system in Fig. 1.3, given $\sum_{i=1}^{m} C_i = 1$ with $C_i > 0$ for $i \in [m]$, the channel allocation that minimizes τ^* is the uniform allocation, i.e., $C_i = m^{-1}$ for all $i \in [m]$.*

Proof. Please see Appendix H.2. □

Given a sum-capacity constraint, Corollary C.1.1 indicates that the lowest end-to-end latency with $\mathcal{M}_{\text{global}}$ can be achieved via the uniform allocation, i.e., via splitting the bandwidth equally.

Based on Corollary C.1.1, the minimum end-to-end latency with $C_i = m^{-1}$ for any $i \in [m]$ is

$$\tau^* = \left(C_0\left(1 - \left(1 + (mC_0)^{-1}\right)^{-m}\right)\right)^{-1}, \tag{12}$$

achieved by

$$\alpha_i = \frac{\left(1 + (mC_0)^{-1}\right)^{i-1} (mC_0)^{-1}}{\left(1 + (mC_0)^{-1}\right)^{m} - 1}. \tag{13}$$

We next investigate the monotonicity of τ^* in (12) and the asymptotic performance as $m \to \infty$.

Corollary C.1.2. *For the two-hop system in Fig. 1.3, assuming $C_i = m^{-1}$ for all $i \in [m]$, τ^* monotonically decreases with m and $\tau^* \to \left(C_0\left(1 - \exp\left(-C_0^{-1}\right)\right)\right)^{-1}$, when $m \to \infty$, and the limiting traffic allocation tends to be the uniform allocation.*

Proof. The monotonic decrease of τ^* with respect to m follows since $\left(1 + x^{-1}\right)^x$ increases with x. For the asymptotic latency, i.e., as $m \to \infty$, we have

$$\begin{aligned}
\lim_{m\to\infty} \tau^* = & \lim_{m\to\infty} C_0^{-1}\left(1 - \left(1 + (mC_0)^{-1}\right)^{-m}\right)^{-1} \\
= & C_0^{-1} \lim_{m\to\infty}\left(1 - \left(\left(1 + (mC_0)^{-1}\right)^{mC_0}\right)^{-\frac{1}{C_0}}\right)^{-1} \\
= & C_0^{-1}\left(1 - \left(\lim_{m\to\infty}\left(1 + (mC_0)^{-1}\right)^{mC_0}\right)^{-\frac{1}{C_0}}\right)^{-1} \\
= & \left(C_0\left(1 - \exp\left(-C_0^{-1}\right)\right)\right)^{-1}.
\end{aligned} \tag{14}$$

With (13), we notice that $\lim_{m\to\infty} \frac{\alpha_{i+1}}{\alpha_i} = \lim_{m\to\infty}\left(1 + (mC_0)^{-1}\right) = 1$ for all $i \in [m]$, which indicates that the limiting traffic allocation reduces to the uniform allocation. □

With a fixed sum capacity for channels between the source and the relay, it is demonstrated from Corollary C.1.2 that it is always beneficial to increase the number of channels. This observation coincides with the well-known fact that dividing the bandwidth into as many fractions as possible is delay-optimum for full-duplex channels (see results, e.g., in [NJS17]). However, this advantage diminishes diminishes as m grows, and the resulting end-to-end latency with $\mathcal{M}_{\text{global}}$ approaches a constant limit, which only depends on C_0. This corollary indicates that, with respect to a given sum capacity, it is better to have multiple channels with smaller capacity rather than one single channel with larger capacity, or to split the channel into sub-channels using frequency division. Also, it is worth noting that the asymptotic result is a lower bound on the latency for all cases where the sum capacity over the first hop is fixed, and the lower bound is determined by C_0 when applying $\mathcal{M}_{\text{global}}$.

Comparison with time division

In what follows, we consider a scenario that has only a single channel with capacity 1 (rather than multiple sub-channels shown in Fig. C.2) in the source-relay hop (corresponding to the systems that have only a single server at the source node). The file is partitioned into fractions via the traffic allocation in the time domain, which are sequentially delivered from the source node. Following the method for the proof of Lemma C.2, the minimum latency can be achieved if the condition $\alpha_i = C_0^{-1}\alpha_{i-1}$ holds for all $i \in [m] \setminus \{1\}$. The performance for traffic allocation in the time domain with $\mathcal{M}_{\text{global}}$ is given as below.

Lemma C.3. *Using time division, the minimum end-to-end latency τ^* with $\mathcal{M}_{\text{local}}$ is*

$$\tau^* = C_0^{-1}\frac{\sum_{i=0}^{m} C_0^i}{\sum_{i=0}^{m-1} C_0^i} = \frac{1 - C_0^{m+1}}{C_0 - C_0^{m+1}}, \tag{15}$$

achieved by

$$\alpha_i = C_0^{1-i}\sum_{k=0}^{m-1} C_0^{-k}. \tag{16}$$

τ^* monotonically decreases with m, and $\tau^* \to \max\left\{C_0^{-1}, 1\right\}$ when $m \to \infty$.

Proof. The detailed derivation for (15) is omitted since the method is similar to that in Appendix H.1. The monotonic decease of τ^* as m increases is evident by observing observing (15). For the limiting latency as $m \to \infty$, we have

$$\lim_{m\to\infty} \tau^* = C_0^{-1} \lim_{m\to\infty} \frac{\sum_{i=0}^{m} C_0^i}{\sum_{i=0}^{m-1} C_0^i} = \begin{cases} 1, & C_0 \geq 1 \\ C_0^{-1}, & C_0 < 1, \end{cases} \tag{17}$$

which can be summarized as $\lim_{m\to\infty} \tau^* = \max\left\{C_0^{-1}, 1\right\}$. Therefore, the proof is completed. □

We can see that the bottleneck channel (with smaller capacity) in the two-hop system determines the limiting latency. Unlike Corollary C.1.2, the file fractions are not simultaneously delivered from the source in Lemma C.3. This traffic allocation is performed using time division techniques. It is worth mentioning that we aim to study the minimum latency among all feasible traffic allocations in the time domain. We find that the optimal traffic allocation (with $\mathcal{M}_{\text{global}}$) based on the time division follows a geometric progression with the scale factor C_0^{-1}, while the uniform traffic allocation (with $\mathcal{M}_{\text{local}}$), i.e., $\alpha_i = m^{-1}$ for $i \in [m]$, cannot achieve the optimum.

For notational simplicity, we denote by τ_{f}^* and τ_{t}^* the minimum latency in (12) and (15), corresponding to the traffic allocations in the frequency domain and the time domain, respectively. We compare τ_{f}^* and τ_{t}^* for any positive integer m in the following corollary.

Corollary C.1.3. *$\tau_{\text{f}}^* \geq \tau_{\text{t}}^*$ holds for all $m \in \mathbb{N}$.*

Proof. Note that for $C_0 > 0$ we have

$$\left(C_0 + m^{-1}\right)^m = \sum_{i=0}^{m} m^{-i} \binom{m}{i} C_0^{m-i} \leq \sum_{i=0}^{m} C_0^i, \tag{18}$$

where the property $\binom{m}{i} \leq m^i$ for all $i \in \{0\} \cup [m]$ is applied, and the equality holds if $m = 1$ or $i = 0, 1$. Then we can obtain that

$$\left(1 + (mC_0)^{-1}\right)^{-m} \geq \frac{C_0^m}{\sum_{i=0}^{m} C_0^i}, \tag{19}$$

which leads to

$$\left(1 - \left(1 + (mC_0)^{-1}\right)^{-m}\right)^{-1} \geq \frac{\sum_{i=0}^{m} C_0^i}{\sum_{i=0}^{m-1} C_0^i}, \tag{20}$$

thereby concluding $\tau_{\text{f}}^* \geq \tau_{\text{t}}^*$. □

Corollary C.1.3 demonstrates that, if the total channel capacity in the source-relay hop is fixed, the time-division traffic allocation outperforms the frequency-division scheme in achieving the lower end-to-end latency.

$\mathcal{M}_{\text{global}}$ for general network

Based on Lemma C.2, for the multi-hop buffer-aided network illustrated in Fig. 1.1, the optimal traffic allocation at the source node and the resulting minimum end-to-end latency are given for $\mathcal{M}_{\text{global}}$ in the following theorem.

Theorem C.2. *For the tandem network in Fig. 1.1 with n relay nodes and m_h channels in the h^{th} hop, the minimum end-to-end latency with $\mathcal{M}_{\text{global}}$ is*

$$\tau_n^* = \frac{\left(C_{n,m_n}^{-1} + \tau_{n-1}^*\right) \prod\limits_{k=1}^{m_n-1} C_{n,k+1}\left(C_{n,k}^{-1} + \tau_{n-1}^*\right)}{1 + \sum\limits_{i=2}^{m_n} \prod\limits_{k=1}^{i-1} C_{n,k+1}\left(C_{n,k}^{-1} + \tau_{n-1}^*\right)}, \tag{21}$$

with initial condition $\tau_0^* \triangleq \left(\sum_{i=1}^{m_0} C_{0,i}\right)^{-1}$, achieved by

$$\alpha_{h,k} = \begin{cases} C_{h,k}\left(\sum\limits_{k=1}^{m_h} C_{h,k}\right)^{-1}, & h = 0 \\ \dfrac{\prod\limits_{i=1}^{k-1} C_{h,i+1}\left(C_{h,i}^{-1} + \tau_{h-1}^*\right)}{1 + \sum\limits_{j=2}^{m_h}\prod\limits_{i=1}^{j-1} C_{h,i+1}\left(C_{h,i}^{-1} + \tau_{h-1}^*\right)}, & h \geq 1. \end{cases} \tag{22}$$

Proof. For $h \in \{0\} \cup [n]$, we denote by τ_h^* the minimum end-to-end latency from hop 0 to hop h. In addition, we define the effective capacity as the reciprocal of minimum end-to-end latency, i.e., $\mathcal{E}_h \triangleq (\tau_h^*)^{-1}$. According to Lemma C.1, the initial effective capacity $\mathcal{E}_0$ is given as $\mathcal{E}_0 = \sum_{i=1}^{m_0} C_{0,i}$.

For the h^{th} hop with $h \geq 1$, we lump hops 0 to $h-1$ together, with effective capacity $\mathcal{E}_{h-1}$. By Lemma C.2, we know that at the relay node the local effective capacity depends on the its connected channels on two hops. More precisely, if the resulting effective capacity by lumping hops 0 to $h-1$ is updated to C_0 in Fig. 1.3, then the effective capacity at the $(h+1)^{\text{th}}$ node can be obtained by treating all channels in the h^{th} hop equal to those in the source-relay hop in Fig. 1.3. Thus, the effective capacity for the concatenated system with h hops is expressed as

$$\mathcal{E}_h = \frac{1 + \sum\limits_{i=2}^{m_h} \prod\limits_{k=1}^{i-1} C_{h,k+1}\left(C_{h,k}^{-1} + \mathcal{E}_{h-1}^{-1}\right)}{\left(C_{h,m_h}^{-1} + \mathcal{E}_{h-1}^{-1}\right) \prod\limits_{k=1}^{m_h-1} C_{h,k+1}\left(C_{h,k}^{-1} + \mathcal{E}_{h-1}^{-1}\right)}. \tag{23}$$

With the recursive expression of $\mathcal{E}_h$, we can finally obtain $\tau_n^* = \mathcal{E}_n^{-1}$ when h reaches n. □

We know from Theorem C.2 that in the middle phase of the transmission the file is distributed in all buffer-aided relay nodes rather than stacked in one buffer, and all relay nodes can simultaneously forward the buffered fractions to the subsequent nodes. The advantage of applying $\mathcal{M}_{\text{global}}$ is that a longer queue at a single relay node (i.e., a bottleneck) can be avoided. Meanwhile, concurrent transmissions at all relay nodes enable even utilizations among all buffers, thereby resulting in higher transmission efficiency.

In Theorem C.2, we find that the traffic allocation with $\mathcal{M}_{\text{global}}$ is found through a recursion. In contrast to the exhaustive method for searching the optimal solution, this recursive scheme significantly decreases the computational complexity. Intuitively, the traffic allocation at each node in the network is a function of the delay in the sub-network consisting of the subsequent nodes, while disregarding exact traffic allocations over the following hops. Hence, the resulting latency from the present node to the end can be treated as a whole and adopted for computing the latency from its previous node, thereby indicating the recursive property when applying $\mathcal{M}_{\text{global}}$.

C.3 Discussion of Computational Complexity

Note that the analysis in this paper is conducted for a given channel capacity. Assuming that the cost for acquiring the information of each channel, i.e., channel estimation, signaling, or beamforming, is fixed (or upper bounded), then the overall cost can be characterized by the number of channels in total for performing the optimized traffic allocation, i.e., the computational complexity.

In what follows, we determine the computational complexity for $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$. With respect to the multi-hop network in Fig. 1.1, we know that:

- Using $\mathcal{M}_{\text{local}}$, there are m_h channels considered for computing the traffic allocation at node $h+1$ for any $h \in \{0\} \cup [n]$ within each hop. Therefore, for the whole network, the overall computational complexity is in $\mathcal{O}\left(\sum_{h=0}^{n} m_h\right)$.
- Using $\mathcal{M}_{\text{global}}$, within each hop there are $\sum_{i=0}^{h} m_i$ channels considered for computing the traffic allocation with $\mathcal{M}_{\text{global}}$ at node $h+1$ for any $h \in \{0\} \cup [n]$. Thus, for the whole network, the overall computational complexity is in $\mathcal{O}\left(\sum_{h=0}^{n} \sum_{i=0}^{h} m_i\right)$.

Due to the fact that $\sum_{h=0}^{n} m_h \leq \sum_{h=0}^{n} \sum_{i=0}^{h} m_i$ (the equality holds only if $n = 0$), we have $\mathcal{O}\left(\sum_{h=0}^{n} m_h\right) \subset \mathcal{O}\left(\sum_{h=0}^{n} \sum_{i=0}^{h} m_i\right)$, which indicates that the overall computational complexity for $\mathcal{M}_{\text{local}}$ is lower. The comparison of the overall complexity to perform two traffic allocation schemes will be elaborated on in the next section.

D Performance Gain by Global Allocation

In this section, we study the performance gain by adopting $\mathcal{M}_{\text{global}}$, relative to that by adopting $\mathcal{M}_{\text{local}}$. For the sake of fairness and convenience, we here restrict ourselves to a homogeneous version of the multi-hop networks in Fig. 1.1. That is, for all $h \in \{0\} \cup [n]$ and $k \in [m]$, each hop has the same number of channels, i.e., $m_h = m$, and the capacity of all channels are identical, i.e., $C_{h,k} = C$.

D.1 Relative Performance Gain and Its Asymptote

Given n relay nodes and m channels per hop, taking $\mathcal{M}_{\text{global}}$ as the reference, we define the relative performance gain by $\mathcal{M}_{\text{global}}$ as

$$\rho\left(n,m\right) \triangleq \frac{\tau_n^*|_{\text{local}}}{\tau_n^*|_{\text{global}}}, \tag{24}$$

where $\tau_n^*|_{\text{global}}$ and $\tau_n^*|_{\text{local}}$ represent the minimum end-to-end latencies obtained in Theorem C.2 and Theorem C.1, respectively.

We next present a result on the relative performance gain $\rho\left(n,m\right)$.

Theorem D.1. *Given n relay nodes and m channels in each hop, the relative performance gain $\rho\left(n,m\right)$ is*

$$\rho\left(n,m\right) = \frac{n+1}{m}\left(u_n^*\right)^{-1}, \tag{25}$$

where u_k^* for all $k \in [n]$ is given as $u_k^* = \left(1-\left(1+u_{k-1}^*\right)^{-m}\right)^{-1} u_{k-1}^*$, with initial condition $u_0^* = m^{-1}$.

Proof. Please see Appendix H.3. □

Based on Theorem D.1, it is interesting to study the relative performance gain when the number of channels per hop increases. With any given number of relay nodes n, the asymptotic relative performance gain $\bar{\rho}\left(n\right)$ is defined as

$$\bar{\rho}\left(n\right) \triangleq \lim_{m\to\infty} \rho\left(n,m\right). \tag{26}$$

The following Theorem D.2 gives an expression for $\bar{\rho}\left(n\right)$.

Theorem D.2. *Given n relay nodes, the asymptotic performance gain $\bar{\rho}\left(n\right)$ for any $n \geq 0$ is recursively given as*

$$\bar{\rho}\left(n\right) = \frac{n+1}{n}\left(1-\exp\left(-\frac{n}{\bar{\rho}\left(n-1\right)}\right)\right)\bar{\rho}\left(n-1\right), \tag{27}$$

with initial condition $\bar{\rho}\left(0\right) = 1$.

Proof. Please see Appendix H.4. □

Theorem D.2 demonstrates that the relative performance gain approaches a constant that only depends on the number of relay nodes, when the number of channels in each hop goes to infinity. This limiting performance quantifies the maximum relative performance gain, achieved by letting the number of channels per hop grow to infinity.

D.2 Gain-Complexity Trade-off

Considering the overall computational complexity for the different traffic allocation schemes, for the homogeneous setting in this section, it is easy to obtain that the number of channels for traffic allocation with $\mathcal{M}_{\text{local}}$, denoted as $f_{\text{local}}(m,n)$, is given by

$$f_{\text{local}}(m,n) = m(n+1), \tag{28}$$

while that with $\mathcal{M}_{\text{global}}$, denoted as $f_{\text{global}}(m,n)$, is given by

$$f_{\text{global}}(m,n) = \frac{m(n+1)(n+2)}{2} = \frac{n+2}{2} \cdot f_{\text{local}}(m,n). \tag{29}$$

Thus, evidently, the relative overall computational complexity, which can be defined via comparing $\mathcal{M}_{\text{global}}$ to $\mathcal{M}_{\text{global}}$, grows with the number of relay nodes, linearly, i.e.,

$$\frac{f_{\text{global}}(m,n)}{f_{\text{local}}(m,n)} = \frac{n+2}{2} \in \mathcal{O}(n), \tag{30}$$

which only depends on the number of relay nodes. Jointly with Theorem D.2, we can see that $\mathcal{M}_{\text{global}}$ achieves a relative performance gain of $\bar{\rho}(n)$ at the expense of an n times higher computational complexity. Thus, there exists a trade-off between the relative performance gain and the overall computational complexity by global allocation, i.e., between $\bar{\rho}(n)$ and $\mathcal{O}(n)$.

E Average Latency for Two-Hop Linear mm-wave Networks with Nakagami-m Fading

In this section, we focus on the average end-to-end latency for a two-hop mm-wave network as shown in Fig. 1.3, with two independently fading channels between the source and the relay node. We consider small-scale fading for all channels in the two-hop network and assume independent block fading. That is, for each fraction and each hop, the channel is independent and identically distributed (i.i.d.) but constant during the delivery of the file fraction over each corresponding channel. We assume that the instantaneous channel state information (CSI) is known at the node that makes the resource allocation. Hence, the traffic allocation is performed after acquiring the CSI[2].

Observing the form of the end-to-end latency in Theorem C.1 or Theorem C.2, we notice that it is rather difficult to derive a closed-form expression, since the end-to-end latency is a reciprocal of the end-to-end effective capacity. For the sake

[2] In this paper we only consider the case of perfect CSI to study the best possible performance. If only the statistical CSI or no CSI is available, the performance by using $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ is inevitably degraded.

of tractability, we aim at lower bounds to characterize the average latency performance. In what follows, we first give the average capacity for mm-wave channels with Nakagami-m fading, and we subsequently derive the lower bounds on the end-to-end latency when using $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, respectively.

E.1 Average Channel Capacity

It has been reported in [RSM+13] that, unlike the channel characteristics in sub-6 GHz bands, the small-scale fading in mm-wave channels is not significant, due to the adoption of highly directional antennas and the weak capability of reflection/diffraction. For tractability, in this paper, we assume that the amplitude of mm-wave channel coefficient follows Nakagami-m fading, as in [BH15]. Hence, for a given signal-to-noise (SNR) ξ, the normalized capacity of a mm-wave channel with Nakagami-m fading can be written as $C = \log_2(1 + g \cdot \xi)$, where the random variable g represents the channel power gain, which follows the gamma distribution, i.e., $g \sim \Gamma\left(M, M^{-1}\right)$ with a positive Nakagami parameter M. The variance is M^{-1}, hence the randomness decreases with M, and the channel becomes deterministic as $M \to \infty$. $M = 1$ corresponds to Rayleigh fading.

We assume that mm-wave channels C_i, $i \in \{0, 1, 2\}$, have the identical Nakagami parameter[3] M. Then, with the aid of Meijer G-function, the average capacity for channels with Nakagami-m fading can be obtained as

$$\begin{aligned}\mathbb{E}[C_i] &= \int_0^\infty \log_2(1 + x\xi_i)\, f(x; M)\, dx \\ &= \frac{M^M}{\xi_i^M \Gamma(M) \ln(2)} \cdot G_{2,3}^{3,1}\left(\begin{matrix} -M, 1-M \\ 0, -M, -M \end{matrix} \middle| \frac{M}{\xi_i} \right),\end{aligned} \tag{31}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator, and ξ_i is the SNR on C_i. Here, $G_{p,q}^{m,n}\left(\begin{smallmatrix} a_1, a_2, \ldots, a_p \\ b_1, b_2, \ldots, b_q \end{smallmatrix} \middle| z \right)$ denotes the Meijer G-function [Olv10], where $0 \le m \le q$ and $0 \le n \le p$, and parameters a_j, b_j and $z \in \mathbb{C}$.

E.2 Lower Bounds on Average Latency

In this subsection, for different traffic allocation schemes, we derive lower bounds on the average end-to-end latency, in terms of $\mathbb{E}[C_i]$ for $i \in \{0, 1, 2\}$ (as in the previous subsection).

Using $\mathcal{M}_{\text{local}}$

Given C_0, C_1 and C_2, according to Theorem C.1 it is easy to obtain that the minimum end-to-end latency with $\mathcal{M}_{\text{local}}$ is

$$\tau_2^* = C_0^{-1} + (C_1 + C_2)^{-1}. \tag{32}$$

[3] Normally, the randomness in mm-wave channels is relatively weak, such that the Nakagami parameter $M \ge 3$ as in [BH15, YZHL17].

Then, a lower bound on the expectation of τ_2^* is presented in the following proposition.

Proposition E.1. *A lower bound on the minimum average end-to-end latency for $\mathcal{M}_{\text{local}}$ is*

$$\mathbb{E}\left[\tau_2^*\right] \geq \left(\mathbb{E}\left[C_0\right]\right)^{-1} + \left(\mathbb{E}\left[C_1 + C_2\right]\right)^{-1}. \tag{33}$$

Proof. Based on τ_2^* in (32), the minimum average end-to-end latency can be expressed as $\mathbb{E}\left[\tau_2^*\right] = \mathbb{E}\left[C_0^{-1}\right] + \mathbb{E}\left[\left(C_1 + C_2\right)^{-1}\right]$. Regarding the existence of $\mathbb{E}\left[C_0^{-1}\right]$ for $M > 1$, we note that

$$\begin{aligned}\mathbb{E}\left[C_0^{-1}\right] \leq& \ln\left(2\right)\left(\frac{1}{2} + \xi_0^{-1}\mathbb{E}\left[g^{-1}\right]\right)\\ =& \ln\left(2\right)\left(\frac{1}{2} + \xi_0^{-1}\left(1 + \left(M-1\right)^{-1}\right)\right) < \infty,\end{aligned} \tag{34}$$

where the first line applies the following inequality for $x \geq 0$: $\ln\left(1+x\right) \geq \frac{2x}{2+x}$. Therefore, the existence $\mathbb{E}\left[C_0^{-1}\right]$ is guaranteed.

Finally, by applying Jensen's inequality, i.e., $\mathbb{E}\left[X\right] \cdot \mathbb{E}\left[X^{-1}\right] \geq 1$ for positive random variables X, we can obtain that $\mathbb{E}\left[\tau_2^*\right] \geq \left(\mathbb{E}\left[C_0\right]\right)^{-1} + \left(\mathbb{E}\left[C_1 + C_2\right]\right)^{-1}$, which completes the proof. □

Using $\mathcal{M}_{\text{global}}$

According to Theorem C.2, given C_0, C_1 and C_2, the minimum end-to-end latency with $\mathcal{M}_{\text{global}}$ is

$$\tau_2^* = \frac{\left(C_0^{-1} + C_1^{-1}\right)\left(C_0^{-1} + C_2^{-1}\right)}{C_0^{-1} + C_1^{-1} + C_2^{-1}}. \tag{35}$$

Then, a lower bound on the expectation of τ_2^* is given in the following proposition.

Proposition E.2. *A lower bound on the minimum end-to-end latency for $\mathcal{M}_{\text{global}}$ is*

$$\mathbb{E}\left[\tau_2^*\right] \geq \left(\mathbb{E}\left[C_0\right]\right)^{-1} + \left(\mathbb{E}\left[C_1 + C_2\right] + \epsilon_0 \mathbb{E}\left[C_1 C_2\right]\right)^{-1}, \tag{36}$$

where ϵ_0 for $M > 1$ is defined as

$$\epsilon_0 \triangleq \ln\left(2\right)\left(\frac{1}{2} + \xi_0^{-1}\left(1 + \left(M-1\right)^{-1}\right)\right). \tag{37}$$

Proof. With τ_2^* given in (35), we know that

$$\begin{aligned}\mathbb{E}\left[\tau_2^*\right] =& \mathbb{E}\left[C_0^{-1}\right] + \mathbb{E}\left[\left(C_1 + C_2 + C_0^{-1}C_1C_2\right)^{-1}\right]\\ \geq& \left(\mathbb{E}\left[C_0\right]\right)^{-1} + \mathbb{E}^{-1}\left[C_1 + C_2 + C_0^{-1}C_1C_2\right],\end{aligned} \tag{38}$$

where the second line is achieved by Jensen's inequality. We notice that $\mathbb{E}\left[C_0^{-1}\right]$ is upper bounded as $\mathbb{E}\left[C_0^{-1}\right] \leq \ln(2)\left(\frac{1}{2} + \xi_0^{-1}\left(1 + (M-1)^{-1}\right)\right) \triangleq \epsilon_0$. □

For $M = 1$, $\mathbb{E}\left[C_0^{-1}\right]$ does not exist since $\int_0^\infty \left(\log_2(1+\xi x)\right)^{-1} \exp(-x)\,dx$ does not converge. Hence, the bounding techniques in Proposition E.1 and Proposition E.2 are not applicable to the scenarios with Rayleigh fading channels ($M = 1$). Fortunately, this case is less relevant for mm-wave communications.

It is worth mentioning that the method for analyses above can be extended to mm-wave networks with more relay nodes and more channels in each hop: The lower bound $\mathcal{M}_{\text{local}}$ can be obtained by performing Jensen's inequality on the component latency in each hop, while the lower bound with respect to $\mathcal{M}_{\text{global}}$ can be obtained by following the recursive expression for end-to-end latency presented in Theorem C.2.

F Performance Evaluation

In this section, we will evaluate the end-to-end latency for $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$. The performance evaluation consists of the following two parts:

(i) We focus on the allocation schemes $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ and investigate their corresponding performance. With simulations, we first validate allocation schemes developed in Theorem C.1 and Theorem C.2 for the minimum end-to-end latency. After validating the allocation schemes, we subsequently provide numerical results focusing on Theorem D.1 and Theorem D.2 and assess the performance achieved by two distinct schemes.

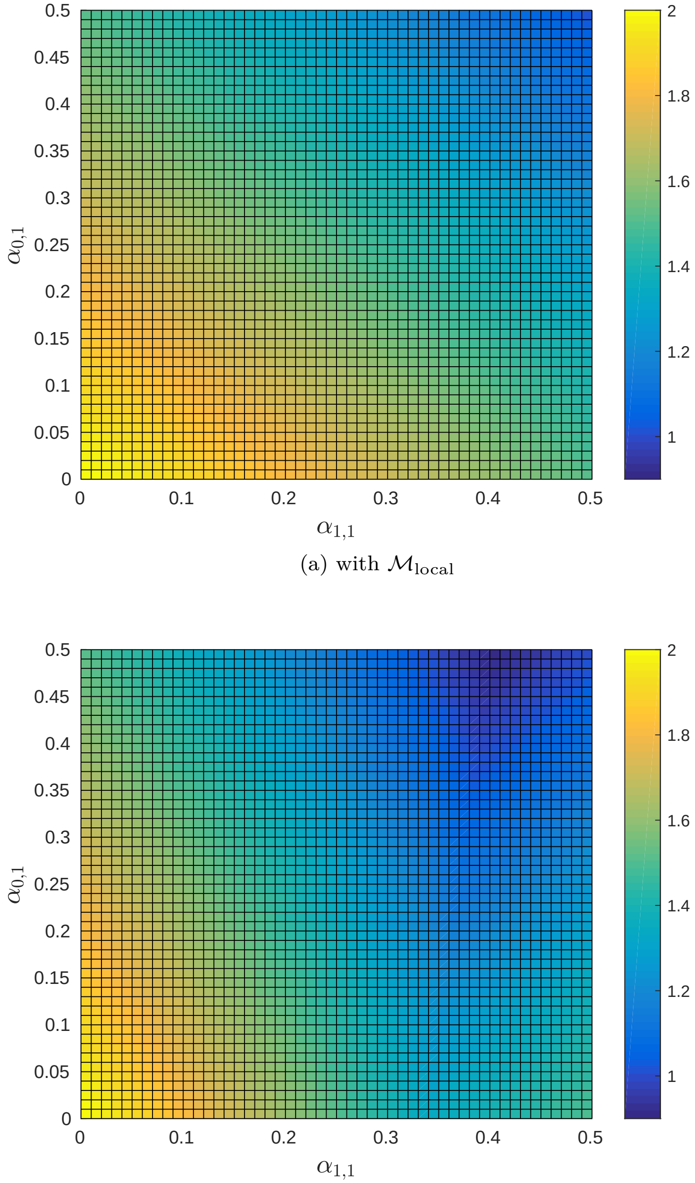
(ii) Following the two-hop network adopted in Sec. E, we evaluate the average end-to-end latency performance in the presence of Nakagami-m fading in mm-wave channels. We first show the tightness of the lower bounds derived in Proposition E.1 and Proposition E.2. Further discussions related to the average performance are presented afterwards.

We assume that the size of the transmitted file is normalized to 1 without loss of generality. Other system settings for the above two assessments will be elaborated on.

F.1 Performance of the Two Allocation Schemes

Focusing on the performance of two traffic allocation schemes, we assume deterministic channels, such that the channel capacity is treated as constants. Furthermore, for fairness and simplicity, we follow the homogeneous setting used in Sec. D, i.e., $m_h = m$ and $C_{h,k} = C$ for all $h \in \{0\} \cup [n]$ and $k \in [m]$.

We simulate the end-to-end latency of a two-hop system with two channels per hop, i.e., $n = 1$ and $m = 2$, and the performance for $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ is shown



(a) with $\mathcal{M}_{\text{local}}$

(b) with $\mathcal{M}_{\text{global}}$

Figure 1.4: End-to-end latency performance with $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, where the number of relay nodes is $n = 1$, the number of channels per hop is $m = 2$, the capacity of each channel is $C = 1$, and traffic allocation $\alpha_{0,1}$ and $\alpha_{1,1}$ both vary from 0 to 0.5.

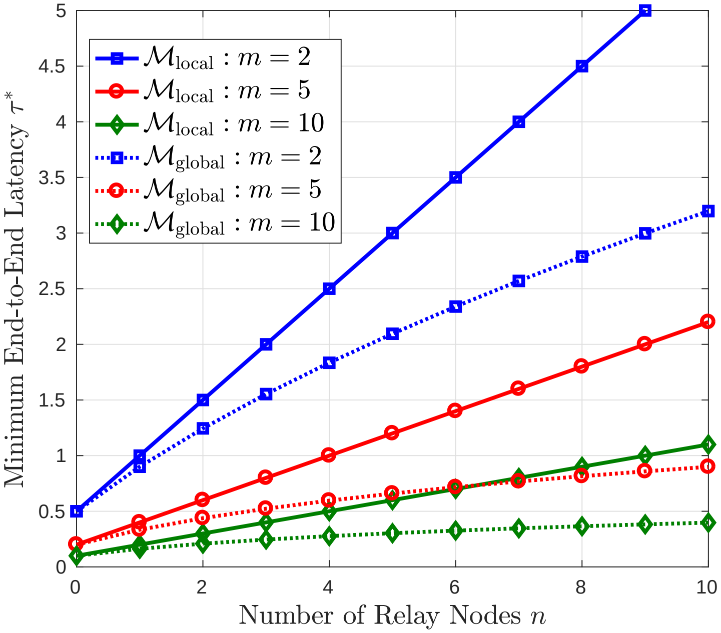


Figure 1.5: Minimum end-to-end latency τ^* vs. number of relay nodes n, where the number of channels per hop is $m = 2$, 5 or 10, the capacity of each channel is $C = 1$, and the size of transmitted file is 1.

in Fig. 1.4. For traffic allocations $\underline{\alpha}_0$ (at the relay node) and $\underline{\alpha}_1$ (at the source), we consider variables $\alpha_{0,1} \in [0, 0.5]$ and $\alpha_{1,1} \in [0, 0.5]$, and the remaining allocations can be characterized in terms of $\alpha_{0,1}$ and $\alpha_{1,1}$, i.e., $\alpha_{0,2} = 1-\alpha_{0,1}$ and $\alpha_{1,2} = 1-\alpha_{1,1}$, respectively, due to the fact $m = 2$. In both Fig. a and Fig. b, we vary $\alpha_{0,1}$ and $\alpha_{1,1}$, jointly. In general, it is evident that the resulting latencies by distinct allocation schemes are different. We can see in Fig. a that the minimum end-to-end latency is 1 when applying $\mathcal{M}_{\text{local}}$, which is achieved at $\alpha_{0,1} = 0.5$ and $\alpha_{1,1} = 0.5$. However, in Fig. b, the minimum end-to-end latency is 0.9 when applying $\mathcal{M}_{\text{global}}$, which is achieved at $\alpha_{0,1} = 0.5$ and $\alpha_{1,1} = 0.4$. The optimal traffic allocations enabling the minimum end-to-end latency observed from Fig. 1.4 are in accordance with our analytical results derived in Theorem C.1 and Theorem C.2.

In Fig. 1.5, we investigate the minimum end-to-end latency τ^* against the number of relay nodes n, where different numbers of per-hop channels m are considered. For both $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, we find that τ^* can be significantly reduced when elevating $m = 2$ to $m = 10$. This coincides with the intuition that increasing the number of channels is equivalent to producing a larger effective channel capacity in each hop, which in turn leads to a lower latency for file delivery. Besides, we

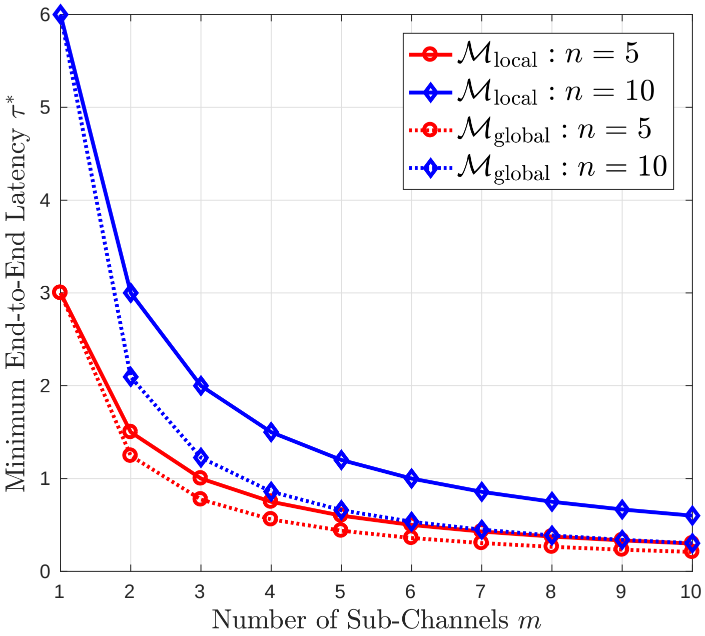


Figure 1.6: Minimum end-to-end latency τ^* vs. number of channels m, where the number of relay nodes is $n = 5$ or 10, the capacity of each channel is $C = 1$, and the size of transmitted file is 1.

notice that the benefit by adopting $\mathcal{M}_{\text{global}}$ becomes remarkable as n increases. For instance, for $m = 2$, compared to $\mathcal{M}_{\text{local}}$, applying $\mathcal{M}_{\text{global}}$ reduces the latency by 25% at $n = 3$, while the reduction is enlarged to 40% at $n = 9$. The performance improvement shown above stems from the efficient utilization of buffers at relay nodes in $\mathcal{M}_{\text{global}}$, since long queues are avoided by performing the optimal traffic allocations globally at all relay nodes. This observation reveals the great advantage of $\mathcal{M}_{\text{global}}$ in networks with more relay nodes. Furthermore, we can see that there is an intersection at $n = 6$, between the curve with $m = 5$ for $\mathcal{M}_{\text{local}}$ and the curve with $m = 10$ for $\mathcal{M}_{\text{global}}$. This finding indicates that $\mathcal{M}_{\text{global}}$ with fewer channels is still competitive in outperforming $\mathcal{M}_{\text{local}}$ with more channels as long as there are sufficiently many relay nodes, which again highlights the benefit of global allocation.

Given n relay nodes, the minimum end-to-end latency τ^* against the number of channels m is illustrated in Fig. 1.6. For both $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$, τ^* is dramatically reduced at the beginning of increasing m, while the decaying rates slow down when m becomes large, i.e., when $m \geq 5$. This finding indicates that it is definitely beneficial to have multiple channels for reducing the end-to-end latency, but

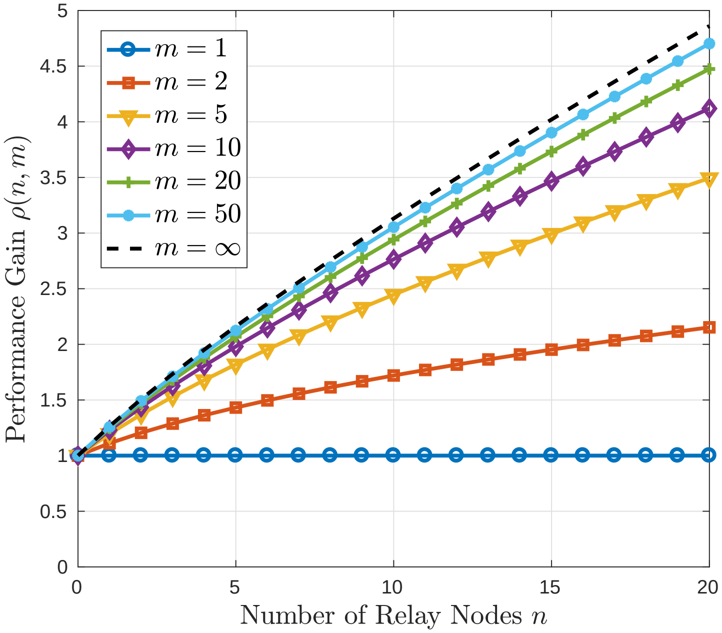


Figure 1.7: Relative performance gain $\rho\,(n, m)$ vs. number of relay nodes n, where the number of channels per hop is $m = 1$, 2, 5, 10, 20, 50 or ∞.

the benefit diminishes as the number of channels increases. Therefore, in practice, considering the cost of system implementations, it is not necessary to increase the number of channels above about 8.

In Fig. 1.7, we investigate the relative performance gain $\rho\,(n, m)$ with respect to the number of relay nodes n, where the number of per-hop channels m varies from $m = 1$ to ∞. We find that there is no difference between $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ when $m = 1$, i.e., $\rho\,(n, 1) = 1$ for all n, since neither the local allocation nor the global allocation is actually performed if there is only one single channel per hop. However, $\rho\,(n, m)$ obviously increases when m or n grows. This indicates the substantial advantage of $\mathcal{M}_{\text{global}}$ compared to $\mathcal{M}_{\text{local}}$, especially when the number of channels per hop or the number of relay nodes is large. In addition, when $m \to \infty$, asymptotic performance gain $\bar{\rho}\,(n)$ characterizes the upper bound of the relative benefits. For instance, at $n = 20$, the latency can be reduced to 20% of $\mathcal{M}_{\text{local}}$ at the most, when applying $\mathcal{M}_{\text{global}}$. The asymptote in Fig. 1.7 is obtained from Theorem D.2.

F.2 Average Latency in Millimeter-wave Networks

To investigate the average end-to-end latency of the two-hop mm-wave network in Sec. E (see Fig. 1.3), we consider a network where the relay node is deployed on the line between the source and the destination (All networks considered here are linear in the sense of the network topology.). Starting from the source, we denote by r and L the (normalized) distances to the relay node and to the destination, respectively. Besides, we assume that the transmit power at the source and the relay node are both γ, and the power of the background noise is set to 1 without loss of generality. Applying the path loss model for line-of-sight (LOS) mm-wave communications [RSM+13, XMH+17], the SNR ξ_i in $C_i = \log_2(1+g_i\xi_i)$ can be written as

$$\xi_i = \begin{cases} \gamma\left(L-r\right)^{-\alpha}, & i=0 \\ \gamma r^{-\alpha}, & i\in\{1,2\}, \end{cases} \tag{39}$$

where α denotes the path loss exponent. The gamma-distributed random variables g_i for $i \in \{0,1,2\}$ are independent and identically distributed with Nakagami parameter M.

With Nakagami fading in mm-wave channels, the average end-to-end latency and the lower bounds are illustrated in Fig. 1.8. From the simulation results for the two traffic allocation schemes, it can be seen that the lower bounds given in Proposition E.1 and Proposition E.2 are quite tight. Furthermore, the average latency first decreases to the minimum and subsequently increases, when r grows from 40 to 160. This observation indicates the critical role of relay deployment in minimizing the end-to-end latency. We can also see that the minimum average latency for $\mathcal{M}_{\text{local}}$ is obtained roughly at $r = 120$, while the minimum average latency for $\mathcal{M}_{\text{global}}$ is obtained roughly at $r = 115$. This slight difference tells that different relay deployments may be needed for different allocation schemes.

In Fig. 1.9 we show the average end-to-end latency $\mathbb{E}\left[\tau_2^*\right]$ against varying Nakagami parameter M. As aforementioned, a larger M corresponds to a more deterministic channel. We see from Fig. 1.9 that for both $\mathcal{M}_{\text{local}}$ and $\mathcal{M}_{\text{global}}$ the average end-to-end latency $\mathbb{E}\left[\tau_2^*\right]$ decreases as M grows from 3 to 15, while the reduction in latency by increasing M gradually diminishes. Moreover, the simulation results gradually approach the corresponding lower bounds when M increases. This is due to the fact that Jensen's inequality in Proposition E.1 or Proposition E.2 gives a tighter lower bound $\left(\mathbb{E}\left[X\right]\right)^{-1}$ for $\mathbb{E}\left[X^{-1}\right]$ when mm-wave channels become more deterministic (higher M). In addition, we find that the tightness of the lower bounds for both allocation schemes are improved when γ increases from 55 dB to 65 dB. Thus, the lower bounds get even tighter when the transmit powers are high.

G Conclusions

We have studied the end-to-end latency in multi-hop mm-wave networks by applying two traffic allocation schemes, namely local allocation and global allocation. In

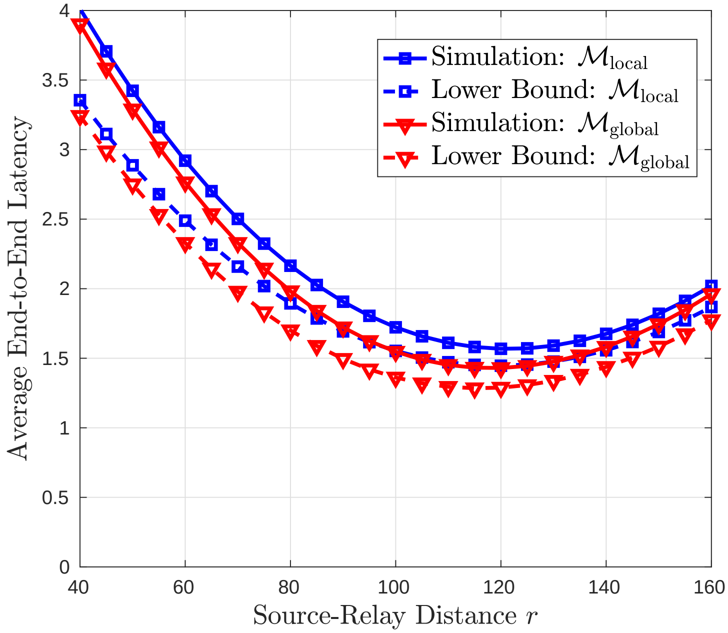


Figure 1.8: Average end-to-end latency $\mathbb{E}[\tau_2^*]$ and the lower bounds, against source-relay distance r, where (normalized) source-destination distance $L = 200$, transmit power $\gamma = 60$ dB, path loss exponent $\alpha = 3$, and Nakagami parameter $M = 5$.

our networks, buffers are equipped at the source node and the relay nodes, and multiple independent channels exist in each hop. For given channel capacities, we have provided closed-form expressions of the end-to-end latency for the two allocation schemes and quantified the advantages of the global allocation scheme relative to the local one. Some asymptotic analyses have also been performed. Compared to local allocation, the advantage of global allocation grows as the number of relay nodes n increases, at the expense of an n times higher computational complexity. Besides, increasing the number of channels monotonically decreases the latency, and it asymptotically reaches a constant that depends only on the number of relay nodes. Furthermore, taking a specific two-hop linear mm-wave network as an example, we have derived tight lower bounds on the average end-to-end latency for two traffic allocation schemes with Nakagami-m fading incorporated. We have also noticed the great importance of proper deployment of the relay node. These results can provide insights for designing or implementing low-latency multi-hop mm-wave networks.

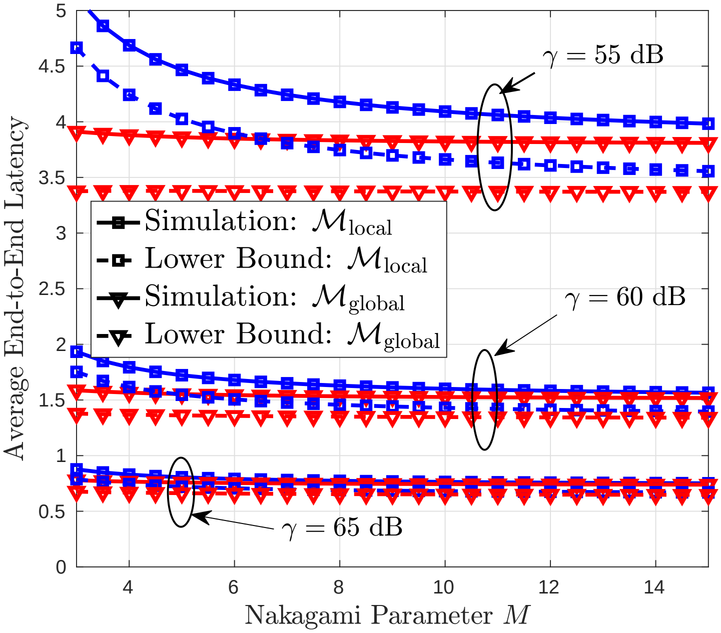


Figure 1.9: Average end-to-end latency $\mathbb{E}[\tau_2^*]$ and the lower bounds against Nakagami parameter M, where (normalized) source-destination distance $L = 200$, (normalized) source-relay distance $r = 100$, transmit power $\gamma = 55$ dB, 60 dB or 65 dB, and path loss exponent $\alpha = 3$.

H Appendices

H.1 Proof of Lemma C.2

Due to the buffer at the relay node, fractions from distinct channels are first stacked in the queue and subsequently pushed on the channel connecting the destination. To simplify the notation, we assume that fraction α_j arrives at the buffer-aided relay node prior to fraction α_k if $j \leq k$, for all $j, k \in [m]$, without loss of generality. Letting w_i for all $i \in [m]$ denote the latency of fraction α_i traversing from the source to the destination, according to the arrival orders at the buffer-aided relay node, we have $w_1 = \alpha_1 \left(C_1^{-1} + C_0^{-1}\right)$ and $w_2 = \max\left\{w_1, \alpha_2 C_2^{-1}\right\} + \alpha_2 C_0^{-1}$. In light of above, we can express the component delay w_i for $i \in [m]$ in general as $w_i = \max\left\{w_{i-1}, \alpha_i C_i^{-1}\right\} + \alpha_i C_0^{-1}$ with the initial condition $w_0 = 0$.

It is evident that $w_j < w_k$ for any $j < k$, since $w_k \geq w_{k-1} + \alpha_k C_0^{-1} > w_{k-1} > \ldots > w_j$. Thus, latency τ is reduced to w_m, i.e., $\max_{1 \leq i \leq m}\{w_i\} = w_m$, and the minimum latency τ^* can be expressed as $\tau^* = \min_{\|\alpha\|_1 = 1} w_m$. Applying the recursion for w_m, we equivalently have

$$\begin{aligned}\tau^* &= \min_{\|\alpha\|_1=1} \max\left\{w_{m-1}, \alpha_m C_m^{-1}\right\} + \alpha_m C_0^{-1} \\ &= \min_{\alpha_m \in (0,1)} \min_{\sum_{i=1}^{m-1} \alpha_i = 1-\alpha_m} \max\left\{w_{m-1}, \alpha_m C_m^{-1}\right\} + \alpha_m C_0^{-1} \\ &= \min_{\alpha_m \in (0,1)} \max\left\{ \min_{\sum_{i=1}^{m-1} \alpha_i = 1-\alpha_m} \left\{w_{m-1}\right\}, \alpha_m C_m^{-1} \right\} + \alpha_m C_0^{-1},\end{aligned} \tag{40}$$

where the last line is obtained based on the fact that w_{m-1} depends on $\{\alpha_i\}$ for all $i \in [m-1]$, while $\alpha_m C_0^{-1}$ and $\alpha_m C_m^{-1}$ can be treated as constants with respect to a given α_m.

For notational simplicity, we define $\xi^* \triangleq \min_{\sum_{i=1}^{m-1} \alpha_i = 1} w_{m-1}$, which denotes the optimized latency of delivering one normalized-size file in the network with $m-1$ channels between the source and the relay node. Thanks to the linear mapping between the allocated traffic loads and the resulting latency, with respect to any $z > 0$, we can easily obtain that

$$\min_{\sum_{i=1}^{m-1} \alpha_i = z} w_{m-1} = \min_{\sum_{i=1}^{m-1} \alpha_i = 1 \cdot z} w_{m-1} = z \cdot \xi^*. \tag{41}$$

Therefore, τ^* can be further reduced to

$$\begin{aligned}\tau^* &= \min_{\alpha_m \in (0,1)} \max\left\{(1-\alpha_m)\,\xi^*, \alpha_m C_m^{-1}\right\} + \alpha_m C_0^{-1} \\ &= \min_{\alpha_m \in (0,1)} \max\left\{\xi^* - \alpha_m\left(\xi^* - C_0^{-1}\right), \alpha_m\left(C_m^{-1} + C_0^{-1}\right)\right\}.\end{aligned} \tag{42}$$

Note that ξ^* denotes the latency for delivering one normalized-size file from the source to the destination via the relay node, while C_0^{-1} denotes the latency for delivering the file from the relay node to the destination. The former is strictly greater than the latter, i.e., $\xi^* > C_0^{-1}$. Then, we can see that $\xi^* - \alpha_m\left(\xi^* - C_0^{-1}\right)$ is monotonically increasing with α_m, while $\alpha_m\left(C_m^{-1} + C_0^{-1}\right)$ is monotonically decreasing with α_m. Hence, τ^* is obtained whenever $\xi^* - \alpha_m\left(\xi^* - C_0^{-1}\right) = \alpha_m\left(C_m^{-1} + C_0^{-1}\right)$.

One particular solution that meets the condition shown above for minimizing the latency is to $w_m = \alpha_m C_m^{-1}$. Iteratively, $w_{i-1} = \alpha_i C_i^{-1}$ should hold for all $i \in [m]$. In this case, we can obtain that $\alpha_2 C_2^{-1} = \alpha_1\left(C_1^{-1} + C_0^{-1}\right)$, and the general expression for $2 \le i \le m$ is

$$\alpha_i = \alpha_1 \prod_{k=1}^{i-1} C_{k+1}\left(C_k^{-1} + C_0^{-1}\right). \tag{43}$$

Paired with the constraint $\|\alpha\|_1 = 1$, we can immediately solve α_1 as

$$\alpha_1 = \left(1 + \sum_{i=2}^{m} \prod_{k=1}^{i-1} C_{k+1}\left(C_k^{-1} + C_0^{-1}\right)\right)^{-1}. \tag{44}$$

Thus, applying the recursive expression of α_i, we have

$$\alpha_m = \frac{\prod_{k=1}^{m-1} C_{k+1}\left(C_k^{-1}+C_0^{-1}\right)}{1+\sum_{i=2}^{m}\prod_{k=1}^{i-1} C_{k+1}\left(C_k^{-1}+C_0^{-1}\right)}, \tag{45}$$

which further gives the minimum latency τ^* as

$$\tau^* = \frac{\left(C_m^{-1}+C_0^{-1}\right)\prod_{k=1}^{m-1} C_{k+1}\left(C_k^{-1}+C_0^{-1}\right)}{1+\sum_{i=2}^{m}\prod_{k=1}^{i-1} C_{k+1}\left(C_k^{-1}+C_0^{-1}\right)}. \tag{46}$$

H.2 Proof of Corollary C.1.1

We define the multinominal $\mathcal{C}(i,m)$ as

$$\mathcal{C}(i,m) \triangleq \sum_{k_1,\ldots,k_i\in[m]}^{\neq} \prod C_{k_j}, \tag{47}$$

where the $\neq$ indicates that k_u and k_v are not equal for any $u,v\in[m]$. Rewriting the expression of τ^* in Lemma C.2, we can obtain that

$$\tau^* = \frac{\sum\limits_{i=0}^{m} C_0^{m-i}\cdot\mathcal{C}(i,m)}{C_0\sum\limits_{i=1}^{m} C_0^{m-i}\cdot\mathcal{C}(i,m)} = \frac{C_0^m+\sum\limits_{i=1}^{m} C_0^{m-i}\cdot\mathcal{C}(i,m)}{C_0\sum\limits_{i=1}^{m} C_0^{m-i}\cdot\mathcal{C}(i,m)} = C_0^{-1}+\left(\sum_{i=1}^{m} C_0^{1-i}\cdot\mathcal{C}(i,m)\right)^{-1}. \tag{48}$$

Subsequently, the minimization of τ^* with respect to the constraint $\|\underline{C}\|_1 \triangleq \sum_{i=1}^{m} C_i = 1$ can be reformulated as

$$\min_{\|\underline{C}\|_1=1}\tau^* = C_0^{-1}+\left(\sum_{i=1}^{m} C_0^{1-i}\max_{\|\underline{C}\|_1=1}\left\{\mathcal{C}(i,m)\right\}\right)^{-1}. \tag{49}$$

Paired with the symmetry of the multinominal $\mathcal{C}(i,m)$, we apply the Lagrange multiplier optimization and obtain that

$$\mathcal{C}(i,m) \le \binom{m}{i} m^{-i}, \tag{50}$$

where the equality is achieved when having $C_i = m^{-1}$ for all $i\in[m]$. Then, we have

$$\min_{\|\underline{C}\|_1=1}\tau^* = \left(1-\left(1+(mC_0)^{-1}\right)^{-m}\right)^{-1}. \tag{51}$$

H.3 Proof of Theorem D.1

With $\mathcal{M}_{\text{local}}$, from Theorem C.1, we can easily obtain that $\tau_n^* = (n+1) \cdot (mC)^{-1}$. Applying a change of variables, i.e., $v_n^* = C\tau_n^*$, we can equivalently write the latency above as $v_n^* = m^{-1}(n+1)$, which is treated as the normalized latency with $\mathcal{M}_{\text{local}}$.

With $\mathcal{M}_{\text{global}}$, from Theorem C.2, we obtain that

$$\begin{aligned}\tau_k^* =& \frac{\left(C^{-1} + \tau_{k-1}^*\right)\left(C\left(C^{-1} + \tau_{k-1}^*\right)\right)^{m-1}}{1 + \sum\limits_{i=2}^{m}\left(C\left(C^{-1} + \tau_{k-1}^*\right)\right)^{i-1}} \\ =& \frac{\left(C\left(C^{-1} + \tau_{k-1}^*\right)\right)^{m}}{C\sum\limits_{i=0}^{m-1}\left(C\left(C^{-1} + \tau_{k-1}^*\right)\right)^{i}} = \frac{\tau_{k-1}^*}{1 - \left(1 + C\tau_{k-1}^*\right)^{-m}},\end{aligned} \tag{52}$$

associated with the initial condition $\tau_0^* = m^{-1}C^{-1}$. Applying a change of variables, i.e., $u_k^* = C\tau_k^*$ for all $k \in [n]$, we equivalently have the recursive expression as

$$u_k^* = \left(1 - \left(1 + u_{k-1}^*\right)^{-m}\right)^{-1} u_{k-1}^* \tag{53}$$

with $u_0^* = m^{-1}$, which is treated as the normalized latency with $\mathcal{M}_{\text{global}}$, likewise.

According to the definition of $\rho(n, m)$, we can obtain that

$$\rho(n,m) = \frac{v_n^*}{u_n^*} = \frac{n+1}{m}\left(u_n^*\right)^{-1}. \tag{54}$$

H.4 Proof of Theorem D.2

For all $k \in \{0\} \cup [n]$, following the expression of u_k^* used in Theorem D.1, we define

$$z_k \triangleq \frac{\bar{\rho}(k)}{k+1} = \lim_{m\to\infty}\left(mu_k^*\right)^{-1}. \tag{55}$$

When $k \leq 1$, we can easily obtain that $z_0 = 1$ and

$$z_1 = \lim_{m\to\infty}\left(m\left(1 - \left(1 + m^{-1}\right)^{-m}\right)^{-1} m^{-1}\right)^{-1} = 1 - e^{-1} \tag{56}$$

For any given $k \geq 2$, since it is known that z_{k-1} is finite and $\lim\limits_{m\to\infty} m^{-1}z_{k-1} = 0$, we have

$$\begin{aligned} z_k =& \lim_{m\to\infty} m^{-1}\left(1 - \left(1 + u_{k-1}^*\right)^{-m}\right)\left(u_{k-1}^*\right)^{-1} \\ =& \lim_{m\to\infty}\left(1 - \left(1 + \lim_{m\to\infty}\frac{mu_{k-1}^*}{m}\right)^{-m}\right)\lim_{m\to\infty}\left(mu_{k-1}^*\right)^{-1} \\ =& \lim_{m\to\infty}\left(1 - \left(1 + \frac{z_{k-1}^{-1}}{m}\right)^{-m}\right) z_{k-1} = \left(1 - e^{-z_{k-1}^{-1}}\right) z_{k-1}.\end{aligned} \tag{57}$$

Thus, we can recursively obtain z_n as $z_n = \left(1 - e^{-z_{n-1}^{-1}}\right) z_{n-1}$, associated with initial condition $z_0 = 1$. Recovering $\bar{\rho}\left(k\right)$ in terms of z_k, we obtain that

$$\bar{\rho}\left(n\right) = \frac{n+1}{n}\left(1 - \exp\left(-\frac{n}{\bar{\rho}\left(n-1\right)}\right)\right)\bar{\rho}\left(n-1\right), \tag{58}$$

with initial condition $\bar{\rho}\left(0\right) = 1$.

Bibliography

[3GP17] 3GPP. Study on scenarios and requirements for next generation access technologies. Tech. Report 3GPP TR 38.913 V14.3.0, European Telecommunications Standards Institute (ETSI), Oct. 2017. `http://www.etsi.org/deliver/etsi_tr/138900_138999/138913/14.03.00_60/tr_138913v140300p.pdf`.

[ABC$^+$14] Jeffrey G Andrews, Stefano Buzzi, Wan Choi, Stephen V Hanly, Angel Lozano, Anthony CK Soong, and Jianzhong Charlie Zhang. What will 5G be? *IEEE Journal on selected areas in communications*, 32(6):1065–1082, 2014.

[AEAH12] Salam Akoum, Omar El Ayach, and Robert W Heath. Coverage and capacity in mmWave cellular systems. In *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 688–692. IEEE, 2012.

[AEALH14] Ahmed Alkhateeb, Omar El Ayach, Geert Leus, and Robert W Heath. Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE Journal of Selected Topics in Signal Processing*, 8(5):831–846, 2014.

[AH16] Hatem Abbas and Khairi Hamdi. Full duplex relay in millimeter wave backhaul links. In *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2016.

[AJS98] Bruce Anderson, Jeffrey Jackson, and Meera Sitharam. Descartes' rule of signs revisited. *American Mathematical Monthly*, pages 447–451, 1998.

[ALS$^+$14] Mustafa Riza Akdeniz, Yuanpeng Liu, Mathew K Samimi, Shu Sun, Sundeep Rangan, Theodore S Rappaport, and Elza Erkip. Millimeter Wave Channel Modeling and Cellular Capacity Evaluation. *IEEE Journal on Selected Areas in Communications*, 32(6):1164–1179, Jun. 2014.

[AS00] Mohamed-Slim Alouini and Marvin K Simon. An MGF-based performance analysis of generalized selection combining over Rayleigh fading channels. *IEEE Transactions on Communications*, 48(3):401–415, 2000.

[ASP$^+$09] Xueli An, Chin-Sean Sum, R.V. Prasad, Junyi Wang, Zhou Lan, Jing Wang, R. Hekmat, H. Harada, and I. Niemegeers. Beam switching support to resolve link-blockage problem in 60 GHz WPANs. In *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 390–394, Sept. 2009.

[AZLB13] Hussein Al-Zubaidy, Jörg Liebeherr, and Almut Burchard. A $(\min,\times)$ network calculus for multi-hop fading channels. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, pages 1833–1841. IEEE, 2013.

[AZLB16] Hussein Al-Zubaidy, Jörg Liebeherr, and Almut Burchard. Network-layer performance analysis of multihop fading channels. *IEEE/ACM Transactions on Networking*, 24(1):204–217, Feb. 2016.

[Azz86] RMA Azzam. Relationship between the p and s Fresnel reflection coefficients of an interface independent of angle of incidence. *JOSA A*, 3(7):928–929, 1986.

[BA08] Hacène Belbachir and ALGERIA ALGER. A multinomial extension of an inequality of Haber. *J. Ineq. Pure Appl. Math*, 9(4), 2008.

[BA09] Nabhendra Bisnik and Alhussein A Abouzeid. Queuing network models for delay analysis of multihop wireless ad hoc networks. *Ad Hoc Networks*, 7(1):79–97, 2009.

[BB06] Hacene Belbachir and Farid Bencherif. Linear recurrent sequences and powers of a square matrix. *Integers*, 6:A12, 2006.

[BB$^+$10] François Baccelli, Bartłomiej Błaszczyszyn, et al. Stochastic geometry and wireless networks: Volume II Applications. *Foundations and Trends® in Networking*, 4(1–2):1–312, 2010.

[BDF63] D.E. Barton, F.N. David, and E. Fix. Random points in a circle and the analysis of chromosome patterns. *Biometrika*, pages 23–29, 1963.

[BDRQL11] Eshar Ben-Dor, Theodore S Rappaport, Yijun Qiao, and Samuel J Lauffenburger. Millimeter-wave 60 GHz outdoor and vehicle AOA propagation measurements using a broadband channel sounder. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–6. IEEE, 2011.

[Ber98] Dimitri P Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific Belmont, 1998.

[BGH87] Dimitri P Bertsekas, Robert G Gallager, and Pierre Humblet. *Data Networks*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1987.

[BH15] Tianyang Bai and Robert W Heath. Coverage and rate analysis for millimeter-wave cellular networks. *IEEE Transactions on Wireless Communications*, 14(2):1100–1114, 2015.

[BMP15] Mohammed Baz, Paul D Mitchell, and Dave AJ Pearce. Analysis of queuing delay and medium access distribution over wireless multi-hop PANs. *IEEE Transactions on Vehicular Technology*, 64(7):2972–2990, Jul. 2015.

[BMY15] Li Bing, Daniel Månsson, and Guang Yang. An Efficient Method for Solving Frequency Responses of Power-Line Networks. *Progress In Electromagnetics Research B*, 62(1):303–317, 2015.

[Bru92] Richard A Brualdi. *Introductory combinatorics*. New York, 1992.

[Buc13] James Bucklew. *Introduction to Rare Event Simulation*. Springer Science & Business Media, 2013.

[CBL06] Florin Ciucu, Almut Burchard, and Jörg Liebeherr. Scaling Properties of Statistical End-to-end Bounds in the Network Calculus. *IEEE/ACM Transactions on Networking*, 14(SI):2300–2312, Jun. 2006.

[CCS01] Cheng-Shang Chang, Yuh-ming Chiu, and Wheyming Tina Song. On the performance of multiplexing independent regulated inputs. In *Proc. ACM SIGMETRICS Performance Evaluation Review*, volume 29, pages 184–193. ACM, 2001.

[CCSM10] Lin X Cai, Lin Cai, Xuemin Shen, and Jon W Mark. REX: a randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna. *IEEE Transactions on Wireless Communications*, 9(1):113–121, 2010.

[Cha94] Cheng-Shang Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, 1994.

[Cha00] Cheng-Shang Chang. *Performance Guarantees in Communication Networks*. Springer, 2000.

[Che11] Ray Jinzhu Chen. An upper bound solution for homogeneous fork/join queuing systems. *IEEE Transactions on Parallel and Distributed Systems*, 22(5):874–878, 2011.

[CHS⁺09] Lin X Cai, HY Hwang, Xuemin Shen, Jon W Mark, and Lin Cai. Optimizing Geographic Routing for millimeter-wave wireless networks with directional antenna. In *Proc. Broadband Communications, Networks, and Systems (BROADNETS)*, pages 1–8. IEEE, 2009.

[CJS⁺10] Jung Il Choi, Mayank Jain, Kannan Srinivasan, Phil Levis, and Sachin Katti. Achieving single channel, full duplex wireless communication. In *Proc. Annual International Conference on Mobile Computing and Networking*, pages 1–12. ACM, 2010.

[CLCS09] Matteo Cypriani, Frédéric Lassabe, Philippe Canalda, and François Spies. Open wireless positioning system: a Wi-Fi-based indoor positioning system. In *Proc. IEEE Vehicular Technology Conference Fall (VTC-Fall)*, pages 1–5, Sept. 2009.

[CLSC16] Xianghui Cao, Lu Liu, Wenlong Shen, and Yu Cheng. Distributed scheduling and delay-aware routing in multihop MR-MC wireless networks. *IEEE Transactions on Vehicular Technology*, 65(8):6330–6342, Aug. 2016.

[CLW16] Hung-Yi Cheng, Ching-Chun Liao, and An-Yeu Andy Wu. Progressive channel estimation for ultra-low latency millimeter-wave communications. In *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 610–614. IEEE, 2016.

[CLY15] Ying Cui, Vincent KN Lau, and Edmund Yeh. Delay optimal buffered decode-and-forward for two-hop networks with random link connectivity. *IEEE Transactions on Information Theory*, 61(1):404–425, Jan. 2015.

[Cru91a] Rene L. Cruz. A calculus for network delay. I. Network elements in isolation. *IEEE Transactions on Information Theory*, 37(1):114–131, Jan. 1991.

[Cru91b] Rene L. Cruz. A calculus for network delay. II. Network analysis. *IEEE Transactions on Information Theory*, 37(1):132–141, Jan. 1991.

[CY13] Hong Chen and David D Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, volume 46. Springer Science & Business Media, Apr. 2013.

[CZZ04] Sylvain Collonge, Gheorghe Zaharia, and G EL Zein. Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel. *IEEE Transactions on Wireless Communications*, 3(6):2396–2406, 2004.

[DA15] Dibakar Das and Alhussein A Abouzeid. Spatial–temporal queuing theoretic modeling of opportunistic multihop wireless networks with and without cooperation. *IEEE Transactions on Wireless Communications*, 14(9):5209–5224, Sept. 2015.

[DCK16] Tolga Dinc, Anandaroop Chakrabarti, and Harish Krishnaswamy. A 60 GHz CMOS full-duplex transceiver and link with polarization-based antenna and RF cancellation. *IEEE Journal of Solid-State Circuits*, 51(5):1125–1140, 2016.

[DH07] Robert C Daniels and Robert W Heath. 60 GHz wireless communications: emerging requirements and design recommendations. *IEEE Vehicular Technology Magazine*, 2(3):41–50, 2007.

[DHHV08] Jürgen Deißner, Johannes Hübner, Dietrich Hunold, and Jens Voigt. *RPS Radiowave Propagation Simulator user manual version 5.4*, 2008.

[DLZ12] Ke Dong, Xuewen Liao, and Shihua Zhu. Link blockage analysis for indoor 60 GHz radio systems. *Electronics Letters*, 48(23):1506–1508, Nov. 2012.

[DS10] Melissa Duarte and Ashutosh Sabharwal. Full-duplex wireless communications using off-the-shelf radios: Feasibility and first results. In *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 1558–1562. IEEE, 2010.

[DSZC13] Linhao Dong, Sumei Sun, Xu Zhu, and Yeow-Khiang Chia. Power efficient 60 GHz wireless communication networks with relays. In *Proc. IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 2808–2812. IEEE, 2013.

[Epp06] Bernhard Epple. Using a GPS-aided inertial system for coarse-pointing of free-space optical communication terminals. In *SPIE Optics & Photonics*, pages 630418–630418. International Society for Optics and Photonics, 2006.

[Fan08] Zhong Fan. Wireless networking with directional antennas for 60 GHz systems. In *Proc. European Wireless Conference (EW)*, pages 1–7, Jun. 2008.

[Fid06a] Markus Fidler. A Network Calculus Approach to Probabilistic Quality of Service Analysis of Fading Channels. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–6, Nov. 2006.

[Fid06b] Markus Fidler. An end-to-end probabilistic network calculus with moment generating functions. In *Proc. IEEE International Workshop on Quality of Service (IWQoS)*, pages 261–270. IEEE, 2006.

[Fid10] Markus Fidler. Survey of deterministic and stochastic service curve models in the network calculus. *IEEE Communications Surveys & Tutorials*, 12(1):59–86, 2010.

[FJ16] Markus Fidler and Yuming Jiang. Non-asymptotic delay bounds for (k, l) fork-join systems and multi-stage fork-join networks. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, pages 1–9. IEEE, Apr. 2016.

[FR15] Markus Fidler and Amr Rizk. A guide to the stochastic network calculus. *IEEE Communications Surveys & Tutorials*, 17(1):92–105, 2015.

[FZM$^+$17] Russell Ford, Menglei Zhang, Marco Mezzavilla, Sourjya Dutta, Sundeep Rangan, and Michele Zorzi. Achieving ultra-low latency in 5G millimeter wave cellular networks. *IEEE Communications Magazine*, 55(3):196–203, 2017.

[Gag12] Robert M Gagliardi. *Satellite communications*. Springer Science & Business Media, 2012.

[GKBS16] Hadi Ghauch, Taejoon Kim, Mats Bengtsson, and Mikael Skoglund. Subspace estimation and decomposition for large millimeter-wave MIMO systems. *IEEE Journal of Selected Topics in Signal Processing*, 10(3):528–542, 2016.

[GKZV09] Suiyan Geng, Jarmo Kivinen, Xiongwen Zhao, and Pertti Vainikainen. Millimeter-wave propagation channel characterization for short-range wireless communications. *IEEE Transactions on Vehicular Technology*, 58(1):3–13, 2009.

[Gol05] Andrea Goldsmith. *Wireless communications*. Cambridge University Press, 2005.

[GOON10] Zulkuf Genc, Gencay M Olçer, Ertan Onur, and Ignas Niemegeers. Improving 60 GHz indoor connectivity with relaying. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1–6, May 2010.

[GQMT07] Nan Guo, Robert C Qiu, Shaomin S Mo, and Kazuaki Takahashi. 60-GHz millimeter-wave radio: Principle, technology, and new results. *EURASIP Journal on Wireless Communications and Networking*, 2007(1):48–48, 2007.

[GRON10] Zulkuf Genc, Umar H Rizvi, Ertan Onur, and Ignas Niemegeers. Robust 60 GHz Indoor Connectivity: Is It Possible with Reflections? In *Proc. IEEE Vehicular Technology Conference Sprint (VTC-Spring)*, pages 1–5, May 2010.

[GT12] Carl Gustafson and Fredrik Tufvesson. Characterization of 60 GHz shadowing by human bodies and simple phantoms. In *Proc. European Conference on Antennas and Propagation (EuCAP)*, pages 473–477, Mar. 2012.

[HCCP16] Jianping He, Lin Cai, Peng Cheng, and Jianping Pan. Delay minimization for data dissemination in large-scale VANETs with buses and taxis. *IEEE Transactions on Mobile Computing*, 15(8):1939–1950, 2016.

[HDL11] Mahdi Hajiaghayi, Min Dong, and Ben Liang. Jointly optimal channel pairing and power allocation for multichannel multihop relaying. *IEEE Transactions on Signal Processing*, 59(10):4998–5012, 2011.

[HOI$^+$16] Katsuyuki Haneda, Nobutaka Omaki, Tetsuro Imai, Leszek Raschkowski, Michael Peter, and Antti Roivainen. Frequency-agile pathloss models for urban street canyons. *IEEE Transactions on Antennas and Propagation*, 64(5):1941–1951, 2016.

[HSA$^+$13] Ken Hiraga, Kazumitsu Sakamoto, Maki Arai, Tomohiro Seki, Tadao Nakagawa, and Kazuhiro Uehara. An SDM method utilizing height pattern due to two-ray fading characteristics. *IEEE Antennas and Wireless Propagation Letters*, 12:1622–1626, 2013.

[HWGL09] K Haddadi, MM Wang, D Glay, and Tuami Lasri. A 60 GHz six-port distance measurement system with sub-millimeter accuracy. *IEEE Microwave and Wireless Components Letters*, 19(10):644–646, 2009.

[HZ12] Jalil Seifali Harsini and Michele Zorzi. Effective capacity for multirate relay channels with delay constraint exploiting adaptive cooperative diversity. *IEEE Transactions on Wireless Communications*, 11(9):3136–3147, 2012.

[JCK$^+$11] Mayank Jain, Jung Il Choi, Taemin Kim, Dinesh Bharadia, Siddharth Seth, Kannan Srinivasan, Philip Levis, Sachin Katti, and Prasun Sinha. Practical, real-time, full duplex wireless. In *Proc. Annual International Conference on Mobile Computing and Networking*, pages 301–312. ACM, 2011.

[JJS13] Bo Ji, Changhee Joo, and Ness B Shroff. Delay-based back-pressure scheduling in multihop wireless networks. *IEEE/ACM Transactions on Networking*, 21(5):1539–1552, Oct. 2013.

[JL08] Yuming Jiang and Yong Liu. *Stochastic Network Calculus*. Springer, 2008.

[JLGH09] Yi Jiang, Keren Li, Jing Gao, and Hiroshi Harada. Antenna space diversity and polarization mismatch in wideband 60GHz-Millimeter-wave wireless system. In *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1781–1785. IEEE, 2009.

[JPK$^+$13] Martin Jacob, Sebastian Priebe, Thomas Kürner, Michael Peter, Mike Wisotzki, Robert Felbecker, and Wilhelm Keusgen. Fundamental analyses of 60 GHz human blockage. In *Proc. European Conference on Antennas and Propagation (EuCAP)*, pages 117–121, Apr. 2013.

[JPM$^+$11] Martin Jacob, Sebastian Priebe, Alexander Maltsev, Artyom Lomayev, Vinko Erceg, and Thomas Kürner. A ray tracing based stochastic human blockage model for the IEEE 802.11ad 60 GHz channel model. In *Proc. European Conference on Antennas and Propagation (EuCAP)*, pages 3084–3088, Apr. 2011.

[JWL15] Zhe Ji, Youzheng Wang, and Jianhua Lu. Delay-aware resource control and routing in multihop wireless networks. *IEEE Communications Letters*, 19(11):2001–2004, Nov. 2015.

[JZSS15] Vahid Jamali, Nikola Zlatanov, Hebatallah Shoukry, and Robert Schober. Achievable rate of the half-duplex multi-hop buffer-aided relay channel with block fading. *IEEE Transactions on Wireless Communications*, 14(11):6240–6256, Nov. 2015.

[KCH15] Amin Abdel Khalek, Constantine Caramanis, and Robert W Heath. Delay-constrained video transmission: Quality-driven resource allocation and scheduling. *IEEE Journal of Selected Topics in Signal Processing*, 9(1):60–75, 2015.

[KPL03] Intae Kang, Radha Poovendran, and Richard Ladner. Power-efficient broadcast routing in adhoc networks using directional antennas: technology dependence and convergence issues. Technical Report UWEETR-2003-0015, University of Washington, Washington, USA, 2003.

[KR09] James F Kurose and Keith W Ross. *Computer Networking: A Top-down Approach*, volume 4. Addison Wesley, Boston, USA, 2009.

[LBL07] Chengzhi Li, Almut Burchard, and Jörg Liebeherr. A network calculus with effective bandwidth. *IEEE/ACM Transactions on Networking*, 15(6):1442–1453, 2007.

[LBT01] Jean-Yves Le Boudec and Patrick Thiran. *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. Springer, 2001.

[LDBL07] Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 37(6):1067–1080, 2007.

[LGAZ$^+$16] Peter Larsson, James Gross, Hussein Al-Zubaidy, Lars Kildehøj Rasmussen, and Mikael Skoglund. Effective capacity of retransmission schemes: A recurrence relation approach. *IEEE Transactions on Communications*, 64(11):4817–4835, 2016.

[LH08] Long Le and Ekram Hossain. Tandem queue models with applications to QoS routing in multihop wireless networks. *IEEE Transactions on Mobile Computing*, 7(8):1025–1040, 2008.

[LJT14] Liangbin Li, Kaushik Josiam, and Rakesh Taori. Feasibility study on full-duplex wireless millimeter-wave systems. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2769–2773. IEEE, 2014.

[LLNS04] Charles SC Leong, Beng-Sin Lee, Andrew R Nix, and Paul. Strauch. A robust 60 GHz wireless network with parallel relaying. In *Proc. IEEE International Conference on Communications (ICC)*, volume 6, pages 3528–3532, Jun. 2004.

[LPCF12] Zhiwei Lin, Xiaoming Peng, Francois Chin, and Wenquan Feng. Outage performance of relaying with directional antennas in the presence of co-channel interferences at relays. *IEEE Wireless Communications Letters*, 1(4):288–291, 2012.

[LSW$^+$09] Zhou Lan, Chin-Sean Sum, Junyi Wang, Tuncer Baykas, Fumihide Kojima, Hiroyuki Nakase, and Hiroshi Harada. Relay with deflection routing for effective throughput improvement in Gbps millimeter-wave WPAN systems. *IEEE Journal on Selected Areas in Communications*, 27(8):1453–1465, 2009.

[LvdBS$^+$16] Anders Landstrom, Jaap van de Beek, Arne Simonsson, Magnus Thurfjell, and Peter Okvist. Measurement-Based Stochastic mmWave Channel Modeling. In *Proc. IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2016.

[MC04] Nektarios Moraitis and Philip Constantinou. Indoor channel measurements and characterization at 60 GHz for wireless local area network applications. *IEEE Transactions on Antennas and Propagation*, 52(12):3180–3189, 2004.

[MC06] Nektarios Moraitis and Philip Constantinou. Measurements and characterization of wideband indoor radio channel at 60 GHz. *IEEE Transactions on Wireless Communications*, 5(4):880–889, 2006.

[MEP$^+$10] Alexander Maltsev, V Erceg, E Perahia, C Hansen, R Maslennikov, A Lomayev, A Sevastyanov, A Khoryaev, G Morozov, M Jacob, et al. Channel models for 60 GHz WLAN systems. *IEEE Document 802.11-09/0334r6*, 2010.

[MHR$^+$17] Antonio J Morgado, Kazi Mohammed Saidul Huq, Jonathan Rodriguez, Christos Politis, and Haris Gacanin. Hybrid Resource Allocation for Millimeter-Wave NOMA. *IEEE Wireless Communications*, 24(5):23–29, 2017.

[MiW14] MiWEBA. D5.1: Channel Modeling and Characterization, Jun. 2014. EU Contract No. FP7-ICT-608637.

[MJA$^+$17] Shahid Mumtaz, Josep Miquel Jornet, Jocelyn Aulin, Wolfgang H Gerstacker, Xiaodai Dong, and Bo Ai. Terahertz communication for vehicular networks. *IEEE Transactions on Vehicular Technology*, 66(7):5617–5625, 2017.

[MKJ$^+$16] Vien V Mai, Juyeop Kim, Sang-Woon Jeon, Sang Won Choi, Beomjoo Seo, and Won-Yong Shin. Degrees of freedom of millimeter wave full-duplex systems with partial CSIT. *IEEE Communications Letters*, 20(5):1042–1045, 2016.

[MMT16] Mihalis G Markakis, Eytan Modiano, and John N Tsitsiklis. Delay stability of back-pressure policies in the presence of heavy-tailed traffic. *IEEE/ACM Transactions on Networking*, 24(4):2046–2059, Aug. 2016.

[MPH06] Shiwen Mao, Shivendra S Panwar, and Y Thomas Hou. On minimizing end-to-end delay with optimal traffic partitioning. *IEEE Transactions on Vehicular Technology*, 55(2):681–690, 2006.

[MRJ11] Kashif Mahmood, Amr Rizk, and Yuming Jiang. On the Flow-Level Delay of a Spatial Multiplexing MIMO Wireless Channel. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, Jun. 2011.

[NJS17] Marzieh Najafi, Vahid Jamali, and Robert Schober. Optimal Relay Selection for the Parallel Hybrid RF/FSO Relay Channel: Non-Buffer-Aided and Buffer-Aided Designs. *IEEE Transactions on Communications*, 65(7):2794–2810, Jul. 2017.

[OBB$^+$14] Afif Osseiran, Federico Boccardi, Volker Braun, Katsutoshi Kusume, Patrick Marsch, Michal Maternia, Olav Queseth, Malte Schellmann, Hans Schotten, Hidekazu Taoka, Hugo Tullberg, Mikko A. Uusitalo, Bogdan Timus, and Mikael Fallgren. Scenarios for 5G mobile and wireless communications: The vision of the METIS project. *IEEE Communications Magazine*, 52(5):26–35, May 2014.

[Olv10] Frank WJ Olver. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[PAZKG15] Neda Petreska, Hussein Al-Zubaidy, Rudi Knorr, and James Gross. On the recursive nature of end-to-end delay bound for heterogeneous wireless networks. In *Proc. IEEE International Conference on Communications (ICC)*, pages 5998–6004. IEEE, 2015.

[PFLZ12] Yali Peng, Wei Fan, Jiayao Liu, and Fan Zhang. The research of traffic flow assignment model based on the network calculus of computer networks. *Information Technology Journal*, 11(3):307, 2012.

[PN02] Joongsuk Park and Cam Nguyen. A new millimeter-wave step-frequency radar sensor for distance measurement. *IEEE Microwave and Wireless Components Letters*, 12(6):221–222, 2002.

[PPT17] Manolis Ploumidis, Nikolaos Pappas, and Apostolos Traganitis. Flow allocation for maximum throughput and bounded delay on multiple disjoint paths for random access wireless multihop networks. *IEEE Transactions on Vehicular Technology*, 66(1):720–733, Jan. 2017.

[PR07] Cheolhee Park and Theodore S Rappaport. Short-range wireless communications for Next-Generation Networks: UWB, 60 GHz millimeter-wave WPAN, and ZigBee. *IEEE Wireless Communications*, 14(4):70–78, 2007.

[QCSM11] Jian Qiao, Lin X Cai, Xuemin Shen, and Jon W Mark. Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks. *IEEE Transactions on Wireless Communications*, 10(11):3824–3833, 2011.

[R$^+$96] Theodore S Rappaport et al. *Wireless communications: principles and practice*, volume 2. Prentice Hall PTR New Jersey, 1996.

[Ram01] Ram Ramanathan. On the performance of ad hoc networks with beamforming antennas. In *Proc. ACM international Symposium on Mobile Ad hoc Networking & Computing*, pages 95–105. ACM, 2001.

[RFC12] Allen L Ramaboli, Olabisi E Falowo, and Anthony H Chan. Bandwidth aggregation in heterogeneous wireless networks: A survey of current approaches and issues. *Journal of Network and Computer Applications*, 35(6):1674–1690, 2012.

[RJLMPG17] Juan M Romero-Jerez, F Javier Lopez-Martinez, José F Paris, and Andrea J Goldsmith. The fluctuating two-ray fading model: Statistical characterization and performance analysis. *IEEE Transactions on Wireless Communications*, 16(7):4420–4432, 2017.

[RLMAG15] Madhav Rao, F Javier Lopez-Martinez, Mohamed-Slim Alouini, and Andrea Goldsmith. MGF approach to the analysis of generalized two-ray fading models. *IEEE Transactions on Wireless Communications*, 14(5):2548–2561, 2015.

[RMIR$^+$14] Juan Reig, M-T Martínez-Inglés, L Rubio, V-M Rodrigo-Peñarrocha, and J-M Molina-García-Pardo. Fading Evaluation in the 60 GHz Band in Line-of-Sight Conditions. *International Journal of Antennas and Propagation*, 2014, 2014.

[RMSS15] Theodore S Rappaport, George R Maccartney, Mathew K Samimi, and Shu Sun. Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design. *IEEE Transactions on Communications*, 63(9):3029–3056, 2015.

[Ron11] Yue Rong. Multihop nonregenerative MIMO relays–QoS considerations. *IEEE Transactions on Signal Processing*, 59(1):290–303, 2011.

[RPC16] Amr Rizk, Felix Poloczek, and Florin Ciucu. Stochastic bounds in fork–join queueing systems under full and partial mapping. *Queueing Systems*, 83(3-4):261–291, 2016.

[RRE14] Sundeep Rangan, Theodore S Rappaport, and Elza Erkip. Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges. *Proceedings of the IEEE*, 102(3):366–385, Mar. 2014.

[RRS$^+$05] Ram Ramanathan, Jason Redi, Cesar Santivanez, David Wiggins, and Stephen Polit. Ad hoc networking with directional antennas: a complete system solution. *IEEE Journal on Selected Areas in Communications*, 23(3):496–506, 2005.

[RSM$^+$13] Theodore S Rappaport, Shu Sun, Rimma Mayzus, Hang Zhao, Yaniv Azar, Kangping Wang, George N Wong, Jocelyn K Schulz, Mathew Samimi, and Felix Gutierrez. Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access*, 1:335–349, 2013.

[RST14] Waheed Ur Rehman, Tabinda Salam, and Xiaofeng Tao. Receiver based distributed relay selection scheme for 60-GHz networks.

In *Proc. IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pages 1585–1590. IEEE, 2014.

[Saa14] Mohamed Saad. Joint optimal routing and power allocation for spectral efficiency in multihop wireless networks. *IEEE Transactions on Wireless Communications*, 13(5):2530–2539, 2014.

[SGFF+15] Hossein Shokri-Ghadikolaei, Carlo Fischione, Gabor Fodor, Petar Popovski, and Michele Zorzi. Millimeter wave cellular networks: A MAC layer perspective. *IEEE Transactions on Communications*, 63(10):3437–3458, 2015.

[SH14] Kostas Stamatiou and Martin Haenggi. Delay characterization of multihop transmission in a Poisson field of interference. *IEEE/ACM Transactions on Networking*, 22(6):1794–1807, Dec. 2014.

[SKGA15] Sarabjot Singh, Mandar N Kulkarni, Amitava Ghosh, and Jeffrey G Andrews. Tractable model for rate in self-backhauled millimeter wave cellular networks. *IEEE Journal on Selected Areas in Communications*, 33(10):2196–2211, 2015.

[SLC11] Chao-Fang Shih, Wanjiun Liao, and Hsi-Lu Chao. Joint routing and spectrum allocation for multi-hop cognitive radio networks with route robustness consideration. *IEEE Transactions on Wireless Communications*, 10(9):2940–2949, 2011.

[SMGL12] In Keun Son, Shiwen Mao, Michelle X Gong, and Yihan Li. On frame-based scheduling for directional mmWave WPANs. In *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, pages 2149–2157. IEEE, 2012.

[SMM11] Sumit Singh, Raghuraman Mudumbai, and Upamanyu Madhow. Interference analysis for highly directional 60-ghz mesh networks: The case for rethinking medium access control. *IEEE/ACM Transactions on Networking*, 19(5):1513–1527, 2011.

[SMS+17] Mansoor Shafi, Andreas F Molisch, Peter J Smith, Thomas Haustein, Peiying Zhu, Prasan De Silva, Fredrik Tufvesson, Anass Benjebbour, and Gerhard Wunder. 5G: A Tutorial Overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, 35(6):1201–1221, Jun. 2017.

[Smu09] Peter FM Smulders. Statistical characterization of 60-GHz indoor radio channels. *IEEE Transactions on Antennas and Propagation*, 57(10):2820–2829, 2009.

[SPB16] Eleni Stai, Symeon Papavassiliou, and John S Baras. Performance-aware cross-layer design in wireless multihop networks via a weighted backpressure approach. *IEEE/ACM Transactions on Networking*, 24(1):245–258, Feb. 2016.

[SRH$^+$14] Shu Sun, Theodore S Rappaport, RW Heath, Andrew Nix, and Sundeep Rangan. MIMO for millimeter-wave wireless communications: beamforming, spatial multiplexing, or both? *IEEE Communications Magazine*, 52(12):110–121, 2014.

[SSRRM15] Mathew K. Samimi, Theodore S. Rappaport, and George R. MacCartney. Probabilistic Omnidirectional Path Loss Models for Millimeter-Wave Outdoor Communications. *IEEE Wireless Communications Letters*, 2015.

[SZ09] Hang Su and Xi Zhang. Joint link scheduling and routing for directional-antenna based 60 GHz wireless mesh networks. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–6, Nov. 2009.

[SZM$^+$09] Sumit Singh, Federico Ziliotto, Upamanyu Madhow, E Belding, and Mark Rodwell. Blockage and directivity in 60 GHz wireless personal area networks: from cross-layer model to multihop MAC design. *IEEE Journal on Selected Areas in Communications*, 27(8):1400–1413, 2009.

[TBHJ16] Andrew Thornburg, Tianyang Bai, and Robert W Heath Jr. Performance Analysis of Outdoor mmWave Ad Hoc Networks. *IEEE Transactions Signal Processing*, 64(15):4065–4079, 2016.

[TDHA14] Hina Tabassum, Zaher Dawy, Ekram Hossain, and Mohamed-Slim Alouini. Interference statistics and capacity analysis for uplink transmission in two-tier small cell networks: A geometric probability approach. *IEEE Transactions on Wireless Communications*, 13(7):3837–3852, 2014.

[TH15] Andrew Thornburg and Robert W Heath. Ergodic capacity in mmWave ad hoc network with imperfect beam alignment. In *Proc. IEEE Military Communications Conference (MILCOM)*, pages 1479–1484. IEEE, 2015.

[TS15] Rakesh Taori and Arun Sridharan. Point-to-multipoint in-band mmwave backhaul for 5G networks. *IEEE Communications Magazine*, 53(1):195–201, 2015.

[TSI06] I Toyoda, T Seki, and K Iiguse. Reference antenna model with side lobe for TG3c evaluation. Technical report, IEEE document, 2006.

[TZ08] Jia Tang and Xi Zhang. Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks. *IEEE Transactions on Wireless Communications*, 7(6), 2008.

[VEDS88] Edmond J Violette, Richard H Espeland, ROBERT O DeBOLT, and FK Schwering. Millimeter-wave propagation at street level in an urban environment. *IEEE Transactions on Geoscience and Remote Sensing*, 26(3):368–380, 1988.

[VH16] Kiran Venugopal and Robert W Heath. Millimeter wave networked wearables in dense indoor environments. *IEEE Access*, 4:1205–1221, 2016.

[VLB$^+$17] Trung Kien Vu, Chen-Feng Liu, Mehdi Bennis, Mérouane Debbah, Matti Latva-aho, and Choong Seon Hong. Ultra-Reliable and Low Latency Communication in mmWave-Enabled Massive MIMO Networks. *IEEE Communications Letters*, 21(9):2041–2044, 2017.

[WAN97] MR Williamson, GE Athanasiadou, and AR Nix. Investigating the effects of antenna directivity on wireless indoor communication at 60 GHz. In *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, volume 2, pages 635–639, Sept. 1997.

[WiG10] WiGig. Defining the future of multi-gigabit wireless communications, Jul. 2010.

[WKW10] Jue Wang, Linghe Kong, and Min-You Wu. Capacity of wireless ad hoc networks using practical directional antennas. In *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2010.

[WN03] Dapeng Wu and Rohit Negi. Effective capacity: A wireless link model for support of quality of service. *IEEE Transactions on Wireless Communications*, 2(4):630–643, 2003.

[WNE02] Jeffrey E Wieselthier, Gam D Nguyen, and Anthony Ephremides. Energy-aware wireless networking with directional antennas: The case of session-based broadcasting and multicasting. *IEEE Transactions on Mobile Computing*, 1(3):176–191, 2002.

[WNLa$^+$14] Jeffrey Wildman, Pedro Nardelli, Matti Latva-aho, Simon Weber, et al. On the joint impact of beamwidth and orientation error on throughput in directional wireless Poisson networks. *IEEE Transactions on Wireless Communications*, 13(12):7072–7085, 2014.

[Wol18] Wolfam. WolframAlpha@ computational knowledge engine, 2018.

[WPK$^+$15] Richard J Weiler, Michael Peter, Wilhelm Keusgen, Andreas Kortke, and Mike Wisotzki. Millimeter-wave channel sounding of outdoor ground reflections. In *Proc. IEEE Radio and Wireless Symposium (RWS)*, pages 95–97. IEEE, 2015.

[WSW11] Wei Wang, Kang G Shin, and Wenbo Wang. Joint spectrum allocation and power control for multihop cognitive radio networks. *IEEE Transactions on Mobile Computing*, 10(7):1042–1055, 2011.

[WZS$^+$15] Zhongxiang Wei, Xu Zhu, Sumei Sun, Yi Huang, Linhao Dong, and Yufei Jiang. Full-Duplex Versus Half-Duplex Amplify-and-Forward Relaying: Which is More Energy Efficient in 60-GHz Dual-Hop Indoor Wireless Systems? *IEEE Journal on Selected Areas in Communications*, 33(12):2936–2947, 2015.

[WZS$^+$16a] Zhongxiang Wei, Xu Zhu, Sumei Sun, Yi Huang, Ahmed Al-Tahmeesschi, and Yufei Jiang. Energy-efficiency of millimeter-wave full-duplex relaying systems: Challenges and solutions. *IEEE Access*, 4:4848–4860, 2016.

[WZS$^+$16b] Zhongxiang Wei, Xu Zhu, Sumei Sun, Yi Huang, and Hai Lin. Cross-layer energy-efficiency optimization for multiuser full-duplex decode-and-forward indoor relay networks at 60 GHz. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

[WZSH16] Zhongxiang Wei, Xu Zhu, Sumei Sun, and Yi Huang. Energy-Efficiency-Oriented Cross-Layer Resource Allocation for Multiuser Full-Duplex Decode-and-Forward Indoor Relay Systems at 60 GHz. *IEEE Journal on Selected Areas in Communications*, 34(12):3366–3379, 2016.

[XMH$^+$17] Ming Xiao, Shahid Mumtaz, Yongming Huang, Linglong Dai, Yonghui Li, Michail Matthaiou, George K Karagiannidis, Emil Björnson, Kai Yang, I Chih-Lin, and Amitabha Ghosh. Millimeter Wave Communications for Future Mobile Networks. *IEEE Journal on Selected Areas in Communications*, 35(9):1909–1935, Sept. 2017.

[YA12] Ferkan Yilmaz and Mohamed-Slim Alouini. A unified MGF-based capacity analysis of diversity combiners over generalized fading channels. *IEEE Transactions on Communications*, 60(3):862–875, 2012.

[YDX15] Guang Yang, Jinfeng Du, and Ming Xiao. Maximum throughput path selection with random blockage for indoor 60 GHz relay networks. *IEEE Transactions on Communications*, 63(10):3511–3524, 2015.

[YHX18] Guang. Yang, Martin Haenggi, and Ming Xiao. Traffic Allocation for Low-Latency Multi-Hop Networks with Buffers. *ArXiv e-prints*, 2018.

[YX17a] Guang Yang and Ming Xiao. Blockage robust millimeter-wave networks. *Science China Information Sciences*, 60(8):080307, 2017.

[YX17b] Guang Yang and Ming Xiao. Interference statistics of regular ring-structured networks with 60 GHz directional antennas. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[YX18] Guang Yang and Ming Xiao. Performance Analysis of Millimeter-Wave Relaying: Impacts of Beamwidth and Self-Interference. *IEEE Transactions on Communications*, 66(2):589–600, Feb. 2018.

[YXAH18] Guang Yang, Ming Xiao, Muhammad Alam, and Yongming Huang. Low-Latency Heterogeneous Networks with Millimeter-Wave Communications. *ArXiv e-prints*, 2018.

[YXAZ$^+$18] Guang Yang, Ming Xiao, Hussein Al-Zubaidy, Yongming Huang, and James Gross. Analysis of Millimeter-Wave Multi-Hop Networks With Full-Duplex Buffered Relays. *IEEE/ACM Transactions on Networking*, 26(1):576–590, Feb. 2018.

[YXG$^+$16] Guang Yang, Ming Xiao, James Gross, Hussein Al-Zubaidy, and Yongming Huang. Delay and backlog analysis for 60 GHz wireless networks. In *Proc. IEEE Global Communications Conference (GLOBECOM)*, pages 1–7, Dec. 2016.

[YXP18] Guang Yang, Ming Xiao, and H. Vincent Poor. Low-Latency Millimeter-Wave Communications: Traffic Dispersion or Network Densification? *IEEE Transactions on Communications*, PP(99):1–1, 2018.

[YYMZ06] Jin Yu, Yu-Dong Yao, Andreas F Molisch, and Jinyun Zhang. Performance evaluation of CDMA reverse links with imperfect beamforming in a multicell environment using a simplified beamforming model. *IEEE Transactions on Vehicular Technology*, 55(3):1019–1031, 2006.

[YZHL17] Xianghao Yu, Jun Zhang, Martin Haenggi, and Khaled B Letaief. Coverage Analysis for Millimeter Wave Networks: The Impact of Directional Antenna Arrays. *IEEE Journal on Selected Areas in Communications*, 35(7):1498–1512, Jul. 2017.

[YZYX17] Yu Ye, Zhengquan Zhang, Guang Yang, and Ming Xiao. Minimum cost based clustering scheme for cooperative wireless caching network with heterogeneous file preference. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017.

[ZZM$^+$15] Kan Zheng, Long Zhao, Jie Mei, Mischa Dohler, Wei Xiang, and Yuexing Peng. 10 Gb/s hetsnets with millimeter-wave communications: access and networking-challenges and protocols. *IEEE Communications Magazine*, 53(1):222–231, 2015.