Buffer-aided relay networks in 5G and beyond-5G mobile systems

Buffer-Aided Relay Networks in 5G and Beyond 5G Mobile Systems

by

Mohammad Alkhawatrah

A Doctoral Thesis submitted in partial fulfilment
of the requirements for the award of the degree of
Doctor of Philosophy
of Loughborough University

July 2020

Signal Processing and Networks Research Group
Wolfson School of Mechanical, Electrical and Manufacturing Engineering

To my parents, family and friends

Declaration

I, with this, declare that I am responsible for the work submitted in this thesis. The original work is my own except as specified in acknowledgments or footnotes. Neither the thesis nor the original work therein has been submitted to this university or any other institution for a degree.

_____

Mohammad Alkhawatrah

July 2020

Acknowledgements

I am very grateful to my main supervisor Dr. Alex Gong for his substantive support and positive comments, which have helped me during the program. Also, I want to thank Prof. Sangarapillai Lambotharan for his valuable comments and suggestions. I would also like to thank Dr. Gan Zheng and Dr. Mahsa Derakhshani for their helpful suggestions to enhance the research quality.

I would like to thank Al-Ahliyya Amman University (AAU) for granting me a doctoral scholarship, and supporting me with generous financial assistance during the entire course.

Also, I would like to express my gratitude to those who have helped me in whatever form during this long journey. A big thank you to my colleagues Ashraf, Ramadan, Abdullahi, Amjad, Mike and many other names, your support is highly appreciated. Thank you all.

Lastly, I wish to express my deep and most sincere appreciation and gratitude to my parents for their enormous and continual support. A special thanks to my family and friends for their support and prayers.

Abstract

Wireless communication in the 5G era has to meet challenging goals such as huge target data rates, ultra-low latency, massive connectivity, and several other requirements. In the literature, several techniques were suggested to fulfill these goals. One of the promising techniques to achieve higher data rates, which has received considerable attention recently, is buffer-aided cooperative relay networks.

This thesis aims to study how to exploit buffer-aided relays in the 5G cooperative networks more effectively to get closer to achieve the goals of the 5G and beyond. Specifically, the proposed techniques in this thesis are directed to improve the most critical performance metrics of the buffer-aided cooperative relay network: the system throughput, the diversity gain and the average packet delay.

Firstly, a novel prioritization-based buffer-aided relay selection scheme, which is able to combine non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA) transmission in buffer-aided cooperative relay networks is proposed in this thesis. Opposing to the available buffer-aided relays in cooperative NOMA schemes, which are only valid for high signal-to-noise ratio (SNR) ranges, the result shows that the proposed scheme significantly improves the data throughput and the diversity gain in both low and high SNR ranges. While all the available schemes for buffer-aided relays in cooperative NOMA have considered a single relay, the proposed scheme has shown its excellence for the multiple relays scenario. The closed-form analytical expression for the average throughput of the proposed scheme is successfully derived and verified by numerical simulations. Besides, the analytical and simulation analysis show that the diversity gain of the proposed scheme is equal three times the number of relays in the system, which was twice the number of relays in previous studies. In addition, the impact of setting different values for the target

length is discussed. Result shows that in delay-unconstrained applications, it is better to set the target length based only on avoiding empty and full buffers.

Secondly, buffer-aided relays can lengthen the packet delay if queues in the buffers are not controlled. In order to make a fair delay performance comparison between the non-buffer-aided relays and the buffer-aided relays, a new factor, the delay that packets encounter at the source (the source delay) which is not considered in the available literature, is thoroughly discussed in this thesis. Buffer-aided relays have better outage performance than non-buffer-aided relays, therefore, packets tend to leave the source faster. The result shows that buffer-aided relays have shorter source delay than non-buffer-aided relays. Hence, buffer-aided relays can beat non-buffer-aided relays in the packet delay in some cases especially at low SNR. The closed-form expression for the source delay is successfully derived. Some of the delay reduction techniques such as broadcasting and small buffer size are tested while considering the source delay. The result shows that this technique has a positive impact on the delay performance of buffer-aided relays. Thirdly, a novel relay selection scheme, which introduces the idea of an adaptive target length based on the status of the relay transmission channel, is proposed. Simulation results show that the proposed scheme has better average packet delay when compared to other schemes.

Finally, in delay-constrained applications such as tactile internet, which requires the latency below 1ms, studying the distribution of the delay is necessary because every packet with delay higher than the target delay will be re-transmitted or discarded. This has an impact on the system performance metrics such as throughput. In this thesis, Trellis state diagram and Markov chain are used to analyse the delay-constrained outage probability which caused by both the channel outage and the delay exceeding the target delay. The closed-forms of the delay-constrained outage probability is successfully derived for three benchmark selection schemes for the 3-node relay network. This thesis proposes an adaptive buffer-size relay selection scheme which is applied on the available schemes and achieves significant better delay-constrained outage probability than their fixed buffer size counterparts.

Contents

List of Publications

The following publications are the list of the author's publications which have been produced during the PhD degree research.

Journals

1. **M. Alkhawatrah**, Y. Gong, G. Chen, S. Lambotharan and J. A. Chambers, "Buffer-Aided Relay Selection for Cooperative NOMA in the Internet of Things," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 5722-5731, June 2019.

2. **M. Alkhawatrah**, Y. Gong and S. Lambotharan. "The Impact of Source Delay on End-to-End Average Packet Delay in Buffer-Aided Relay Networks," in IEEE Open Journal of the Communications Society. **In Submission**

3. **M. Alkhawatrah**, Y. Gong and G. Chen. "Delay-Constrained Adaptive Link Selection in Buffer-Aided Relay Networks," in IEEE Internet of Things Journal. **In Submission**

Conference Papers

4. **Alkhawatrah, M.**, Gong, Y., Aldabbas, O., and Hammoudeh, M. (2019, July). Buffer-aided 5G cooperative networks: Considering the source delay. In Proceedings of the 3rd International Conference on Future Networks and Distributed Systems (p. 13). ACM.

List of Acronyms

IoT	Internet of Things
4G	Fourth Generation
5G	Fifth Generation
OMA	Orthogonal Multiple Access
NOMA	Non-Orthogonal Multiple Access
LTE	Long-Term Evolution
IP	Internet Protocol
ms	Millisecond
ITU-WRC	International Communication Union World Radio Conference
MIMO	Multiple Input Multiple Output
OFDMA	Orthogonal Frequency Division Multiple Access
eMBB	enhanced Mobile Broadband
URLLC	Ultra-Reliable and Low Latency Communication
mMTC	Massive Machine Type Communication
MA	Multiple Access
DL	Down Link
UL	Up Link
i.i.d.	Independent and Identically Distributed
i.n.i.d.	Independent and Non-Identically Distributed
TDMA	Time Division Multiple Access
FDMA	Frequency Division Multiple Access

CDMA	Code Division Multiple Access
SIC	Successive Interference Cancelation
BS	Base Station
MS	Mobile Station
IMT	International Mobile Telecommunications
ATSC	Advanced Television System Committee
3GPP	3rd Generation Partnership Project
MC	Markov Chain
DF	Decode and Forward
AF	Amplify and Forward
HD	Half Duplex
FD	Full Duplex
MRC	Maximal Ratio Combining
CSI	Channel State Information
CSIT	Channel State Information at Transmitter
CSIR	Channel State Information at Receiver
BSI	Buffer State Information
SNR	Signal to Noise Ratio
SINR	Signal to Noise and Interference Ratio
BSR	Best Relay Selection
MMRS	Max-Max Relay Selection
HRS	Hybrid Relay Selection
ACK	Acknowledgement
NACK	Negative-Acknowledgement
FR	Fast Relay
SR	Slow Relay
M2M	Machine to Machine
D2D	Device to Device

ML Machine Learning

List of Figures

List of Tables

Chapter 1

Introduction

The increasing demand on higher data rates in wireless communication, the requirement of having everything online all the time such as the case with the internet of things IoT technology, and the ultra low latency requirement in applications such as autonomous vehicle, are hard to be met in the current 4G [1]. The requirements of the upcoming high demanding applications such as ultra high definition video streaming and many other applications are expected to be met in the next-generation 5G. These high demands led to intensifying the research in the 5G area. Multiple techniques have been proposed to support these applications. One of these techniques which was part of 4G and it is suggested for 5G is cooperation; this chapter explains this technique and how cooperation can also be combined with other 5G techniques such as non-orthogonal multiple access NOMA to achieve better performance.

1.1 Fifth Generation 5G

1.1.1 Evolution in Wireless Technologies

Wireless communication has evolved over the past few decades from analog voice calls to modern technologies that provide high quality mobile services with up to 100 Mbps user data rate [14]. The first generation was introduced 1980's. It has up to 2.4kbps data rate. It has disadvantages like low capacity. The second generation

was announced in 1990's. Digital technology is used in mobile telephones. Global systems for mobile communications was the first 2nd generation system used for voice communication and it has up to 64kbps data rate. A 2.5G system generally uses 2G system frameworks, but it applies packet switching along with circuit switching. It has up to 144kbps data rate. The third generation was presented in 2000. It has transmission rate up to 2Mbps. The evolving technologies has made an intermediate wireless generation between 3G and 4G named as 3.5G and 3.75G with improved data rate of 5-30 Mbps. In 4G, voice, data and multimedia are delivered to users on every time and everywhere and at quite higher data rates compared to earlier generations [112]. Fig. 1.1 summarizes the evolution in wireless generations (so far) in terms of data rate, mobility and coverage. It is clear that as wireless technologies are growing, the data rate, mobility, and coverage are also growing as well [52].

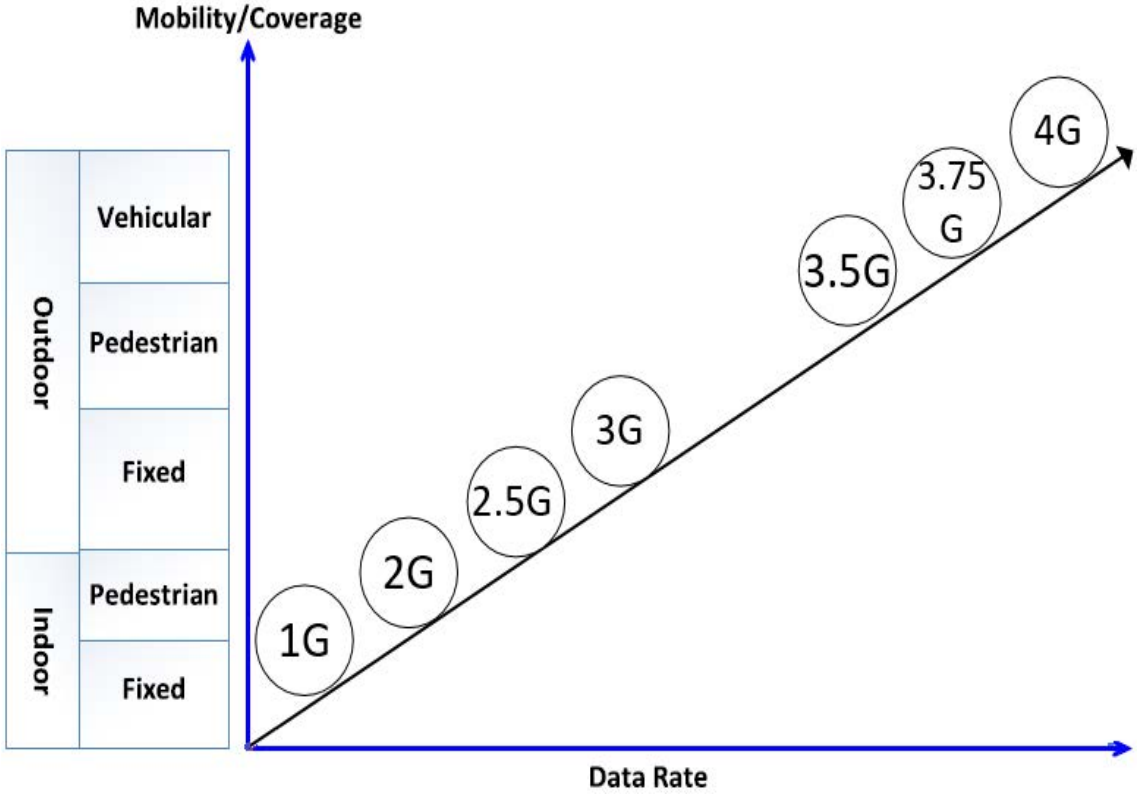

Figure 1.1 Wireless generations [52].

The aiming for future mobile networks is to provide users with unbounded access to information and sharing information every time and everywhere for

everyone and everything. The maturity in 4G (e.g. long-term evolution (LTE)) makes any improvement to be limited and quite complicated. At the same time, data explosion in wireless communication will continue. In the last decade, the number of IP devices increased by over a factor of 100, and the data required by each device has risen. Therefore, the volume of data, the number of connected devices and the data rates are growing in an unprecedented pace. Meeting all these requirements can not be done with LTE, and a significant shift in performance is needed to meet these demands. To this end, practical and efficient techniques are suggested in the literature under the term 5G. 5G techniques focus on three main communication enhancing categories [138]:

- Enhanced mobile broadband (eMBB): this includes extended coverage area, higher mobility, higher data rate and high user density.
- Ultra-reliable and low latency communications (URLLC) applications such as intelligent transport systems and remote medical surgery.
- Massive machine-type communication (mMTC) consists of a massive number of low data rate devices such as IoT. Table 1.1 summarizes the 5G requirements.

Table 1.1 Minimum 5G requirements in IMT 2020 [114]

Peak data rate	DL: 20Gbps, UL: 10 Gbps
Peak spectral efficiency	DL: 30 bit/sec/Hz, UL: 15 bit/sec/Hz
User experienced data rate	DL: 100 Mbps, UL: 50 Mbps
Area traffic capacity (indoor)	DL: 10 Mbps/m^2
Latency	1-4 ms
Connection density	10^6 devices/km^2
Reliability	1^{-5} error probability for 32 byte/ms transmission
Mobility	up to 500 km/h
Bandwidth	0.1-1 GHz

1.1.2 5G Techniques

Understanding the available 5G techniques is vital as well as integrating them for achieving the maximum performance and minimum overhead. This section

discusses multiple existing and future techniques necessary for 5G deployment.

A. Millimeter Wave

Wireless systems operate in a thin range of microwave frequencies that begins from several hundred MHz to a few GHz which corresponds to wavelengths from centimeters up to about a meter. This spectral band has become nearly fully occupied, therefore, much wider bandwidth is required. To get wide new bandwidth, there is only one way, using higher frequencies. Luckily, large amounts of not used spectrum is in the mm wave range of 30–300 GHz, where wavelengths are 1–10 mm. The mm wave spectrum had been thought of as unsuitable for mobile communications because of its low propagation qualities, including high path loss, atmospheric and rain absorption, low diffraction around obstacles and penetration through objects [85, 105]. However, these propagation problems can be solved with cooperation techniques such as relay cooperation. Very high data rate can be met for example, with wider bandwidth offered by millimeter waves. Since the available band (lower than 6 GHz) is matured, ITU-WRC-15 has suggested higher bands 24-100 GHz, which are suitable for high data rate and short-range indoor applications, which can be extended with **cooperative** solutions.

B. Massive MIMO

Based on research started in 1990s, MIMO communication was introduced into 3G cellular around 2006. Multiple-input multiple-output (MIMO) can exploit the spatial dimension of the communication channel. In single-user MIMO (SU-MIMO), the number of links are limited by the number of antennas on a mobile device. However, if each BS communicate with several users concurrently, the multiuser version of MIMO (MU-MIMO) number of links is given by the smallest between the number of antennas at users and the number of antennas at the BS. Furthermore, in coordinated multi-point (CoMP), multiple BSs can cooperate and act as a single effective MIMO transceiver and turning some of the interference in the system into signals. MIMO was a part of 4G standard with two-to-four antennas per mobile device and as many as eight per BS sector. For 5G, massive MIMO aims to increase

the number of antennas per BS into the hundreds which offers noticeable benefits: 1) Enhancements in spectral efficiency. 2) Smoothed out channel responses. 3) Simple transmit/receive structures because of the quasi-orthogonal nature of the channels between each BS and the set of active users, orthogonality increases as the number of BS antennas grows and simple linear transceivers perform close-to-optimally [76, 43].

Theses benefits brought massive MIMO to a central position in preliminary 5G discussions related to providing a high-capacity and wider coverage. However, massive MIMO has several challenges to overcome such as interference in different cells and the requirement of extensive field measurements. In addition, small cell would not be equipped with massive MIMO due to their smaller form factor, and applying massive MIMO at mm wave frequencies requires finding the correct balance between power gain and interference reduction [26].

C. Spectrum Sharing

As mentioned in mm wave technique, wider bandwidths as compared to the current available spectrum is required for realizing the target performance of 5G applications. So to overcome this difficulty, spectrum sharing is a promising technique. As it makes the available spectrum more accessible. One of the spectrum sharing techniques that can utilize the available spectrum effectively is cognitive radio. Cognitive radios are fully programmable wireless devices and has a perfect adaptation property for achieving a better network and application performance. It is able sense the environment and dynamically performs adaptation in the networking protocols, spectrum utilization methods, channel access methods and transmission waveform used . Non orthogonal multiple access (NOMA) is another technique to achieve spectrum sharing. While ultra-low latency requires shorter transmit time which may be accomplished by new waveform [4, 53, 115], or by small packet size. The currently available multiple access techniques such as orthogonal frequency multiple access (OFDMA) has a significant overhead on packet size associated with the scheduling for orthogonal transmission. New multiple access techniques such as non-orthogonal multiple access (NOMA) may help in achieving the ultra-low

latency goal for 5G systems by relaxing the requirement for scheduling the transmission as all the available spectrum is shared by all users [10, 77]. NOMA is discussed in detail next [133].

D. Densification

A straightforward but effective way to increase the network capacity is to make the cells smaller, which means more BSs (densification) are required to preserve coverage. This approach has been demonstrated over several cellular generations. Started with size of order of hundreds of square kms. Since then, those sizes keep shrinking and now they are fractions of a square km. Small cells are picocells (range under 100 meters) and femtocells (WiFi-like range). Cell shrinking has several benefits, like the reuse of spectrum across a geographic area and the ensuing reduction in the number of users competing for resources at each BS. With densification some challenges arise such as supporting mobility through such a highly heterogeneous network, and affording the rising costs of installation and maintenance [27].

Densification with mm wave adds additional complexity, since the cell boundary is blurry at mm wave frequencies because of the strong impact of blockages, which results in nearby BSs being bypassed in favor of farther ones that are unblocked. On the other hand, interference is much less important in mm wave. But in mm wave, hand-offs will be particularly challenging since transmit and receive beams must be aligned to communicate, this puts restrictions on mobility. Smaller cells requires lower-power and cheaper BSs. 5G and all networks beyond it are expected to be extremely dense and heterogeneous [103].

E. Device to Device (D2D) and Machine to Machine (M2M)

D2D enables the devices in proximity to communicate directly bypassing the BS for sharing relevant contents. There are three main types of device-level communications:

- Device relaying with BS. In this type, the devices communicate with the BS by relaying their information through other devices. This is helpful for the device to attain a higher quality of service.

- Direct D2D with BS controlled link formation: source and destination devices exchange data with each other without the involvement of a BS, but they are supported by the BS for link formation.

- Device relaying with device: BS is neither involved in link formation nor for communication purpose. So, source and destination devices are totally responsible for synchronizing communication using relays (other devices in the network) amid each other.

Some technical issues with D2D need to be addressed like security, since devices in the middle may violate privacy. This can be solved by making each device has a list of certain reliable devices, and uses an appropriate encryption [119, 11]. Unlike D2D communications, machine to machine (M2M) communications connect massive number of devices, like meters, sensors and other smart equipments in wide coverage areas. Major features of intelligent machines M2M communications are automated data generation, processing, transfer and exchange. M2M communications have small data, high reliability, low latency and real time operation. Like D2D communications, M2M communications are also expected to have to be part of 5G [66]. To summaries, the main 5G techniques [10, 114] are

- Wide bandwidth: 4G users are heavily using the lower than 6 GHz band, this encourages the movement towards new bands like millimeter wave, which can provide wider bandwidth. However, all the mentioned bands can be used to support 5G. For instance, a base station may work in lower than 6 GHz band carries user control signaling, and another base station may work in the millimeter band carries users traffic.

- More spatial diversity: one technology is massive multiple-input multiple-output (MIMO) arrays at the base station BS. With higher carrier frequencies

massive number of antennas on a relatively small size array becomes permissible. This technology overcomes the path loss and gives spatial multiplexing gain. Another approach that achieves the spatial diversity with much lower complexity is relay selection in cooperative relay networks.

- High spectral efficiency, this means raising the number of bit/sec/Hz per node, this can be done by several techniques such as NOMA. NOMA may also help in delay reduction. Another technology to improve spectral efficiency is cognitive radio (CR) networks, which allow spectrum sharing to avoid leaving part of the available spectrum unused.

- Massive densification by adding more active nodes such as small cells per unit area.

- D2D and M2M reduce the overhead on BSs, and provide reliable link for communication. A comprehensive overview for the available 5G techniques is in [26, 101, 35].

1.1.3 5G Applications

A wide variety of new applications is the main force behind the commercial roll out of 5G wireless systems. 5G techniques are expected to provide network solutions for a wide range sectors like energy, agriculture, city management, health care, manufacturing and transport. Two of the challenging applications are:

A. Internet of Things (IoT)
As shown in Fig. 1.2, IoT aims to millions of simultaneous connections, involving a variety of devices, connected homes, smart grids and smart transportation systems. This aim can be eventually realized only with high bandwidth 5G wireless networks. IoT enables internet connections and data interchange for numerous smart objects and applications. Implementation of IoT is complex, as it includes **cooperation** among massive, distributed, autonomous and heterogeneous components. Relay cooperation is a capable candidate to achieve IoT aims. The concept of cloud,

offering large storage, computing and networking capabilities, can be integrated with diverse IoT devices [33].

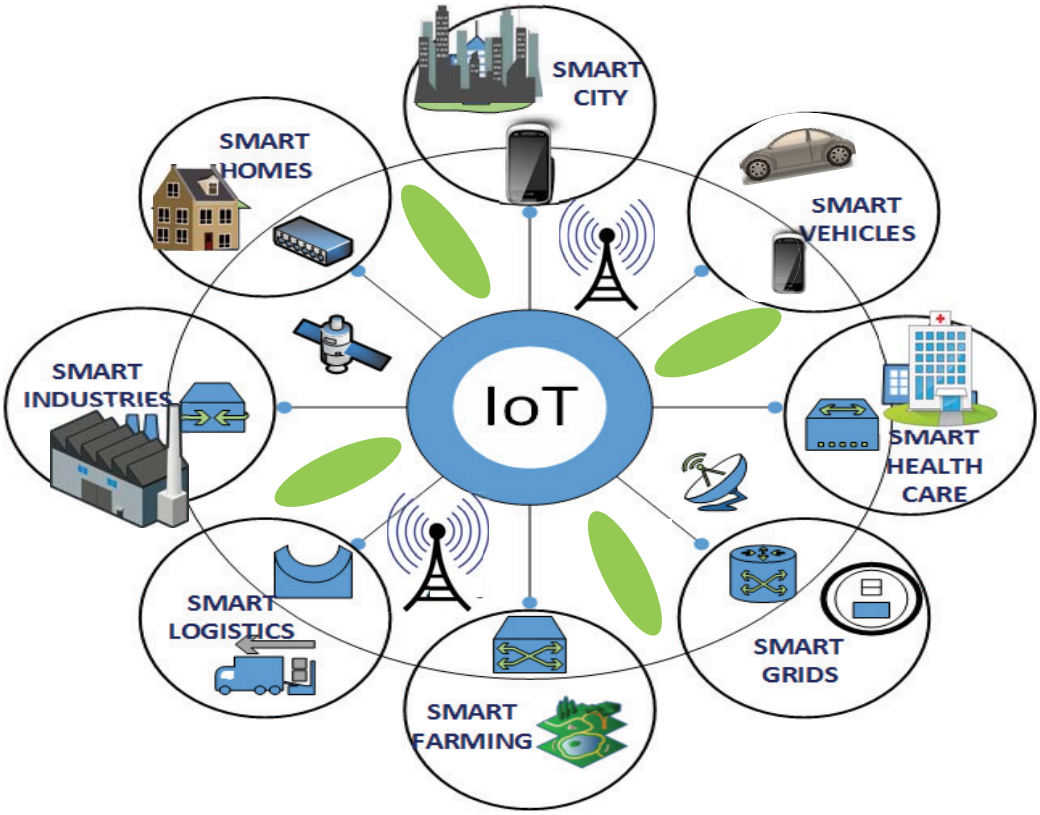

Figure 1.2 IoT connecting anything, anyone, anytime, anywhere [1].

B. Tactile Internet—Real-Time Context

Real-time interaction with the environment is important to humans. The real-time interaction is when the communication response delay is negligible. Tactile internet is an ultra-reliable and ultra-responsive network connectivity will enable it to deliver real-time control. Tactile internet will provide paradigm shift from content-delivery to skill-delivery networks. Tactile internet addresses areas with reaction times in the order of a millisecond, example areas are real-time gaming, overseas online surgeries. Because tactile internet will be servicing critical applications, it has to be ultra-reliable, with a second of outage per year. and it has to support very low latencies, and have sufficient capacity to allow large numbers of devices to communicate with each other simultaneously. 5G mobile communications systems

are expected to support the tactile internet challenging requirements [118]. Several 5G applications are covered in [1].

Integrating 5G technologies is a trending research manner nowadays because of its ability to achieve better performance. For example, NOMA is combined with MIMO to achieve better spectral efficiency and spatial diversity simultaneously. Specifically, authors in [75], have proved the superiority of MIMO-NOMA compared to MIMO-OMA in terms of the sum channel capacity. While a hybrid system combining both MIMO-NOMA and MIMO-OMA was suggested in [15] to maximize the bandwidth efficiency.

Finally, cooperative network is an effective solution in 4G, hence, it is part of many wireless standards such as 3GPP LTE-Advanced [122]. During the last decade, several enhancements on the cooperation performance have been suggested, which makes cooperation a promising 5G technique, simultaneously, other 5G techniques and applications such as mm wave and IoT require cooperation to enhance their performance.

1.2 Cooperative Networks

Due to time-variant fading, wireless system typically include some degree of diversity to provide the receiver with several realizations of the signal, which increases the probability of a successful transmission. Many forms of diversity are possible, time diversity consists of transmitting replicas with enough separation in time. Frequency diversity relies on multiple carriers, and space diversity systems sufficiently spaced paths for the same signal. Wireless user devices tend to be compact in size with low complexity and power which makes previous diversity methods unfeasible. To deal with this problem, the spatial diversity has effectively exploited recently. Networks which achieve spatial diversity with single antenna device are called cooperative networks [113].

Although mobile cellular networks are the main target of cooperative diversity techniques, any wireless network affected by fading can use them. Since the benefit increases as more intermediate devices are in the network, dense sensor networks

represent a good application scenario for low complexity cooperative techniques. There are multiple options from information theory to improve cooperative diversity transmission systems. The first option is to increase the number of intermediate devices. Beside, the restriction of orthogonal transmission may be removed using MIMO coding techniques, it is possible to allow several transmissions simultaneously, or the source to transmit new information while previous information is being relayed by intermediate nodes [47].

There are three levels of cooperation (applying cooperative diversity): cooperation can be between base stations to serve users more efficiently, like in the coordinated multi-point cooperative (CoMP). The CoMP is an attractive technique to enhance the service offered for far users (cell edge users), by allowing multiple BSs to perform coordinated beamforming towards these users [5]. Another form of cooperation is between BSs and users, where BSs or users can act as relays between source and destination. The third level is cooperation between users like the case in D2D, where users can communicate directly without the BS, and intermediate users act as relays.

Cooperative communication is well-studied and has capabilities to enhance the performance of wireless networks [125]. The cooperation can be in power, computation and other forms. One form of cooperation that has gathered significant attention is relay cooperation, which is the main focus of this thesis work.

1.2.1 Cooperative Relay Networks

In non-relay-aided communication networks, sources (e.g. BS) transmit signals directly to destinations (e.g. mobile station (MS)) without assistance. In general, intermediate nodes may enhance performance by tackling path loss and shadowing. In the 1970s, the idea of cooperation was presented under the name relay channel [126], where an intermediate node act as a relay receives data from the source and re-sends it to the destination, Fig. 1.3 illustrates a typical 3-node cooperative relay network consisting of a source S, a relay R and a destination D. Relaying can take place either by user cooperation, where strong users act as relays or through dedicated relays. Relay nodes are essential in both wired (repeaters) and wireless

networks. In wireless networks, any node decodes nearby communication can act as a relay.

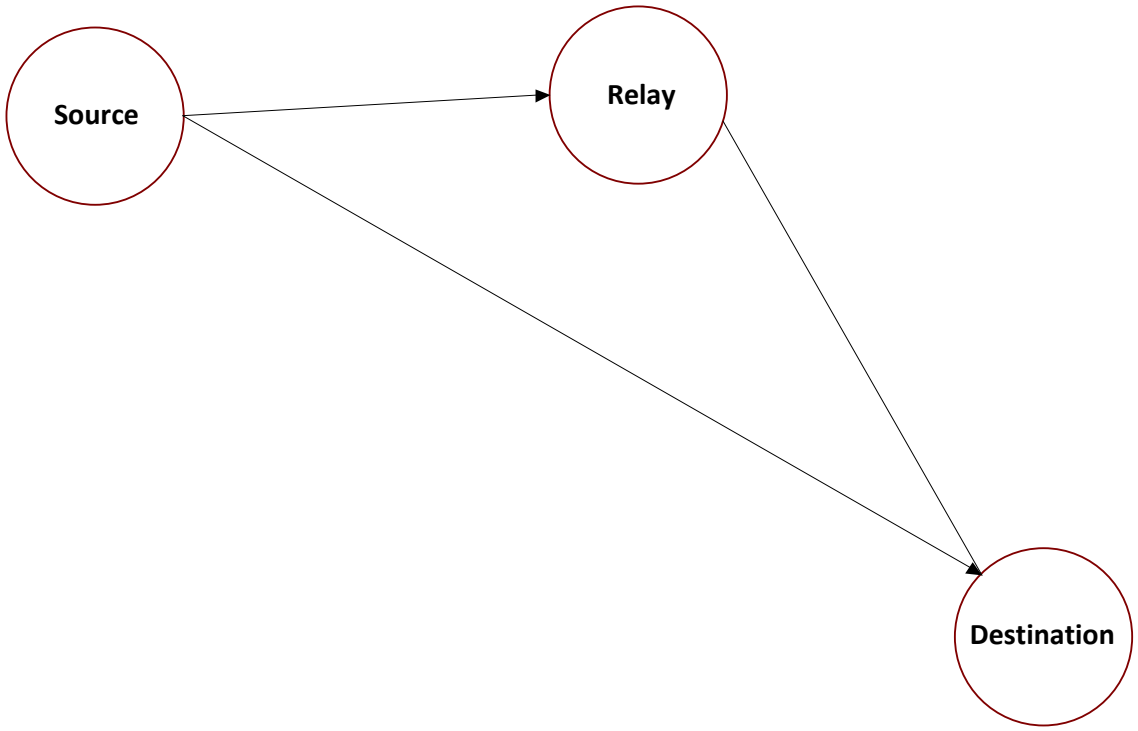

Figure 1.3 Typical 3-node cooperative relay network.

The main advantages of relay cooperation [95] are as follows:

1. Increasing diversity since more independent paths are available, and data can be transmitted over any of these paths.

2. Reducing path loss, as the distance between source and destination is reduced.

3. Mitigating shadowing, since placing relays in the right position could help in avoiding obstacles.

A comparison in the outage probability at different target rates between the source-to-destination path (direct) and the relay path is depicted in Fig. 1.4. In [126], outage probability is defined as the probability that channel capacity is lower than the target data rate. Inserting a relay R between source S and destination D makes

the path loss in each the resulting two links S→R and R→D, lower than the path loss of S→D. This is why the relay path has a better outage performance.

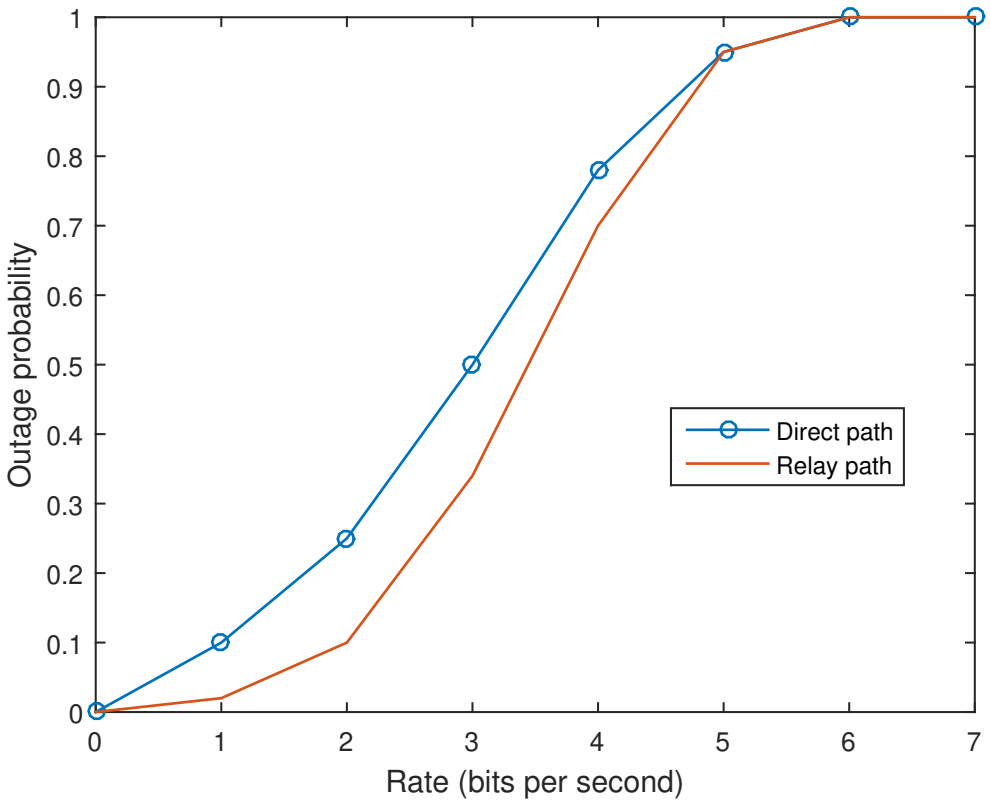

Figure 1.4 Direct path vs relay path $SNR = 10$dB for all links [126].

If the channel gains for all links in Fig. 1.3 network are known. In direct communication, S uses the full slot to transmit its packet to S over link S $\rightarrow$ D . In conventional relaying (no buffering), S uses the first half of the slot to transmit its packet to R over link S $\rightarrow$ R. If R can successfully decode the packet, it re-encodes and transmits it to D in the second half of the slot over link R $\rightarrow$ D.

In this thesis, transmission time duration is partitioned into time-slots with equal length of t. In each time-slot, the source or the relay is selected to transmit packets. The source or the relay assembles information symbols intended for transmission into a packet with size of ηBt bits and transmits it to the relay or the destination where η denotes the target transmission rate and B denotes the bandwidth of the system. For example, if $\eta = 1$bps, $B = 1$Hz and time slot duration $t = 1$s, then the packet size is 1 bit.

When multiple relays are employed, due to the broadcast nature of wireless transmissions, other relay nodes may receive the signal from the transmission by S and can cooperatively relay it to D. The destination now receives multiple copies/signals and can use all of them jointly to decode the packet. Since these signals have been transmitted over independent paths, the probability that all of them have poor quality is significantly smaller. Cooperative communication protocols take advantage of this spatial diversity gain by making use of multiple relays for cooperative transmissions to increase reliability and/or reduce energy costs. Compared to the single relay case, multiple relays guarantee better outage performance [64, 141, 124].

One way of multiple relay cooperation is done by making all relays participate in cooperation using space-time coding, analogous to maximal-ratio-combine (MRC) in MIMO transmission. Multiple single antenna relays achieve spatial diversity similar to MIMO [90]. The difficulties this method causes [142] are:

- It requires orthogonal transmission, which is inefficient in bandwidth usage.
- It demands that channel state information (CSI) and control information of each link being exchanged between all nodes (each relay requires this information about other relays).
- It requires that perfect synchronization among relays has to be preserved.

Another way to perform the relay cooperation, which overcomes the mentioned difficulties, is relay selection.

1.2.1.1 Relay Selection

Relay selection was proposed as an efficient and practical scheme to exploit spatial diversity in [25]. In relay selection, only one relay (the best relay) is selected to perform the cooperation. The impact of the relay selection cooperation on outage probability (outage definition similar to [126]) is depicted in Fig. 1.5. The relay selection is usually performed based on CSI. Recently, buffering capability was added to relays. This has improved the performance of cooperative relay

network. The superiority of buffer-aided relays compared to non-buffer-aided relays in throughput and diversity gain was shown in [61]. This thesis extends that superiority even further. In buffer-aided relays, another factor, other than the CSI, is considered into relay selection that is the buffer state information (BSI). Recent relay selection schemes are discussed in Chapter 2.

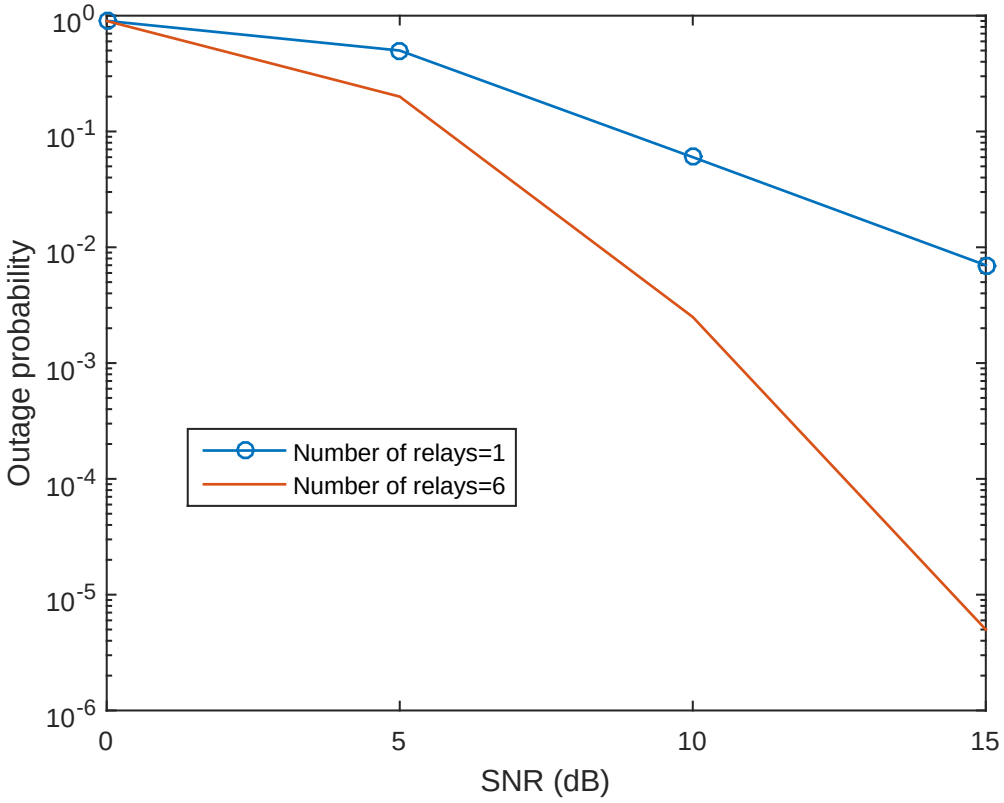

Figure 1.5 Outage probability for selection cooperation, rate$= 1bit/sec/Hz$ and channels gain$= 1$ [25].

1.2.1.2 Relay Protocols

Based on how the relay process the source data upon receiving it, the relay has two main protocols: decode-and-forward (DF) and amplify-and-forward (AF). Authors in [113] investigated the DF technique in the form of user cooperation. After that, the DF heavily studied in the literature. The DF relay acts as a repeater: first, the relay receives data from the source then the relay decodes the received data. Next, the relay regenerates and retransmits the original data towards the

destination; this requires nodes with sufficient processing power to decode and re-encode received signals. The DF technique is popular because of its immunity against error propagation.

On the other hand, the AF technique was suggested in [70]. As the name suggests, the AF relay amplifies the received data from the source and then retransmits it to the destination. In this technique, the noise is amplified with the signal; however, in some cases, where nodes have low-level processors, the AF still can be the right choice.

Two other relay protocols are based on how a relay transmits and receives data, it is either termed as full-duplex (FD) or half-duplex (HD) relay. In the FD relays, the same channel is used for reception and transmission, which leads to higher throughput. Nevertheless, this causes loop interference from the output of the relay to its input, which increases the implementation complexity [106, 107]. On the other hand, the HD relay is transmitting and receiving on two orthogonal channels. Although orthogonal transmission reduces spectral efficiency, it makes the HD relay implementation much more straightforward compared to the FD [13, 104, 42]. Relay cooperation can be performed under different multiple access (MA) schemes.

1.2.2 Relay Cooperation in Multiple Access

Multiple access allows multiuser to use resources more efficiently and effectively [39]. A significant part of the available relay cooperation studies was performed with traditional orthogonal multiple access (OMA) schemes.

1.2.2.1 OMA

Traditional multiple access schemes require orthogonality among signals, which prevents mutual interference and lowers the complexity of receivers. As shown in Table **??**, several types of orthogonal multiple access (OMA) schemes evolved with wireless generations. Frequency division multiple access (FDMA) presented in 1G, and time division multiple access (TDMA) is 2G, while code division multiple

access (CDMA) is 3G, and orthogonal frequency division multiple access (OFDMA) is 4G.

All mentioned advantages of relay cooperation in Subsection 1.2.1, were exploited with OMA schemes during the last decade. However, adding buffering capabilities to relays has revived the research in this area. Several relay selection schemes for buffer-aided relay with OMA have been proposed recently.

As mentioned in Section 1.1, the orthogonality constraint of OFDMA (4G) limits the number of users to the number of orthogonal resources, and it causes high latency [34, 146, 147, 127, 91]. Therefore, other MA schemes such as NOMA have been suggested for 5G. Relay cooperation with NOMA is discussed in Chapter 3; thus, a discussion about NOMA is presented next.

1.2.2.2 NOMA

NOMA is fundamentally different from conventional OMA schemes, in NOMA, multiple users are allowed to transmit at the same code, time and frequency, but with different power levels. Specifically, NOMA assigns less power to users with better channel state, andsuch users decode their own information by applying successive interference cancellation (SIC). As a result, such users will know the messages directed to other users; such prior information can be exploited to improve to other users performance through cooperation [40].

NOMA connects more users in the same resource block, such as frequency band or time-slot, but with different other aspects such as levels of power [81]. Exploiting channel differences amongst users makes NOMA a promising technique to improve spectral efficiency [38]. Duo to the advancement in the signal processing field, NOMA becomes achievable with complicated receivers, which can remove mutual interference caused by sharing the same resources such as SIC. One of the well-studied types of NOMA is the power-domain NOMA. Power-domain NOMA is performed by assigning higher levels of power to users with poorer channel conditions (weak user) [54]. In the rest of this thesis, the term NOMA refers to power-domain NOMA.

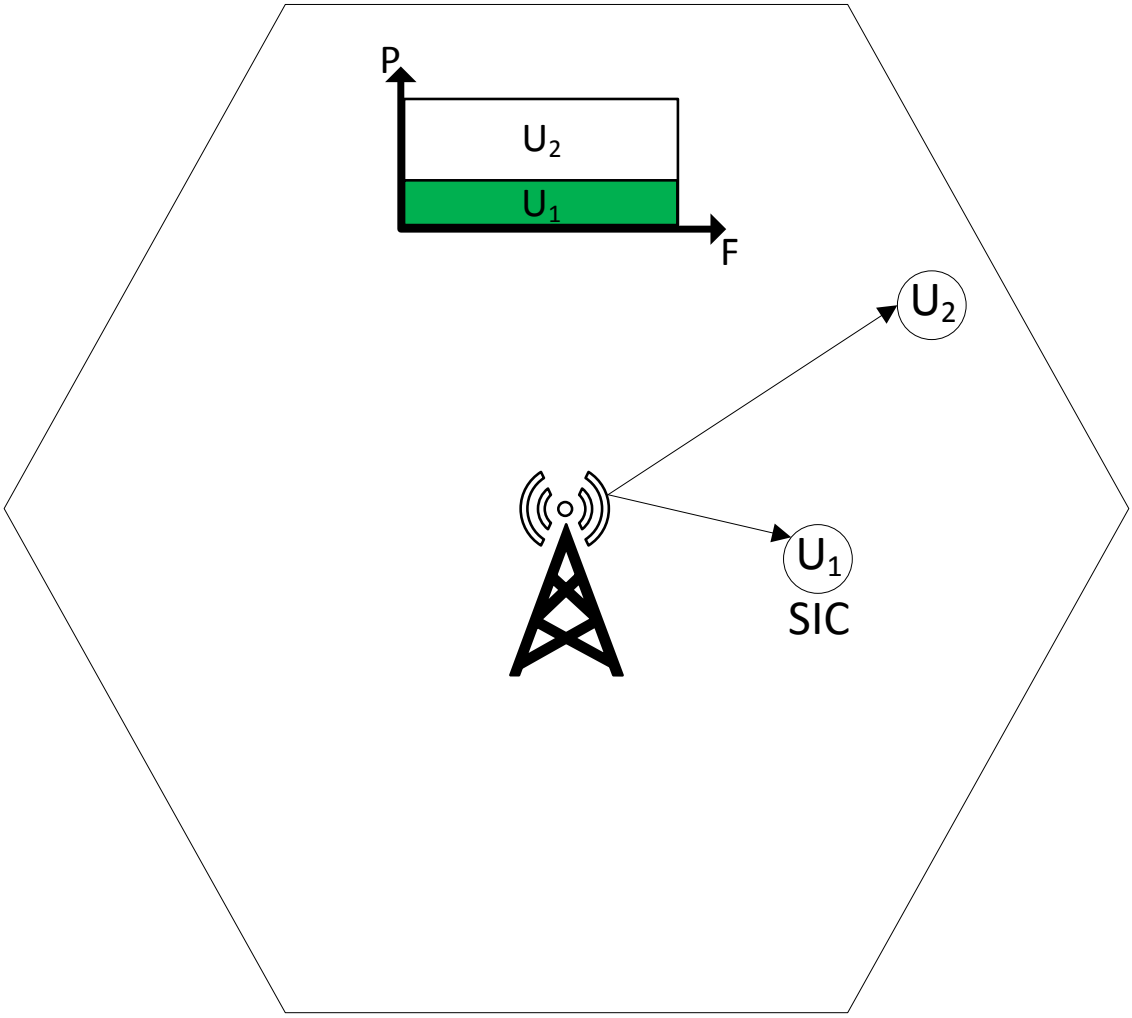

Figure 1.6 Down-link NOMA for two-users case.

Fig. 1.6 illustrates the down-link (DL) NOMA network for two users U_1 and U_2. The SIC is added to the strong user (U_1) receiver. In a two users DL NOMA system, decoding in the SIC begins with the signal corresponding to the weak user (U_2), which has smaller $\frac{|h|^2}{N}$, where h is the channel gain, and N denotes the noise and the interference power. After decoding U_2 signal, the SIC subtracts U_2 signal from the original signal, and the remaining part is U_1 signal. Since higher power is allocated to the weak user, the weak user can cancel the strong user signal (treating it as interference) and retrieve its signal successfully.

For a comparison with OMA, authors in [19] gave an example of two users, DL system comparisons, the difference between the strong user and the weak user in terms of their $\frac{|h|^2}{N}$ is 20 dB. In OMA, both users get the same power level P, with 0.5

Hz of bandwidth is assigned to each users. While in NOMA, $0.8P$ is allocated to the weak user and $0.2P$ for the strong user, and both users share the full bandwidth (1 Hz), both the strong and the weak users achieve 32% and 48% higher rates with NOMA. Since the strong user is bandwidth-limited, doubling the bandwidth raises its rate. In contrast, the weak user is power-limited, hence giving it a lower power ($0.8P$) is not preferable unless higher bandwidth is available. In some scenarios, applying NOMA is more significant such as:

- If one of the users is in deep fading; hence, if OMA is used, one subcarrier is wasted.
- Another scenario is the IoT devices. The IoT devices require a low data rate, so assigning one subcarrier to each device is not efficient.

The superiority of NOMA sum rate (of all users) over OMA, when the channels of all users are asymmetric, was also mathematically proved in [32]. Fig. 1.7 and Fig. 1.8 show the user data rate comparison between applying OMA and the DL NOMA for a two-users case, with symmetric and asymmetric channels, respectively.

For up-link (UL) NOMA, the SIC is installed at the BS. Both strong and weak users transmit their signals to the BS at the same time and frequency. The strong user has better channel conditions and lower path loss, hence, its signal arrives at the BS with higher power than the weak user signal (transition power of all nodes are equal). Therefore, SIC at BS starts with decoding the strong user signal then subtract it from the superposed signal; hence, the remaining part represents the weak user signal [137, 12].

Experimental trials in [19, 16] has also proved that DL and UL NOMA achieve higher throughput than OMA by exploiting the available spectrum more efficiently. Several other experiments on NOMA performance evaluation have been done in different labs such as Huawei Technologies and others; their experiments supported the superiority of NOMA over OMA, further details are in [18, 17, 110, 111, 108, 109]. It is worth mentioning that a new 5G broadcasting standard in the USA ATSC 3.0 has applied the concept of power-domain NOMA successfully in the physical layer technologies [34].

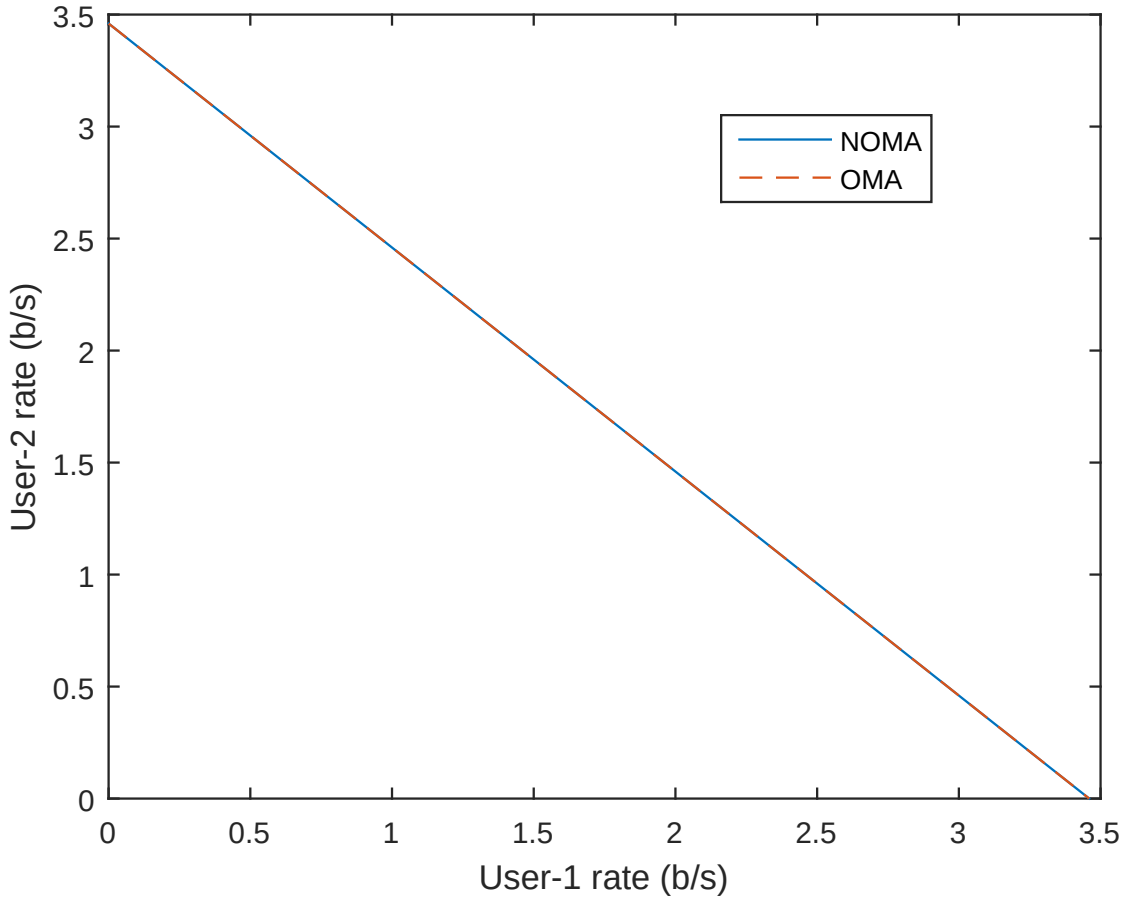

Figure 1.7 User data rate comparison between applying OMA and the DL NOMA, SNR= 10dB for both users (symmetric)[55].

The main advantages of NOMA over OMA can be summarized as follows:

- NOMA improves spectral efficiency, which causes higher throughput. It achieves that by allowing different users to share the same time and frequency resources non orthogonally. Simulation and excremental results show that NOMA may achieve about 30% higher throughput than the OFDMA [18, 17, 34].

- NOMA achieves massive connectivity because it relaxes OMA constraint that the number of users is restricted to the number of the orthogonal channels.

- NOMA reduces transmission latency. In contrast to NOMA, OMA users have to send a scheduling request to the BS, and data transmission can not start

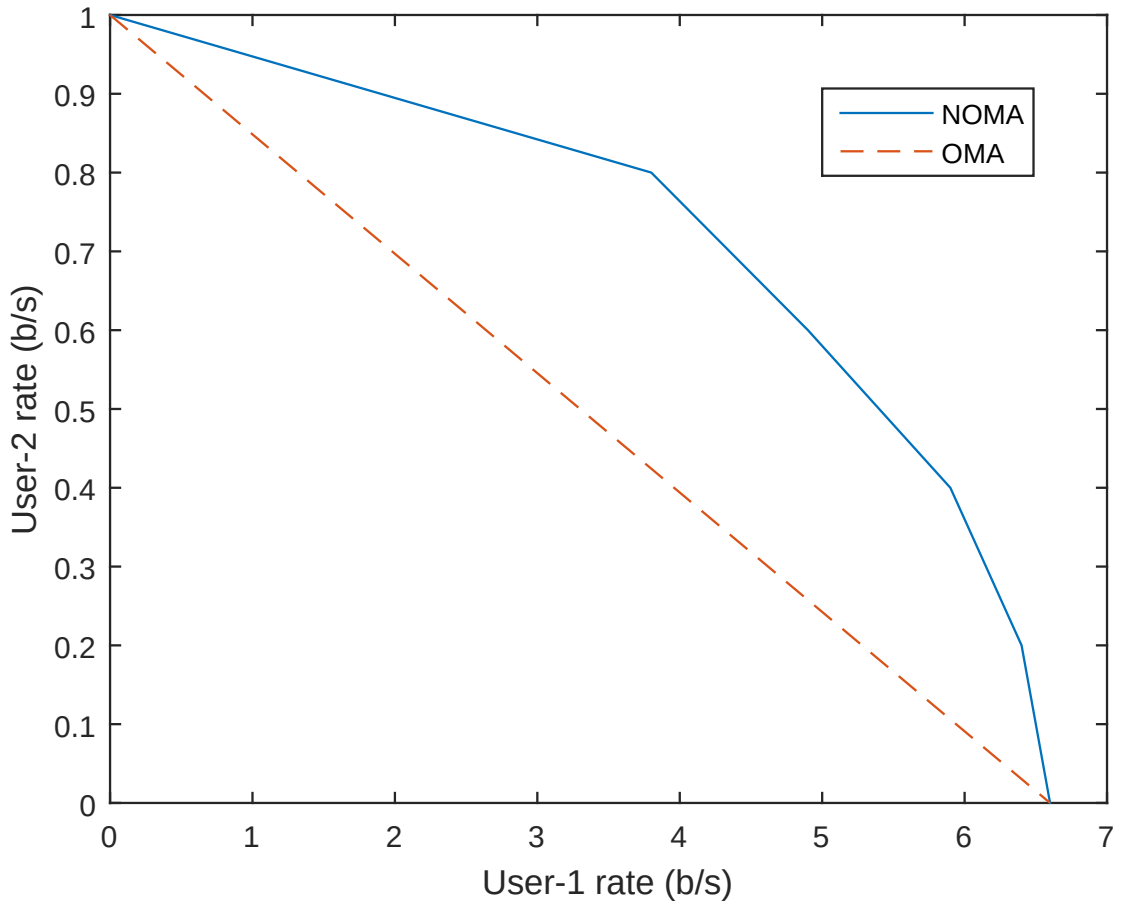

Figure 1.8 User data rate comparison between applying OMA and the DL NOMA, SNR= 10dB for both users (asymmetric)[55].

before receiving the schedule. This process causes a delay of 15.5 ms in the 4G LTE [10]. Such scheduling can be avoided with NOMA.

In order to grasp better understanding of buffer-aided relays cooperative networks, wireless channels which connect all communicating nodes need to be discussed. Also, Markov process which is used in modeling various buffer states and how they are related is also discussed next.

1.2.3 Wireless Channel

A key characteristic of the mobile wireless channel is the variations of the channel strength with time and over frequency. This variations are known as fading and modeled as a random process. Fading is caused by various variables such as time, position, and frequency. The presence of reflectors surrounding a transmitter and

receiver create multiple paths for signals. And the receiver gets the superposition of multiple copies of the transmitted signal. Each signal copy experiences different attenuation, delay and phase shift in its path to the receiver. This results in either constructive or destructive interference. Strong destructive interference is frequently referred to as a deep fading and causes failure in communication due to severe drop in the SNR. Fading channel models are often used to model the effects of electromagnetic signal transmission over the air in cellular networks [123].

Slow fading occurs when the coherence time of the channel is large compared to the application delay requirement. The variations imposed by the channel is considered constant during transmission. Slow fading can be caused by shadowing, where a large obstacle such as a hill or large building object the main signal path between the transmitter and the receiver. While fast fading occurs when the coherence time of the channel is small compared to the delay requirement of the application. In this case, the signal variations imposed by the channel varies considerably during transmission. In a fast-fading channel, the transmitter may take advantage of the variations in the channel conditions using time diversity to help increase robustness of the communication to a temporary deep fade, because the transmitter exploits multiple realizations of the channel within its delay constraint. In a slow-fading channel, it is not possible to use time diversity because the transmitter sees only a single realization of the channel within its delay constraint. A deep fade therefore lasts the entire duration of transmission and cannot be mitigated using coding [123, 2].

The coherence bandwidth measures the separation in frequency after which the signal experiences uncorrelated fading passing the channel. In flat fading, the coherence bandwidth of the channel is larger than the bandwidth of the signal. Therefore, all frequencies of the signal experiences the same magnitude of fading. In frequency-selective fading, the coherence bandwidth of the channel is smaller than the bandwidth of the signal. Therefore, different frequency components of the signal experience uncorrelated fading. OFDM divides the wide-band signal into many narrow-band sub-carriers, each exposed to flat fading rather than frequency selective fading [86].

When the decoding error probability can not be made small to a level where error correction methods can help in retrieving the original signal, the system is said to be in outage. Reliable communication can be achieved otherwise. There is a conceptual difference between the AWGN channel and the slow fading channel. In the AWGN, one can send data at any rate less than the channel capacity while making the error probability as small as required. This cannot be done with the slow fading channel as long as the probability the channel is in deep fade is above zero. Thus, the capacity of the slow fading channel in zero outage is zero. For fading distributions for which the fading coefficient can be arbitrarily small (such as for Rayleigh, Rician, or Nakagami fading), the probability of an outage is positive. For applications in which outage probability larger than zero is acceptable, the maximal achievable rate as a function of the outage probability (also known as outage capacity) may be a more relevant performance metric than Shannon capacity. The outage capacity is the largest achievable rate under the assumption that the outage probability is less than a certain threshold [46].

In a slow fading channel, the duration of each of the transmitted packets is smaller than the coherence time of the channel, so it is reasonable to assume that the random fading coefficients stay constant (flat) over the duration of each packets. This slow and flat fading channel during the packet transmission (and varies from one packet to another) is known as quasi-static fading channel. If the quasi-static channel model is reasonable, the outage capacity is a meaningful performance metric. In this thesis, the assumption is that the channels are quasi-static fading channels. This is a valid assumption because this channel model characterizes practical settings that experience slow fading conditions, such as fixed wireless access point [28]. Rayleigh channel means that the channel follows Rayleigh probability density function which may be flat or selective within each packet transmission. Quasi-static Rayleigh channel means that the channel has flat fading during each packet transmission and this flat fading during each packet vary independently from one packet to another. It is worth mentioning that Rayleigh is reasonably models wireless system when no dominant path (LOS) is available [134, 51].

Many real-time applications (e.g., voice) have stringent delay constraints and fixed rate requirements. In slow fading environments (where decoding delay is of the order of the channel coherence time), it may not be possible to meet these delay constraints for every packet. However, these applications can often tolerate a certain fraction of lost packets or outages. A variety of techniques are used to combat fading and meet this target outage probability such as exploiting diversity through cooperation [124]. Finally, the transmitter can track CSI. There are several ways in which such channel information can be obtained at the transmitter (CSIT). In a time-division duplex (TDD) system, the transmitter can exploit channel reciprocity and make channel measurements based on the signal received along the opposite link. In a frequency-division duplex (FDD) system, there is no reciprocity and the transmitter will have to rely on feedback information from the receiver, which means more overhead on the communication system. However, full CSI (CSI at both transmitter and receiver) enhances the system achievable rate [123].

1.2.4 Markov Process

The Markov process is a probabilistic model attractive for analyzing complex systems. The main concepts of Markov process models are the concepts of state and state transition. The explanation of state is as follows, a physical system can be described by a number of variables that describe the system. For example, a chemical system can often be described by the values of temperature, pressure, and volume. Such critical variables of a system are called state variables. When the values of all state variables of a system are known, then its state has been specified. The state of a system thus represents all we need to know to describe the system at any instant. In the course of time a system passes from state to state and thus exhibits dynamic behavior. In the chemical system such changes are caused by the application of heat, an increase in volume, etc. Such changes of state are called state transitions. The most general state transition model would allow states described by continuous variables and transitions that could occur at any time. While practical buffer (i.e. finite size buffer) content belongs to systems that have only a finite number of states. The time interval separating transitions in

buffer-aided relay is assumed to be fixed, it can be generalized to a random process. However, we are interested not in the time pattern of transitions, but in the buffer state after successive transitions. Noticing that being at any state is random [57].

The Markovian assumption greatly simplifies both the possible behavior of the process and the problem of specifying the process. That is only the last state occupied by the process is relevant in determining its future behavior. Thus, the probability of making a transition to each state of the process depends only on the state presently occupied. The Markov process is an extremely useful model for wide classes of systems ranging from genetics to inventory control. When Markovian assumption can be justified, then researchers can enjoy analytical and computational convenience not often found in complex models [21].

To define a Markov process we must specify for each state in the process and for each transition time the probability of making the next transition to each other. A matrix whose elements cannot lie outside the range $[0, 1]$ and whose columns sum to one is called a stochastic matrix; thus the transition probability matrix that defines a Markov process is a stochastic matrix. The interesting thing about Markov process that the possibility of knowing the process is at which state after a specific number of transitions [68].

The Markov chains (discrete-time process) are stochastic processes defined only at integer values of time. At each integer time the process is at a specific state. For the countably infinite case, the most common applications come from queueing theory, where the state often represents the number of waiting customers, which might be zero. A popular application of Markov chain is birth–death Markov chain. Birth–death is a Markov chain in which the state space is the set of non-negative integers, the transition probabilities are non-negative. A transition from state to a larger state is regarded as a birth and to smaller state as a death. Thus the restriction on the transition probabilities means that only one birth or death can occur in one unit of time. Many applications of birth–death processes arise in queueing theory, where the state is the number of customers, births are customer arrivals, and deaths are customer departures [45].

1.3 Buffer-Aided Relay in 5G and Beyond 5G

Since the superiority of buffer-aided relays compared to non-buffer aided relays has been proved in several studies like [61], any benefits from applying non-buffer aided relays in 5G are expected to become even better with buffer-aided relays. Therefore, any promising application of non-buffer aided relay in 5G is considered as a promising application of buffer-aided relay in 5G.

In the LTE-Advanced protocols, relay standards were detailed. However, the role of relays in these networks was limited. For example, cognitive relaying and D2D relaying were partly investigated but their implementation was left for the next generations of wireless networks 5G and beyond 5G [58]. There are several future communication areas where buffer-aided relay can play an important role: High mobility should be supported in 5G networks. As the position of the relay and the users changes fast, channel estimation is performed harder and relays with accurate and fast sensing offer superior performance. Moreover, the increased coverage offered relays result in less handovers and possible outages. In many 5G and beyond 5G applications, delay is the most critical concern e.g. emergency applications like remote surgery. The delay is usually defined as the time required for a packet to reach the destination after it is transmitted by the source. To this end, relays must be capable of handling delays by increasing the throughput. Regarding economic and environmental sustainability, green communications is a major research topic in the next wireless generations. In this field, relay aiming at efficient power usage, leads to reduced public exposure to electromagnetic fields [96].

Buffer-aided relay can be integrated with other 5G techniques to achieve higher performance. D2D and millimeter wave are two important 5G techniques which buffer-aided relays can improve their performance. As discussed in Section 1.1.2, in 5G D2D communication, devices are able to communicate either directly or by partial involvement of BS, if both devices are within the proximity area. D2D communication enhances cell throughput, especially at cell edge users where signals are much weaker and devices are not able to communicate directly with

the BS. This can offload data from base stations by direct transmission between mobile devices, which makes D2D a promising technique for 5G and beyond 5G wireless networks. When the devices are not in their proximity area and they want to communicate, they choose other devices (relays) for their communication, this is known as multi-hop communication. In multi-hop communication, the devices communicate with the help of relays. Relays in D2D networks can further reduce the energy consumption of mobile devices, enhance the quality of data transmission, assist connection establishment among devices, and increase the range of D2D communication [88, 83]. These benefits that relay brings to D2D are crucial in the next wireless generations.

And one of the most important techniques for 5G and beyond 5G is millimeter wave. The main challenge for millimeter wave is that it is severely effected by path-loss and shadowing. Relays can be used to route around blockages and extend the millimeter wave link. Several studies have shown the positive impact of applying relay with millimeter wave on coverage and spectral efficiency [89].

1.4 Thesis Outline

The outline and the organization of this thesis are as follows:

In Chapter 2, some of the main challenges that buffer-aided relay faces are introduced. Meanwhile, this chapter presents a literature review of different state-of-the-art schemes that have been employed to tackle the challenges of using the buffer-aided relays. Additionally, this chapter highlights some main shortcomings and weaknesses of the available schemes.

It has been shown in the available literature, applying buffer-aided relays can further increase the throughput and diversity order in NOMA cooperative network. However, the full potential of using the buffer-aided relay is not achieved. Therefore, in Chapter 3, we propose a novel prioritization-based buffer-aided relay selection scheme, which is able to seamlessly combine NOMA and OMA transmission in the buffer-aided cooperative relay network. The proposed scheme shows improvement in the data throughput of the system, and it increases the diversity order.

Although exploiting buffering capabilities can tremendously enhance the performance of the cooperative relay networks, this enhancement can be at the expense of lengthening the packet delay. In Chapter 4, a new factor, the source delay affecting the average packet delay, which was not considered in previous studies, is thoroughly studied. The importance of considering the source delay is making the comparison between the buffer-aided and the non-buffer-aided relays more accurate. Also, accurate delay calculation is very important in the 5G delay sensitive applications. The buffer-aided relays show a better source delay performance than the non-buffer-aided relays, hence, it has a better average packet delay in some cases, especially at low SNR range.

Studying the degradation on the buffer-aided performance caused by constraining the packet delay by a specific target delay is necessary. Because all the packets that exceed the target delay are discarded or retransmitted, which causes performance degradation. In Chapter 5, we introduce the delay-constrained outage probability to study the impact of imposing a target delay on the buffer-aided relay outage performance. In addition, we propose an adaptive buffer-size algorithm to enhance the outage performance under delay constraints.

Finally, Chapter 6 concludes the thesis with a discussion of the main remarks and possible future research challenges.

1.5 Original Contributions

The main contributions of this thesis are focussed on the improvement of buffer-aided relay application in 5G and beyond 5G cooperative networks. The specific contributions of each chapter are listed below and supported either by international journal or conference paper, all publications can be found in the publication list.

Chapter 3

This chapter features a novel buffer-aided relay selection scheme for multiple relay cooperative NOMA networks. The new selection scheme led to a better system

throughput, and the diversity order of the new scheme is higher than the available schemes in the literature. It also shows the importance of selecting a proper value for target length based on the application requirements [8].

Chapter 4

This chapter shows the impact of considering the source delay on the average end-to-end delay. Numerical simulations show that buffer-aided relays can beat non-buffer-aided relays in average packet delay in some cases, especially at low SNR range. It also shows the positive impact of two delay reduction techniques. In addition, a novel relay selection rule is proposed, where the idea of adaptive target length is presented in this rule [9, 6].

Chapter 5

This chapter studies the impact of constraining the delay to a certain target delay on the system performance. Specifically, delay-constrained outage probability is derived by Trellis diagram and Markov chain. It also suggests a novel modification on the available relay selection schemes to enhance their performance in the delay-constrained applications. Numerical simulations show that the altered schemes are superior compared to their original forms [7].

Chapter 2

Literature Review

The objective of this thesis is to study the buffer-aided cooperative relay networks. Specifically, the goal is to suggest improvements for the main performance metrics. In this chapter, the challenges associated with buffer-aided relays that degrading their performance and the relevant attempts to tackle these challenges are discussed. This chapter also discusses certain shortcomings and limitations corresponding to the available solutions in the literature.

2.1 Cooperative Relay Selection

As presented in Chapter 1, cooperative relay has gains in communication systems which was proved experimentally in [84]. Relay selection is one of the simplest methods to achieve these gains; hence, many relay selection schemes were suggested in the literature.

Relay selection started first with non-buffer-aided relays, which are also known as conventional relays. In conventional relays, the relay receives a packet from a source in one time-slot and then transmits the received packet to a destination in the next time-slot. Several conventional relay selection schemes were proposed in [56, 69, 98, 60, 13, 92, 23, 67, 71, 22, 151]. For example, [22, 151] suggest making the selection based on distances between nodes, which requires the geographical knowledge of each node. After that, authors in [87] suggested a different scheme. This scheme is considered the optimal scheme for conventional relay selection; it is

known as the best relay selection (BSR), also known as the max-min scheme [87]. The max-min requires the exchange of the CSI between all nodes and the control node, then the relay with the best end-to-end (from source to destination) channel conditions is selected for receiving data from the source and transmitting it to the destination. This procedure happens in two time-slots. The max-min algorithm is analytically represented by

$$R_{Best} = \arg\max_{R_k} min\{\gamma_{sr_k}, \gamma_{r_k d}\},\ k = 1, 2, ..., K \tag{2.1}$$

where R_{Best} is the selected relay, K is the number of relays, γ_{sr_k} denotes source-to-relay link SNR and $\gamma_{r_k d}$ is the relay-to-destination link SNR. To clarify this scheme, Fig. 2.1 shows a 3 relays network example.

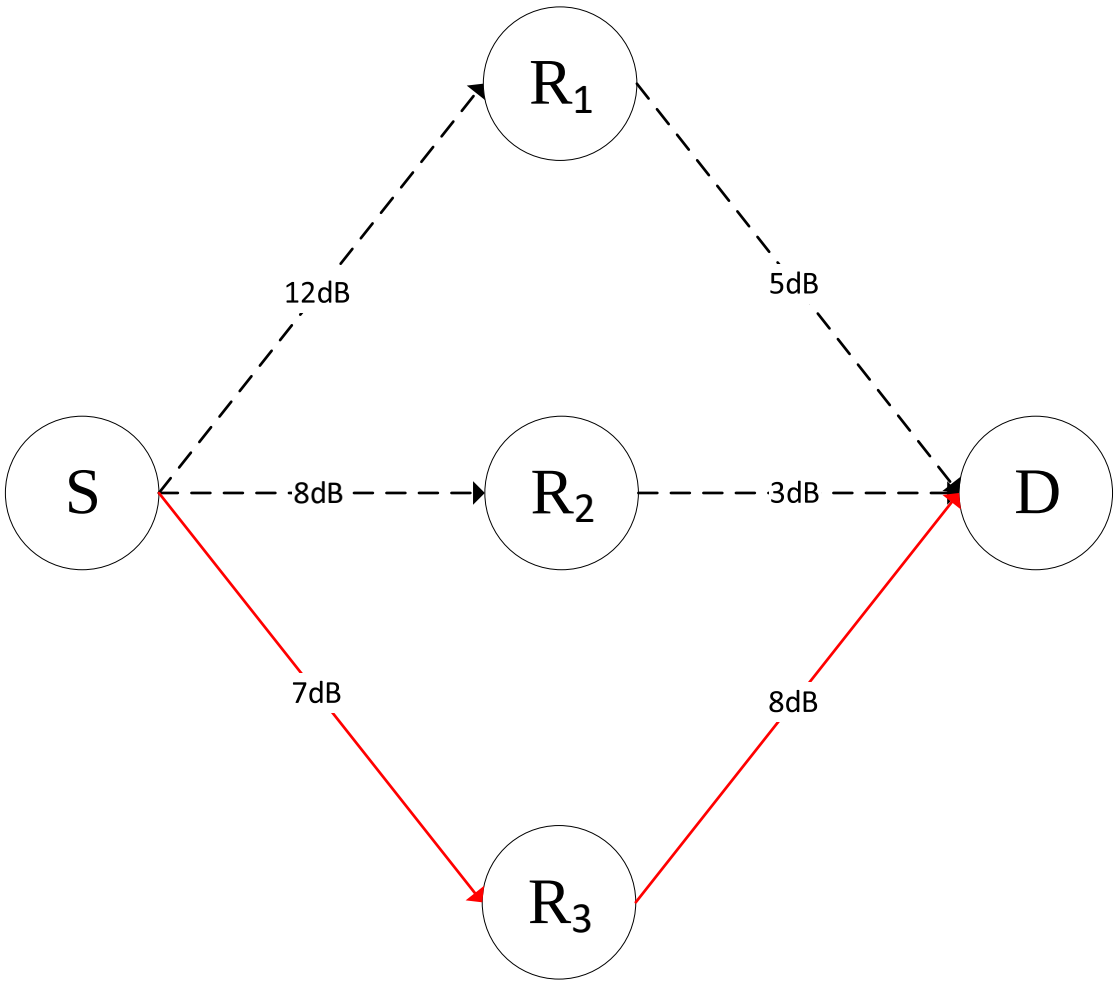

Figure 2.1 Cooperative relay network example with three non-buffer aided relays.

Although $S \rightarrow R_1$ link has the highest SNR 12dB, R_1 is not selected by the max-min scheme. The minimum SNR of $S \rightarrow R_1 \rightarrow D$, $S \rightarrow R_2 \rightarrow D$ and $S \rightarrow R_3 \rightarrow D$ pathes are 5, 3 and 7 respectively, hence, the highlighted path $S \rightarrow R_3 \rightarrow D$ has the highest minimum SNR, thus, R_3 is the selected relay.

The max-min scheme achieves a diversity gain of the number of relays K [25, 23]. Nonetheless, the max-min scheme suggests that the $R \rightarrow D$ link stays fixed for two time-slots, which can not be guaranteed. Hence, if the $R \rightarrow D$ link SNR reduced for any reason (e.g., moving object) and it becomes unable to support the transmission, then the received packet in the first time-slot is dropped.

2.2 Buffer-Aided Cooperative Relay Selection

To relax conventional relay constraint of receiving and transmitting consequently, and provide more flexibility in relay selection, authors in [129] suggested giving relays buffering capability. This improves the performance of the cooperative relay networks since buffer-aided relays can store data and transmit it in favorable channel conditions [61].

Fig. 2.2 shows a typical buffer-aided cooperative relay network with S, K relays R_k and D, it can be seen that the receiving relay does not have to be the transmitting one.

Recently, buffer-aided relays have been considered in many 5G applications for different purposes. Cognitive network is an example. At first, conventional relays were considered in cognitive network, where the relay which causing minimum interference from the secondary network to the primary network is selected. This increases the available frequency for secondary network [139]. After that, further enhancement on cognitive networks performance was achieved by introducing buffer-aided relays. Because of buffering capabilities, relays can keep data when the interference is high rather than dropping it [31]. Another buffer-aided relay application is physical layer security in wireless networks. Buffer-aided relays were suggested for the MIMO network in [78] to maximize the secrecy rate.

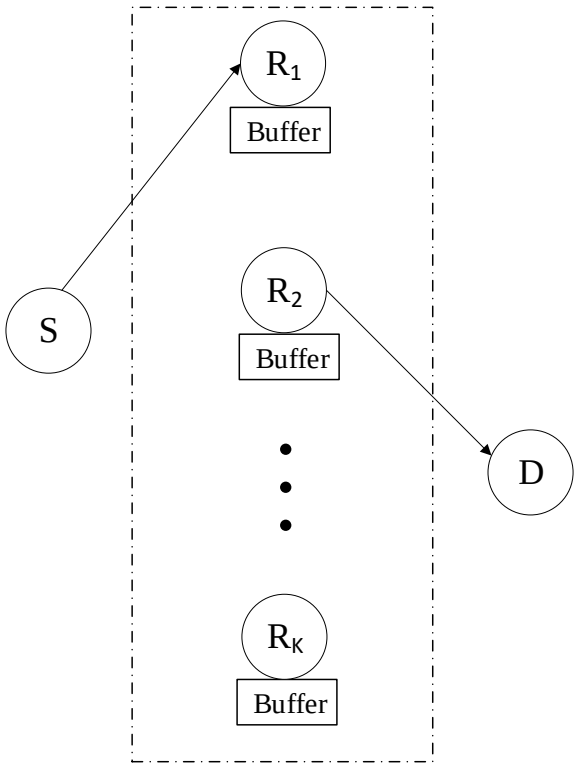

Figure 2.2 Typical buffer-aided cooperative relay network.

Lastly, the majority of the available studies have proved the effectiveness of relay cooperation with OMA. Because NOMA is superior to OMA, researchers have been motivated to study the impact of combining relay cooperation with NOMA. First relay cooperation with NOMA was performed with a single relay in [130]. Results show better throughput and outage performance compared to the non-cooperative NOMA. After that, multiple relays was the logical next step; hence, other studies like [135] have considered multiple relays, which resulted in better performance. After the famous of buffer-aided relay, combining buffer-aided relay with NOMA was another application. Cooperative NOMA with a single buffer-aided relay was proposed in [140, 80, 136]. Result shows performance improvement compared to results in combining NOMA with a conventional relay. It is worth noting that the aforementioned studies focus on the relay cooperation with DL NOMA. Studies on the relay cooperation with UL NOMA have also shown a positive impact on the performance as shown in [65, 132]. Further details about buffer-aided relays in cooperative NOMA is in Chapter 3. With buffer-aided relays, a new era of relay selection schemes has begun, and several schemes have been suggested.

2.2.1 Max-Max Relay Selection (MMRS)

The first relay selection scheme for buffer-aided relays is in [62]; the authors suggested the max-max relay selection (MMRS) scheme. In the MMRS, during the first time-slot, the relay with the best available $S \rightarrow R_k$ link (has the highest SNR) is selected to receive a packet from the source S. In the next time-slot, the relay with the best available $R_k \rightarrow D$ link is chosen to transmit a packet to the destination D. The $S \rightarrow R_k$ link is considered available if the corresponding buffer is not full. The $R_k \rightarrow D$ link is deemed to be available if the corresponding buffer is not empty. The MMRS has more flexibility compared to the max-min since the relay which received a packet does not have to be the same relay for transmission. The MMRS kept the conventional two time-slots scheduling, where the first slot is always for the reception and the second one for the transmission. The MMRS scheme can be represented as:

$$R_{Best}^{receive} = \arg\max_{R_k}\{\gamma_{sr_k}\} \tag{2.2}$$

$$R_{Best}^{transmit} = \arg\max_{R_k}\{\gamma_{sr_k}\} \tag{2.3}$$

where $R_{Best}^{receive}$ denotes the relay selected for the reception, and $R_{Best}^{transmit}$ is the selected relay for the transmission.

Result in [62, 148] shows that similar to the max-min, the MMRS achieves full diversity order (K), but the MMRS achieves higher coding gain than the max-min. Higher coding gain means delivering the same outage at lower SNR. To achieve this gain, the MMRS requires buffers of relays to be neither empty nor full. Because if any relay has a full buffer, this reduces the number of independent receiving paths, so, the diversity gain is reduced. Similarly, when a relay has an empty buffer, the number of independent transmitting paths is reduced. The MMRS deepens the problem of a full and empty buffer, as the buffer corresponding to the best $S \rightarrow R_k$ link tends to overflow because the MMRS keeps selecting it, and vice versa, the buffer of the relay with best $R_k \rightarrow D$ link tends to be empty more often. The case of neither empty nor full buffer can be guaranteed with infinite buffers. However,

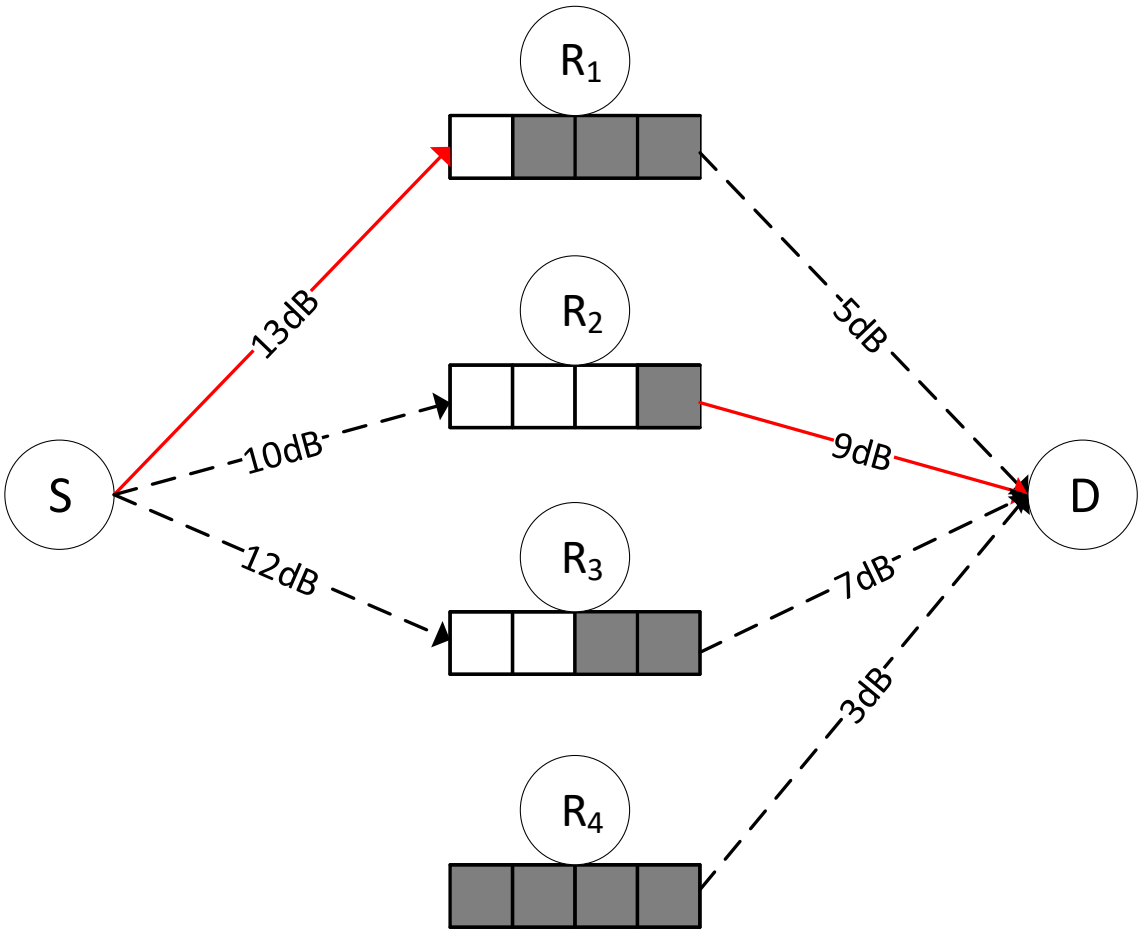

Figure 2.3 The MMRS example.

practical buffers are finite. Hence, this problem needs to be avoided by modifying the selection scheme.

To demonstrate how the MMRS works, Fig. 2.3 is a four relays network example; each buffer can hold four packets. This example is also used in the following schemes. Since $S \rightarrow R_1$ has the highest SNR among $S \rightarrow R_k$ links, R_1 is selected for receiving a packet from S. In the next time-slot, R_2 is the best $R_k \rightarrow D$ link, accordingly, it is selected for transmission. Subsequently, R_1's buffer is full, and R_2's buffer is empty, which reduces the number of available links.

2.2.2 Hybrid Relay Selection (HRS)

To overcome the MMRS limitations, authors in [61] addressed the empty and full buffer challenge by proposing the hybrid relay selection (HRS) scheme. The HRS

applies the MMRS on all links (even links correspond to a full or empty buffer are available for selection). If the selected relay by the MMRS has a full or empty buffer, the max-min is applied on the selected relay, which means the same selected relay has to receive and transmit consequently.

$$R_{Best}^{receive} = \begin{cases} MMRS - R_{Best}^{receive}, & \text{if } q < L. \\ BSR - R_{Best}, & \text{otherwise.} \end{cases} \tag{2.4}$$

$$R_{Best}^{transmit} = \begin{cases} MMRS - R_{Best}^{transmit}, & \text{if } q > 0. \\ BSR - R_{Best}, & \text{otherwise.} \end{cases} \tag{2.5}$$

where q is the buffer length and L is the buffer size. the HRS achieves full diversity order (K), but it does not fully exploit the flexibility of the buffer-aided relays. The HRS follows the traditional two time-slots scheduling. Also, the HRS treats empty or full buffers as non-buffer cases, rather than avoiding them.

An example of the HRS scheme is shown in Fig. 2.4. Because $S \rightarrow R_4$ link has the highest SNR 15dB, R_4 is chosen for reception by the MMRS, but it has a full buffer, hence, R_4 is treated as a conventional relay. So, R_4 has to receive and transmit consequently.

2.2.3 Max-Link Selection

Authors in [68] proposed a relay selection scheme known as the max-link. The max-link selects the best available link, which has the highest SNR, whether it is a $S \rightarrow R_k$ or $R_k \rightarrow D$ link. Hence, the max-link relaxes the two time-slots scheduling and fully exploits the flexibility offered by buffers. The max-link is the starting point for adaptive link selection. The max-link is expressed by:

$$R_{Best} = \arg\max_{R_k} \{ \bigcup_{R_k:q<L} \{\gamma_{sr_k}\}, \bigcup_{R_k:q>0} \{\gamma_{r_k d}\} \} \tag{2.6}$$

In the max-link scheme, selection is performed only on the available $S \rightarrow R_k$ and $R_k \rightarrow D$ links. The total number of the available links may be added to 2K if all

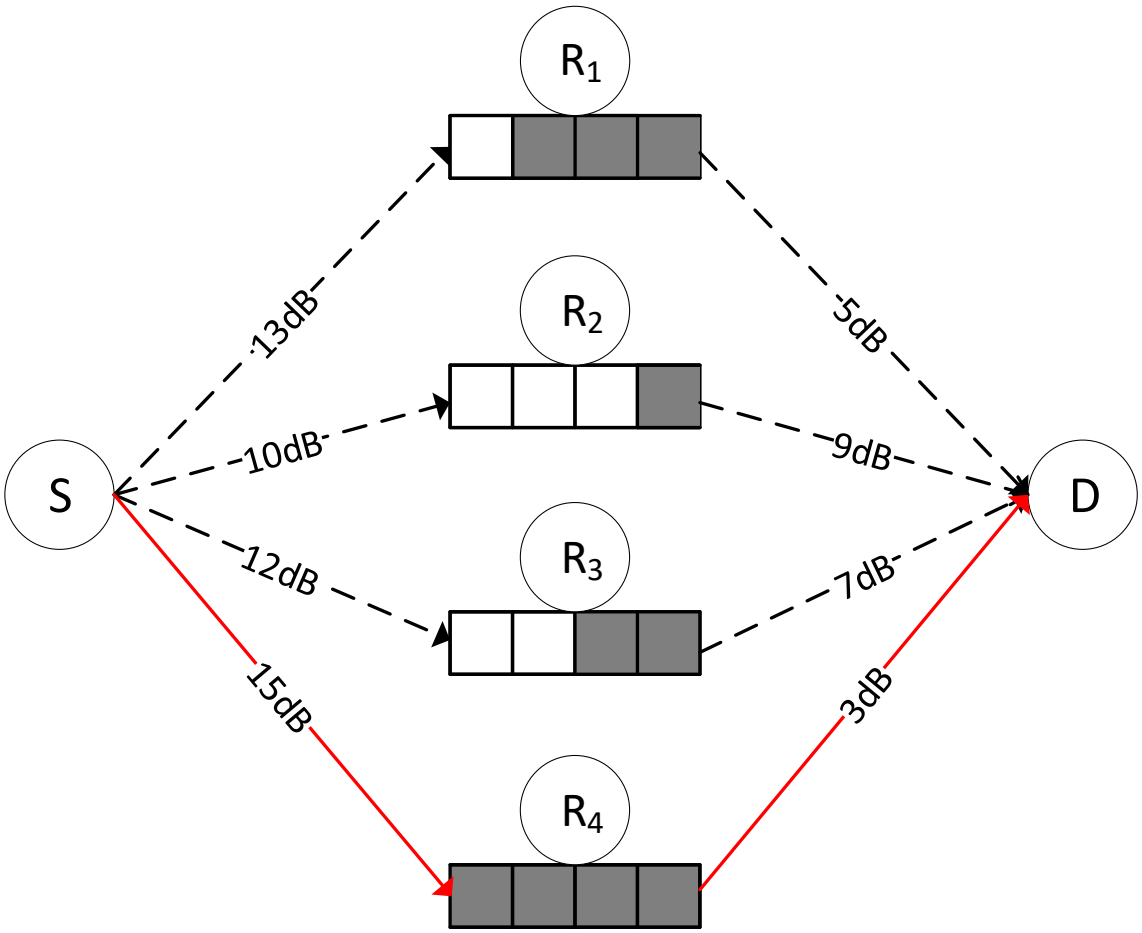

Figure 2.4 The HRS scheme example.

links are available (all buffers are neither full nor empty). Still, it can not be lower than K. Hence, there are $K \rightarrow 2K$ choices. Consequently, the diversity order of the max-link may double the diversity order of previous schemes to 2K. Therefore, the max-link scheme outperforms all the schemes mentioned above in terms of the diversity order. Yet, the empty or full buffer problem is much worse with the max-link. Specifically, if one link has the highest SNR for a while, this leads to an empty or full buffer faster than the MMRS or the HRS.

Although the buffer-aided relays can significantly improve the throughput and the diversity gain compared to the conventional relays, it can worsen the delay by giving packets the ability to reside in the buffer. So, if long queues are not controlled, unacceptable delays can occur. The most critical challenge faces the max-link is the delay. All the aforementioned relay selection schemes have focused on achieving higher diversity or coding gain without considering the delay. However, the max-

link causes longer queues because it gives relays the ability to keep receiving for a while before transmitting. The max-link delay is calculated as follows: by Little's law [72], average packet delay at any relay R_k can be obtained as

$$\bar{D_{r_k}} = \frac{\bar{L_{r_k}}}{\bar{\xi_{r_k}}} \tag{2.7}$$

where $\bar{L_{r_k}}$ and $\bar{\xi_{r_k}}$ are the notations of average queue length and average throughput at the relay R_k respectively. Because all packets are transmitted from one source node, the average throughput at the source node ($\bar{\xi_s}$) is similar to that for the overall system (because what goes in goes out) which is given by $\bar{\xi_s} = \bar{\xi} = 0.5$, where $\bar{\xi}$ is the average throughput of the overall network..

The average queue length at R_k is calculated by averaging the queue lengths over all buffer states

$$\bar{L_{r_k}} = \sum_{i=1}^{(L+1)} \pi_i q_k^{(i)} \tag{2.8}$$

where π is the stationery probability. Because selecting any of the relays has the same probability, the average throughput at the relay R_k is given by

$$\bar{\xi_{r_k}} = \frac{\bar{\xi}}{K} = \frac{1}{2K} \tag{2.9}$$

The delay at any relay

$$\bar{D_r} = 2K \sum_{i=1}^{(L+1)} \pi_i q_k^{(i)}. \tag{2.10}$$

For any fixed-size buffers, the number of packets arrive at all the relays equal to that leave these relays, because no data packet can stay in a relay buffer forever and fail to reach the destination. Thus, we have $P(S \to R) = P(R \to D) = 0.5$, and average queuing length at the source $\bar{L_s} = 0.5$. Hence,

$$\bar{D_s} = \frac{\bar{L_s}}{\bar{\xi_s}} = 1 \tag{2.11}$$

At high SNR, all buffer states are equally likely, this makes $\bar{L_{r_k}} = \frac{1}{L+1}0 + \frac{1}{L+1}1 .... + \frac{1}{L+1} = \frac{L}{2}$, so $\bar{D_r} + \bar{D_s} = KL + 1$, which is proportional to K and L [120].

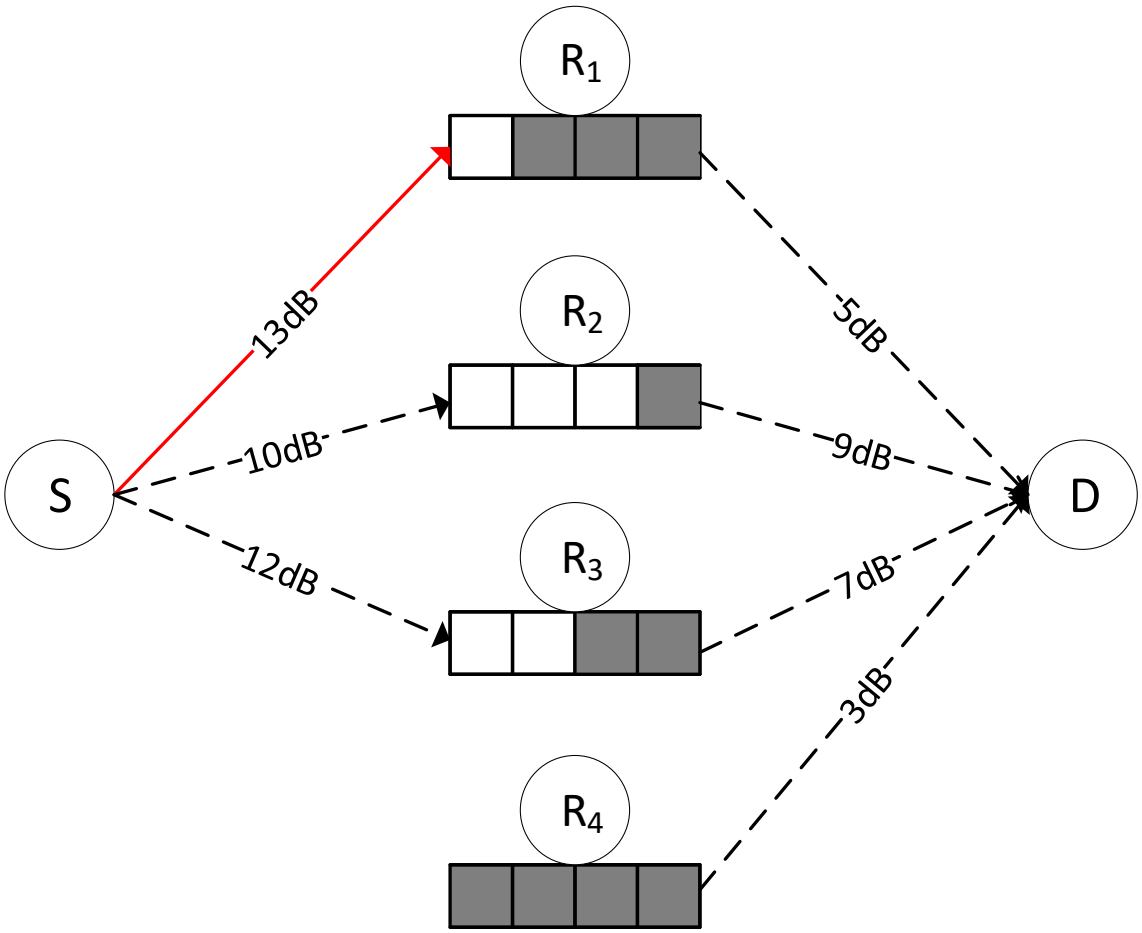

Figure 2.5 Max-link relay selection example.

The example in Fig. 2.5, shows how the max-link selects only one of the available links (with the highest SNR), which is $S \rightarrow R_1$. This makes R_1 buffer full, hence, $S \rightarrow R_1$ link becomes not available. This example represents the worst-case scenario for the max-link, where $S \rightarrow R_k$ links are stronger than $R_k \rightarrow D$ links. Therefore, buffers tend to become full, and long queues are more likely to occur.

Authors in [100], proposed a modified version of the max-link by priorities transmission in odd time-slots and receiving in even time-slots. This equivalent to the MMRS at sufficient SNR levels. While giving relays the ability to transmit in odd time-slot if none of the $S \rightarrow R_k$ links has adequate SNR to the reception, and vice versa. This keeps the max-link diversity gain while minimizing the imbalance between reception and transmission. So, long queues are less likely to occur. However, pre-fixed two time-slots scheduling is needed.

2.2.4 Buffer-State Based Relay Selection (State-Based)

Long queues in the max-link raise the need for queue-control schemes, where the buffer-state information BSI is considered in the selection procedure. The first scheme to consider the BSI is state-based in [79]. The state-based scheme is performed into two stages, in stage one: both the CSI and the BSI are jointly considered in the selection according to the method in Table 2.1.

Table 2.1 Decisions by relays [79]

Case	SR_k link	R_kD link	q_k	Decision
1	out	out	any q_k	Silent
2	not available	out	$= L$	Silent
3	out	not available	$= 0$	Silent
4	suc	out	$< L$	Receive
5	out	suc	> 0	Transmit
6	suc	suc	≥ 2	Transmit
7	suc	suc	≤ 1	Receive

In Table 2.1, "out" denotes the link is in outage, "suc" means the link is not in outage, q_k is the kth relay buffer content, and the relay can receive, transmit or remain in silent. For example, when both links are in outage, the relay remains silent (case 1). If any relay has a full (empty) buffer and it can not transmit (receive), then that relay keeps silent, this is case 2 (case 3). Case 4 (case 5) is a non-full (non-empty) buffer that receives (transmits) a packet. The most remarkable cases are case 6 and case 7. In case 6, a higher priority is given to transmit even with the ability to receive, this case reduces the packet delay. In case 7, receiving has a higher priority than transmitting to avoid empty buffer, which maintains diversity.

In the second stage, the control node (any R that can communicate with other nodes) divides all relays into two groups: transmitting and receiving groups based on stage one. Then the control node finds the relay the has the highest number of packets in the transmitting group. This relays are denoted as L^t_{max}. Similarly, the relays of the lowest content in the receiving group are denoted as L^r_{min}. As a result, there are four cases as follows

1. If $L^t_{max} = L$, then transmit.
2. If $L^t_{max} < L$ and $L^r_{min} = 0$, then receive.
3. If $2 \leq L^t_{max} < L$ and $L^r_{min} > 0$, then transmit
4. If $L^t_{max} = 1$ and $L^r_{min} > 0$, then receive. If more than one relay can be selected in any case, the control node selects one randomly.

The Result in [79] show the superiority of the state-based over the max-link. By reducing the probability of an empty or full buffer, the diversity of the state-based is maintained at $2K$ with $L \geq 3$. The state-based also reduces the packet delay by prioritizing transmission in cases 1 and 3. The state-based delay has similar calculations to those in max-link except that at high SNR, if the system is in the state that all relays have a packet in their buffers. Then the system chooses one relay randomly to receive a packet (Case 7 of stage one and Case 4 of stage two). In result, the selected relay has two packets in its buffer and all the other relays have one packet in their buffers. Next, the relay with two packets in its buffer is chosen to transmit a packet in the following time slot (Case 6 of stage one and Case 3 of stage two). Thus, all the relay buffers again have only one packet. This process repeats and thus when the SNR is high enough, the system is in the state in which all relays have one packet in the buffer (K packets in total) with a probability of 1/2. The probability for the buffers to have $K + 1$ packets in total is also 1/2. Based on this result, the average number of packets stored in the buffers is $\frac{K}{2} + \frac{K+1}{2} = K + \frac{1}{2}$, hence,

$$\bar{D_r} = \frac{K + \frac{1}{2}}{\frac{1}{2}} = 2K + 1 \tag{2.12}$$

$\bar{D_r} + \bar{D_s} = 2K + 2$, which is proportional to K only [79].

Fig. 2.6 shows how the state-based selects one link based on its buffer state. R_4 has a full buffer, so $R_4 \rightarrow D$ link is selected, although it has the lowest SNR (assuming 3dB is sufficient for transmission).

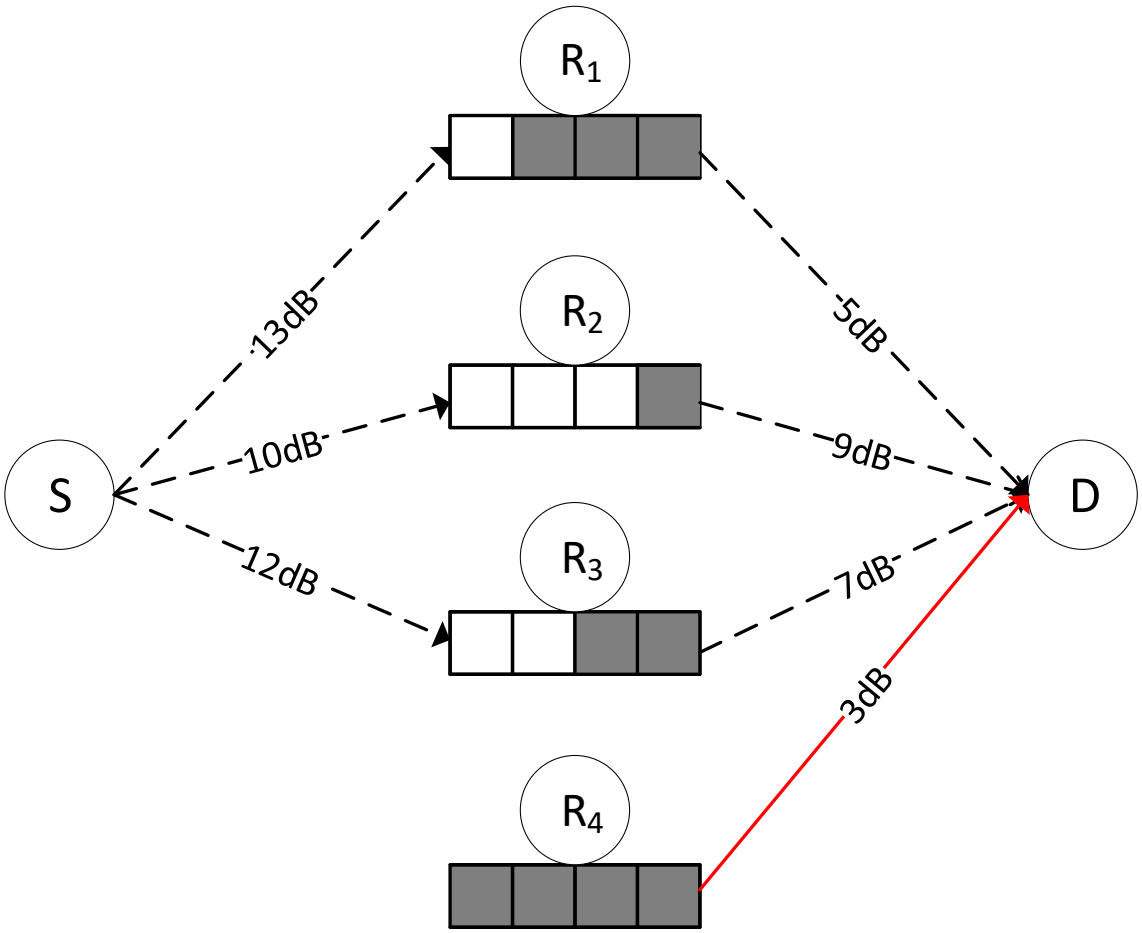

Figure 2.6 The state-based scheme example.

2.2.5 Minimum Delay Relay Selection (Delay-Reduced)

Although the state-based has achieved better performance than the max-link, still delay-sensitive applications in 5G may not tolerate the delays offered by the state-based scheme. This encouraged the authors in [121] to prioritize transmission always by suggesting the delay-reduced. In the delay-reduced, the transmission has a higher priority than the receiving unless the transmission is not possible. Hence, the queues in buffers become shorter.

For instance, if all K buffers are empty, then a packet is sent to a relay after one time-slot. In the next time-slot, the received packet is forwarded to the destination with probability $\frac{1}{K+1}$ with the traditional max-link if all links are available. So, the packet is more likely to remain in the buffer with probability $\frac{K}{K+1}$, which increases

the delay by one time-slot. On the other hand, this extra delay is avoided by prioritizing transmission.

In the delay-reduced, selecting $R_k \rightarrow D$ link is based on the CSI, where the link with the highest SNR is selected. If no $R_k \rightarrow D$ link can be selected, then $S \rightarrow R_k$ with the highest SNR is selected. Otherwise, the system is in outage. The results in [121] show enhancements in average packet delay compared to the state-based. At high SNR, the delay of the delay-reduced scheme is 2, which is equivalent to the conventional relays. However, these improvements are at the price of lower diversity order ($\geq K$ and $\leq K+1$).

In Fig. 2.7, the delay-reduced selects the best available $R_k \rightarrow D$ link which is $R_2 \rightarrow D$. This kept R_4 with full buffer, and now R_2 has an empty buffer, which reduces the diversity. If all $R_k \rightarrow D$ links stay available for a while, all packets will be transmitted before receiving any new packet.

It is worth noting that a trade-off has to be done between the diversity order and the delay based on the application requirements. The dilemma is as follows: multiple relays and large buffer sizes increase the diversity order. Nevertheless, multiple relays and large buffer sizes may cause long delays. Thus, for delay-sensitive applications, some studies have suggested to reduce the number of relays and to use small buffers [131].

2.2.6 Priority-Based Relay Selection

For the trading-off between the delay and the diversity order, authors in [48], presented the target buffer length (denoted as θ_k) at every relay. In addition, the authors have defined $\Delta_k = q_k - \theta_k$ as the difference between the buffer content q_k and the target length θ_k of the relay R_k. Each relay has to keep its buffer content closer to θ. So, the further the buffer content from θ (higher in $|\Delta|$), the higher the priority for its corresponding link. After obtaining the priorities for all the available links, the priority-based selects the link with the highest priority. If more than one link has the same priority, the link with the highest SNR is selected.

To enhance the delay performance, the PBRS prioritizes the transmission in case of equal $|\Delta_k|$. This is done by giving higher priority to buffers with positive

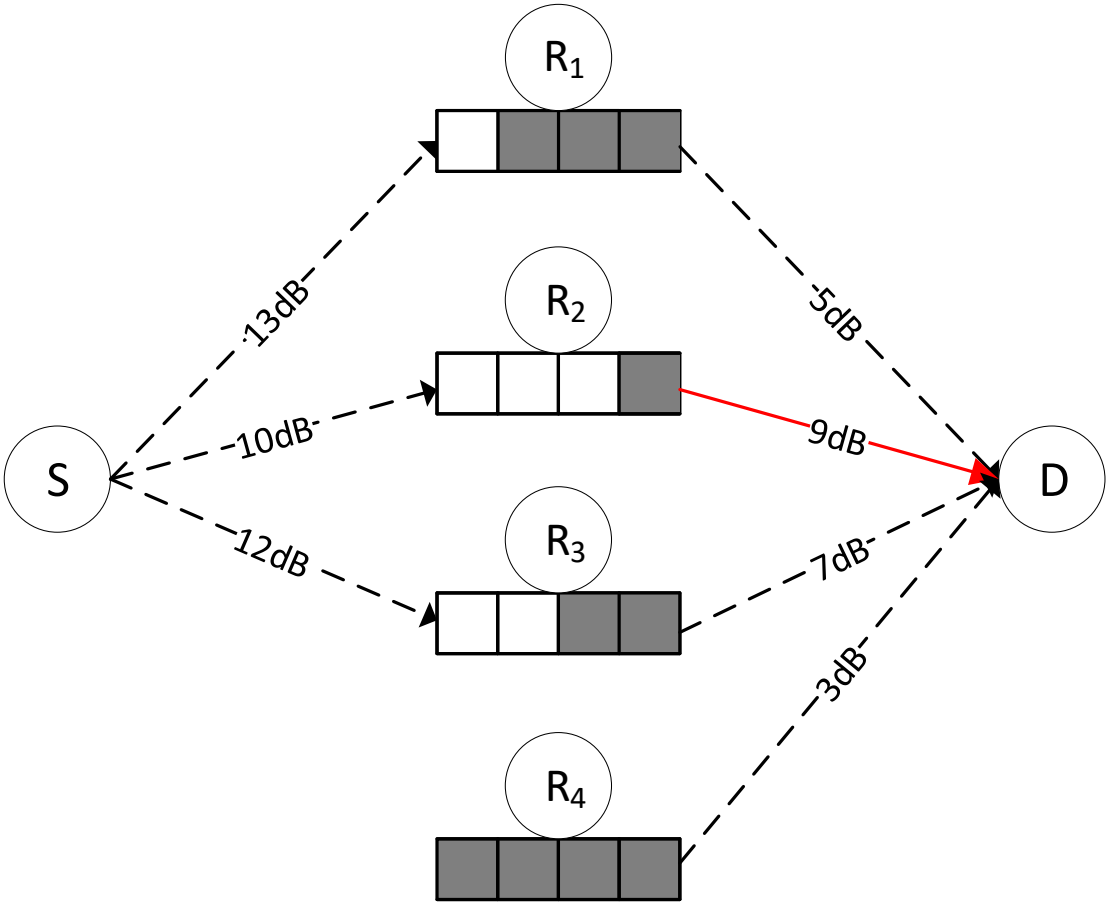

Figure 2.7 The delay-reduced scheme example.

Δ. For example, if $\theta_1 = \theta_2 = 3$ in a two relays network. Let the buffer of the first relay have one packet, while the other buffer has 5 packets. Hence, both buffers are 2 packets away from their θ_k. However, the first relay has a negative value of $\Delta_1 = -2$, so R_1 needs to receive a packet to get closer to θ_1. The second relay has $\Delta_2 = +2$, so it has extra packets which have to be transmitted. The priority-based gives $R_2 \rightarrow D$ link higher priority than $S \rightarrow R_1$.

To clarify the priority-based, Fig. 2.8 shows how the priority-based works in the four relays network example with $\theta_k = 2$ for all relays. Since R_4 buffer is the furthest from target $\Delta_4 = +2$, the priority-based gives it the highest priority. The rest of the links are ordered based on their priority as follows: $R_1 \rightarrow D$, $S \rightarrow R_2$, $R_3 \rightarrow D > S \rightarrow R_3 > R_2 \rightarrow D > S \rightarrow R_1$.

The state-based and the delay-reduced are special cases of the priority-based. Specifically, at target length $\theta = 2$, the priority-based is equivalent to the state-

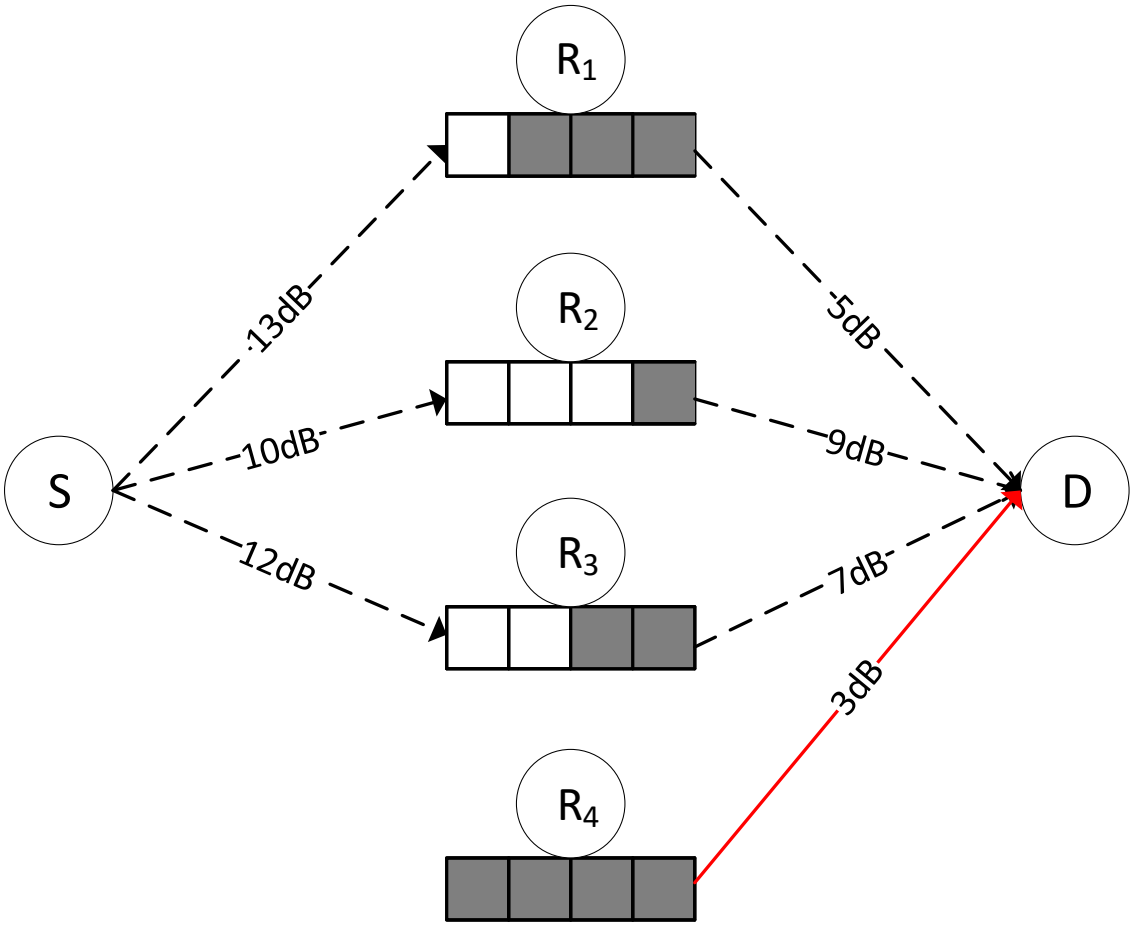

Figure 2.8 priority-based example.

based. And at $\theta = 0$, the priority-based is equivalent to the delay-reduced. Finally, finding the optimal target length for all buffers to achieve the optimal trade-off between the delay performance and the diversity order, is still an open problem.

2.3 Challenges and Opportunities

Buffer-aided cooperative relay network is a promising technique for 5G networks and beyond. This motivated researchers to investigate new methods to enhance buffer-aided cooperative relay network performance either by proposing new relay selection schemes or by combining it with other available 5G technologies.

This thesis suggests novel enhancements that can be done on the buffer-aided cooperative relay network performance. Accordingly, solutions for the challenges and limitations that come with the available solutions are proposed. Firstly, part of

the available studies on buffer-aided relays in cooperative NOMA assumes infinite buffer size in their relay selection scheme, which is impractical. And the other part of the available studies considers a finite buffer size, but it did not consider combining the NOMA and the OMA in the relay selection. In addition, the available studies were done on a single relay. Therefore, in Chapter 3, a novel prioritization-based buffer-aided relay selection scheme, which is able to combine the NOMA and the OMA transmission in multiple relay cooperative networks is proposed.

Secondly, the main challenge for applying buffer-aided relays is that the buffers may lengthen the packet delay. And non of the available studies has considered the source delay, which is the delay that packets encounter at the source. This motivates us to study how the buffer-aided relay deals with the source delay in Chapter 4. To check if the source delay is worsening the delay problem in the buffer-aided relay. This consideration of the source delay makes the delay comparison between buffer-aided and non-buffer-aided relays more accurate. In addition, we propose a new relay selection scheme to reduce the packets delay.

Finally, the degradation on the buffer-aided relay performance caused by constraining the delay to a certain target delay is studied in Chapter 5. This is important in the 5G applications, which have stringent delay constraints, where every packet exceeds the target delay will be re-transmitted or discarded, which is expected to cause huge degradation in the performance. Therefore, we presented the delay-constrained outage probability to study this degradation. Then we propose an adaptive buffer-size algorithm to maintain the performance of the buffer-aided relay under delay constraints.

2.4 Summary

Conventional cooperative relay enhances the quality of the communication system by exploiting the spatial diversity and extending the coverage area. This is done while maintaining the simplicity of the system by selecting the best relay. A better form of cooperative relay is buffer-aided cooperative relay. Several relay selection schemes have been proposed to maximize the benefits of buffer-aided cooperative

relay and make it more suitable for the 5G applications. Part of the available schemes focuses on improving the diversity gain and the throughput of the buffer-aided cooperative relay network. Another part of the available schemes focus on delay reduction. The rest of the available schemes suggest the trade-off between different performance metrics. Still, the necessity for new schemes is high, to overcome the shortcomings of the available schemes and to exploit buffer-aided relays more efficiently. In the next chapter, we propose a novel buffer-aided relay selection scheme to combine both NOMA and OMA transmission in the 5G cooperative relay network.

Chapter 3

Buffer-Aided Relay Selection for Cooperative NOMA in 5G systems

Non-orthogonal multiple access (NOMA) improves the spectral efficiency by allowing more than one user to share the same resources. Which is particularly essential in the fifth generation (5G) systems, such as the Internet of Things (IoT), which involves massive number of connections.

It has been theoretically shown that using buffer-aided relays can further increase the throughput in NOMA relay network. This is however valid only when the channels SNR's are large enough to support NOMA transmission. Although it would be straightforward for the cooperative network to switch between NOMA and the traditional orthogonal multiple access (OMA) transmission modes based on the channel SNR, the best potential throughput would not be achieved. In this chapter, a novel prioritization-based buffer-aided relay selection scheme which is able to combine NOMA and OMA transmission effectively in the relay network is proposed.

3.1 Introduction

As introduced in Chapter 1, 5G systems such as the IoT aims to connecting large number of devices, which imposes great challenges in mobile network design [30, 29]. NOMA theoretically improves transmission efficiency by allowing multiple

devices to share the same spectrum resources [50, 37, 41, 59], which provides an attractive solution to achieve massive connectivity required for 5G applications such as the IoT [82, 74, 102].

NOMA has been successfully applied in cooperative relay networks. The importance of the relay selection (as described in Subsection 1.2.1) is the simplicity of implementation, since it does not involve complex physical layer transmission techniques such as synchronization processes or distributed space-time codes like the codes used in MIMO [68].

Cooperative NOMA is briefly introduced in Section 2.2. In this section, we extend the discussion about cooperative NOMA. Firstly, several studies have suggested conventional (non-buffer) relay selection for cooperative NOMA. In [38], a two-stage relay selection scheme is described to maximize the data rate for one NOMA user opportunistically upon satisfying the target transmission for other users which has lower data rate requirements. The analytical and simulations results show that the suggested scheme outperforms the traditional max-min scheme when combined with NOMA in maximizing diversity gain. Other conventional relay selection schemes were suggested in [36, 143].

On the other hand, another recent development in cooperative networks is buffer-aided relay [94]. With buffer-aided relay, the transmission can be better aligned with strong links than traditional schemes such as the max-min relay selection [24]. Thus, buffer techniques have been applied in the cooperative NOMA networks. In [80], adaptive link scheme for a single-relay NOMA network with an infinite buffer size is proposed, the analysis show that the proposed system has higher throughput compared with conventional relaying NOMA.

In [80], NOMA and OMA transmission can be optimally chosen by letting the buffer operate at the edge of non-absorbing mode, which is a necessary condition for optimality in link selection as proved in [149]. So, the queue of each buffer has to be at the boundary of absorbing and non-absorbing. In other words, the number of arriving packets at buffers have to equal the number of departing packets. However, to get the optimal selection rule, the authors in [149] and [80] assumed an infinite buffer size and an infinite delay, which is impractical as described in

[149], also, long queues are not acceptable in 5G. Therefore, the authors in [140] proposed another buffer-aided cooperative NOMA link selection scheme, where the system model is the same as that in [80], but with finite buffer size.

Because of the limited buffer size, it is usually not possible to have the buffer operating at the non-absorbing edge since full or empty buffer are more likely to occur, this complicating the optimization problem in [149]. In addition, as stated in [63], designing an optimal protocol which can achieve the maximum throughput at a given delay constraint, is still an open problem even in the most elementary buffer-aided relay network. As a result, optimal selection scheme is still an open problem for relays with finite buffer size.

In the link selection scheme, which was proposed in [140] for a two users single buffer-aided relay system, the relay always applies NOMA to serve the two users. The result shows that diversity order of two can be achieved with buffer size larger than or equal to two. The proposed scheme, however, only has higher throughput than its OMA counterpart in the high SNR range (low outage), this is because the throughput is defined as the successful (no outage) data transmission rate per unit time and NOMA doubles the data rate of OMA. This is similar to the delay-limited throughout which equals $\eta(1-P_{out})$ [124] where η is the data rate and P_{out} denotes outage probability. Compared with OMA, although NOMA doubles OMA's data rate, it increases the outage probability P_{out}.

The suggested scheme in [140], has defined P_{out} as the probability that neither the source-to-relay link can achieve NOMA data rate nor the relay-to-users links can support NOMA data rate. It is worth noting that switching to OMA is better than applying NOMA when only one user can be served. Because in NOMA, part of the power is wasted on unavailable user and more processing is required.

At low SNR range, because P_{out} is close to one, the throughput is dominated by P_{out} and OMA has higher throughput than NOMA. At high SNR range, on the other hand, P_{out} approaches zero when the SNR goes to infinity. Then the throughput is determined by η and NOMA has higher throughput than OMA. Therefore, when the SNR is not large enough to support NOMA, instead of stop transmitting (as in [140]), OMA may still be applied.

Authors in [140], highlighted that combining NOMA and OMA will make performance analysis “very complicated" (Remark 3, [140]). To avoid this scenario, the authors in [140], have suggested a compromise approach by switching between NOMA and OMA based on the outage events, this is done by setting a threshold SNR. When the SNR is larger than the threshold, NOMA is used, and otherwise, OMA is used. As will be shown later in this chapter, this compromised approach cannot achieve the full potential of the system.

The performance of the buffer-aided cooperative rely networks depends on buffer states which are determined by the number of packets in the buffers. If a relay buffer is full or empty, the corresponding source-to-relay or relay-to-destination link is not available for reception or transmission respectively. The early proposed buffer-aided max-link relay selection [68] may achieve full diversity order (i.e. twice the number of relay nodes) when the buffers have infinite size and balanced input/output data rates which is however not always the case in practice.

In [79], the state-based was proposed, in which the link selection is based on buffer states. As discussed in Chapter 2, state-based achieves better outage performance than max-link scheme, but the improvement becomes less significant for unbalanced channels since full or empty buffer are more probable. This becomes more serious in NOMA cooperative network: even when the source-to-relay and relay-to-user links have the same average gains, the buffer input/output rate may still be unbalanced because the source-to-relay and relay-to-user apply different transmission modes. It is interesting to note that the buffer-aided relay for cooperative NOMA link selection scheme in [140] is similar to state-based but for NOMA transmission.

As aforementioned, the optimum link selection in [80] applies to the relay network with infinite buffer sizes, which is often impractical. On the other hand, the link selection in [140] considers finite buffer size, but it does not include OMA transmission and the selection rule is not always optimum. Neither [80] nor [140] considers the multiple relay scenario. This motivates us to investigate the finite size buffer-aided relay selection for cooperative NOMA 5G networks. The main contributions of this chapter are listed as follows:

- Proposing a novel buffer-aided relay selection scheme for multiple relay cooperative NOMA networks.
- Composing a prioritization-based selection rule to combine both NOMA and OMA transmission.
- Analyzing the average throughput of the proposed scheme. Turns out combining NOMA and OMA makes the performance analysis very complicated, and considering multiple relays also further complicates the analysis.
- Obtaining the diversity order of the proposed scheme as $3K$, where K is the number of relays. In contrast, if the link selection in [140] is generalized to multiple relays, the diversity order would be $2K$.

The remainder of this chapter is organised as follows: Section 3.2 describes the system model; Section 3.3 covers the performance analysis of the proposed system; Section 3.4 shows simulation results; Finally, Section 3.5 concludes this chapter.

3.2 System Model

The system model of the buffer-aided cooperative NOMA in 5G is shown in Fig. 4.1, where there are one source node S, K half-duplex decode-and-forward (DF) relay nodes denoted as R_k, $k = 1, \cdots, K$ and two users U_1 and U_2, respectively. The channel coefficients for $S \to R_k$, $R_k \to U_1$ and $R_k \to U_2$ links are denoted as h_{sr_k}, $h_{r_k u_1}$ and $h_{r_k u_2}$ respectively. All channels have flat Rayleigh fading coefficients that remain constant within the time-slot and change independently from one slot to another. Every relay R_k is equipped with two L-size buffers for data transmissions to users U_1 and U_2 respectively (two buffers for organisation and simpler notation). We assume that source always has enough information to send to relays in all time-slots. In each time-slot, a packet can be transmitted by the source or a relay, information symbols intended for the two users are assembled into packets of equal size. In addition, we assume that the source and the users are not directly connected. Without losing generality, we assume that the transmit powers at all transmit nodes are P_t, and the noise variances at all receiving nodes are σ^2.

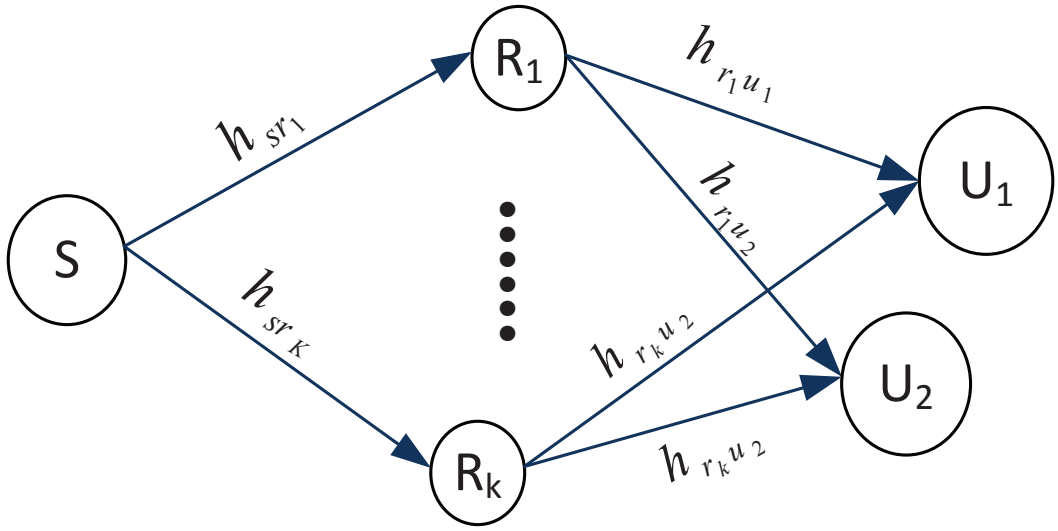

Figure 3.1 System model for the cooperative relay with NOMA network.

When OMA transmission is applied, at time-slot t, the link capacity for channel $h_{d_k}(t)$ is given by

$$\begin{aligned} C_{d_k}(t) &= \log_2\left(1+\gamma_{d_k}(t)\right), \\ d_k &\in \{sr_k, r_ku_1, r_ku_2\},\ k=1,\cdots,K, \end{aligned} \tag{3.1}$$

where $\gamma_{d_k}(t) = (P_t/\sigma^2)|h_{d_k}(t)|^2$. Assuming $|h_{d_k}(t)|^2$ is exponentially distributed with the average $\Omega_{d_k} = E[|h_{d_k}(t)|^2]$, where $E[.]$ is the expectation. $\gamma_{d_k}(t)$ is also exponentially distributed with average $\bar{\gamma}_{d_k} = (P_t/\sigma^2)\Omega_{d_k}$. Thus $\gamma_{d_k}(t)$ and $\bar{\gamma}_{d_k}$ are the instantaneous and average SNR for channel $h_{d_k}(t)$ respectively.

3.2.1 Transmission Mode

At every time-slot, both of source-to-relay $S \rightarrow R_k$ and relay-to-users $R_k \rightarrow U_m$ transmissions may operate in two modes: double and single packet transmission. For the $S \rightarrow R_k$ link, if it satisfies

$$C_{sr_k}(t) \geq 2\eta, \tag{3.2}$$

where η is the target data rate, the source S is able to transmit two packets to both buffers at R_k. This is achieved based on TDMA (time-division-multiple-access) principle by applying half of the time-slot to transmit each packet. Otherwise, if (3.2) does not hold but $C_{sr_k}(t) \geq \eta$, a single packet can be transmitted to either of the buffers at R_k. TDMA is chosen for comparison with the available study [140]. Also, fixed rate transmission is used because we are not studying the achievable rate, instead, other performance metrics like throughput. Noting that fixed transmission is simpler than adaptive transmission since it does not require the availability of CSI at the transmitter CSIT.

On the other hand, for the $R_k \rightarrow U_m$ ($m = 1$ or 2) link, NOMA can be applied to transmit packets to U_1 and U_2 simultaneously. The superimposed NOMA symbol at R_k is given by

$$x_{r_k}(t) = \sqrt{\alpha} x_{r_{k,1}}(t) + \sqrt{1-\alpha} x_{r_{k,2}}(t), \tag{3.3}$$

where $x_{r_{k,1}}(t)$ and $x_{r_{k,2}}(t)$ are data for users U_1 and U_2 respectively, and $0 \leq \alpha \leq 1$ is the power allocation factor. Then the received signal at U_m is given by

$$y_m(t) = \sqrt{\alpha P_t} h_{r_k u_m}(t) x_{r_{k,1}}(t) + \sqrt{(1-\alpha)P_t} h_{r_k u_m}(t) x_{r_{k,2}}(t) + n_m(t), \quad m = 1, 2, \tag{3.4}$$

where $n_m(t)$ is the noise at user U_m. When NOMA is applied, the link capacity is not given by (3.1) but must include the interference within the superimposed symbol. To be specific, when $\gamma_{r_k u_1}(t) > \gamma_{r_k u_2}(t)$, the SNR to decode $x_{r_{k,2}}(t)$ at U_2 is given by

$$SINR(x_{r_{k,2}}(t)) = \frac{(1-\alpha)\gamma_{r_k u_2}(t)}{\alpha \gamma_{r_k u_2}(t) + 1}. \tag{3.5}$$

Because $\gamma_{r_k u_1}(t) > \gamma_{r_k u_2}(t)$, $x_{r_{k,2}}(t)$ can also be decoded at U_1 if it can be decoded at U_2. Removing $x_{r_{k,2}}(t)$ from the received signal at U_1 by SIC, the required SNR to decode $x_{r_{k,1}}(t)$ at U_1 is given by

$$SNR(x_{r_{k,1}}(t)) = \alpha \gamma_{r_k u_1}(t). \tag{3.6}$$

Following similar procedures as those in [140], the condition that there exists an α to support NOMA transmission to both U_1 and U_2 (i.e. $\log_2(1 + SINR(x_{r_{k,2}}(t)) \geq \eta$ and $\log_2(1 + SNR(x_{r_{k,1}}(t)) \geq \eta$) is given by

$$\frac{(1-\alpha)\gamma_{r_k u_2}(t)}{\alpha\gamma_{r_k u_2}(t)+1} \geq 2^{\eta} - 1, \tag{3.7}$$

$$\alpha\gamma_{r_k u_1}(t) \geq 2^{\eta} - 1, \tag{3.8}$$

from (3.7) and (3.8)

$$\frac{2^{\eta}-1}{\gamma_{r_k u_1}(t)} \leq \alpha \leq \frac{1}{2^{\eta}}(1 - \frac{2^{\eta}-1}{\gamma_{r_k u_2}(t)}), \tag{3.9}$$

$$\gamma_{r_k u_2}(t) \geq \frac{(2^{\eta}-1)\gamma_{r_k u_1}(t)}{\gamma_{r_k u_1}(t) - 2^{\eta}(2^{\eta}-1)}, \qquad \text{if } \gamma_{r_k u_1}(t) > \gamma_{r_k u_2}(t). \tag{3.10}$$

Similarly, if $\gamma_{r_k u_1}(t) < \gamma_{r_k u_2}(t)$, NOMA condition becomes

$$\gamma_{r_k u_1}(t) \geq \frac{(2^{\eta}-1)\gamma_{r_k u_2}(t)}{\gamma_{r_k u_2}(t) - 2^{\eta}(2^{\eta}-1)}. \tag{3.11}$$

If the SNR for the $R_k \rightarrow U_m$ ($m = 1$ or 2) links is not large enough to satisfy (3.10) or (3.11), NOMA transmission is not possible or not efficient. In this case, if $C_{r_k u_m}(t) > \eta$, OMA can be used to transmit one packet to U_m.

3.2.2 Selection Rule

Recently, as presented in Chapter 2, relay selection has been done based on CSI or buffer state BSI. Based on this, each relay has one of three decisions to make: transmits, receives or remains silent. As mentioned in Section 3.1, the selection rule proposed in the previous work [140] is similar to the state-based, which was described in detail in Chapter 2.

On the other hand, the selection rule is more complicated in our proposed system. In particular, for the relay R_k, the transmission may be chosen from the

following six candidates

$$\{(sr_{k,1}),\ (sr_{k,2}),\ (TDMA_k),\ (r_{k,1}u_1),\ (r_{k,2}u_2),\ (NOMA_k)\}, \tag{3.12}$$

where $(sr_{k,m})$ indicates the single packet transmission from S to the m-th buffer at R_k, $(TDMA_k)$ indicates the double packet transmission based on TDMA from S to both buffers at R_k, $(r_{k,m}u_m)$ is the single transmission from the m-th buffer at R_k to U_m, and $(NOMA_k)$ is NOMA based double transmission from R_k to both U_1 and U_2. In total, there are $6K$ candidates. The relay selection is to select not only a relay link but also a transmission mode, among all available transmission candidates.

At any time, the numbers of data packets in relay buffers (i.e. the buffer length) form the buffer states. While each relay has two buffers, if the relay number is K and buffer size is L, there are $(L+1)^{2K}$ states in total. The l-th state vector is defined as

$$\mathbf{q}^{(l)} = [q_{1,1}^{(l)}, q_{1,2}^{(l)}, \cdots, q_{K,1}^{(l)}, q_{K,2}^{(l)}], \qquad l = 1, \cdots, (L+1)^{2K}, \tag{3.13}$$

where $q_{k,m}^{(l)}$ is the buffer length for the m-th buffer at R_k at state $\mathbf{q}^{(l)}$. At any time-slot, given the BSI and CSI of all channels, the relay selection is carried out as following:

- First, selection priorities are given to all available transmission candidates. This will be described later.
- All candidates are then checked, from the highest to lowest priorities, whether they can support the target data rate or not. This is meant to check whether (3.2) is satisfied for candidate $(TDMA_k)$, (3.10) or (3.11) for candidate $(NOMA_k)$, and $C_{d_k} > \eta$ for single transmission candidates.
- The candidate with the highest priority which can support the target transmission rate is selected for data transmission.
- Outage occurs if no candidate can be selected.

In order to give priority orders to select the available transmission candidates, we introduce the “target buffer length", $\Theta_{k,m}$, for the m-th buffer ($m = 1$ or 2) at relay R_k. Supposing the buffer state vector is $\mathbf{q}^{(i)}$, the distance between the buffer length and the corresponding target length is defined as

$$\Delta_{k,m}^{(i)} = |q_{k,m}^{(i)} - \Theta_{k,m}|, \qquad m = 1, 2, \quad k = 1, \cdots, K, \tag{3.14}$$

Then we can give higher priorities to candidates corresponding to buffers further away from the target length as following:

- The double transmission candidates always have higher priority than the single transmission candidates. If an available double transmission candidate $cand_b$ is selected, the buffer lengths of both buffers at relay R_{k_b} are changed by one, and the buffer state becomes $\mathbf{q}^{(i,cand_b)}$. Then for $m = 1$ and 2, we obtain

$$\begin{aligned} \Delta_{k_b,m}^{(i,cand_b)} &= |q_{k_b,m}^{(i,cand_b)} - \Theta_{k_b,m}|, \\ & cand_b \in \{(TDMA_{k_b}),\ (NOMA_{k_b})\} \end{aligned} \tag{3.15}$$

  While selecting $cand_b$ leads to buffer length change of two buffers at relay R_{k_b}, the buffer with higher $\Delta_{k_b,m}^{(i,cand_b)}$ is used for prioritization. Then the priority measurement for selecting candidate $cand_b$ at state $\mathbf{q}^{(i)}$ is defined as

$$\mathcal{M}^{(i,cand_b)} = \text{sign}\left(\Delta_{k_b,m_b}^{(i,cand_b)} - \Delta_{k_b,m_b}^{(i)}\right) \cdot \Delta_{k_b,m_b}^{(i,cand_b)}, \tag{3.16}$$

  where $m_b = \arg\max_m \left(\Delta_{k_b,m}^{(i,cand_b)} |\ m = 1, 2\right)$. It is clear that, if $\mathcal{M}^{(i,cand_b)} < 0$, selecting $cand_b$ will decrease the distance between the corresponding buffer and target lengths, and otherwise will increase it. Thus higher priority is given to candidates with smaller $\mathcal{M}^{(i,cand_b)}$.

  Specifically, in (3.15), $\Delta_{k_a,m}^{(i,cand_b)}$ would be the distance between the buffer length (if the corresponding candidate would be selected) and the target buffer length. Thus the larger the $\Delta_{k_a,m}^{(i,cand_b)}$ is, the further buffer length is away from the target. Then in (3.16), if $\mathcal{M}^{(i,cand_b)} < 0$, selecting $cand_b$ will

decrease the distance between the corresponding buffer and target lengths, and otherwise will increase it. Thus higher priority is given to candidates with smaller $\mathcal{M}^{(i,cand_b)}$. Nevertheless, the example in Fig. 3.2 shows an illustrative example on how (3.15) and (3.16) are used in setting the priority orders.

- Similarly, the priorities for single transmission candidates are ordered as follows. If an available single transmission candidate $cand_a$ is selected, the buffer length of the m_a-th buffer at relay R_{k_a} is changed by one so that the buffer state becomes $\mathbf{q}^{(i,cand_a)}$, and then the new distance between the buffer length and the target is given by

$$\begin{aligned} &\Delta_{k_a,m_a}^{(i,cand_a)} = |q_{k_a,m_a}^{(i,cand_a)} - \Theta_{k_a,m_a}|, \\ &\qquad cand_a \in \{(sr_{k_a,1}),\ (sr_{k_a,2}),\ (r_{k_a,1}u_1),\ (r_{k_a,2}u_2)\}. \end{aligned} \tag{3.17}$$

  The priority measurement for selecting candidate $cand_a$ is then obtained as

$$\mathcal{M}^{(i,cand_a)} = \text{sign}\left(\Delta_{k_a,m_a}^{(i,cand_a)} - \Delta_{k_a,m_a}^{(i)}\right) \cdot \Delta_{k_a,m_a}^{(i,cand_a)}, \tag{3.18}$$

  higher priority is then given to candidates with smaller $\mathcal{M}^{(i,cand_a)}$. It is worth noting that with the mentioned rule, the relay-to-users links have higher priority orders than source-to-relay links when they are at the same distance from the target buffer length, this applicable only if the two compared candidates are from two different buffers, the example in Fig. 3.2 shows this clearly.

High throughput relies on large data rate and low outage probability. In the proposed scheme, the large data rate is achieved by giving higher priority to select double-packet transmission modes, and the low outage probability is achieved by setting appropriate target lengths so that the buffer lengths are kept away from empty or full as much as possible.

In general, for buffers at relay R_k, if the input data rate is higher than the output rate, the buffers are likely to be saturated and thus the target length shall be set close to zero. Otherwise, if the input rate is smaller than the output rate,

the buffers tend to be empty and the target buffer length shall be close to the full buffer size. Particularly, if a buffer's input and output rates are the same, the target buffer length can be set as 2 (where we assume the buffer size is larger than or equal 3), because this not only keeps buffer lengths away from empty or full but also leads to small packet delay.

In the proposed scheme, however, the input and output rates at buffers depend on not only channel gains but also transmission modes. Therefore, even if the $S \to R_k$ and $R_k \to U_m$ links have the same average SNR, setting the target length to 2 may not be the best choice. It is interesting to note that the selection rules in [140], except that it does not include OMA transmission, is equivalent to the proposed selection rule with the target buffer lengths being set to 2. On the other hand, in order to achieve minimum transmission delay, the target buffer length shall be set as zero so that the data in the buffers can be transmitted out as quickly as possible (same as setting the target to 0 in the priority-based).

Before leaving this section, we show an example of giving priority orders to all available candidates in Fig. 3.2, where the relay number $K = 2$, the buffer size $L = 4$, the target buffer lengths for all buffers are set as 2, and the buffer state is $\mathbf{q} = [4, 1, 3, 0]$. From (3.14), the distance between the buffer length and the target for the four buffers can be obtained as $(2, 1, 1, 2)$ respectively. There are two available double transmission candidates at this state, which are $(TDMA_2)$ and $(NOMA_1)$ respectively. From (3.16), their priority measurements are obtained as $+2$ for both, since transmission is prioritized over reception, their priorities are given as

$$\mathcal{O}(NOMA_1) > \mathcal{O}(TDMA_2), \tag{3.19}$$

where $\mathcal{O}(.)$ is the selection priority for the enclosed candidate. On the other hand, there are six single transmission candidates, which are $(sr_{1,2})$, $(r_{1,1}u_1)$, $(r_{1,2}u_2)$, $(sr_{2,1})$, $(sr_{2,2})$ and $(r_{2,1}u_1)$ respectively. From (3.18), the priority measurements are obtained as $(0,\ -1,\ +2,\ +2,\ -1,\ 0)$ respectively. Thus the six candidates are prioritized as $\mathcal{O}(r_{1,1}u_1) > \mathcal{O}(sr_{2,2}) > \mathcal{O}(r_{2,1}u_1) > \mathcal{O}(sr_{1,2}) > \mathcal{O}(r_{1,2}u_2) > \mathcal{O}(sr_{2,1})$.

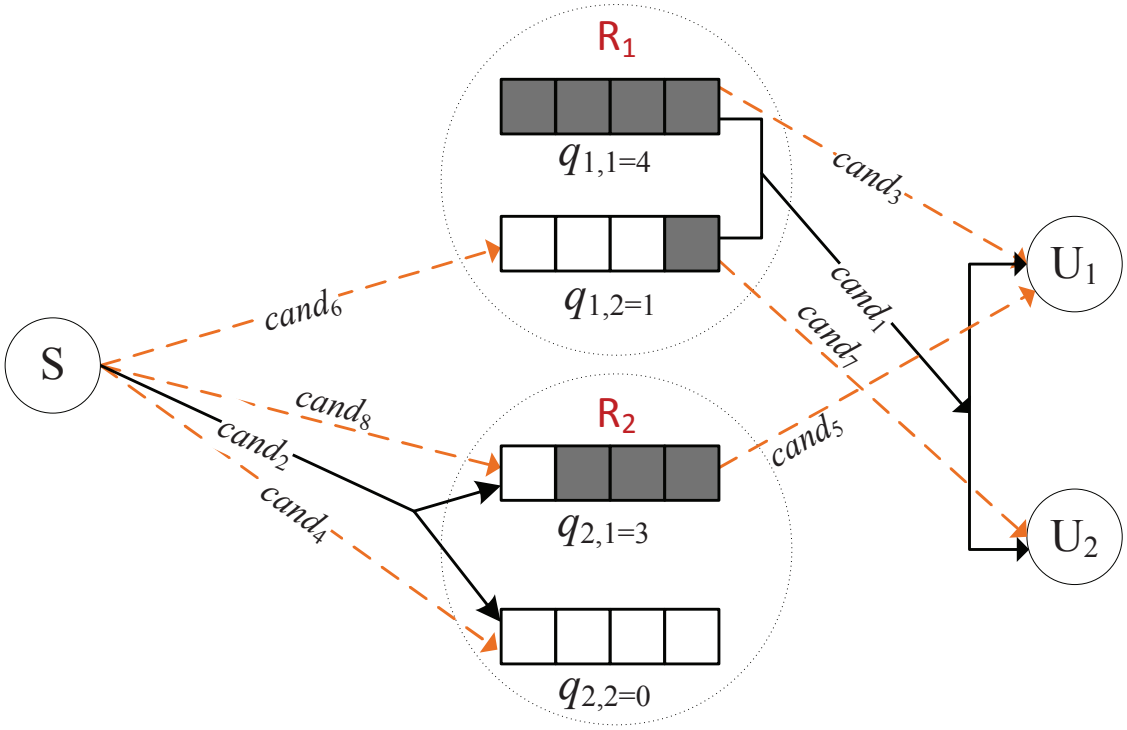

Figure 3.2 An example of giving priori orders to available candidates at state $\mathbf{q} = [4, 1, 3, 0]$, where the target buffer length is 2, $cand_m$ indicates the priority order of the corresponding candidate is m, the links with the solid-lines are for the double transmission, and the links with dash-lines are for the single transmission.

The priorities for all available candidates are illustrated in Fig. 3.2, where $cand_m$ indicates the priority order m of the corresponding candidate.

Note that, although both relays are at the same distances from target buffer length (R_1 with distances $(2, 1)$ and R_2 with distances $(1, 2)$), NOMA is prioritized over TDMA and single transmission from rely-to-user has higher priority like the case with $cand_3$ and $cand_4$

3.3 Performance Analysis

Let $\mathbf{A}$ be the $(L+1)^{2K} \times (L+1)^{2K}$ state transition matrix, where the entry $A_{i,j}$ as the transition probability from state $\mathbf{q}^{(j)}$ to $\mathbf{q}^{(i)}$. Particularly $A_{i,i}$ is the outage probability at state $\mathbf{q}^{(i)}$. We assume that at buffer state $\mathbf{q}^{(i)}$, there are L_i available candidates for selection at state $\mathbf{q}^{(i)}$, denoted as $cand_1, \cdots, cand_{L_i}$ from the highest to the lowest priority order respectively.

Every double transmission candidate is associated with a pair of single transmission candidates: candidate $(TDMA_k)$ is associated with $(sr_{k,1})$ and $(sr_{k,2})$, and candidate $(NOMA_k)$ is associate with $(r_{k,1}u_1)$ and $(r_{k,2}u_2)$. We have the following remarks:

Remark 1 *A double transmission candidate and its two associated single transmissions are not independent, because they correspond to the same link(s).*

Remark 2 *If a single transmission candidate is in outage, its associated double transmission candidate must also be in outage.*

Below we derive the transition probability $A_{i,j}$ for $i = j$ and $i \neq j$, from which the average throughput is obtained. For better exposition, we will show the analysis for the example in Fig. 3.2. As shown in Fig. 3.2, there are eight available candidates for selection at state $\mathbf{q}^{(i)} = [4, 1, 3, 0]$, in which candidates $\{cand_1, cand_3, cand_7\}$ are associated, so are the candidates $\{cand_2, cand_4, cand_8\}$, but $cand_5$ and $cand_6$ are not associated with any other candidates. We denote $P(\overline{cand_l})$ and $P(cand_l)$ as the probabilities that the candidate $cand_l$ is and not in outage, respectively.

3.3.1 Outage Probability

The outage probability at state $\mathbf{q}^{(i)}$ is the probability that all available candidates are in outage as

$$P_{out}^{\mathbf{q}^{(i)}} = A_{i,i} = P(\overline{cand}_1, \cdots, \overline{cand}_{L_i}). \tag{3.20}$$

For the example in Fig. 3.2, from remark 1, we have

$$\begin{aligned} P_{out}^{\mathbf{q}^{(i)}=[4,1,3,0]} &= P((\overline{NOMA_1}), \overline{cand}_3, \overline{cand}_7) \\ &\quad \times P(\overline{TDMA_2}, \overline{cand}_4, \overline{cand}_8) P(\overline{cand}_5) P(\overline{cand}_6), \end{aligned} \tag{3.21}$$

where candidates $cand_1$ and $cand_2$ are represented as $(NOMA_1)$ and $(TDMA_2)$ respectively for better exposition. From remark 2, we have

$$\begin{aligned} P((\overline{NOMA_1}), \overline{cand}_3, \overline{cand}_7) &= P(\overline{cand}_3, \overline{cand}_7) \\ &= P(\overline{cand}_3)P(\overline{cand}_7), \end{aligned} \tag{3.22}$$

where the second equation comes from the fact that, if candidate $(\overline{NOMA_1})$ is removed, $cand_3$ and $cand_7$ become independent as they correspond to two independent channels. We also have

$$P(\overline{TDMA_2}, \overline{cand}_4, \overline{cand}_8) = P(\overline{cand}_4, \overline{cand}_8) = P(\overline{cand}_4), \tag{3.23}$$

where the second equation follows from the fact that both $cand_4$ and $cand_8$ correspond to channel h_{sr_2}, leading to duplicate $S \to R_k$ terms in (3.23).

Substituting (3.22) and (3.23) into (3.21) gives

$$P_{out}^{\mathbf{q}^{(i)}=[4,1,3,0]} = P(\overline{cand}_3)P(\overline{cand}_7)P(\overline{cand}_4)P(\overline{cand}_5)P(\overline{cand}_6). \tag{3.24}$$

Every term in (3.24) corresponds to one single packet transmission. This can be straightforwardly extended to general cases: i.e. the outage probability at state $\mathbf{q}^{(i)}$ can be obtained by removing all double-transmission and removing duplicated $S \to R_k$ link terms in (3.20). Candidates $(sr_{k,m})$ and $(r_k u_m)$ correspond to channels h_{sr_k} and $h_{r_k u_m}$ respectively. Supposing $cand_w$ corresponds to channel h_{d_k}, from (3.1), we have

$$\begin{aligned} P(\overline{cand}_w) = P\{\log_2(1+\gamma_{d_k}(t)) < \eta\} = 1 - e^{\left(-\frac{2^\eta - 1}{\bar{\gamma}_{d_k}}\right)}, \\ d_k \in \{sr_k, r_{k,1}u_1, r_{k,2}u_2\}. \end{aligned} \tag{3.25}$$

For the example in Fig. 3.2, $cand_3, \cdots, cand_7$ correspond to channels $h_{r_1 u_{1,1}}$, h_{sr_2}, $h_{r_2 u_{2,1}}$, h_{sr_1} and $h_{r_1 u_{1,2}}$ respectively. The above analysis leads to the following remark:

Remark 3 *The outage probability at any state depends only on the available single packet transmission candidates and not the double packet transmissions.*

3.3.2 Transition Probability

We suppose that if $cand_l$ is selected, the buffer state transits from $\mathbf{q}^{(i)}$ to $\mathbf{q}^{(i_l)}$, which occurs when all candidates with higher priority order than $cand_l$ are in outage and $cand_l$ is not in outage. Thus we have

$$A_{i_l,i} = P(\overline{cand}_1, \cdots, \overline{cand}_{l-1}, cand_l). \tag{3.26}$$

3.3.3 Double Transmission

Because double transmission candidates have higher priority than the single transmission candidates, no single transmission term is included in (3.26). In the example shown in Fig. 3.2, we have

$$\begin{aligned} A_{i_1,i} =& P(cand_1) = P((NOMA_1)) = 1 - P((\overline{NOMA_1})) \\ A_{i_2,i} =& P(\overline{cand}_1, cand_2) = P(\overline{cand}_1)P(cand_2) \\ =& P((\overline{NOMA_1}))(1 - P((\overline{TDMA_2})), \end{aligned} \tag{3.27}$$

where

$$P((\overline{NOMA_k})) = 1 - P_{k,(1,2)} - P_{k,(2,1)}, \tag{3.28}$$

where $P_{k,(1,2)}$ and $P_{k,(2,1)}$ are the probabilities that NOMA can be supported for (3.10) and (3.11) respectively. Following the similar procedures as those in [140], we have

$$\begin{aligned} P_{k,(m,n)} =& \frac{1}{\bar{\gamma}_{r_k u_m}} e^{\left(-\frac{(2^\eta-1)\bar{\gamma}_{r_k u_m} + (2^{2\eta}-2^\eta)\bar{\gamma}_{r_k u_n}}{\bar{\gamma}_{r_k u_m}\bar{\gamma}_{r_k u_n}}\right)} \\ & \times \int_{2^\eta-1}^{\infty} e^{\left(-\frac{x}{\bar{\gamma}_{r_k u_m}} - \frac{2^\eta(2^\eta-1)^2}{\bar{\gamma}_{r_k u_n} x}\right)} \mathrm{d}x \\ & - \frac{\bar{\gamma}_{r_k u_n}}{\bar{\gamma}_{r_k u_m} + \bar{\gamma}_{r_k u_n}} e^{\left(-\frac{(2^\eta-1)(\bar{\gamma}_{r_k u_m} + \bar{\gamma}_{r_k u_n})(2^{2\eta}+2^\eta)}{\bar{\gamma}_{r_k u_m}\bar{\gamma}_{r_k u_n}}\right)}, \end{aligned} \tag{3.29}$$

where $(m, n) \in \{(1, 2), (2, 1)\}$. On the other hand, we have

$$P((\overline{TDMA_k})) = P\{\log_2(1 + \gamma_{sr_k}(t)) < 2\eta\} = 1 - e^{\left(-\frac{2^{2\eta}-1}{\bar{\gamma}_{sr_k}}\right)}. \tag{3.30}$$

3.3.4 Single Transmission

For the example in Fig. 3.2, the transition probabilities when candidates $cand_3, \cdots, cand_6$ are selected are respectively obtained as

$$\begin{aligned}
A_{i_3,i} &= P(\overline{(NOMA_1)}, cand_3)P(\overline{(TDMA_2)}) \\
A_{i_4,i} &= P(\overline{(NOMA_1)}, \overline{cand_3})P(\overline{(TDMA_2)}, cand_4) \\
&= P(\overline{cand_3})P(\overline{(TDMA_2)}, cand_4) \\
A_{i_5,i} &= P(\overline{(NOMA_1)}, \overline{cand_3})P(\overline{(TDMA_2)}, \overline{cand_4})P(cand_5) \\
&= P(\overline{cand_3})P(\overline{cand_4})P(cand_5) \\
A_{i_6,i} &= P(\overline{(NOMA_1)}, \overline{cand_3})P(\overline{(TDMA_2)}, \overline{cand_4})P(\overline{cand_5}) \\
&\quad \times P(cand_6) \\
&= P(\overline{cand_3})P(\overline{cand_4})P(\overline{cand_5})P(cand_6).
\end{aligned} \tag{3.31}$$

On the other hand, we obtain the transition probabilities when candidates $cand_7$ and $cand_8$ are selected as

$$\begin{aligned}
A_{i_7,i} &= P(\overline{(NOMA_1)}, \overline{cand_3}, cand_7)P(\overline{(TDMA_2)}, \overline{cand_4}) \\
&\quad \times P(\overline{cand_5})P(\overline{cand_6}) \\
&= P(\overline{cand_3}, cand_7)P(\overline{cand_4})P(\overline{cand_5})P(\overline{cand_6}) \\
&= P(\overline{cand_3})P(cand_7)P(\overline{cand_4})P(\overline{cand_5})P(\overline{cand_6}) \\
A_{i_8,i} &= P(\overline{(NOMA_1)}, \overline{cand_3}, \overline{cand_7})P(\overline{cand_5})P(\overline{cand_6}) \\
&\quad \times P(\overline{(TDMA_2)}, \overline{cand_4}, cand_8) \\
&= P(\overline{cand_3}, \overline{cand_7})P(\overline{cand_4}, cand_8)P(\overline{cand_5})P(\overline{cand_6}) \\
&= 0,
\end{aligned} \tag{3.32}$$

where we make use of $P(\overline{cand}_3, cand_7) = P(\overline{cand}_3)P(cand_7)$ and $P(\overline{cand}_4, cand_8) = 0$ in obtaining (3.32). This is because the two associated $S \to R_k$ single transmission candidates $cand_4$ and $cand_8$ correspond to the same channel h_{sr_k}, and it is not possible that one of the candidates is in outage and the other is not. We have obtained all probability terms in (3.31) and (3.32) except $P((\overline{NOMA_k}), (r_{k,m}u_m))$ and $P((\overline{TDMA_k}), (sr_{k,m}))$ which are derived as follows:

$$\begin{aligned} P((\overline{NOMA_k}), (r_{k,m}u_m)) &= 1 - P((\overline{NOMA_k}), \overline{(r_{k,m}u_m)}) \\ &\quad - P((NOMA_k), (r_{k,m}u_m)) - P((NOMA_k), \overline{(r_{k,m}u_m)}). \end{aligned} \tag{3.33}$$

From remark 2, we have $P((\overline{NOMA_k}), \overline{(r_{k,m}u_m)}) = P\overline{(r_{k,m}u_m)}$, $P((NOMA_k), (r_{k,m}u_m)) = P(NOMA_k)$ and $P((NOMA_k), \overline{(r_{k,m}u_m)}) = 0$. Substituting these into (3.33) gives

$$\begin{aligned} P((\overline{NOMA_k}), (r_{k,m}u_m)) &= 1 - P\overline{(r_{k,m}u_m)} - P(NOMA_k) \\ &= P\overline{(NOMA_k)} - P\overline{(r_{k,m}u_m)}, \end{aligned} \tag{3.34}$$

where $P\overline{(NOMA_k)}$ and $P\overline{(r_{k,m}u_m)}$ are obtained in (3.28) and (3.25) respectively. Similarly we have

$$\begin{aligned} &P((\overline{TDMA_k}), (sr_{k,m})) \\ &= 1 - P((\overline{TDMA_k}), \overline{(sr_{k,m})}) - P((TDMA_k), (sr_{k,m})) \\ &\quad - P((TDMA_k), \overline{(sr_{k,m})}) \\ &= 1 - P\overline{(sr_{k,m})} - P(TDMA_k) \\ &= P\overline{(TDMA_k)} - P\overline{(sr_{k,m})}, \end{aligned} \tag{3.35}$$

where $P\overline{(TDMA_k)}$ and $P\overline{(sr_{k,m})}$ are obtained in (3.30) and (3.25) respectively. It is straightforward to extend the above analysis to general cases that the transition probability $A_{j,i}$ can always be decomposed into terms including $P\overline{(NOMA_k)}$, $P\overline{(TDMA_k)}$ and $P(\overline{cand}_w)$, where $cand_w$ is a single transmission candidate. The above analysis leads to the following remark:

Remark 4 *The double packet transmission candidates have higher priority to determine the transmission probabilities than the single packet transmission when both of double and single transmission candidates are not in outage. Only when the double transmission candidates are not available or in outage, do the single transmissions affect the transition probabilities.*

3.3.5 Average Throughput

Average throughput of the system is defined as the number of bits that reach the destination per a time-slot. To proceed with the analysis, buffer states can be casted as a Markov chain where each state describes a possible state of the buffers. The main difference in the proposed system and the traditional system is that the values of the transition probabilities differ from the source to the relays, since the source may transmits two packets to a relay rather than just one packet, and the same holds for the transition probabilities from the relays to the destinations, as two packets may get transmitted when both links are not in outage.

The transition matrix of the Markov chain **A** with the dimension of $(L+1)^n(L+1)^n$, $\mathbf{A}_{mn}$ is the notation for the mth row and nth column entry, which represents the transition probability to move from state q_n at time t to state q_m at time $t+1$:

$$\mathbf{A}_{mn} = P(X_{t+1} = q_m | X_t = q_n) \tag{3.36}$$

The transition probability A_{mn} depends on the state of the two buffers and the channel conditions of the links.

The Markov chain with the transition matrix **A** is irreducible and aperiodic. The Markov chain is said to be irreducible if all states are reachable starting from any state in the chain, and if the probability of staying at any state higher than zero, then the Markov chain is aperiodic, see [97], [20]. As in [68], in irreducible and aperiodic there exists a unique solution for the steady state distribution ($\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_{(L+1)^{2K}}]^T$, π_l is the probability that the buffer state is q_l) where

$$\mathbf{A}\boldsymbol{\pi} = \boldsymbol{\pi}, \tag{3.37}$$

$\boldsymbol{\pi}$ is unchanged by the operation of $\mathbf{A}$

$$\sum_{i=1}^{(L+1)^{2K}} \boldsymbol{\pi_i} = 1, \tag{3.38}$$

$$\mathbf{B}\boldsymbol{\pi} = \mathbf{b}, \tag{3.39}$$

where $\mathbf{b} = [1, \cdots, 1]^T$ and $\mathbf{B}$ denotes a $(L+1)^{2K} \times (L+1)^{2K}$ matrix with all elements of one. From (3.37) and (3.39):

$$\mathbf{A}\boldsymbol{\pi} - \boldsymbol{\pi} + \mathbf{B}\boldsymbol{\pi} = b \implies \boldsymbol{\pi} = (\mathbf{A} - \mathbf{I} + \mathbf{B})^{-1}\mathbf{b}, \tag{3.40}$$

where $\mathbf{I}$ denotes the identity matrix.

A key aspect that governs the system performance is the throughput of the system. In the existing OMA and NOMA buffer-aided relay systems, the throughput is easily calculated because the relay is transmitting a fixed number of packets in case of no outage, but in the hybrid system, in case of no outage, the relay may send one or two packets. For calculating the throughput of the hybrid system, the following need to be considered: if NOMA mode is selected with no outage, then two packets are sent from the relay to the users, while if OMA mode is the chosen mode, one packet is sent to the users.

In other words, at any time-slot, if candidate $(r_{k,1}u_1)$ or $(r_{k,2}u_2)$ is selected, one packet is transmitted to user U_1 or U_2, respectively. While if candidate $(NOMA_k)$ is selected, two packets are transmitted from R_k to the users. At state $\mathbf{q}^{(i)}$, the probabilities to select candidates $(r_{k,1}u_1)$, $(r_{k,2}u_2)$ and $(NOMA_k)$ are denoted as $P^{(i)}_{r_{k,1}u_1}$, $P^{(i)}_{r_{k,2}u_2}$ and $P^{(i)}_{NOMA_k}$ respectively, which are zero if the corresponding candidates are not available at state $\mathbf{q}^{(i)}$ and otherwise are obtained as in (3.26). Considering all buffer states and all relay nodes, the average throughput for user U_m is given by

$$\xi_m = \sum_{i=1}^{(L+1)^2} \pi_i \xi_m^{(i)} = \sum_{i=1}^{(L+1)^2} \pi_i \sum_{k=1}^{K} \left(P^{(i)}_{r_k u_m} + P^{(i)}_{NOMA_k} \right), \tag{3.41}$$

where $m \in \{1, 2\}$, $\xi_m^{(i)} = \sum_{k=1}^{K} \left(P_{r_k u_m}^{(i)} + P_{NOMA_k}^{(i)} \right)$ which is the average throughput for user U_m at state $\mathbf{q}^{(i)}$. And the sum throughput for all users is given by $\xi = \xi_1 + \xi_2$.

For illustration, Fig. 3.3 shows all possible buffer state transition at the state $\mathbf{q}^{(i)} = [4, 1, 3, 0]$ for the example in Fig. 3.2, where the single and double arrows represent the state transitions due to the single and double packet transmission, respectively. The average throughput for users U_1 and U_2 at this state are given by $\xi_1^{(i)} = A_{i_1,i} + A_{i_3,i} + A_{i_5,i}$ and $\xi_2^{(i)} = A_{i_1,i} + A_{i_7,i}$ respectively. In other words, in Fig. 3.2, for single packet transmission, one of the two buffers in a relay node will be changed. To be specific, if a source-to-relay link is selected, the corresponding buffer will be increased by one; if a relay-to-destination link is selected, the corresponding buffer will be decreased by one. While for double packet transmission, both buffers will be changed. To be specific, if a source-to-relay TDMA transmission is selected, both buffers of the corresponding relay will be increased by one; if a relay-to-destination NOMA transmission is selected, both buffers of the corresponding relay will be decreased by one. Accordingly, the change of the buffer length will lead to the corresponding change in buffer state.

3.3.6 Diversity Order

The diversity order is defined as

$$d = -\lim_{\bar{\gamma} \to \infty} \frac{\log P_{out}}{\log \bar{\gamma}}, \tag{3.42}$$

where $\bar{\gamma} = P_t/\sigma^2$ and P_{out} is the outage probability. Outage probability of the system is defined as the probability that all relays neither transmit to the users nor receive from the source. When this happens, the number of packets stored in the buffers remains the same, which means the Markov chain remains in the same state, so outage probability can be calculated as follows:

$$P_{out} = \sum_{i} P_{out}^{\mathbf{q}^{(i)}} \pi_i. \tag{3.43}$$

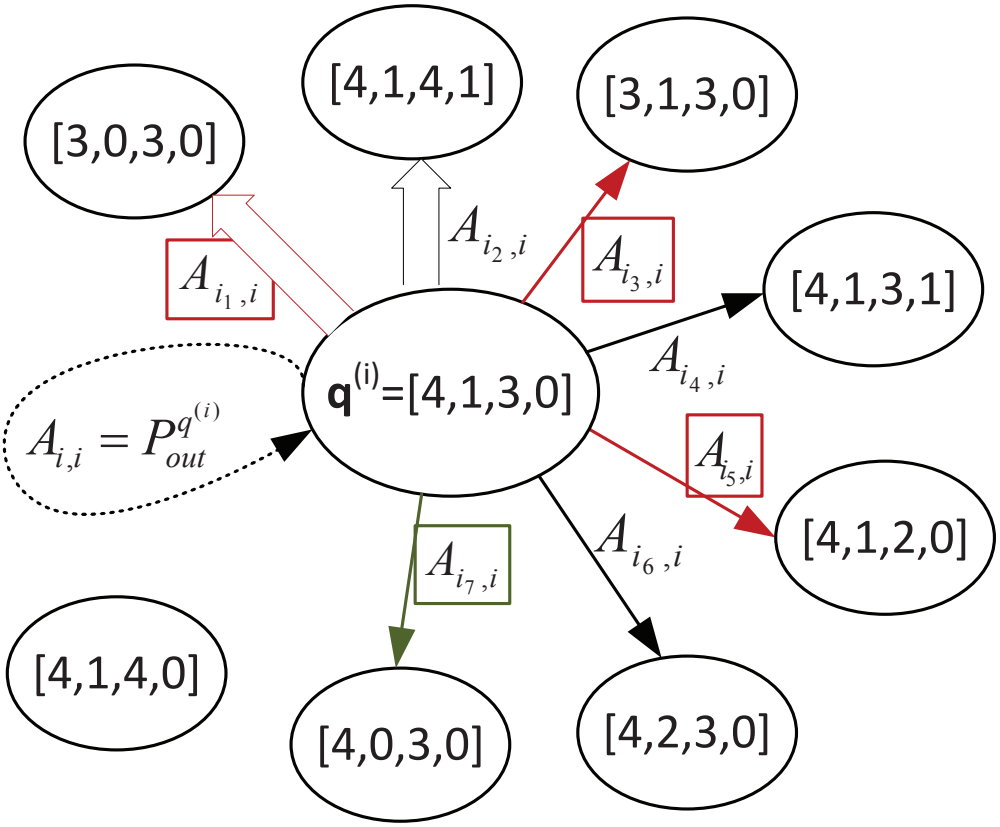

Figure 3.3 State transition diagram at the state $\mathbf{q}^{(i)} = [4, 1, 3, 0]$ for the example in Fig. 3.2, where the single and double arrows represent the state transitions due to the single and double packet transmissions respectively.

The diversity order depends on both the outage probabilities at every state $P_{out}^{\mathbf{q}^{(i)}}$ and the stationary buffer state probabilities π_i.

When $\bar{\gamma} \to \infty$, all transmission candidates are able to support the target rate transmission. Thus if the target buffer length is set as $2 \leq \Theta_i < L$ (where we assume the buffer size $L \geq 3$), according to the proposed prioritization-based selection rule, the buffer lengths at any time-slot are either Θ_i or $\Theta_i - 1$ which are neither empty nor full. From Remark 4, the transition probabilities are then only determined by the double transmission candidates (because they are all available), and the buffers can only be in two states: either all buffer lengthes are Θ_i, or only the pair of buffers for one of the relays have length of $\Theta_i - 1$ and all other buffer lengthes are Θ_i. In both cases, the corresponding $P_{out}^{\mathbf{q}^{(i)}}$ are the same. Further from Remark 3, $P_{out}^{\mathbf{q}^{(i)}}$ only depends on the single transmission candidates which are also

all available. Therefore, if the target buffer length is set as $2 \leq \Theta_i < L$, we have

$$\begin{aligned} d &= -\lim_{\bar{\gamma}\to\infty} \frac{\log P_{out}}{\log \bar{\gamma}} = -\lim_{\bar{\gamma}\to\infty} \frac{\log P_{out}^{\mathbf{q}^{(i)}}}{\log \bar{\gamma}} \\ &= -\lim_{\bar{\gamma}\to\infty} \frac{\log \prod_{k=1}^{K} P(C_{sr_k} < \eta) P(C_{r_k u_1} < \eta) P(C_{r_k u_2} < \eta)}{\log \bar{\gamma}} \\ &= 3K, \end{aligned} \tag{3.44}$$

where (3.25) is substituted in the second equation of above to give the final result. This states that every relay contributes 3 diversity orders to the system, corresponding to $S \to R_k$, $R_k \to U_1$ and $R_k \to U_2$ transmission respectively. It is interesting to note that if only NOMA transmission is applied (as in [140]), the diversity order is $2K$.

3.3.7 Discussion

Below we explain that the proposed scheme has higher sum throughput than both buffer-aided NOMA and OMA schemes. Recall that the network throughput can be regarded as $\eta(1 - P_{out})$, where η is the data rate (without considering the outage). From Remark 3, the outage probability of the proposed scheme depends on the single packet transmission, which is significantly lower than that of the buffer-aided NOMA relay selection (which only applies the double packet transmission).

On the one hand, because the proposed scheme gives higher priority to the double packet transmission than the single packet transmission, the double packet transmission will always be selected first when possible. This implies that the data rate η of the proposed scheme is no less than that of NOMA scheme. Thus we have

$$\xi_{proposed} > \xi_{NOMA}, \tag{3.45}$$

where $\xi_{proposed}$ and ξ_{NOMA} are the sum throughput for the proposed and buffer-aided NOMA schemes respectively.

On the other hand, compared with the buffer-aided OMA scheme (which only applies single packet transmission), the proposed scheme has similar outage proba-

bility but higher data rate. Thus we also have

$$\xi_{proposed} > \xi_{OMA}, \tag{3.46}$$

where ξ_{OMA} is the sum throughput for OMA scheme. In summary, next section verifies all the above mentioned results:

- $\xi_{proposed} > \xi_{NOMA}$.
- Outage probability of the proposed scheme is significantly lower than that of the buffer-aided NOMA .
- $\xi_{proposed} > \xi_{OMA}$.
- The proposed scheme has similar outage probability of the buffer-aided OMA, because it depends on the single packet transmission.

As is mentioned in the introduction section, a simple alternative to combine NOMA and OMA in the buffer-aided relay selection is to set an appropriate threshold SNR, SNR_t, where $\xi_{NOMA} < \xi_{OMA}$ for $SNR \leq SNR_t$, and $\xi_{NOMA} > \xi_{OMA}$ for $SNR > SNR_t$. Then we can simply apply the buffer-aided OMA scheme if $SNR < SNR_t$ and switch to NOMA scheme otherwise. It is clear the throughput of the switch-based scheme satisfies

$$\xi_{switch} = \begin{cases} \xi_{OMA}, & \text{if } SNR \leq SNR_t, \\ \xi_{NOMA}, & \text{if } SNR > SNR_t. \end{cases} \tag{3.47}$$

Using (3.45) and (3.46) in (3.47), it is clear that the proposed scheme has higher throughput than the switch-based scheme as

$$\xi_{proposed} > \xi_{switch}, \tag{3.48}$$

this is also verified in the next section.

3.4 Numerical Simulations

In all simulations below, the target transmission rate for both users is set to $\eta_1 = \eta_2 = 2$ bps, the buffer size is set to $L = 5$ for every buffer, all noise powers σ^2 are normalized to unity and bandwidth is normalized to 1Hz in all links.

First we consider the single relay scenario. This is for easy comparison with the buffer-aided NOMA scheme in [140] which considers the same scenario. The average channel gains are set to $\Omega_{sr_1} = 1.1$ dB, $\Omega_{r_1u_1} = 1.0$ dB and $\Omega_{r_1u_2} = 1.5$ dB.

In Fig. 3.4, we show the outage probability *vs* transmission SNR P_t/σ^2 of the proposed scheme and the buffer-aided NOMA scheme in [140]. It can be seen that the proposed scheme outperforms buffer-aided NOMA system in terms of outage probability. This verifies that systems which depend on a single packet transmission have lower outage probability.

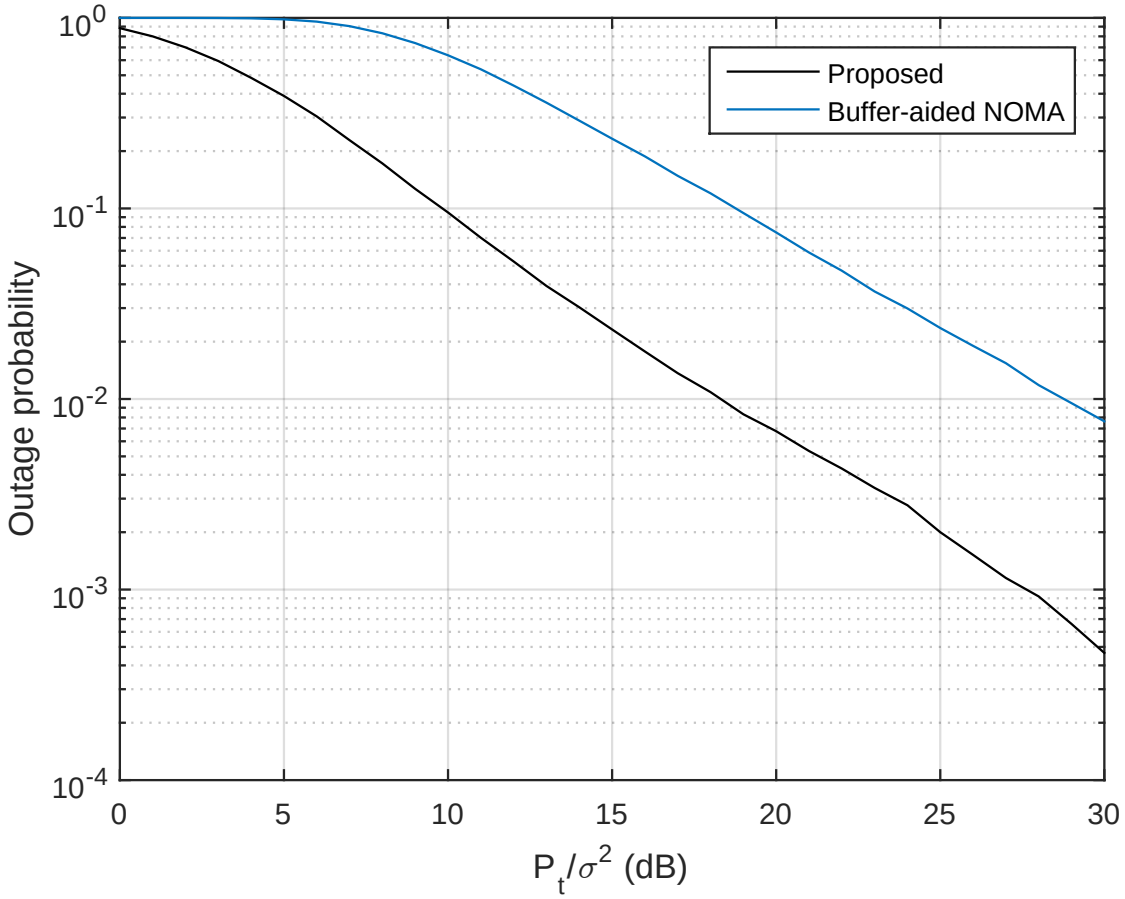

Figure 3.4 Outage probability comparison between the proposed scheme and the buffer-aided NOMA.

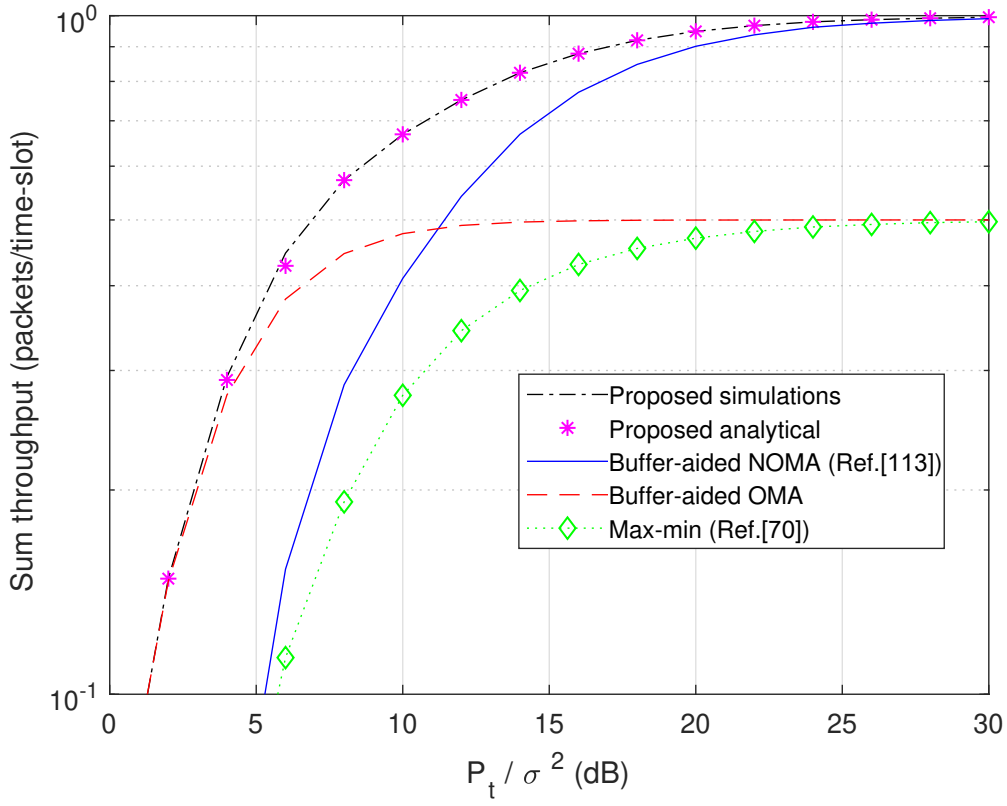

Figure 3.5 Throughput of the buffer-aided NOMA, OMA and proposed schemes, where the relay number $K = 1$, buffer size $L = 5$.

Fig. 3.5 shows the sum throughput *vs* transmission SNR P_t/σ^2 for the proposed scheme, the buffer-aided NOMA scheme in [140], the buffer-aided OMA scheme and the no-buffer traditional max-min scheme. The buffer-aided OMA scheme uses the same selection rule as that in the proposed scheme except that NOMA transmissions are not included in the selection process. The target buffer-lengths in both the proposed and OMA schemes are set to 3, while buffer-aided NOMA scheme has target buffer-length of 2 as proposed in [140].

In Fig. 3.5, the analytical results very well match the simulation results for the proposed scheme, which verifies the analysis in Section 3.3. Fig. 3.5 also shows that both buffer-aided NOMA and the proposed scheme can achieve full throughput rate, i.e. one packet/time-slot, when the SNR is large enough. Where system full throughput rate is one packet/time-slot because we assumed half-duplex relays and data rate is $= 2$ bps. On the other hand, OMA schemes buffer-aided and

non-buffer-aided (max-min) can only achieve the maximum throughput of 1/2 packet/time-slot. This is because NOMA delivers two packets simultaneously.

As expected, NOMA scheme has larger throughput than OMA over the high SNR range (i.e. $P_t/\sigma^2 \geq 12$ dB), but has worse throughput than the latter over the low SNR range (i.e. $P_t/\sigma^2 < 12$ dB). On the contrary, the proposed scheme can achieve significant throughput improvement over both low and high SNR ranges.

It is interesting to observe that, if we simply apply the switch-based scheme in which the buffer-aided OMA scheme is used in the low SNR range (i.e. $P_t/\sigma^2 < 12$ dB) and the buffer-aided NOMA scheme is used in the high SNR range (i.e. $P_t/\sigma^2 \geq 12$ dB), the throughput will still be significantly lower than that in the proposed scheme as can be seen in Fig. 3.6. This verifies (3.48) in the discussions in the last section. In all cases, the non-buffer-aided max-min scheme has the lowest throughput. It is worth mentioning that the switch-based scheme curve is the result of switching from OMA to NOMA at 12 dB in Fig. 3.5

Fig. 3.7 shows the sum throughput for the 2-relay network[1]. Because multiple relays are not considered in [140], the selection rule of the buffer-aided NOMA scheme in Fig. 3.7 is the same as that for the proposed scheme by excluding OMA transmission modes. It is clearly shown in Fig. 3.7 that while all of the three schemes achieve higher throughput than those in Fig. 3.5, the comparison among the three schemes is similar.

Fig. 3.8 and Fig. 3.9 show the throughput and outage probability of the proposed scheme for different relay numbers respectively, where the target buffer length is set to 3 in all cases. Fig. 3.8 shows that higher throughput is achieved with more relay nodes. This is because of the higher diversity order with more relays as shown in Fig. 3.9. According to (3.42), the diversity orders are calculated in Table 3.1. It is clearly shown that the diversity order is approximately $3K$ which well matches the analysis in (3.44).

Figures. 3.10, 3.11 and 3.12 show the throughput *vs* the target buffer lengths for the proposed scheme for the 2-relay network. Three cases are considered. In

[1]To the best of our knowledge, there are not many existing algorithms which can be directly compared with the proposed scheme. Even in NOMA scheme in Ref. [140], only the single relay network was considered. We have to generate it to the multiple-relay case in the following figures.

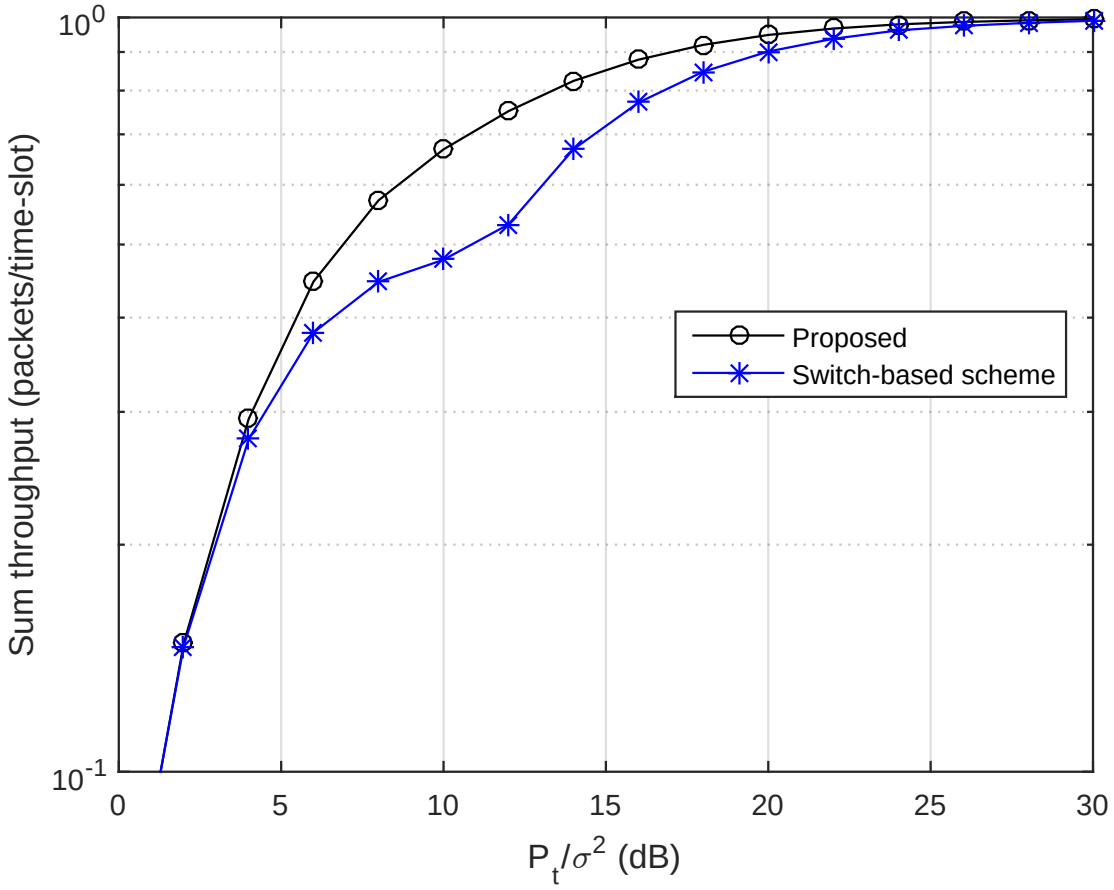

Figure 3.6 Throughput versus SNR in hybrid and switching modes.

case (a), all channels have the same average gains ($\bar{\gamma}_{sr_1} = \bar{\gamma}_{sr_2} = \bar{\gamma}_{r_1u_1} = \bar{\gamma}_{r_1u_2} = \bar{\gamma}_{r_2u_1} = \bar{\gamma}_{r_2u_2} = 7$ dB). Because $S \to R_k$ and $R_k \to U_m$ apply different transmission modes, even with the same average gains for all channels, the input/output rate at the buffers is still not balanced so that the optimum target length is not two. This is clearly shown in Case (a) where the optimum target length which achieves the largest throughput is three. In Case (b), $S \to R_k$ channels are much stronger than the $R_k \to U_m$ channels where $\bar{\gamma}_{sr_1} = \bar{\gamma}_{sr_2} = 10\bar{\gamma}_{r_1u_1} = 10\bar{\gamma}_{r_1u_2} = 10\bar{\gamma}_{r_2u_1} = 10\bar{\gamma}_{r_2u_2} = 10$ dB, so that the buffers are more likely to be saturated. As a result, the optimum target length shall be close to zero, which is clearly verified in Case (b). In Case (c), on the other hand, the $S \to R_k$ channels have much lower average gains than the $R_k \to U_m$ channels where we set as $\bar{\gamma}_{r_1u_1} = \bar{\gamma}_{r_1u_2} = \bar{\gamma}_{r_2u_1} = \bar{\gamma}_{r_2u_2} = 30\bar{\gamma}_{sr_1} = 30\bar{\gamma}_{sr_2} = 13$ dB, and so the buffers tend to be empty. In this case, the optimum

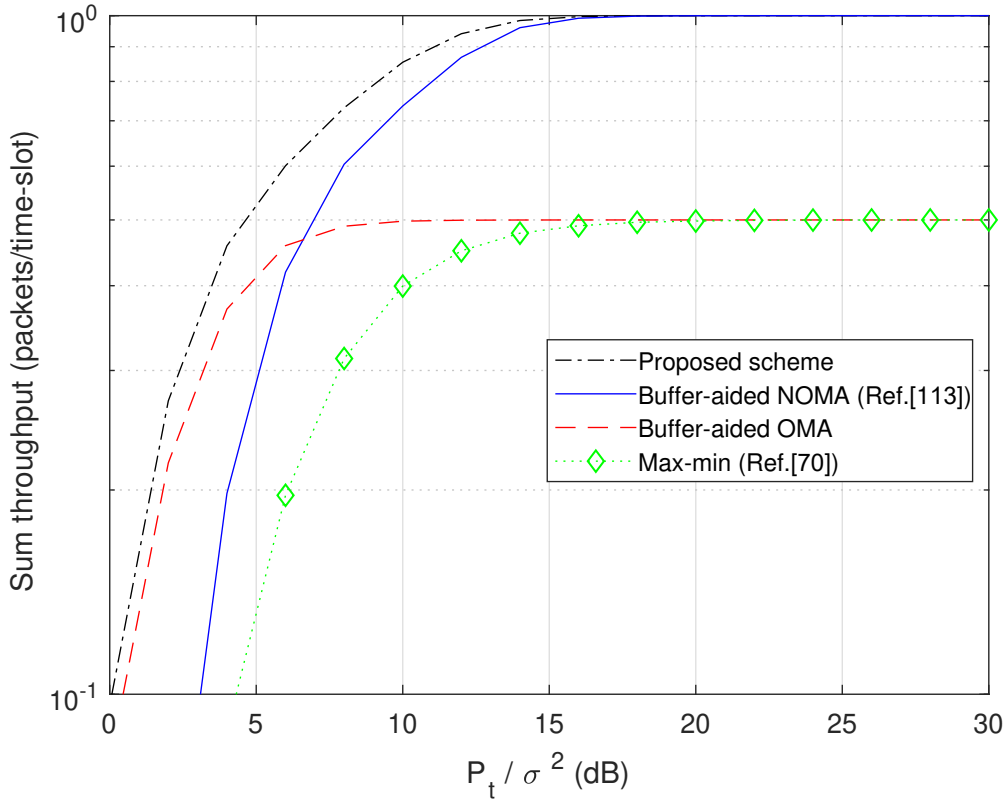

Figure 3.7 Throughput of the buffer-aided NOMA, OMA and proposed schemes, where the relay number $K = 2$, buffer size $L = 5$.

target length shall be chosen close to the full buffer size. These results very well match the statements in Section 3.2.2.

3.5 Summary

This chapter proposes a buffer-aided relay selection scheme to seamlessly include both NOMA and OMA transmission with finite buffer size. The proposed scheme achieves significant improvements in throughput over both low and high SNR ranges. A prioritization-based selection rule is described by introducing the target buffer length for every buffer. The analytical expression of the average throughput is successfully obtained and verified by numerical simulations. Particularly, the diversity order of the proposed scheme is obtained as $3K$, where K is the number of relays. This provides useful insight for designing the cooperative NOMA for 5G

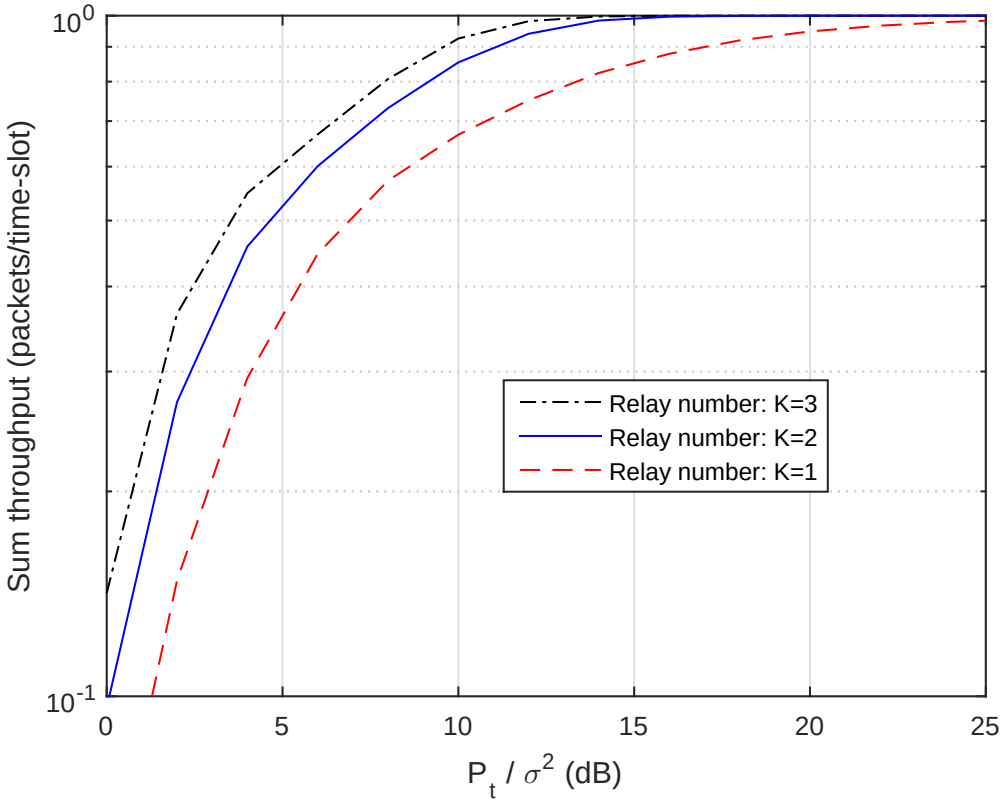

Figure 3.8 Throughput of the proposed scheme for different relay numbers, where all average channel gains are set to 1 and the target buffer length is set to 3.

applications. In the next chapter, we study the impact of considering the delay that the packet encounters while being at the source node on the system delay.

Table 3.1 Diversity orders from Fig. 3.9

K	$P_{out}(P_t/\sigma^2)$ (dB)	$P_{out}(P_t/\sigma^2)$ (dB)	Diversity order
1	$P_{out}(18\text{ dB}) = 38.7$	$P_{out}(20\text{ dB}) = 44.6$	$\frac{44.6-38.7}{20-18} \simeq 3$
2	$P_{out}(10\text{ dB}) = 30.1$	$P_{out}(11\text{ dB}) = 36.2$	$\frac{36.2-30.1}{11-10} \simeq 6$
3	$P_{out}(8\text{ dB}) = 29.6$	$P_{out}(9\text{ dB}) = 38.5$	$\frac{38.5-29.6}{9-8} \simeq 9$

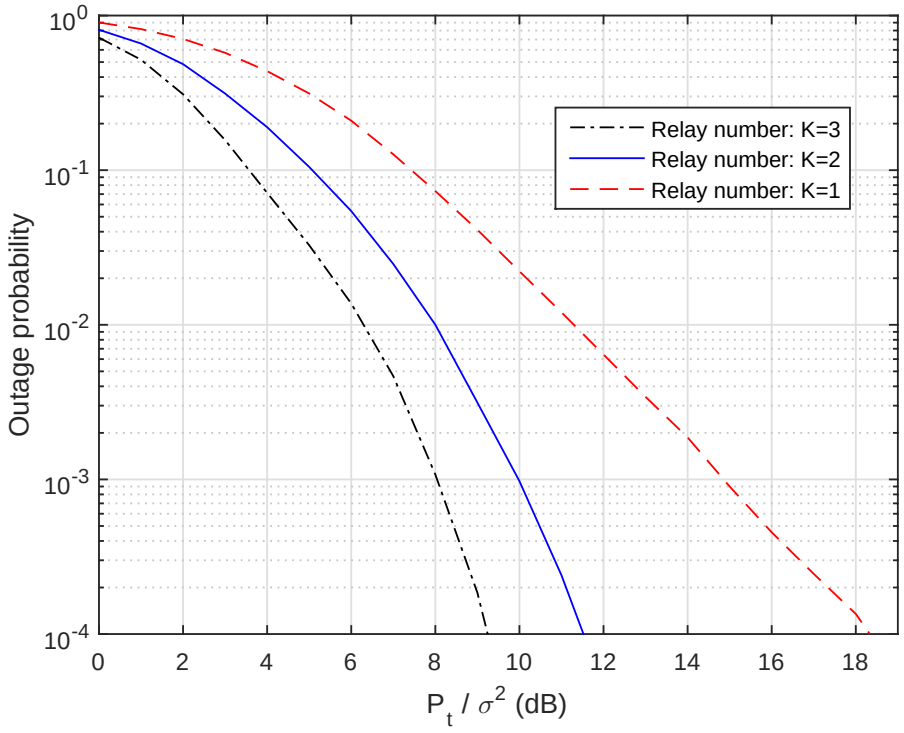

Figure 3.9 Outage probability of the proposed scheme for different relay numbers, where all average channel gains are set to 1 and the target buffer length is set to 3.

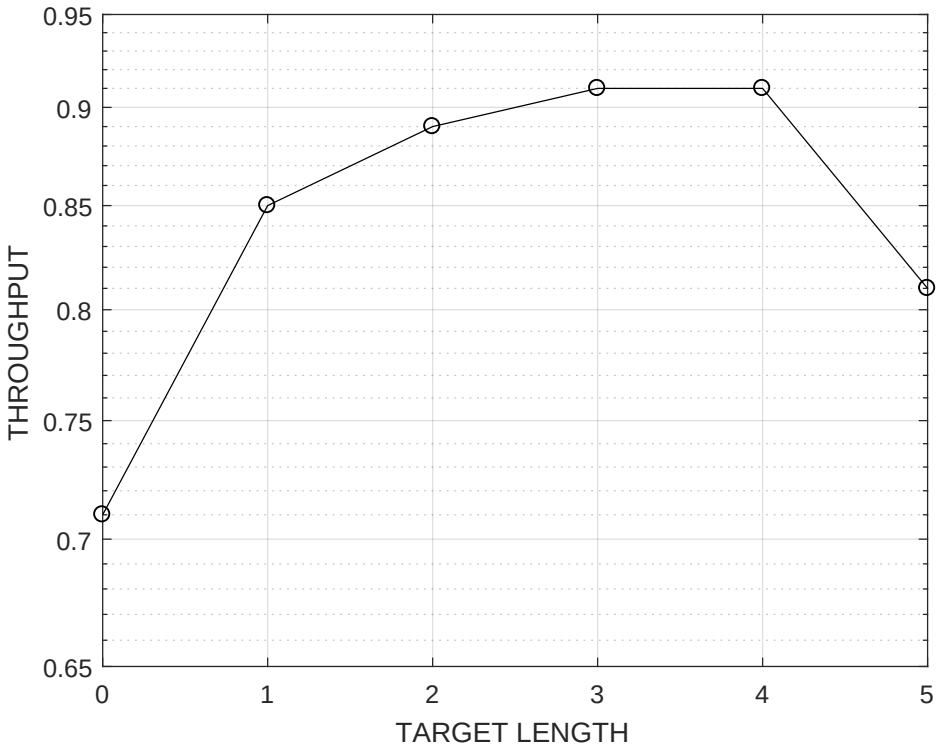

Figure 3.10 Throughput *vs* target buffer lengths for the 2-relay network. Case (a): $\bar{\gamma}_{sr_1} = \bar{\gamma}_{sr_2} = \bar{\gamma}_{r_1u_1} = \bar{\gamma}_{r_1u_2} = \bar{\gamma}_{r_2u_1} = \bar{\gamma}_{r_2u_2} = 7$ dB.

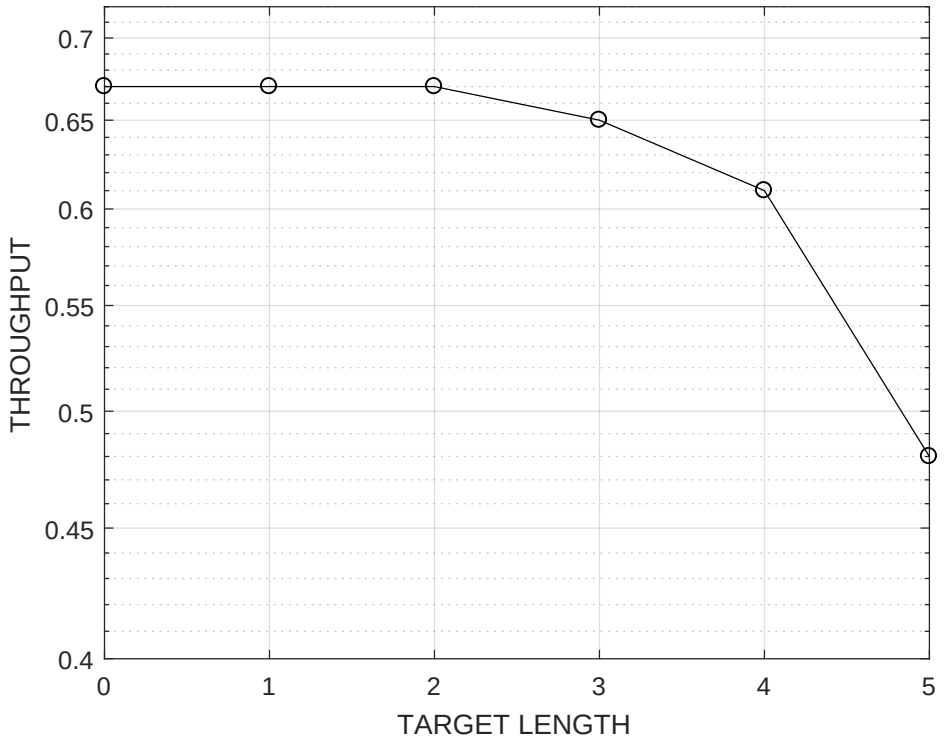

Figure 3.11 Throughput *vs* target buffer lengths for the 2-relay network. Case (b): $\bar{\gamma}_{sr_1} = \bar{\gamma}_{sr_2} = 10\bar{\gamma}_{r_1u_1} = 10\bar{\gamma}_{r_1u_2} = 10\bar{\gamma}_{r_2u_1} = 10\bar{\gamma}_{r_2u_2} = 10$ dB.

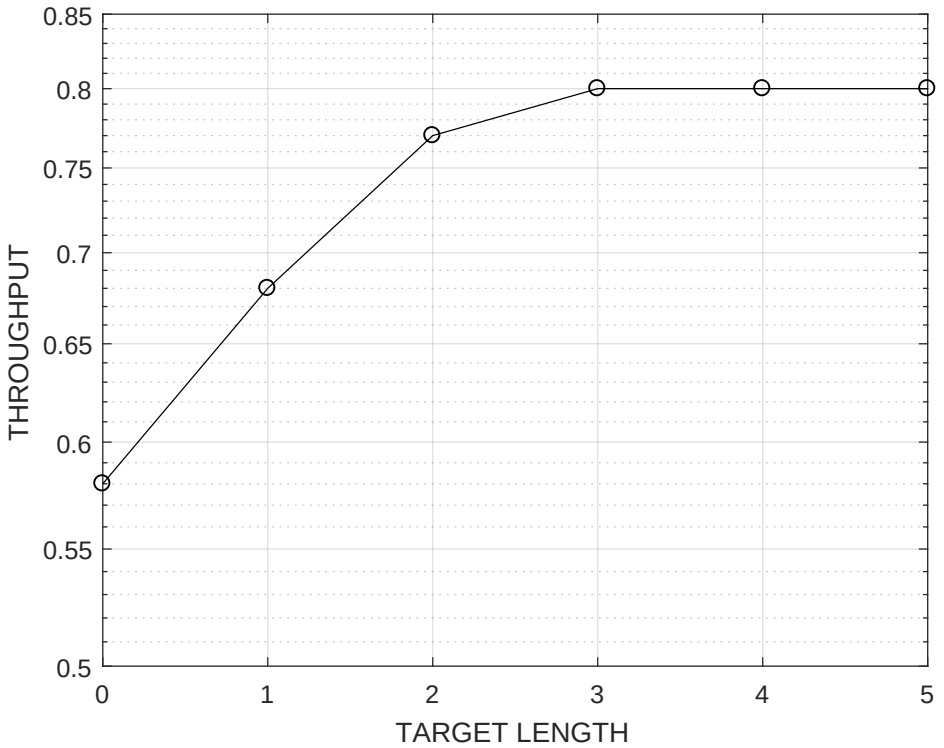

Figure 3.12 Throughput *vs* target buffer lengths for the 2-relay network. Case (c): $\bar{\gamma}_{r_1u_1} = \bar{\gamma}_{r_1u_2} = \bar{\gamma}_{r_2u_1} = \bar{\gamma}_{r_2u_2} = 30\bar{\gamma}_{sr_1} = 30\bar{\gamma}_{sr_2} = 13$ dB.

Chapter 4

The Impact of Source Delay on End-to-End Average Packet Delay in Buffer-Aided Cooperative Relay Networks

Relays with buffering capability have the potential to enhance the performance of a cooperative network significantly. However, such an enhancement comes with difficulties such as data buffering may increase packet delays considerably. In this chapter, a new factor in terms of the delay incurred at the source is considered for the first time in the literature. This new approach enables calculation of end-to-end delay (from source to destination) more accurate, which is essential for delay-sensitive applications. In addition, a novel selection rule is proposed to reduce the end-to-end delay.

4.1 Introduction

In buffer-aided relay schemes, the average packet delay is proportional to the number of relays and the size of buffers [131], so the average packet delay increases as the number of relays increases or with larger buffer size. However, adding more relays or enlarging buffer size may enhance some other performance metrics, such

as diversity gain and throughput. Therefore, a trade-off between delay and other performance metrics has to be considered. This trade-off is challenging in 5G networks due to stringent delay requirements, while requiring an extremely high data rate in order of 10 Gbps (such as the case in tactile internet) [121, 63, 116].

As in [117] and to the best of our knowledge, none of the previous buffer-aided relay works has considered the delay at the source. In delay-sensitive applications, the source delay can not be neglected and has to be counted in the maximum permitted delay [117]. For example, tactile internet requires a real-time response. This requirement makes the end-to-end delay (form the source to the destination) to be in the order of 1 ms starting from the point that the source has data for transmission[44]. The end-to-end delay of any packet is defined as its queuing delay added to its delivery delay [73]

In order to build the case for source delay consideration, we first briefly describe some of the main buffer-aided relay selection schemes followed by a study on the impact of the source delay on the described schemes. Relay selection schemes are discussed in further detail in Chapter 2.

Prior to -aided relays, the optimal relay selection scheme is the max-min [24]. The max-min scheme selects the relay, which has the best weaker link. Then the source transmits a packet to the selected relay in one time-slot, and the relay forwards the packet to the destination in the next time-slot. Thus, the max-min delay without the source delay is two time-slots.

After introducing -aided relays, one of the most popular relay selection schemes is the max-link [68]. The max-link selects the link, which has the highest SNR, which leads to an increase in the throughput and the diversity order compared to the max-min. However, the max-link has longer queues than the max-min because packets in the max-link can reside in s. After the max-link, reducing packet latency became a trend in the latest studies. One of the studies proposed an altered form of the max-link that is the state-based [79]. The authors have considered the BSI in their selection procedure. The results show that the state-based has reduced the values of delay compared to the max-link.

Finally, in [121], the authors have suggested the delay-reduced, which prioritizes transmission always. The delay-reduced reduces the packet delay of -aided relay network effectively compared to the state-based [48]. However, with all these enhancements still, the traditional max-min significantly outperforms -aided relays in terms of the packet delay, especially at low SNR.

This encourages us to consider the source delay in the aforementioned schemes; this makes the comparison more accurate. In this chapter, we suggest considering the source delay, which represents a part of the delay that packets encounter at the source before the transmission. By doing so, the delay definition is generalized to cover all parts of the delay, which leads to a more accurate comparison between -aided and -buffer relays. The results show that -aided relay outperforms -buffer relay in the average packet delay with considering the source delay. Thus, buffer-aided relays are still competitive, even with low latency constraints. The main contributions in this chapter are listed as follows

1. Considering the source delay in -aided relays network, which gives more accurate results for the delays that the packets in the system encounter.

2. Conducting the theoretical outage probability by modeling different buffer states as a Markov chain state.

3. Deriving the analytical expressions for -aided network delay by considering the source delay. In addition, the asymptotic performance of the system by considering the source delay is examined.

4. Proposing a novel selection scheme where the adaptive target length is introduced. The simulations show that the new rule outperforms the delay-reduced in shortening the delay.

The remainder of this chapter is organized as follows. In Section 4.2, the system model is presented. Section 4.3, discuses the outage probability in -aided relay network. Next, Section 4.4 presents the average packet delay with the source delay analysis for -aided relay networks. In Section 4.5, the asymptotic delay performance is discussed. Section 4.6 proposes a new selection rule. Simulation

results thoroughly discussed in Section 4.7. Finally, a summary concludes this chapter in Section 4.8.

4.2 System Model

The system model of -aided relay networks is shown in Fig. 4.1, where there are one source node S, K half-duplex decode-and-forward (DF) relay nodes denoted as R_k, $k = 1, \cdots, K$ and a destination D. The channel coefficients for $S \rightarrow R_k$, $R_k \rightarrow D$ links are denoted as $h_{sr_k}(t)$, $h_{r_kd}(t)$ respectively. All channels have a flat Rayleigh fading coefficient that remains constant within the time slot and change independently from one slot to another.

Every relay R_k is equipped with one L-size buffer for data storing. We assume that the source always has enough information (saturated) to send to relays in all time slots. Due to path loss and shadowing, we assume that the source and the destination are not directly connected. Without losing generality, we assume the transmit power P_t at all transmitting nodes, the assumption for the noise variances at all receiving nodes to be σ^2.

The data rate is assumed to be fixed at the value of η. If the link capacity is greater or equal to η, the link is up, and the transmission is successful. Retransmitting happens based on the ACK/NACK mechanism; this happens between transmitters (source or relay) and receivers (relay or destination). Each receiver broadcasts the ACK/NACK signal to the transmitters: relays $\rightarrow$ source and destination $\rightarrow$ relays. Channel state information at the receivers (CSIR) is assumed to be available.

At time slot t, the link capacity for channels $h_{sr_k}(t)$ and $h_{r_kd}(t)$ are given by

$$C_{sr_k}(t) = \log_2\left(1 + \gamma_{sr_k}(t)\right),\ C_{r_kd}(t) = \log_2\left(1 + \gamma_{r_kd}(t)\right),\ k = 1, \cdots, K, \tag{4.1}$$

where $\gamma_{sr_k}(t) = (P_t/\sigma^2)|h_{sr_k}(t)|^2$ and $\gamma_{r_kd}(t) = (P_t/\sigma^2)|h_{r_kd}(t)|^2$. The channel gains $|h_{sr_k}(t)|^2$ and $|h_{r_kd}(t)|^2$ are exponentially distributed with the average $\Omega_{sr_k} = E[|h_{sr_k}(t)|^2]$ and $\Omega_{r_kd} = E[|h_{r_kd}(t)|^2]$, where $E[.]$ is the expectation. $\gamma_{sr_k}(t)$ and

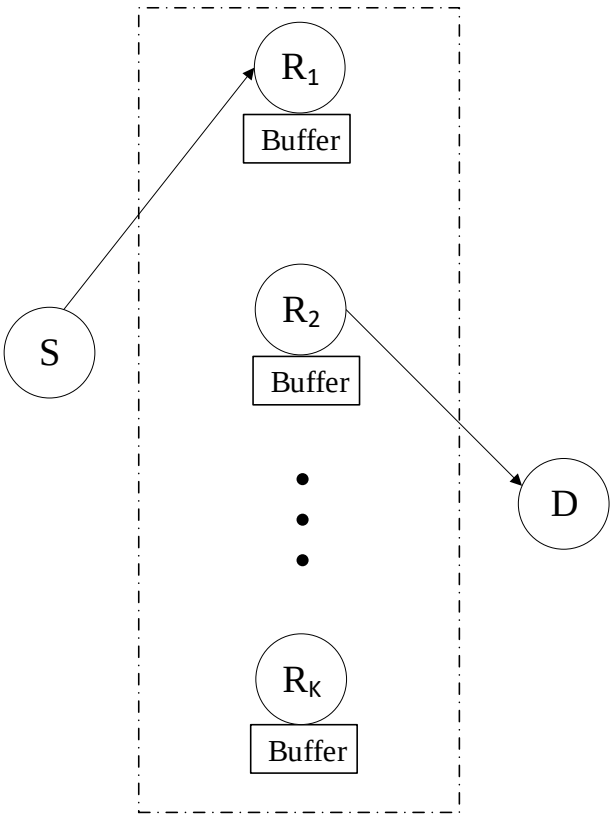

Figure 4.1 System model for -aided relay network.

$\gamma_{r_k d}(t)$ are also exponentially distributed with average $\bar{\gamma}_{sr_k} = (P_t/\sigma^2)\Omega_{sr_k}$ and $\bar{\gamma}_{r_k d} = (P_t/\sigma^2)\Omega_{r_k d}$. Thus $\gamma_{sr_k}(t)$ and $\gamma_{r_k d}(t)$ are the instantaneous SNR, while $\bar{\gamma}_{sr_k}$ and $\bar{\gamma}_{r_k d}$ are the average SNR for channels $h_{sr_k}(t)$ and $h_{r_k d}(t)$ respectively.

The delay-reduced has shown the best delay profile among the available buffer-aided relay selection schemes. Thus, we choose it as the selection scheme for -aided relay network to study the impact of source delay on -aided and -buffer-aided relay networks.

4.3 Outage Probability

Since the source delay is proportional to the outage, we first describe the outage analysis of -aided relay network. The outage occurs if the link capacity is less than the target data rate

$$
\begin{aligned}
P\{\log_2(1+\gamma_{sr_k}(t)) < \eta\} &= 1 - e^{\left(-\frac{2^\eta - 1}{\bar{\gamma}_{sr_k}}\right)} \\
P\{\log_2(1+\gamma_{r_kd}(t)) < \eta\} &= 1 - e^{\left(-\frac{2^\eta - 1}{\bar{\gamma}_{r_kd}}\right)}.
\end{aligned}
\tag{4.2}
$$

In each of the relays buffers, the numbers of data packets represent a state. Since there are K relays with L buffer size, there are $(L+1)^K$ states in total. Every state suggests the numbers of the available $S \to R_k$ and $R_k \to D$ links. Any $S \to R_k$ link is available if the receiving buffer is not full, and any $R_k \to D$ link is available if the transmitting buffer is not empty. The l-th state vector is defined as

$$
\mathbf{q}^{(l)} = [q_1^{(l)}, q_2^{(l)}, \cdots, q_K^{(l)}], \qquad l = 1, \cdots, (L+1)^K, \tag{4.3}
$$

where $q_k^{(l)}$ is length at R_k at state $\mathbf{q}^{(l)}$

By taking all possible states into consideration, the outage probability is the probability that the system remains at the same state, which means that no communication happened during the current time-slot. Hence the outage probability of buffer-aided system can be obtained as

$$
P_{out} = \sum_{i=1}^{(L+1)^K} P_{out}^{\mathbf{q}^{(i)}} \pi_i. \tag{4.4}
$$

where π_i is the stationary probability for state $\mathbf{q}^{(i)}$, and $P_{out}^{\mathbf{q}^{(i)}}$ is the outage probability at state $\mathbf{q}^{(i)}$. Since we are using the delay-reduced, the rest of the analysis is done based on this assumption. Hence, the outage occurs if all $R_k \to D$ links then all $S \to R_k$ links are in outage. So the outage probability at state $\mathbf{q}^{(l)}$ is given by

$$
P_{out} = \overline{p_{sr_k}} . \overline{p_{r_kd}} \tag{4.5}
$$

where

$$
\overline{p_{sr_k}} = (1 - \exp^{-\frac{2^\eta - 1}{\bar{\gamma}_{sr_k}}})^{M_{\mathbf{q}^{(l)}}^{sr_k}} \tag{4.6}
$$

$$\overline{p_{r_k d}} = (1 - \exp^{-\frac{2^{\eta}-1}{\bar{\gamma}_{r_k d}}})^{M_{\mathbf{q}^{(l)}}^{r_k d}} \tag{4.7}$$

where $\overline{p_{sr_k}}$ and $\overline{p_{r_k d}}$ are the probabilities that all available $S \to R_k$ links and $R_k \to D$ links are in outage, respectively. $M_{\mathbf{q}^{(l)}}^{sr_k}$ denotes the number of available $S \to R_k$ links at state $\mathbf{q}^{(l)}$, and $M_{\mathbf{q}^{(l)}}^{r_k d}$ is the number of available $R_k \to D$ links at state $\mathbf{q}^{(l)}$.

In -aided relays, buffer states is modeled as a discrete time Markov chain, the transition matrix of the Markov chain is denoted as **A** representing $(L+1)^K * (L+1)^K$ state transition, $\mathbf{A}_{mn}$ is the notation for the mth row and nth column entry, which expresses the transition probability to move from state $q^{(n)}$ at time t to state $q^{(m)}$ at time $t+1$:

$$\mathbf{A}_{mn} = P(X_{t+1} = q^{(m)} | X_t = q^{(n)}) \tag{4.8}$$

The described Markov chain with the transition matrix **A** has two properties: irreducible and aperiodic. The Markov chain is considered irreducible if all states are reachable by all other states in the chain, and if the probability of staying at any state higher than zero, then the Markov chain is aperiodic, see [97], [20]. As presented in Chapter 3, in irreducible and aperiodic Markov chain, the stationary state probability vector is obtained as

$$\boldsymbol{\pi} = (\mathbf{A} - \mathbf{I} + \mathbf{B})^{-1}\mathbf{b}, \tag{4.9}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_{(L+1)}]$, π_m is the probability that state is $\mathbf{q}_m$, $\mathbf{b} = [1, \cdots, 1]^T$, **I** is the notation of the identity matrix and **B** denotes an $(L+1) \times (L+1)$ matrix with all elements have the value of one.

The definition of the outage probability of the system is the probability that all relays neither transmit to the destination nor receive from the source. When this happens, the number of packets resides in s remains the same, this means the Markov chain remains in the same state, so outage probability can be obtained as follows:

$$P_{out} = \sum_{i=1}^{(L+1)^K} \pi_i \mathbf{A}_{ii} \tag{4.10}$$

where $\mathbf{A}_{ii}$ are the diagonal elements of **A**.

For example, let $L = 3$ and $K = 1$ then the Markov chain which models the system is shown in Fig. 4.2. The transition matrix of the Markov chain is

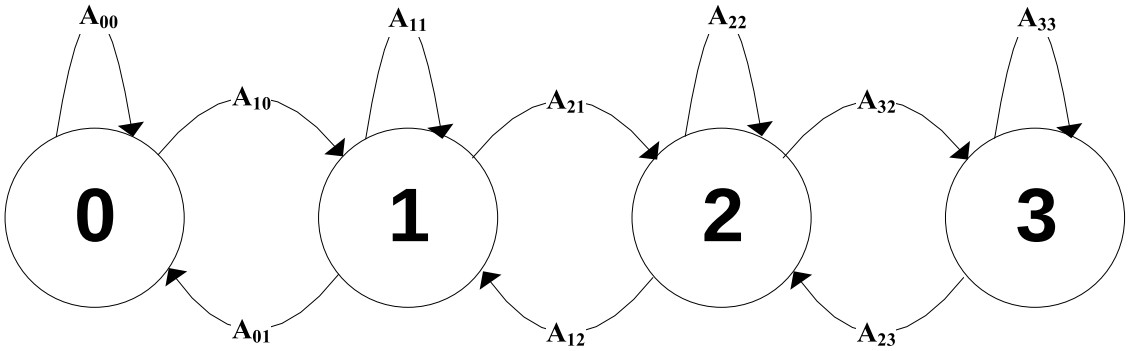

Figure 4.2 Markov chain for $L = 3$ and $K = 1$ system.

$$\mathbf{A} = \begin{pmatrix} \overline{p_{sr_1}} & p_{r_1d} & 0 & 0 \\ p_{sr_1} & \overline{p_{sr_1}}\,\overline{p_{r_1d}} & p_{r_1d} & 0 \\ 0 & \overline{p_{r_1d}} p_{sr_1} & \overline{p_{sr_1}}\,\overline{p_{r_1d}} & p_{r_1d} \\ 0 & 0 & \overline{p_{r_1d}} p_{sr_1} & \overline{p_{r_1d}} \end{pmatrix} \tag{4.11}$$

based on that, the stationary state probabilities can be calculated by (4.9), hence

$$P_{out} = \sum_{i=1}^{4} \pi_i \mathbf{A}_{ii}. \tag{4.12}$$

4.4 Average Packet Delay

The traditional definition of the delay of a packet in -aided scheme is the duration between the packet leaving the source node and arriving to the destination. However, in this chapter, we add the source delay to the definition. The source delay of a packet is the time interval between this packet arrives at the source and its departure from the source. Hence, the average packet delay is given by

$$\bar{D} = \bar{D_s} + \bar{D_{sr}} + \bar{D_r} \tag{4.13}$$

where $\bar{D_s}$ denotes the average source delay, $\bar{D_{sr}}$ is the average delay caused by transmitting the data packet from the source to a relay and $\bar{D_r}$ is the average delay at the relay node. Based on the introduced definition of the source delay:

$$\begin{aligned}
P(D_s = 0) &= (1 - \overline{p_{sr_k}})P(\mathbf{q}^{(1)}), \\
P(D_s = 1) &= (1 - \overline{p_{sr_k}})P(\mathbf{q}^{(1)})\overline{p_{sr_k}}, \\
P(D_s = 2) &= (1 - \overline{p_{sr_k}})P(\mathbf{q}^{(1)})(\overline{p_{sr_k}})^2, \\
&\vdots \\
P(D_s = n) &= (1 - \overline{p_{sr_k}})P(\mathbf{q}^{(1)})(\overline{p_{sr_k}})^n
\end{aligned} \tag{4.14}$$

from (4.14) we get the average of D_s to be:

$$\bar{D_s} = (1 - \overline{p_{sr_k}})P(\mathbf{q}^{(1)}) \sum_{i=0}^{n} i\overline{p_{sr_k}}^i \tag{4.15}$$

as $n \to \infty$:

$$\bar{D_s} = (1 - \overline{p_{sr_k}})P(\mathbf{q}^{(1)}) \left(\frac{\overline{p_{sr_k}}}{(1 - \overline{p_{sr_k}})^2} \right) = \frac{P(\mathbf{q}^{(1)})\overline{p_{sr_k}}}{(1 - \overline{p_{sr_k}})} \tag{4.16}$$

where $(1 - \overline{p_{sr_k}})$ is the probability of no outage in $S \to R_k$ links, $P(\mathbf{q}^{(1)})$ denotes the probability of all buffers are empty and n is the source delay instants. Based on (4.16), the source delay is proportional to the outage probability. Since the source delay is counted only in empty buffers, $R_k \to D$ links are not available, hence, $P_{out} = \overline{p_{sr_k}}$.

It worth mentioning that in -aided relay, the source delay is only calculated when all buffers are empty. In other words, even though some packets may encounter delay at the source (S), this delay is not considered in the source delay because this delay is already considered as $\bar{D_r}$ for other packets. So, each non-empty buffer acts as a source, which gives -aided relay an advantage over -buffer-aided relay.

Algorithm 1 summarises the procedure of calculating the source delay in -aided relay.

Algorithm 1 The proposed algorithm

input $\theta = 0$, γ_{sr_k} and $\gamma_{r_k d}$, $\quad k = 1, \cdots, K$,
find the best $R_k \rightarrow D$ available link:
if $\gamma_{r_k d} >= \eta$ **then**
    the selected link $= \arg\max_k\{\Delta_k\}$,
    a packet is transmitted to the selected relay,
else if $\gamma_{sr_k} >= \eta$ **then**
    the selected link $= \arg\min_k\{h_{sr_k}\}$,
    the selected relay receives a packet,
else
    outage occurs,
end if
if outage occurs and all buffers are empty **then**
    start counting the source delay
end if

After calculating the source delay ($\bar{D_s}$), we move to the second part ($\bar{D_{sr}}$). It takes one time-slot to send a packet from the source to any relay node, so $\bar{D_{sr}}$ is one. Finally, because $\bar{D_r}$ in every relay is the same (identical channels), $\bar{D_r}$ in one relay R_k is analyzed. Based on Little's law [72], the average packet delay at the relay R_k is given by

$$\bar{D_{r_k}} = \frac{\bar{L_{r_k}}}{\bar{\xi_{r_k}}} \tag{4.17}$$

where $\bar{L_{r_k}}$ and $\bar{\xi_{r_k}}$ are the notations of average queue length and average throughput at the relay R_k respectively. The average queue length at R_k is calculated by averaging the queue lengths over all buffer states

$$\bar{L_{r_k}} = \sum_{i=1}^{(L+1)} \pi_i q_k^{(i)} \tag{4.18}$$

Because selecting any of the relays has the same probability, the average throughput at the relay R_k is given by

$$\bar{\xi_{r_k}} = \frac{\bar{\xi}}{K} \tag{4.19}$$

where $\bar{\xi}$ is the average throughput of the overall network. For delay-limited transmission as in [3] and [93], the average throughput $\bar{\xi}$ is obtained as

$$\bar{\xi} = \eta(1 - P_{out}) \tag{4.20}$$

In the selected scheme, every packet needs two time-slots (do not have to be consecutive) to reach the destination, so we have $\eta = 1/2$ packet per time-slot, and thus

$$\bar{\xi_{r_k}} = \frac{(1 - P_{out})}{2K} \tag{4.21}$$

substituting (4.18) and (4.21) into (4.17) gives

$$\bar{D_r} = \frac{2K \sum_{i=1}^{(L+1)} \pi_i q_k^{(i)}}{(1 - P_{out})}. \tag{4.22}$$

4.5 Asymptotic Performance

To gain a better understanding, this section studies the delay performance of the delay-reduced with considering the source delay when the average channels SNR for both of $S \to R_k$ and $R_k \to D$ links goes to infinity. Regarding $\bar{D_s}$

$$\lim_{(\bar{\gamma}_{sr_k}, \bar{\gamma}_{r_k d}) \to \infty} \overline{p_{sr_k}} = 0 \tag{4.23}$$

so outage is impossible to occur when SNR is high enough. Since source delay only occurs when $S \to R_k$ links are in outage, as in (4.16), the source delay approaches zero as SNR goes to infinity:

$$\lim_{(\bar{\gamma}_{sr_k}, \bar{\gamma}_{r_k d}) \to \infty} \bar{D_s} = 0. \tag{4.24}$$

Now we calculate $\bar{D_r}$ under the no outage assumption. Suppose that all s are empty at time t, so that relay R_k is in the state $q_k^{(0)}$. At this point, a packet is assumed to be received by the relay R_k at time $(t + 1)$, and R_k moves to state $q_k^{(1)}$. After that,

the packet in needs to be transmitted to the destination at $(t+2)$ and R_k returns to state $q_k^{(0)}$. And this process continues, thus

$$P(q_k^{(0)}) = P(q_k^{(1)}) = \frac{1}{2} \tag{4.25}$$

if s are empty, the probability that R_k receives a packet is $1/K$, so $P(q_k^{(1)}) = \frac{1}{2K}$ (without the assumption R_k has received the packet)

$$\lim_{(\bar{\gamma}_{sr_k}, \bar{\gamma}_{r_k d}) \to \infty} \bar{L_{r_k}} = 0.P(q_k^{(0)}) + 1.P(q_k^{(1)}) = \frac{1}{2K} \tag{4.26}$$

$$\lim_{(\bar{\gamma}_{sr_k}, \bar{\gamma}_{r_k d}) \to \infty} \bar{\xi_{r_k}} = \frac{\lim_{\gamma_{sr_k}(t) \to \infty}(1 - P_{out})}{2K} = \frac{1}{2K} \tag{4.27}$$

From (4.17), we find $\lim_{\gamma_{sr_k}(t) \to \infty} \bar{D_r} = 1$, and the average delay is

$$\lim_{(\bar{\gamma}_{sr_k}, \bar{\gamma}_{r_k d}) \to \infty} \bar{D} = 0 + 1 + 1 = 2 \tag{4.28}$$

which means in -aided and -buffer-aided relays, the delay has the same value 2 at high values of SNR. This will be verified in Section 4.7.

4.6 Proposed Selection Rule

Before ending this chapter, we are suggesting a novel selection rule. The idea of the adaptive target length θ_k based on the state of the relay channels is introduced in this rule. We kept the idea that the $R_k \to D$ links have higher priority as in the delay-reduced. After that, the relay which has $\Omega_{r_k d}$ above a certain threshold (denoted as Ω_t) is termed as the fast relay (FR), which could be more than one relay. Otherwise, relays are termed as the slow relays (SR)

Since the FR is expected to handle longer queues better than the SR, a longer target length is assigned to the FR. This is expected to reduce the delay compared to give all the relays the same target length. On the other hand, shorter target lengths are assigned to SR's. Then we calculate $\Delta_k = \theta_k - q(k)$ for all links. As in the PBRS, in Chapter 2, higher priorities are given to larger $|\Delta_k|$'s. Among the

same category the FR or the SR, in the case of equal $|\Delta_k|$, positive Δ_k's have higher priority than negative ones (transmission has higher priority).

In case of different category, to avoid long delays caused by the SR, the SR is given a higher priority in the case of equal positive Δ_k because packets are more likely to stuck in its buffer. And the opposite is done in the case of equal negative Δ_k. Otherwise, if more than one link has the same priority, the one with the highest SNR is selected.

For a better exposure, Fig. 4.3 shows how the proposed scheme is applied on a four relay network. In Fig. 4.3, notice the following:

- The $\Omega_{r_k d}$ values for the relays R_1, R_2 and R_4 are larger or equal to Ω_t, so they are the FR's and R_3 is the SR.
- For the FR's $\theta = 2$, and $\theta = 0$ for the SR.
- The priorities for all the available links are shown in the highlighted squares.
- Although R_3 and R_4 have the same $\Delta_k = +2$, the $R_3 \to D$ link is prioritized to reduce the delay.
- $\Delta_1 = +1$ and $\Delta_2 = -1$, so the $R_1 \to D$ link has higher priority than the $S \to R_2$ link.
- In the situation of worsening Δ_k (all the above links are in outage), we prioritize links with smaller $|\Delta_k|$. This is why $R_2 \to D$ and $S \to R_1$ have higher priority than $S \to R_3$.
- Finally, this example can be generalized for more relays with keep increasing the target length as relays get faster (higher $\Omega_{r_k d}$).

4.7 Numerical Simulations

In all the simulations below, the target transmission rate is $\eta = 2$ bps, the bandwidth is 1Hz, size is $L = 5$ for all buffers and all noise powers σ^2 are normalized to unity. The average channel gains are set to $\Omega_{sr_k} = \Omega_{r_k d} = 1$.

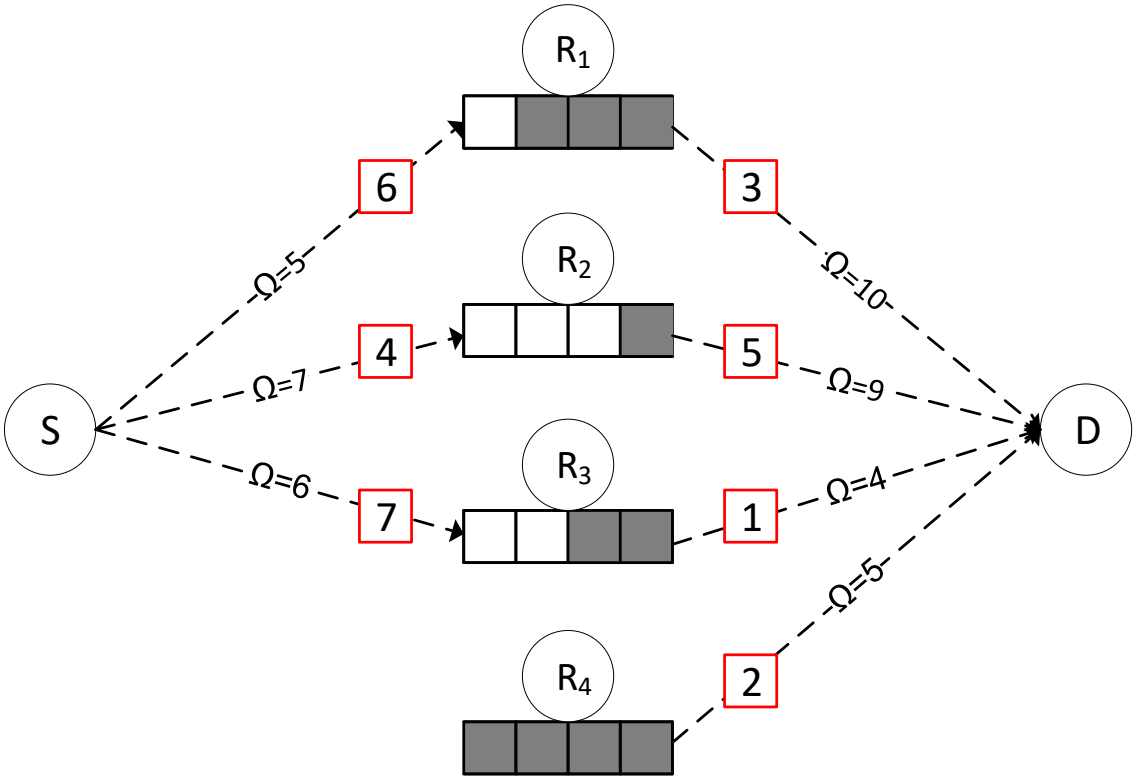

Figure 4.3 Four relay network example, where $\Omega_t = 5$, $L = 4$ and $K = 4$.

In Fig. 4.4, we show the source delay $\bar{D}_s$ *vs* the transmit SNR (in dB) in a single relay system. The selection scheme is the delay-reduced. The simulation results clearly verifies the results derived in Section 4.4, 4.5. The $\bar{D}_s$ starts high and then goes to zero as the SNR reaches high values.

For a better insight, we start with showing how different schemes perform in terms of the outage probability and the throughput. In Fig. 4.5 and Fig. 4.6, buffering capability has improved both the outage probability and the throughput. In particular, buffer-aided schemes the delay-reduced and the state-based outperform -buffer-aided max-min scheme in both of outage probability and throughput. Both figures show that larger target length in the state-based (2) has better outage probability and throughput compared to the delay-reduced.

Fig. 4.7 and Fig. 4.8 show the impact of considering the source delay ($\bar{D}_s$) on the average packet delay. The following can be observed:

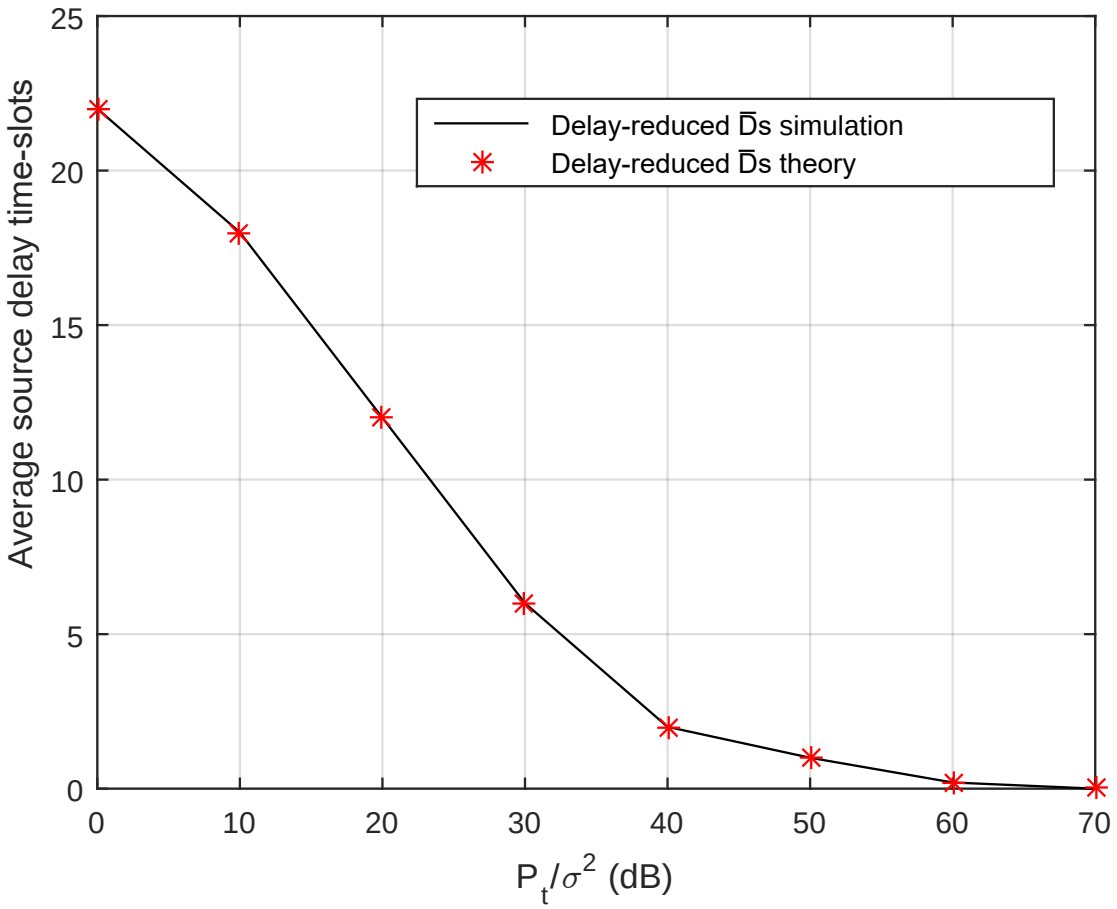

Figure 4.4 The source delay $\bar{D}_s$ with $K = 1$ theoretical *vs* simulation.

- Prior to considering the $\bar{D}_s$, average packet delay in the max-min is always equal to 2 time-slots even at very low SNR which is unbeatable by -aided relay.
- After considering the $\bar{D}_s$, the delay performance in the three schemes has changed. Fig. 4.8 shows that the delay-reduced outperforms the max-min.
- The state-based has longer delay than the delay-reduced because with larger target length, packets tend to stay longer in s.

As stated in Chapter 2, adding more relays boosts the system performance in some performance metrics. However, it can lengthen the $\bar{D}_r$. As the number of relays increases, the outage probability is decreased, and the $\bar{D}_s$ is decreased as well. Fig. 4.9 shows that the delay-reduced and the state-based have longer delay than the max-min.

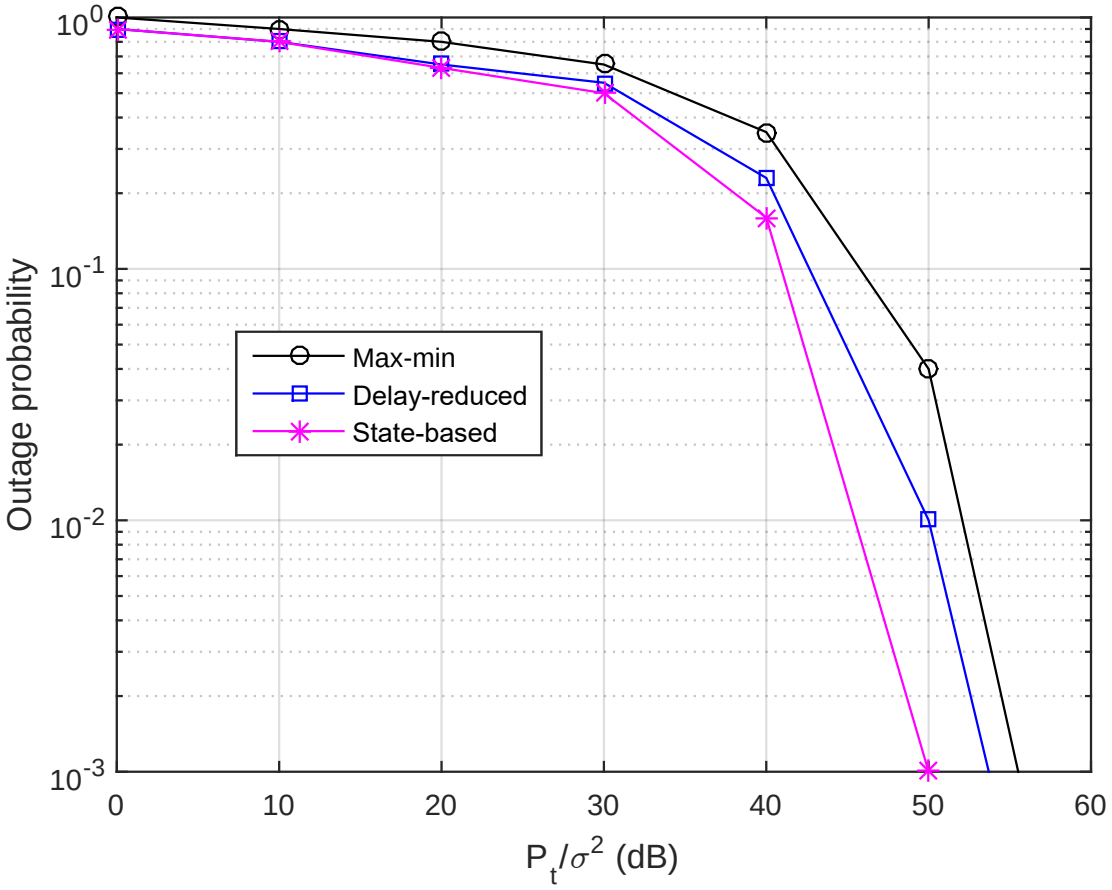

Figure 4.5 System outage probability with $K = 1$.

There are some techniques were suggested to improve the system delay, however, the impact of these techniques were not studied while considering the source delay.

- In [99], the broadcast technique has improved the system delay. Interestingly, when we apply the broadcast technique, which means the source is transmitting each packet to all relays all at once simultaneously, the delay in both cases: the max-min and the delay-reduced get closer as shown in Fig. 4.10.

- The broadcasting enhancement motivates us to look for more enhancement techniques even with more relays. Therefore, using small buffers (e.g. $L = 1$)is another technique which can be combined with the broadcasting. Small buffers reduces the delay at the relay $\bar{D}_r$ because long queues are less likely. So we set sizes to $L = 1$ while applying broadcasting, Fig. 4.11 shows that the delay-reduced again outperforms the max-min especially in low SNR range.

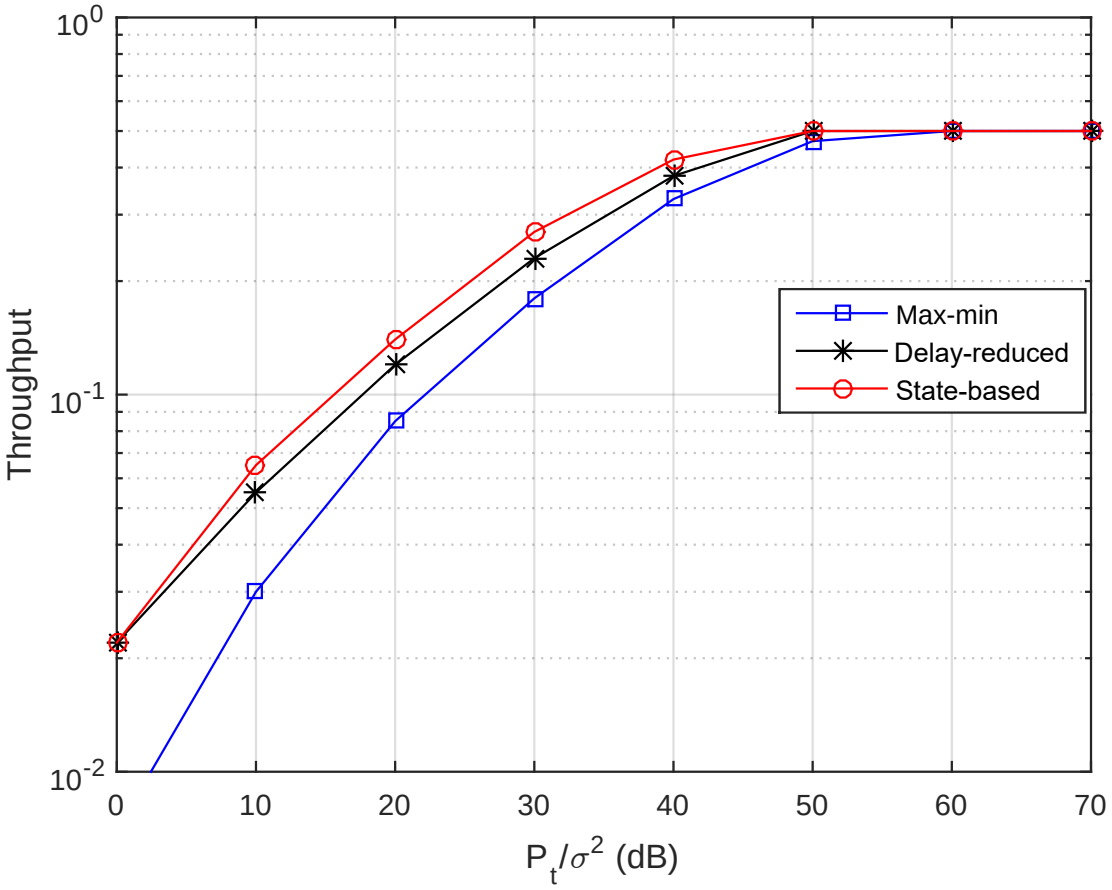

Figure 4.6 System throughput with $K = 1$.

Although using buffers with $L \leq 2$ is harmful for the diversity order, it is beneficial in reducing the delay.

Finally, in Fig. 4.12, we study the proposed selection rule as follows

- The target length $\theta = 2$ for the FR and for the SR $\theta = 0$.
- We set the following: $\Omega_{r_1 d} = 3$, $\Omega_{r_2 d} = 2$, $\Omega_{r_3 d} = 1$, $\Omega_{sr_k} = 1$ and $\Omega_t = 2$.
- The best results for the delay-reduced is with broadcasting and $L = 1$, so we used this results for comparison with matching the $\Omega r_k d$ in both schemes.
- The proposed scheme requires larger buffer size, so we used $L = 3$ for the proposed scheme while keeping $L = 3$ in the delay-reduced. It is noticeable that the new rule has never been worse than the delay-reduced. In particular, the new rule outperforms the delay-reduced in low SNR range, and has similar performance otherwise.

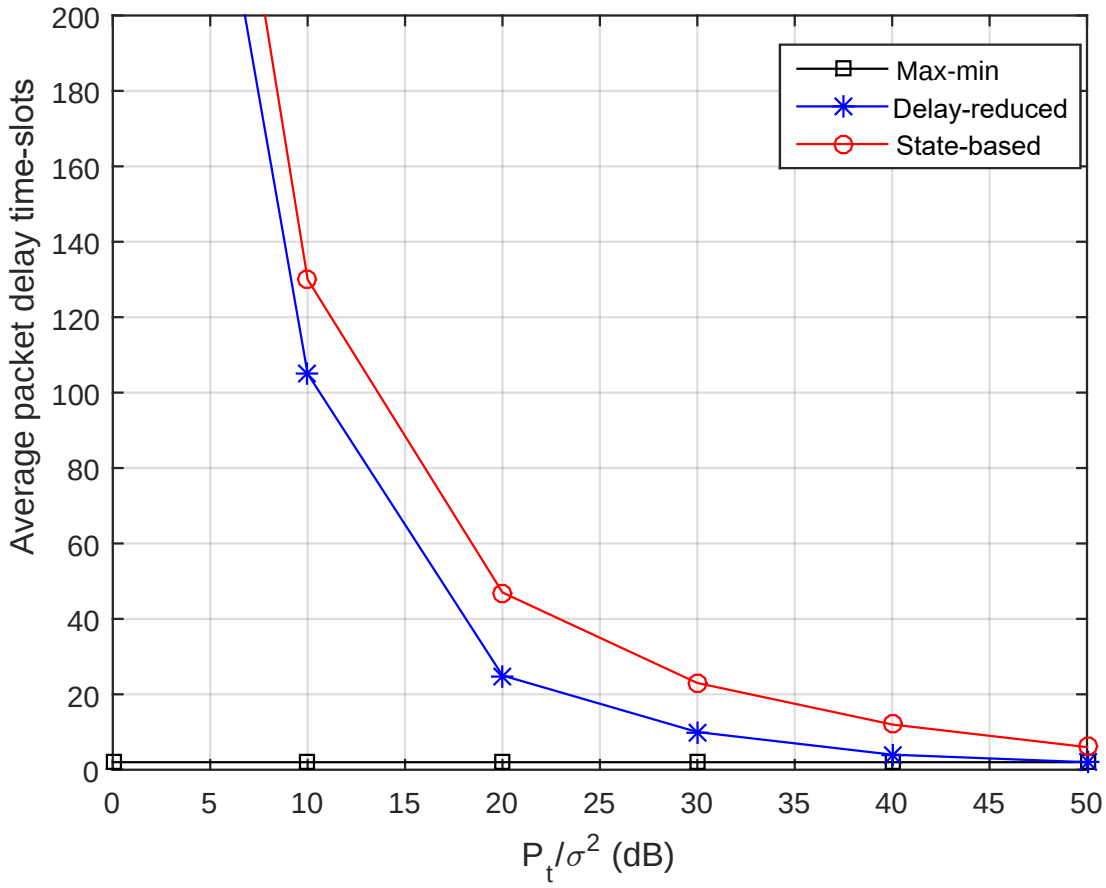

Figure 4.7 Average packet delay with $K = 1$ without considering ($\bar{D}_s$).

4.8 Summury

Longer delay is the main difficultly in applying relays especially in 5G which requires very low values of end-to-end delay (1ms). Buffer-aided relays can have negative impact on system delay. However, buffer-aided relays have shorter source delay. The impact of considering the source delay that each packet encounters before being transmitted is studied in this chapter. This makes delay calculation more accurate. The average packet delay is analyzed asymptotically. Some techniques which may enhance the system performance were discussed. For instance, broadcasting and smaller buffer size have shown positive impact on -aided schemes delay performance. The presented results show that the delay of -aided relay can be shorter than the delay of -buffer-aided relay, this is true in a single and a multiple relay network. Finally, a new selection rule is proposed. The result shows further

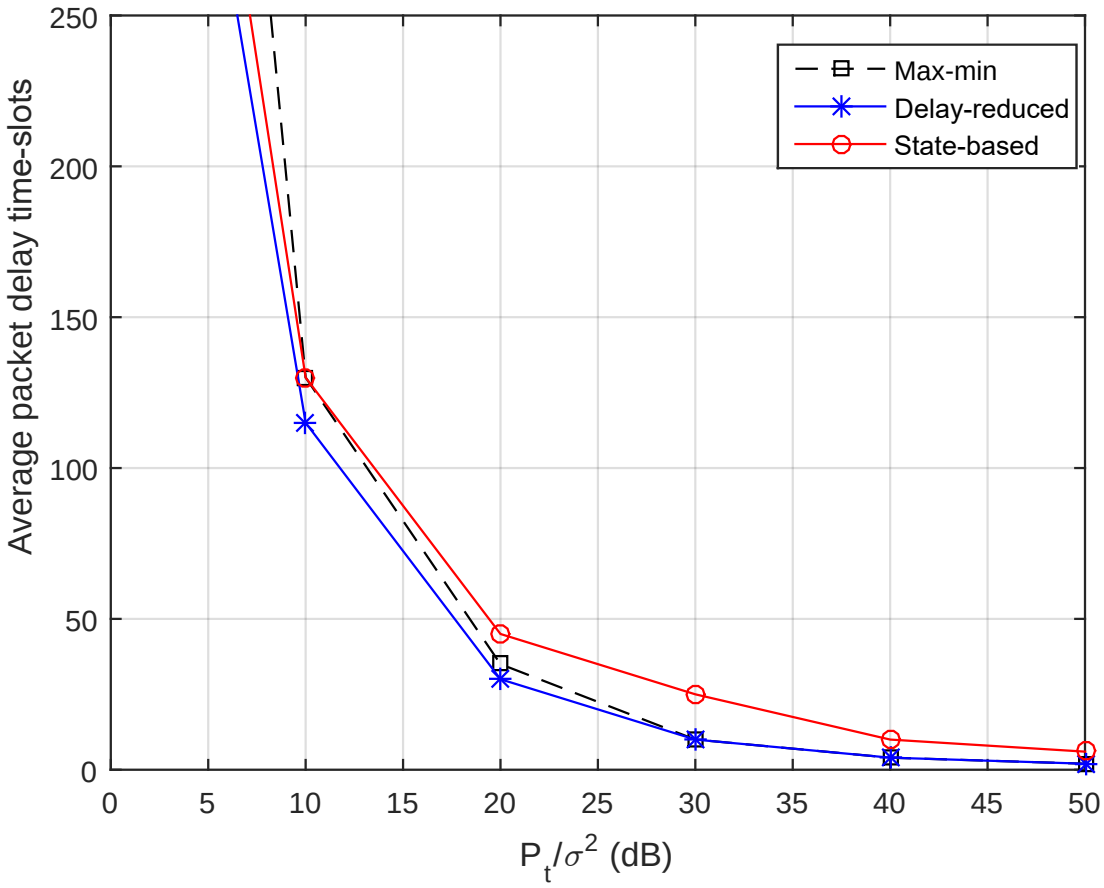

Figure 4.8 Average packet delay at $K = 1$ with $\bar{D}_s$.

enhancement in the delay performance of aided multiple relay network. The next chapter studies the impact of delay constraints on -aided relay performance.

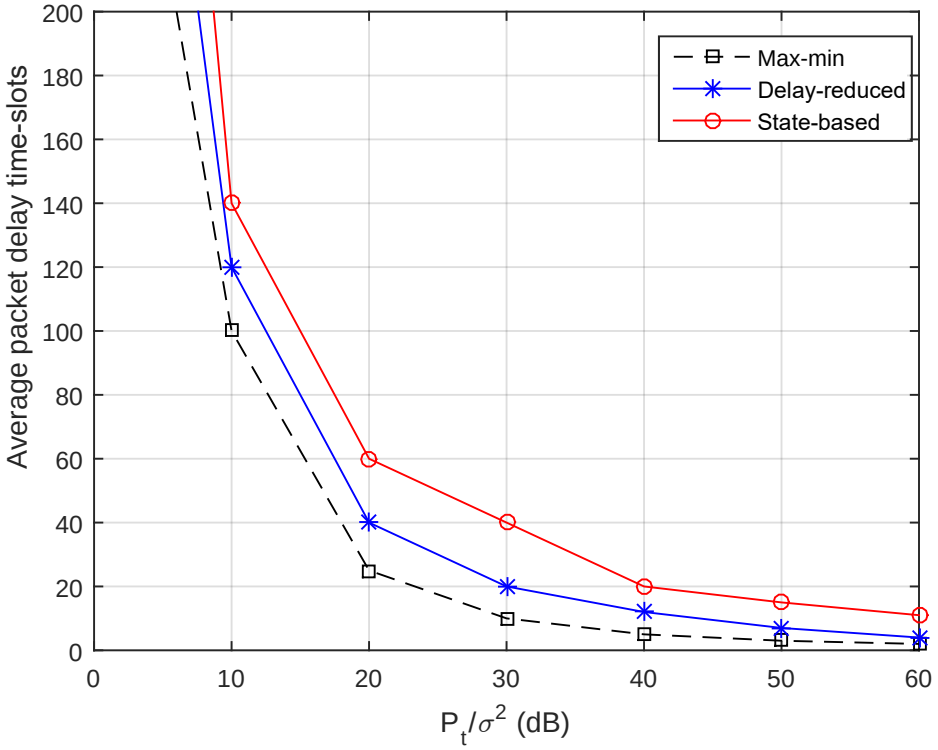

Figure 4.9 Average packet delay in $K = 3$ network with the $\bar{D}_s$.

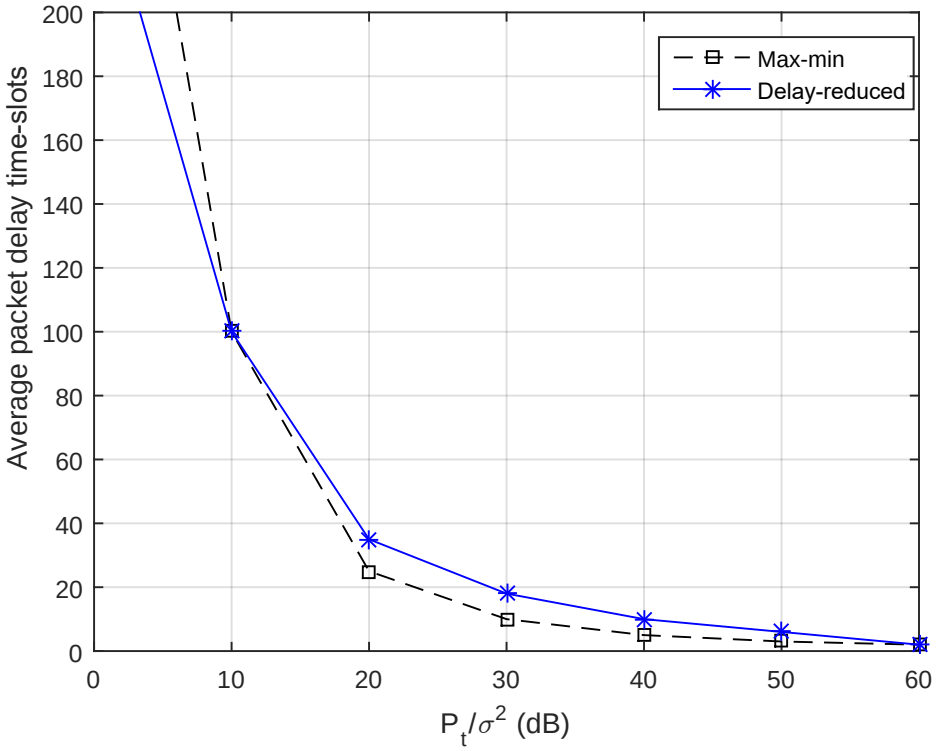

Figure 4.10 The impact of the broadcast technique on average packet delay in $K = 3$ network with $\bar{D}_s$.

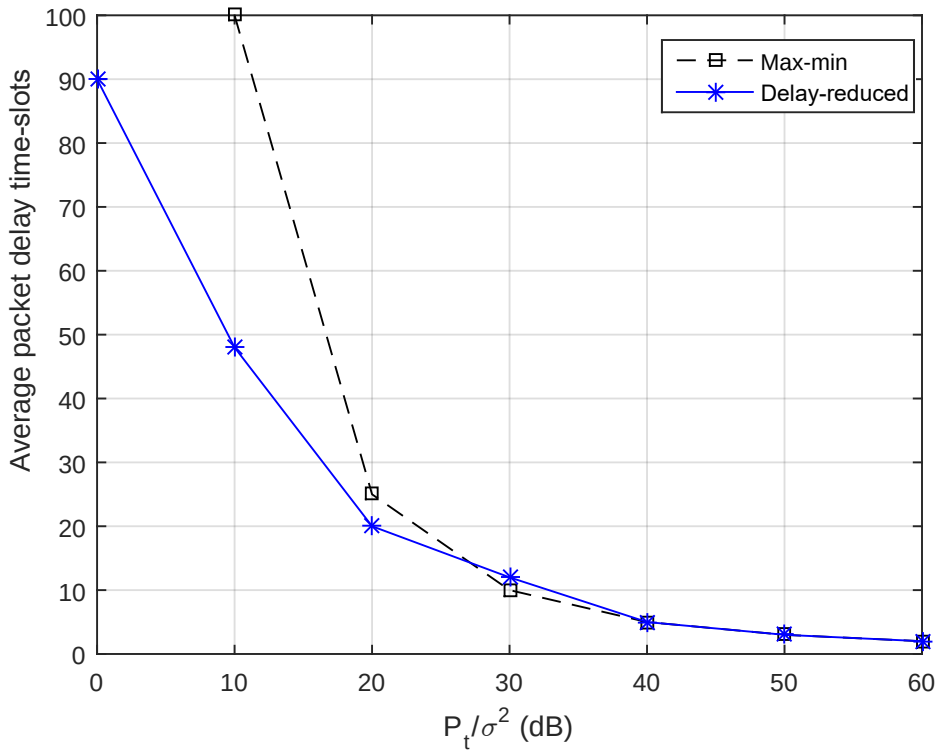

Figure 4.11 The impact of the broadcast technique on average packet delay in $K = 3$ network with $\bar{D}_s$ and $L = 1$.

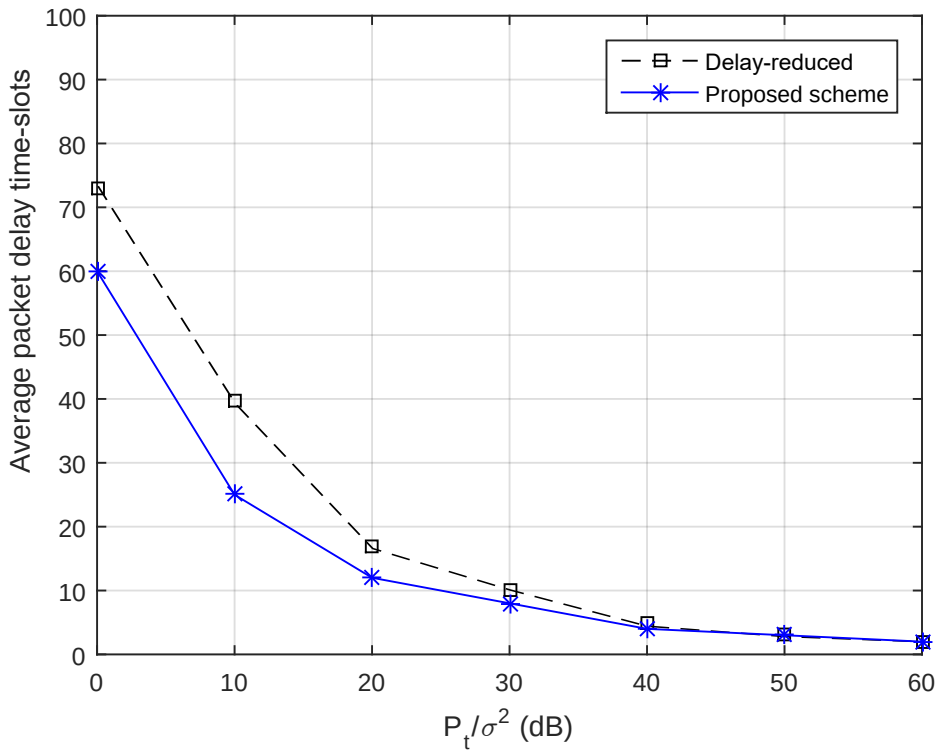

Figure 4.12 The delay including $\bar{D}_s$ in the delay-reduced vs the proposed scheme with $K = 3$.

Chapter 5

Delay-Constrained Adaptive Link Selection in Buffer-Aided Relay Networks

Several 5G applications have strict delay constraints. For example, tactile internet enables real-time interactive, which constraining the delay to ultra low values. This gives rise to the necessity to analyse the impact of constraining the delay to a certain target delay on the buffer-aided cooperative relay network performance. In this chapter, Trellis state diagram and Markov chain are used to analyse the delay-constrained outage probability which caused by both the channel outage and the delay exceeds the target delay.

5.1 Introduction

The cooperative network has been regarded as an important concept in the 5G and beyond mobile networks to improve the wireless connectivity, which is crucial in applications including the internet of things IoT, machine to machine (M2M) and device to device (D2D) communications. A typical 3-node relay network is shown in Fig. 5.1 which consists of a source node S, a relay node R and a destination node D. In the traditional relay scheme, the transmission follows the $S \to R$ and $R \to D$ order consecutively. Therefore, when one of the $S \to R$ and $R \to D$ channels

suffers from deep fading, even if the other channel is strong, the data throughput is severely deteriorated. The buffer-aided cooperative relay is applied to address this issue.

In the buffer-aided relay network, the buffer is applied at the relay node to store the data packets. As a result, the transmission links can be adaptively selected without complying with the order as that in the traditional relay scheme. This brings larger flexibility for the transmission to avoid the deep fading, leading to higher data throughput. Various adaptive link selection has been proposed for the buffer-aided 3-node relay network. Typical examples include the variable rate adaptive link selection in [149], the fixed and mixed rate link selection in [150], and the discrete transmission rate link selection in [128]. While these adaptive link selection can achieve significantly higher throughput than the non-buffer-aided relay schemes, they all assume the relay buffer size is infinite large which is often impractical. The link selection with finite buffer size based on stochastic throughput optimization for two-hop relay is investigated in [145].

On the other hand, a number of buffer-aided relay selection in multiple relay networks have been proposed, which can be directly used in the adaptive link in the 3-node network. In [68], the max-link scheme is proposed to select the transmission link with the highest SNR among all $S \rightarrow R$ and $R \rightarrow D$ links. The max-link can achieve full diversity for the independent-and-identical-distribution i.i.d channels when the buffer-size is large enough. However, the performance of the max-link may deteriorate quickly when the channels become independent-and-non-identical-distribution (i.n.i.d.). This is because the i.n.i.d. channel fading makes buffer be more likely to be empty or full if the max-line selection is applied, leading to lower diversity gain. This issue is addressed in the state-based selection scheme in [79], where the link selection is based on not only the channel status but also the buffer states. The state-based shows better performance both in throughput and delay than the max-link scheme. In [121], the delay-reduced is proposed to reduce the delay by giving higher priority of selection to $R \rightarrow D$ links. The delay-reduced has the minimum delay among all buffer-aided scheme, but may not have throughput as high as that in the state-based. In [49] and

[8], the prioritization-based relay selection is proposed for the social-aware and NOMA networks respectively, in which the target buffer length is introduced for the trade-off between the throughput and delay. Particularly, when the target buffer length is set as zero or two, the prioritization-based scheme is equivalent to the delay-reduced and the state-based selection schemes respectively.

While applying buffers at the relays significantly increases the data throughput, it may lead to higher transmission delay because of the data waiting at the buffers for transmission. Therefore, for communications with delay constraints, many buffer-aided schemes such as those in [149, 79, 68, 61] cannot be used. This can be seen in the benchmark the max-link scheme [68], where the average packet delay increases linearly with the relay number K and buffer size L. For example, when $K = 3$ and $L = 20$, the average packet delay is > 60 time-slots/packet, which will not even satisfy the moderate delay constraints. On the other hand, while some buffer-aided schemes can maximize the throughput with constrained average delays (e.g. [149, 79]), the distribution of the delay is not considered, which is however of particular interest in some applications. For example, in the traditional TCP communications, because every packet with delay higher than the target delay will be re-transmitted or discarded, it is important to obtain the distribution of the packet delay.

In this chapter, we investigate adaptive link selection in the delay-constrained relay network. In [68, 79, 121, 48] and [8], the corresponding outage probability and average packet delay have been derived by using the Markov chain and little law. However the analysis does not apply to the delay-constrained link selection. Moreover, the existing link selection schemes will see performance degradation with delay constraints. New selection rules are necessary. These issues are addressed in this chapter which are summarized as following:

- Analyse the delay-constrained outage probability for both channel outage and delay overtime, which reveals the effective throughput within the delay constraints. Particularly, the Trellis state diagram is introduced to obtain the closed-form expression.

- Obtain the closed-forms of the delay-constrained outage probabilities for the benchmark selection schemes including the max-link, the state-based and the delay-reduced, for the 3-node relay network.
- Propose an adaptive virtue buffer-size relay selection scheme which achieves significant better delay-constrained outage probability than existing schemes.

The rest of this chapter is organised as following. Section 5.2 describes the system model; Section 5.3 analyse the delay-constrained outage probability; Section 5.4 investigate the delay-constrained outage probability for benchmark schemes including the max-link, the state-based and the delay-reduced; Section 5.5 proposes the adaptive virtue buffer-size relay selection scheme; Section 5.6 verify the analysis and the proposed scheme with simulation results; Section 5.7 concludes this chapter.

5.2 System Model

Fig. 5.1 shows the system model for a 3-node buffer-aided relay network, where S is the source node, R is the relay node which works in the half-duplex HD decode-and-forward DF mode and is equipped with a data buffer of size L, and D is the destination node. We assume no direct link between S and D. For simple expression, we use index '1' and '2' to indicate the $S \to R$ and $R \to D$ channels respectively. Both channels are flat Rayleigh fading that the channel coefficients $h_1(t)$ and $h_2(t)$ remain constant within a time slot and change independently from one time slot to another. Without losing generality, we assume that the transmission powers at both transmitting nodes (S and R) are P_t, and the noise variances at both receiving nodes (R and D) are σ^2.

The channel capacity for the i-th channel is given by

$$C_i(t) = \log_2(1 + \gamma_i(t)), \qquad i = 1, 2 \tag{5.1}$$

where $\gamma_i(t) = (P_t/\sigma^2)|h_i(t)|^2$ which is the instantaneous channel SNR at time slot t. For Rayleigh fading channels, the channel gain $|h_i(t)|^2$ is exponentially

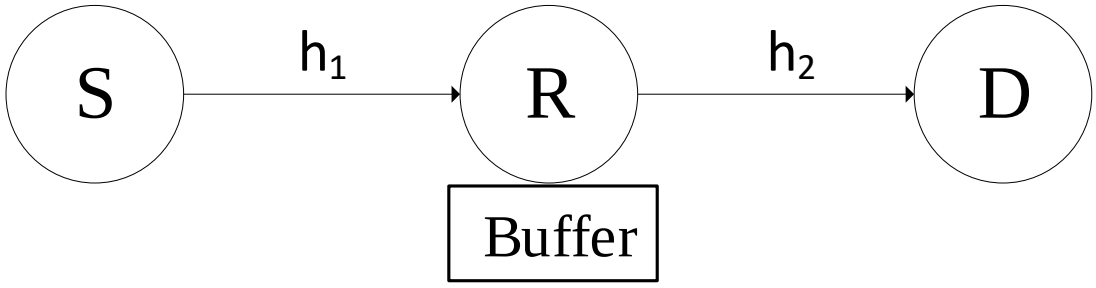

Figure 5.1 System model for the buffer-aided 3-node relay network.

distributed with the averages of $\Omega_i = \mathrm{E}[|h_i(t)|^2]$. The average channel SNR is given by $\bar{\gamma_i} = (P_t/\sigma^2)\Omega_i$. If $C_i(t) < \eta$, the i-th channel is in outage that it cannot support the target data rate η.

Due to the buffer equipped at the relay, either the $S \to R$ or $R \to D$ link may be selected at any time-slot. When the buffer is empty or full, however, the $R \to D$ and $S \to R$ links are not 'available' for selection, respectively. At one time-slot, if no available channel can support target rate transmission, the outage (due to channels) occurs.

On the other hand, when a data packet arrives at the relay node at time slot t, there are three possibilities at the next time slot: 1) the $R \to D$ link is selected so that the packet is forwarded to the destination; 2) the $S \to R$ link is selected for a new packet to arrive at R; 3) the channels are in outage. In both case 2) and 3), the packet arrived at time t will queue at the relay buffer, causing the packet delay. If the packet delay d is larger than the target delay D_0, the outage (due to delays) occurs.

The overall outage probability considering both channel and delay outages can be obtained as

$$P_{out} = P(out_c \text{ or } (d > D_0)), \tag{5.2}$$

where out_c is the event that channels are in outage. If there is no constraint on the packet delay (i.e. $D_0 \to \infty$), we have $P_{out} = P(out_c)$.

5.3 Outage Probability Analysis

From (5.2), the overall outage probability can be expressed as

$$\begin{aligned} P_{out} &= 1 - P(\overline{out}_c, d \leq D_0) \\ &= 1 - P(d \leq D_0 | \overline{out}_c) \cdot P(\overline{out}_c) \\ &= 1 - P(d \leq D_0 | \overline{out}_c) \cdot (1 - P(out_c)) \end{aligned} \tag{5.3}$$

where out_c and $\overline{out}_c$ are the events that channels are in and not in outage, respectively.

As it is shown in (5.3), the overall outage probability depends on the channel outage probability $P(out_c)$ as well as $P(d \leq D_0 | \overline{out}_c)$ which is the delay outage probability conditioned on no channel outage. Therefore, the channel and delay outages are successfully 'detached' in (5.3), making it convenient to obtain the overall outage probability. Below we derive $P(out_c)$ and $P(d \leq D_0 | \overline{out}_c)$ respectively.

5.3.1 $P(out_c)$

The number of data packets in the relay buffer forms a"state". If the buffer size is L, there are $(L+1)$ states in total, denoted as $q_0, \cdots, q_L$ respectively. The state transition matrix $\mathbf{A}$ has dimension $(L+1) \times (L+1)$, where the ij-th entry $a_{ij} = P(q_i | q_j)$ which is the transition probability from state q_j to q_i, $i, j = 0, \cdots, L$. For example, when the buffer size $L = 3$, we have

$$\mathbf{A} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} \\ a_{10} & a_{11} & a_{12} & a_{13} \\ a_{20} & a_{21} & a_{22} & a_{23} \\ a_{30} & a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{00} & a_{01} & 0 & 0 \\ a_{10} & a_{11} & a_{12} & 0 \\ 0 & a_{21} & a_{22} & a_{23} \\ 0 & 0 & a_{32} & a_{33} \end{bmatrix} \tag{5.4}$$

Suppose the state is at q_j at time slot t. If the relay receives or transmit out a data packet at time $(t+1)$, the number of packets in the buffer increases or

decrease by one respectively, and the state transits to another state q_i. The state transition probability depends on both the channels and the link selection rules.

On the other hand, when channels are in outage, the state keeps unchanged. Thus the diagonal element a_{ii} is the channel outage probability at state q_i, $i = 0, \cdots, L$. Particularly, when the state is at q_0 such that the relay buffer is empty, the only available link for selection is $S \to R$. Thus we have

$$a_{00} = P(C_1 < \eta) = 1 - e^{-\frac{2^{\eta}-1}{\bar{\gamma}_1}}. \tag{5.5}$$

When the state is at q_L that the relay buffer is full, the channel outage only depends on the $R \to D$ link so that

$$a_{LL} = P(C_2 < \eta) = 1 - e^{-\frac{2^{\eta}-1}{\bar{\gamma}_2}}. \tag{5.6}$$

When the buffer is neither empty nor full, the channel outage occurs only when both $S \to R$ and $R \to D$ are in outage. Thus we have

$$\begin{aligned} A_{ii} &= P(C_1 < \eta)P(C_s < \eta) \\ &= \left(1 - e^{-\frac{2^{\eta}-1}{\bar{\gamma}_1}}\right)\left(1 - e^{-\frac{2^{\eta}-1}{\bar{\gamma}_2}}\right), \qquad i \neq 0, L \end{aligned} \tag{5.7}$$

The outage probability of the overall system is given by

$$P(out_c) = \sum_{i=0}^{L} \pi_i \cdot a_{ii}, \tag{5.8}$$

where π_i is the stationary probability for state q_i. For column stochastic, irreducible and aperiodic transition matrix, the stationary state probability vector is obtained as

$$\boldsymbol{\pi} = (\mathbf{A} - \mathbf{I} + \mathbf{B})^{-1}\mathbf{b} \tag{5.9}$$

where $\boldsymbol{\pi} = [\pi_0, \cdots, \pi_L]^{\mathrm{T}}$, $\mathbf{b} = (1, 1, ..., 1)^T$, $\mathbf{I}$ and $\mathbf{B}$ are the identity and all one matrices with appropriate dimensions respectively.

5.3.2 $P(d \leq D_0|\overline{out}_c)$

The delay of a packet is the number of time slots for the packet to reach the destination after it leaves the source. Because it always takes one time slot for the $S \to R$ transmission, the packet delay is given by

$$d = 1 + d_r, \tag{5.10}$$

where d_r is the delay for $R \to D$ transmission including the queuing at the relay buffer. Thus we have

$$\begin{aligned} P(d \leq D_0|\overline{out}_c) &= P(d_r \leq D_0 - 1|\overline{out}_c) \\ &= \sum_{i=1}^{D_0-1} P(d_r = i|\overline{out}_c) \end{aligned} \tag{5.11}$$

Because $P(d \leq D_0|\overline{out}_c)$ is conditioned on no channel outage, at any time, the relay always receives or transmits a data packet so that the state transition always occurs. We denote the transition matrix without the channel outage as $\hat{\mathbf{A}}$. The ji-th entry of $\hat{\mathbf{A}}$ is given by

$$\hat{a}_{ji} = \begin{cases} 0, & i = j \\ \frac{a_{ji}}{(\sum_{k=0}^{L} a_{ki}) - a_{ii}}, & i \neq j \end{cases} \tag{5.12}$$

where a_{ji} is the ji-th entry of the original transition matrix $\mathbf{A}$. The stationary state probability without channel outage, $\hat{\pi}$, is similarly obtained as in (5.9) by replacing $\mathbf{A}$ with $\hat{\mathbf{A}}$.

Only when the buffer is not full, it can receive a packet. Therefore, when a packet leaves the source node S, the buffer state is $q_j, j = 0, \cdots, L-1$. After the packet arrives at R and is stored in the buffer, the buffer state transmits from q_j to q_{j+1}. Considering all possible buffer states when a packet transmits from S to R, the probability that it takes i time slot for this packet to go through the buffer (i.e.

$d_r = i$) is given by

$$P(d_r = i|\overline{out}_c) = \sum_{j=0}^{L-1} P(d_r = i|q_{j+1}, \overline{out}_c) \cdot \tilde{\pi}_j, \tag{5.13}$$

where $(d_r = i|q_{j+1}, \overline{out}_c)$ is the probability that it takes i time slots for the last packet in the buffer to be transmitted out when the buffer state is q_{j+1} and no channel outage occurs, and $\tilde{\pi}_j$ is the normalized stationary state probability which is given by

$$\tilde{\pi}_j = \frac{\hat{\pi}_j}{\sum_{l=0}^{L-1} \hat{\pi}_l}, \qquad j = 0, \cdots, L-1 \tag{5.14}$$

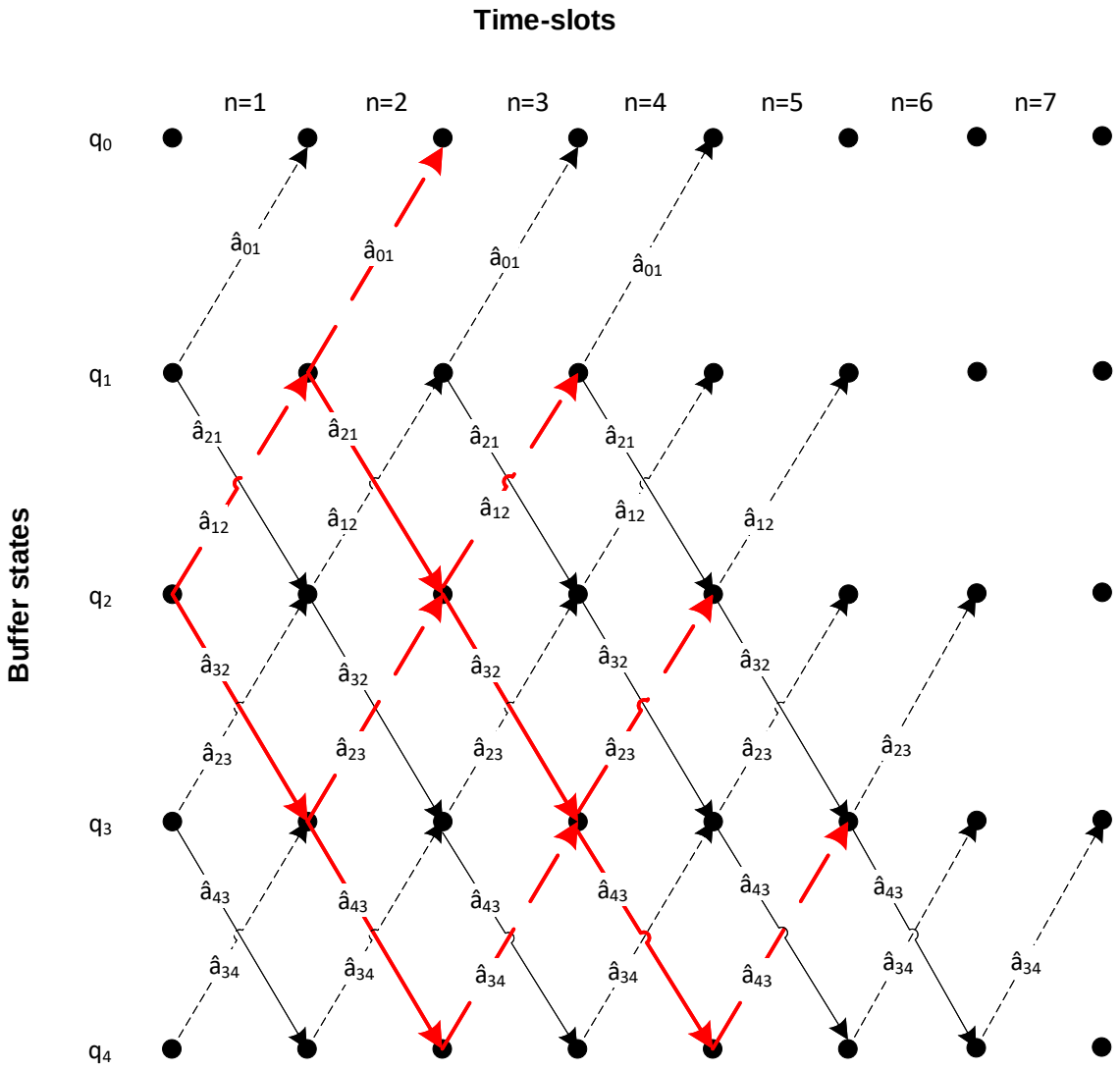

Figure 5.2 State transition Trellis diagram for buffer size $L = 4$.

Below we use the Trellis diagram to calculate $P(d_r = i|q_l, \overline{out}_c)$, $l = 1, \cdots, L$. For better exposition, we consider the buffer size $L = 4$ first. The state transition Trellis diagram is illustrated in Fig. 5.2, where the solid and dash lines correspond to the $S \to R$ and $R \to D$ links selection respectively, and the values for the lines are the corresponding state transition probabilities (which are obtained in (5.12)). From (5.4) and (5.12), the state transition matrix without outages for $L = 3$ is given by

$$\hat{\mathbf{A}} = \begin{bmatrix} 0 & \hat{a}_{01} & 0 & 0 \\ 1 & 0 & \hat{a}_{12} & 0 \\ 0 & \hat{a}_{21} & 0 & 1 \\ 0 & 0 & \hat{a}_{32} & 0 \end{bmatrix}$$

When a new packet leaves S, the buffer state must be either q_0, q_1 or q_2. After the packet arrives at R, the buffer state transits to q_1, q_2 or q_3 accordingly.

When the buffer state is q_l, $R \to D$ link has to be selected l times for the last packet to be transmitted out. For example, we assume the buffer state is q_1 after the new packet arrives R. In order for this packet to be transmitted out, the $R \to D$ needs only to be selected once which can be one of the following cases:

- At first time-slot $n = 1$, $R \to D$ link is selected. This corresponds to path $q_1 \to q_0$, and we have
$$P(d_r = 1|q_1, \overline{out}_c) = \hat{a}_{01} \tag{5.15}$$

- At $n = 1$, $S \to R$ link is selected; at $n = 2$, $R \to D$ link is selected. This corresponds to path $q_1 \to q_2 \to q_1$, and we have
$$P(d_r = 2|q_1, \overline{out}_c) = \hat{a}_{21}\hat{a}_{12} \tag{5.16}$$

- At both $n = 1$ and $n = 2$, $S \to R$ link is selected; at time $n = 3$, $R \to D$ link is selected. This corresponds to path $q_1 \to q_2 \to q_3 \to q_2$, we have
$$P(d_r = 3|q_1, \overline{out}_c) = \hat{a}_{21}\hat{a}_{32}\hat{a}_{23} \tag{5.17}$$

- At $n = 1$, $n = 2$ and $n = 3$, $S \to R$ link is selected; at time $n = 4$, $R \to D$ link is selected. This corresponds to path $q_1 \to q_2 \to q_3 \to q_4 \to q_3$, we have

$$P(d_r = 4|q_1, \overline{out}_c) = \hat{a}_{21}\hat{a}_{32}\hat{a}_{43}\hat{a}_{34} \tag{5.18}$$

We can apply the above procedure on every buffer state q_l, $l = 1, \cdots, 4+$.

The above observation can be generalized. At state q_l ($l = 1, \cdots, L$), without the channel outage, the minimum and maximum time slots to transmit out the last packet are l and $L + l - 1$, respectively. Thus we have

$$P(d_r = i|q_l, \overline{out}_c) = 0, \quad \text{if } i < l \text{ or } i > L + l - 1 \tag{5.19}$$

On the other hand, at state q_l, in order to have delay $d_r = i$ ($l \leq i \leq L + l - 1$), the number of the possible paths in the Trellis diagram need to be counted. For example, the highlighted lines in Fig. 5.2, represent all the possible paths for the buffer state q_2 which are as follows:

- For $i < 2$ or $i > 5$, there is no possible path.
- For $i = 2$, $R \to D$ link is selected twice. This corresponds to path $q_2 \to q_1 \to q_0$, which may happen in only one possibility, $C_{l-1}^{i-1} = C_{2-1}^{2-1} = 1$.
- For $i = 3$, $S \to R$ link is selected once, and $R \to D$ link is selected twice. This corresponds to paths $q_2 \to q_1 \to q_2 \to q_1$ or $q_2 \to q_3 \to q_2 \to q_1$, which may happen in two possibilities, $C_{l-1}^{i-1} = C_{2-1}^{3-1} = 2$.
- For $i = 4$, $S \to R$ link is selected twice, and $R \to D$ link is selected twice. This corresponds to paths $q_2 \to q_1 \to q_2 \to q_3 \to q_2$, $q_2 \to q_3 \to q_2 \to q_3 \to q_2$ or $q_2 \to q_3 \to q_4 \to q_3 \to q_2$, which may happen in three possibilities, $C_{l-1}^{i-1} = C_{2-1}^{4-1} = 3$.
- For $i = 5$, same paths of $i = 4$, and $C_{l-1}^{i-1} - C_{i-L+1}^{i-1} = C_{2-1}^{5-1} - C_{5-4+1}^{5-1} = 3$.

In general, the number of possible paths at any value of l and i ($l \leq i \leq L+l-1$) is given by:

$$N(l,i) = \begin{cases} C_{l-1}^{i-1}, & l \leq i \leq L \\ C_{l-1}^{i-1} - C_{i-L+1}^{i-1}, & L < i \leq L+l-1 \end{cases} \tag{5.20}$$

where $C_n^m = \frac{m!}{n!(m-n)!}$. The reason for subtracting C_{i-L+1}^{i-1} when i is larger than L, is the different shape of the Trellis at $i > L$ and the number of possible paths is reduced. This is illustrated in Fig. 5.3, where the highlighted dashed lines represent the removed part of the Trellis when $i > L$.

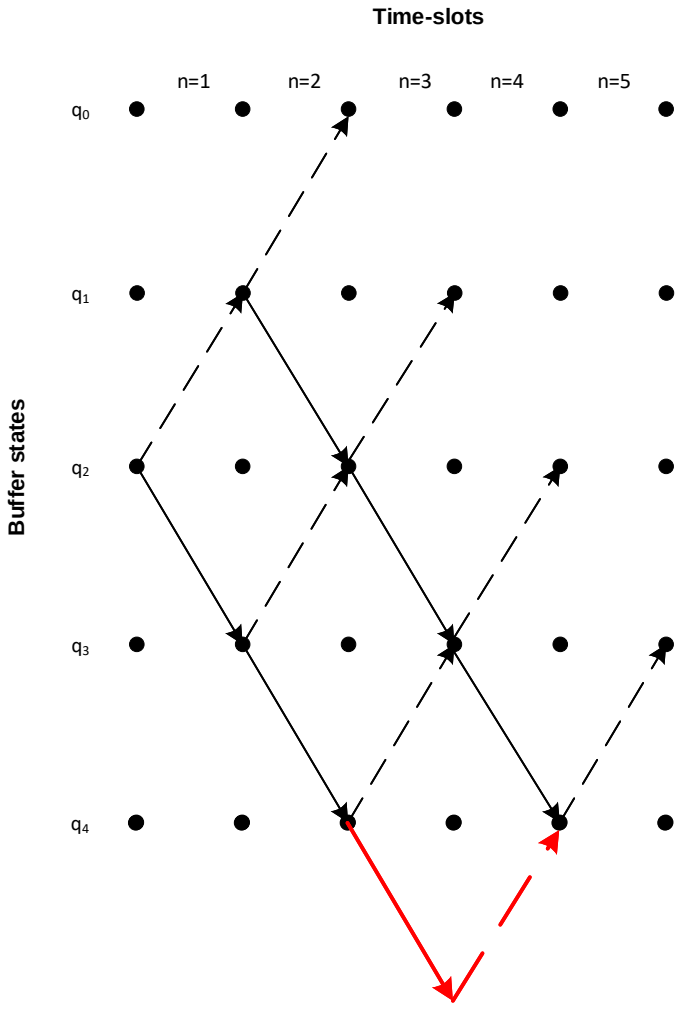

Figure 5.3 The change in Trellis when $i > L$ for $L = 4$.

Denoting $P_{path_m}(l,i)$ as the probability for the m-th path in the Trellis for state q_l and delay $d_r = i$, we have

$$P(d_r = i|q_l, \overline{out}_c) = \sum_{path_m=1}^{N(l,i)} P_{path_m}(l,i), \tag{5.21}$$

where $l \leq i \leq L + l - 1$ and $l = 1, \cdots L$. Substituting (5.21) into (5.13) and then (5.11), we can obtain

$$P(d \leq D_0|\overline{out}_c) = \sum_{j=0}^{L-1} \sum_{i=j+1}^{D(j+1)} \sum_{path_m=1}^{N(j+1,i)} P_{path_m}(j+1,i) \cdot \tilde{\pi}_j \tag{5.22}$$

where $D(j+1) = \min\{D_0, L + j\}$, and $N(j+1,i)$ is given by (5.20).

5.4 Case Studies

5.4.1 Max-link

In the max-link scheme, at any time slot, the available link with the highest SNR is selected. The channel outage probability of the max-link for the i.i.d. channels is obtained in [68]. In this section, we derive the channel outage $P(out_c)$ for the non i.i.d. channels and then $P(d \leq D_0|\overline{out}_c)$, from which the overall outage probability involving both channel and delay is obtained.

In order to obtain $P(out_c)$, we derive the transition matrix $\mathbf{A}$ first. When the buffer is empty or full, the channel outage occurs when either the $S \to R$ or $R \to D$ link is in outage, respectively. In other buffer states, the channel outage occurs when both $S \to R$ or $R \to D$ links are in outage. Thus we have

$$a_{jj} = \begin{cases} 1 - \exp^{-\frac{T}{\bar{\gamma}_1}}, & j = 0 \\ \left(1 - \exp^{-\frac{T}{\bar{\gamma}_1}}\right)\left(1 - \exp^{-\frac{T}{\bar{\gamma}_2}}\right), & j \neq 0,\ L \\ 1 - \exp^{-\frac{T}{\bar{\gamma}_2}}, & j = L \end{cases} \tag{5.23}$$

where $T = 2^{\eta} - 1$.

Similarly, when the buffer is empty or full, the buffer state can only be increased or decreased by one, corresponding to $\gamma_1 > T$ or $\gamma_2 > T$, respectively. At other states, the buffer state is increased by one if $\gamma_1 > \gamma_2$ and $\gamma_1 > T$, or decreased by one if otherwise. Therefore, we have

$$a_{ij} = \begin{cases} \exp^{-\frac{T}{\bar{\gamma_1}}}, & i = 1,\ j = 0 \\ \exp^{-\frac{T}{\bar{\gamma_2}}}, & i = L-1,\ j = L \\ f(\gamma_1, \gamma_2), & i = j+1 \\ f(\gamma_2, \gamma_1), & i = j-1 \\ 0, & \text{otherwise} \end{cases} \tag{5.24}$$

where

$$\begin{aligned} f(\gamma_1, \gamma_2) = P(\gamma_1(t) > \gamma_2(t)) = P(\gamma_1 > T | \gamma_2 < T) \\ + P(\gamma_1(t) > \gamma_2(t) | \gamma_1 > T, \gamma_2 > T) \end{aligned} \tag{5.25}$$

$$\begin{aligned} P(\gamma_1 > T | \gamma_2 < T) &= \int_T^\infty \int_0^T \frac{1}{\bar{\gamma_2}} \exp^{-\frac{y}{\bar{\gamma_2}}} \frac{1}{\bar{\gamma_1}} \exp^{-\frac{x}{\bar{\gamma_1}}}\, dy\, dx \\ &= \left(\frac{1}{\bar{\gamma_1}} \exp^{-\frac{T}{\bar{\gamma_1}}}\right) \left(1 - \frac{1}{\bar{\gamma_2}} \exp^{-\frac{T}{\bar{\gamma_2}}}\right) \end{aligned} \tag{5.26}$$

$$\begin{aligned} P(\gamma_2(t) < \gamma_1(t) | \gamma_1 > T, \gamma_2 > T) = \\ \int_T^\infty \int_T^x \frac{1}{\bar{\gamma_2}} \exp^{-\frac{y}{\bar{\gamma_2}}} \frac{1}{\bar{\gamma_1}} \exp^{-\frac{x}{\bar{\gamma_1}}}\, dy\, dx \end{aligned} \tag{5.27}$$

with mathematical simplification

$$f(\gamma_1, \gamma_2) = \frac{(\bar{\gamma_2}\bar{\gamma_1}^2 - \bar{\gamma_2} - \bar{\gamma_1}) \exp^{-\frac{T(\bar{\gamma_2}+\bar{\gamma_1})}{\bar{\gamma_2}\bar{\gamma_1}}} + \bar{\gamma_1} \exp^{-\frac{T}{\bar{\gamma_2}}} (\bar{\gamma_2} + \bar{\gamma_1})}{(\bar{\gamma_2} + \bar{\gamma_1})\bar{\gamma_2}\bar{\gamma_1}} \tag{5.28}$$

then, similarly

$$\begin{aligned} f(\gamma_2, \gamma_1) =& P(\gamma_2(t) > \gamma_1(t)) \\ =& \frac{(\bar{\gamma_1}\bar{\gamma_2}^2 - \bar{\gamma_1} - \bar{\gamma_2}) \exp^{-\frac{T(\bar{\gamma_1}+\bar{\gamma_2})}{\bar{\gamma_1}\bar{\gamma_2}}} + \bar{\gamma_2} \exp^{-\frac{T}{\bar{\gamma_1}}} (\bar{\gamma_1} + \bar{\gamma_2})}{(\bar{\gamma_1} + \bar{\gamma_2})\bar{\gamma_1}\bar{\gamma_2}} \end{aligned} \tag{5.29}$$

From (5.23) and (5.24), we obtain $\mathbf{A}$, then substituting $\mathbf{A}$ into (5.9) to obtain $P(out_c)$ from (5.8).

To obtain $P(d \leq D_0|\overline{out}_c)$ from (5.22) for max-link, we need to find $P_{path_m}(l,i)$, we assume:

- P_1 denotes $P(q_j \rightarrow q_{j+1})\ 0 < j < L$
- P_2 denotes $P(q_j \rightarrow q_{j-1})\ 0 < j < L$
- P_3 denotes $P(q_j \rightarrow q_{j-1})\ j = L$

For example, Fig. 5.4 shows the Trellis for $L = 5$ and $l = 3$, where $P_{path_m}(l,i)$ for this example is described as follows:

- for $3 \leq i < 5$, links can have one of the two probabilities P_1 or P_2, and links with P_2 have to be chosen 3 times to get the packet transmitted. Similar to P_2 link, every time P_1 link is selected, the delay i is increased by 1, hence, P_1 is selected $i - 3$ times.
- for $i = 5$, a different probability P_3 occurs at full buffer, which is part of one path only $q_3 \rightarrow q_4 \rightarrow q_5 \rightarrow q_4 \rightarrow q_3 \rightarrow q_2$, and the rest of the paths $((N(l,L) - 1))$ are similar to case $3 \leq i < 5$.
- for $L < i \leq L+l-2$, P_3 is involved in multiple paths and it can be encountered more than once within the same path. Since full buffer is more likely to occur when $i > L$. With this values of i, if $l = L$, then all paths have P_3 in it ($v_{l-1} = 0$).
- for $i = L + l - 1$, all paths at $i = L + l - 2$ have to go through one link with probability P_3.

The $P_{path_m}(l,i)$ description for $L = 5$ and $l = 3$ can be generalized to any L and l as:

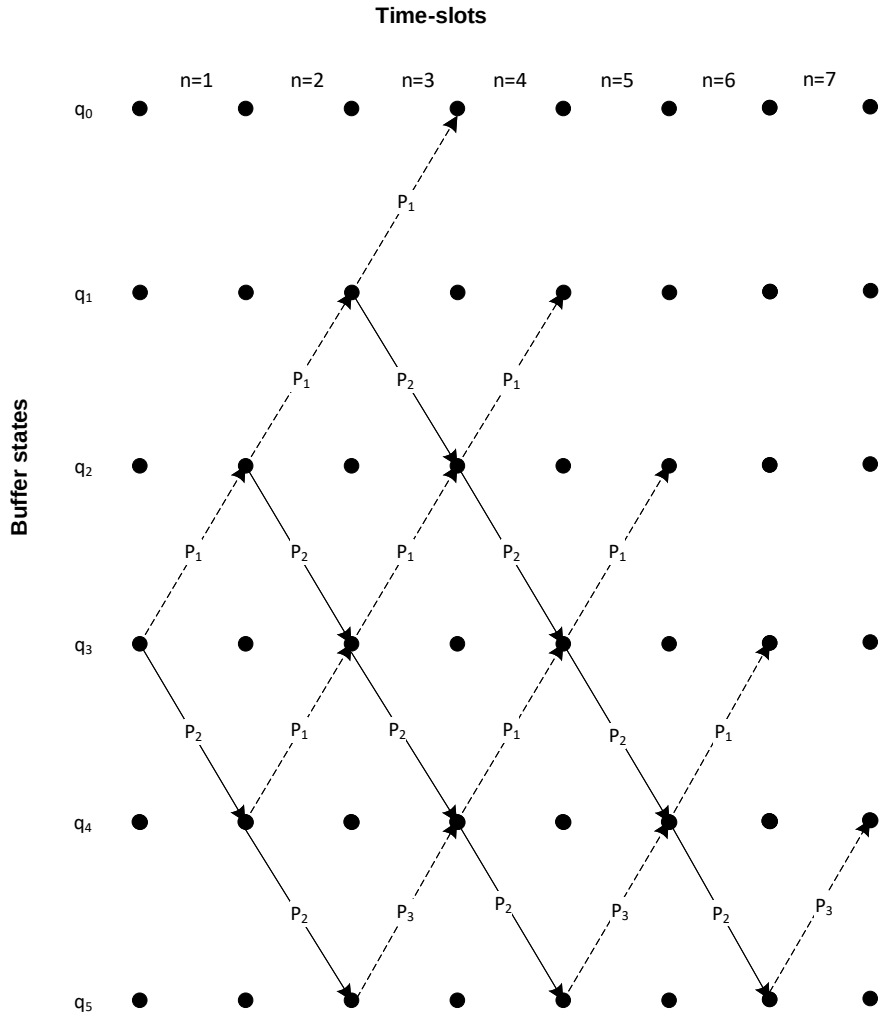

Figure 5.4 Trellis for $L = 5$ and $l = 3$.

$$P_{path_m}(l,i) = \begin{cases} C_{l-1}^{i-1} P_1^{i-l} P_2^{l}, & l \le i < L \\ P_1^{i-l} P_2^{l-1} P_3 + (C_{l-1}^{i-1} - 1) P_1^{i-l} P_2^{l}, & i = L \\ P_1^{i-l} P_2^{l+L-i-1} P_3^{i-L+1} + v_{l-1} P_1^{i-l} P_2^{l} + \\ \sum_{j=1}^{i-L} (v_j P_1^{i-l} \\ P_2^{l+L-i-1+j} P_3^{i-L+1-j}), & L < i \le L+l-2 \\ P_3 P_{path_m}(l, L+l-2), & i = L+l-1 \end{cases} \tag{5.30}$$

where

- $v_j = N(l, L+j) - (s+uz) - 1$, $j = 1, 2 \dots l-2$
- $v_{l-1} = \begin{cases} v_{l-1} = N(l, L+l-2) - (\sum_{i=1}^{l-2} v_i) - 1, & l < L \\ 0, & l = L \end{cases}$
- $u = \begin{cases} 0, & l - j \leq 3 \\ 1, & ow \end{cases}$
- $s = C_{l-1}^{L-1} - 1$
- $z = \sum_{i=1}^{l-3} C_{l-1}^{L-1-i} - 1$.

The path that has the highest power of P_3 occurs only one time ($q_3 \to q_4 \to q_5 \to q_4 \to q_5 \to q_4 \to q_5 \to q_4$ in $L = 5$ and $l = 3$ example), so we subtract 1 from the notations v, s and z. By substituting (5.30) into (5.22), and then substituting $P(out_c)$ and $P(d \leq D_0|\overline{out}_c)$ into (5.3), we get P_{out} for max-link.

5.4.2 State-Based

While max-link gives a better outage performance than max-min, it increases the delay to an unacceptable levels especially for the 5G applications. As a result, the state-based was suggested to consider the buffer state into the selection, which led to a better outage and delay performance. In the state-based, if the buffer content is less than two, the priority is given to the reception, otherwise, the priority is given to the transmission. This modifies both parts of (5.3): $P(out_c)$ and $P(d \leq D_0|\overline{out}_c)$. For $P(out_c)$, a_{jj} of **A** is similar to (5.23), but a_{ij} becomes

$$a_{ij} = \begin{cases} \exp^{-\frac{T}{\bar{\gamma}_1}}, & i = 1,\ j = 0 \\ \exp^{-\frac{T}{\bar{\gamma}_2}}, & i = L-1,\ j = L \\ \left(1 - \exp^{-\frac{T}{\bar{\gamma}_1}}\right) \exp^{-\frac{T}{\bar{\gamma}_2}}, & i = 0,\ j = 1 \\ \exp^{-\frac{T}{\bar{\gamma}_1}}, & i = 2,\ j = 1 \\ \left(1 - \exp^{-\frac{T}{\bar{\gamma}_2}}\right) \exp^{-\frac{T}{\bar{\gamma}_1}}, & i = j+1,\ j > 1 \\ \exp^{-\frac{T}{\bar{\gamma}_2}}, & i = j-1,\ j > 1 \\ 0, & \text{otherwise} \end{cases} \tag{5.31}$$

by obtaining **A**, and substituting **A** into (5.9), we can obtain the state-based $P(out_c)$ from (5.8).

Regarding $P(d \leq D_0|\overline{out}_c)$, similar to max-link we have P_1, P_2 and P_3, but in state-based, there is a special case at q_1, so we have $P_4 = P(q_1 \rightarrow q_2)$ and $P_5 = P(q_1 \rightarrow q_0)$. The effected paths by the new probabilities are highlighted in Fig. 5.5.

This affects $P_{path_m}(l, i)$ for the example in Fig. 5.4, as follows:

- for $i = 3$, when $i = l$, there is only one path $q_3 \rightarrow q_2 \rightarrow q_1 \rightarrow q_0$, which has P_5 once and P_2 is selected 2 times.
- for $3 < i < L$, only one of the paths passes through $q_1 \rightarrow q_2$ which has the probability P_4.

This alters (5.30) as follows

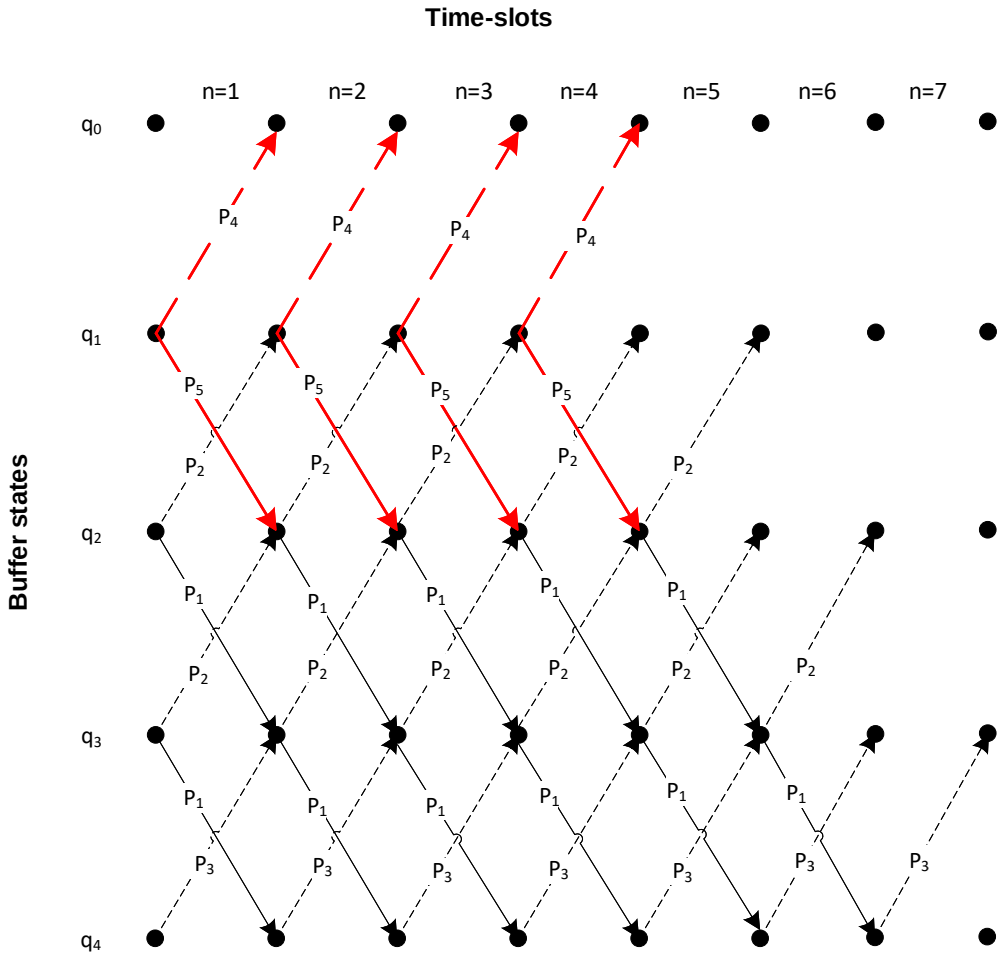

Figure 5.5 Effected paths in the state-based Trellis diagram for $L = 4$.

$$P_{path_m}(l,i) = \begin{cases} P_2^{l-1}P_5, & l = i < L \\ P_1^{i-l-1}P_2^l P_4 + (C_{l-1}^{i-1} - 1)P_1^{i-l}P_2^l, & l < i < L \\ P_2^{l-2}P_3P_5, & l = i = L \\ P_1^{i-l}P_2^{l+L-i-1}P_3^{i-L+1} + \\ +P_1^{i-l-1}P_2^{l-1}P_3P_4 + \sum_{j=1}^{i-L}(v_j \\ P_1^{i-l}P_2^{l+L-i-1+j}P_3^{i-L+1-j}), & L < i \leq L+l-2,\ l = L \\ P_3 P_{path_m}(l,(L+l-2,\ l=L)), & L < i \leq L+l-1,\ l = L \\ \text{Similar to max-link}, & ow \end{cases} \tag{5.32}$$

For $L < i \leq L + l - 2,\ l = L$ same notations of max-link v, u, s and z are used here, but with subtracting 2 instead of 1 in v, s and z, because there are two special paths: the path with the highest power of P_3, and the path involves P_4. The special cases for $l = L$ effect $P_{path_m}(l, i)$ are described as follows:

- for $l = i = L$, there is only one path through P_3 and P_5 ($q_L \rightarrow q_{L-1} \dots q_0$)
- for $L < i \leq L + l - 2$, if $l = L$, there are two special paths, the path with the highest power of P_3, and the path with P_4. Otherwise, this case is similar to max-link.
- for $i = L + l - 1$, if $l = L$, all paths at $i = L + l - 2, l = L$ have to go through P_3 at the end of the Trellis. Otherwise, this case is similar to max-link.

Similar to max-link, by substituting (5.32) into (5.22) and then substituting $P(out_c)$ and $P(d \leq D_0|\overline{out}_c)$ into (5.3) we get P_{out} for the state-based.

5.4.3 Delay-Reduced

In the delay-reduced scheme, higher priorities are given to transmission. This guarantees that each packet leaves the buffer as soon as possible. The delay-reduced is superior in terms of the delay reduction compared to the state-based, however, the delay reduction is on the price of higher outage. In the delay-reduced, a_{jj} is similar to the max-link and the state-based, but a_{ij} is changed as follows:

$$a_{ij} = \begin{cases} \exp^{-\frac{T}{\gamma_2}}, & i = 0,\ j = 1 \\ \exp^{-\frac{T}{\gamma_1}} \left(1 - \exp^{-\frac{T}{\gamma_2}}\right), & i = 2,\ j = 1 \\ \text{Similar to state-based}, & \text{otherwise} \end{cases} \tag{5.33}$$

By substituting $\mathbf{A}$ into (5.9) we can obtain $P(out_c)$ for the delay-reduced from (5.8).

The delay-reduced has $P_{path_m}(l, i)$ similar to (5.30), with different values of P_1, P_2 and P_3. By substituting the new values in (5.30), and then into (5.22), we can get P_{out} for the delay-reduced by substituting $P(out_c)$ and $P(d \leq D_0|\overline{out}_c)$ into

(5.3).

5.5 Link Selection With Adaptive Buffer Size

It is understood from the previous analysis that, with larger buffer size, the channel outage probability becomes smaller but the delay overtime probability is higher. Therefore, there exists an optimum buffer-size corresponding to the smallest delay-constrained outage probability. In this section, we propose the adaptive buffer-size algorithm to search for the optimum buffer-size.

In the adaptive buffer-size algorithm, the delay-constrained outage probability is monitored on a block-to-block time basis. We assume that every block time contains N_t time slots. For the n-th block time, the delay-constrained outage probability is measured as $\hat{P}_{out}(n) = N_e(n)/N_t$, where $N_e(n)$ is the number of the time slots that either the channel outage occurs or the delay is beyond the target delay constraint.

Within one block time, the buffer-size remains unchanged but adapts from one block to another. Suppose that at block times $(n-1)$ and n, the buffer-sizes are $L(n-1)$ and $L(n)$, respectively. If $\hat{P}_{out}(n) < \hat{P}_{out}(n-1)$, this means the buffer-size adaptation from time $(n-1)$ to n is 'correct'. Therefore, we shall have the similar buffer-size adaptation from time n to $(n+1)$. Otherwise, if $\hat{P}_{out}(n) > \hat{P}_{out}(n-1)$, the buffer-size adaptation from n to $(n+1)$ shall be opposite to that from $(n-1)$ to n. With these considerations, we have the buffer-size adaptation rule as

$$L(n+1) = \begin{cases} L(n) + \delta_l \cdot \Delta_L(n), & \text{if } \hat{P}_{out}(n) < \hat{P}_{out}(n-1) \\ L(n) - \delta_l \cdot \Delta_L(n), & \text{otherwise} \end{cases} \tag{5.34}$$

where $\Delta_L(n) = \text{sign}\{L(n) - L(n-1)\}$, and δ_l is the adaptation step-size which is a positive integer. We can express (5.34) in a more compact form as

$$L(n+1) = L(n) + \delta_l \Delta_L(n) \Delta_P(n), \tag{5.35}$$

where $\Delta_P(n) = \text{sign}\{P_{out}(n-1) - P_{out}(n)\}$.

Furthermore, if the maximum and minimum buffer sizes are L_{max} and L_{min} respectively, we have

$$L(n+1) = \min\left\{\max\{L(n) + \delta_l \Delta_L(n)\Delta_P(n),\ L_{min}\},\ L_{max}\right\} \tag{5.36}$$

The adaptive buffer-size algorithm can be used in any buffer-aided link selection schemes including those in Section 5.4.

Algorithm 2 The proposed algorithm

```
Input the number of time-blocks N
for n = 1 : N do
    N_e = 0 :
    the buffer size is adapted according to [5.36], starting with L_max at n = 1
    for i = 1 : N_t do
        if no channel outage and no delay overtime then
            if buffer-length ≥ target-length then
                give R → D link higher priority for selection
            else
                give S → R link higher priority for selection
            end if
        else if channel outage or delay overtime then
            N_e = N_e + 1
        end if
    end for
    P̂_out(n) = N_e/N_t
end for
```

5.6 Numerical Simulations

In all simulations below, the target transmission rate is set to $\eta = 2$ bps/Hz, the buffer size is set to $L = 50$, the target delay is set to $D_0 = 25$ time slots and the average channel gains are set to $\Omega_1 = \Omega_2 = 0.5$ for $S \to R$ and $R \to D$ links respectively except otherwise stated.

Fig. 5.6 compares the channel outage probability $P(out_c)$ and the delay-constrained outage probability $P(out)$ with respect to the SNR for the max-link, delay-reduced and state-based link selection schemes. Particularly, for the delay-

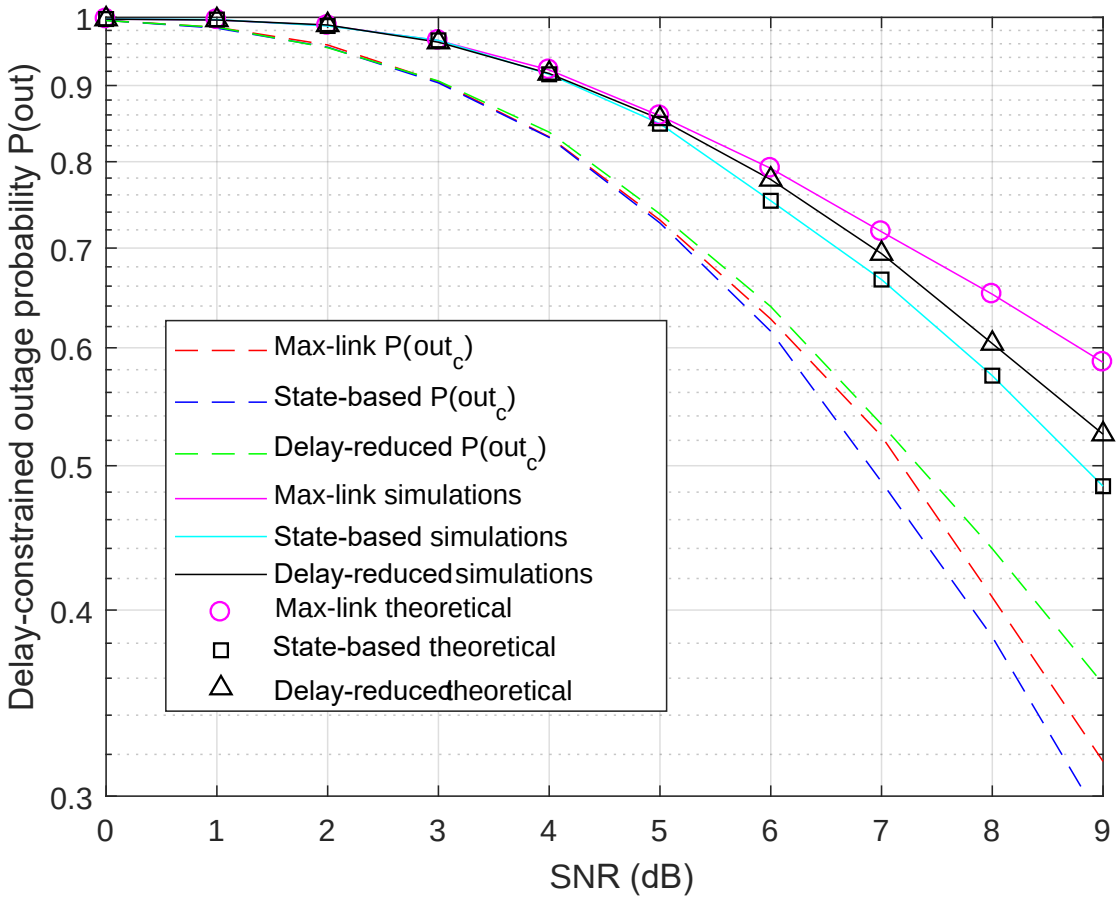

Figure 5.6 The channel outage probability $P(out_c)$ *vs.* the delay constrained outage probability P_{out}, where average channel gains $\Omega_1 = \Omega_2 = 0.5$.

constrained outage probabilities, both the theoretical and simulation results are shown. We have the following observations:

- For the three schemes, the theoretical results for the delay-constrained outage probability well match the simulation results. This verifies the performance analysis in Section 5.4.

- The delay-constrained outage probabilities $P(out)$ are significantly larger than the channel outage probabilities $P(out_c)$ in all schemes. This states that the outage performance deteriorates drastically when the delay constraints are considered.

- For the channel outage probability $P(out_c)$, the state-based and delay-reduced schemes have the best and worst performance, respectively. This matches our expectation, because the state-based scheme can better avoid the full

or empty buffer states than the max-link scheme, while the delay-reduced scheme achieves lower packet delay at the price of higher outage probability.

- On the other hand, for the delay-constrained outage probability $P(out)$, while the state-based scheme still has the best performance, the delay-reduced scheme performs better than the max-link scheme. This is not surprising because the packet delay in the delay-reduced scheme is well lower than that in the max-link, leading to better $P(out)$.

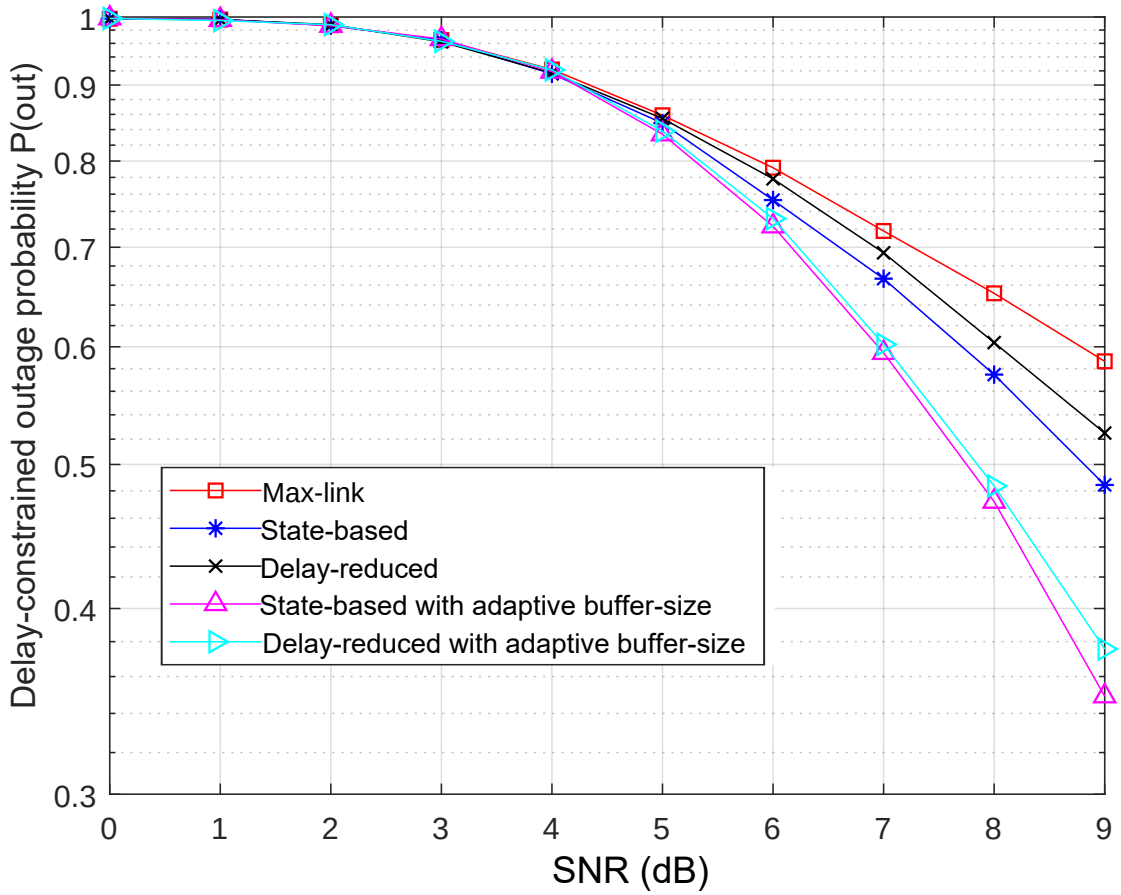

Figure 5.7 Delay-constrained outage probabilities $P(out)$ for the delay-reduced and state-based link selection schemes with and without adaptive buffer-size, where the average channel gains $\Omega_1 = \Omega_2 = 0.5$.

Fig. 5.7 compares the delay-constrained outage probabilities $P(out)$ for the delay-reduced and state-based link selection schemes with and without adaptive buffer-size. Particularly for the adaptive buffer-size, the maximum and minimum buffer-sizes are set as 50 and 0, respectively. For comparison, the result for the

benchmark max-link scheme is also shown. It is clearly shown that the selection schemes with adaptive buffer-size achieve significantly lower outage probabilities than their fixed buffer-size counterpart.

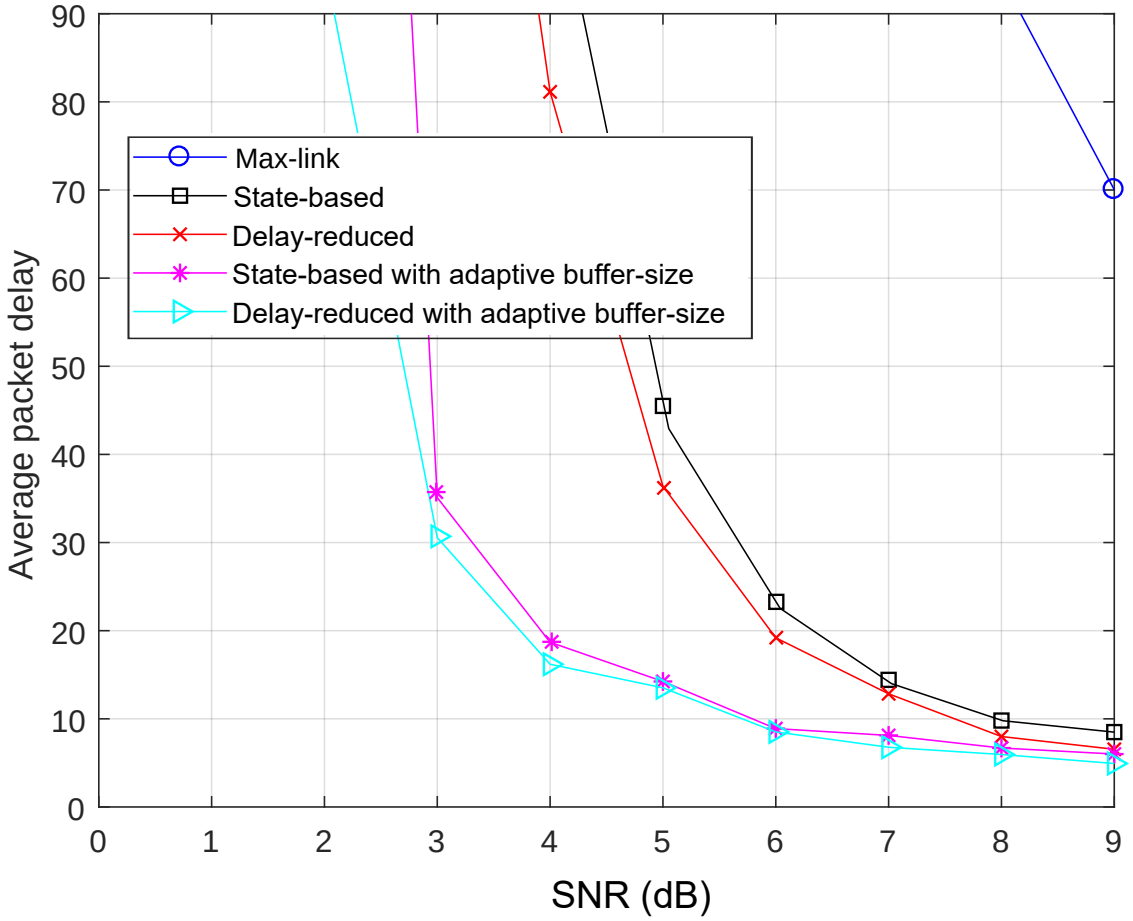

Figure 5.8 Average packet delays for the delay-reduced and state-based link selection schemes with and without adaptive buffer-size, where the average channel gains $\Omega_1 = \Omega_2 = 0.5$.

Fig. 5.8 shows the average packet delays for the schemes in Fig. 5.7. It is clearly shown that the adaptive buffer-size well reduces the average delays, which again leads to lower delay-constrained outage probability as is shown in Fig. 5.7.

Fig. 5.9 is similar to that in Fig. 5.7, except the i.n.i.d. channels are considered such that the average channel gains are set to $\Omega_1 = 0.5$ and $\Omega_2 = 1$. In this case, the $R \to D$ link is stronger than the $S \to R$ link, making the buffer be more likely to be empty than saturated. We have the following observations:

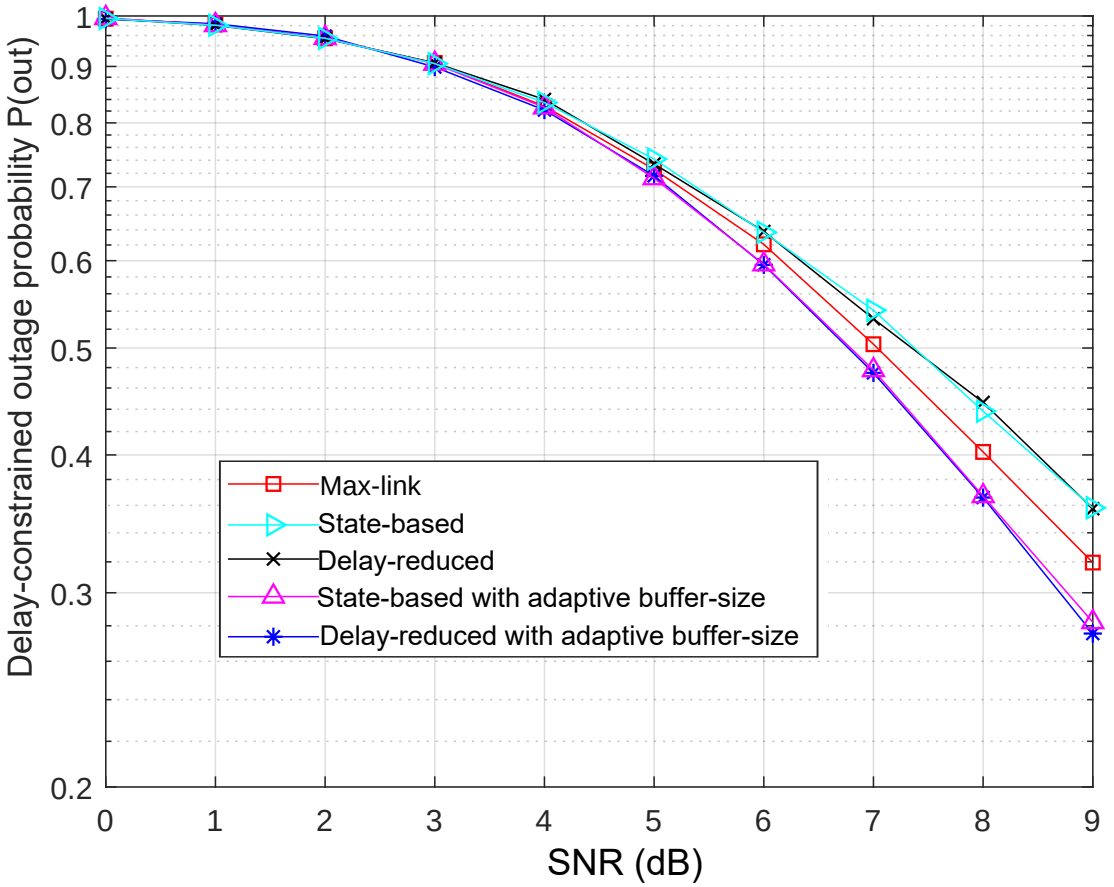

Figure 5.9 Delay-constrained outage probabilities $P(out)$ for the delay-reduced and state-based link selection schemes with and without adaptive buffer-size, where the average channel gains $\Omega_1 = 0.5$ and $\Omega_2 = 1$.

- Without the adaptive buffer-size, the state-based and delay-reduced link selection schemes have close delay-constrained outage probabilities which are both higher than that in the max-link scheme. This is because that the state-based and delay-reduced schemes are equivalent to setting the *target buffer length* to two and zero in our previous proposed prioritization-based relay selection ([49, 8]), respectively. With stronger $R \to D$ link, because the buffer tends to be empty, the target buffer length shall be set close to the full buffer-size. Therefore, because neither the state-based nor delay-reduced scheme has equivalent target buffer length close the buffer size, both have worse outage performance than the max-link scheme.

- With the adaptive buffer-size, the state-based and delay-reduced schemes also have close delay-constrained outage probabilities, and both perform

better than their fixed buffer-size counterpart, though the comparison is not as obvious as that in Fig. 5.7.

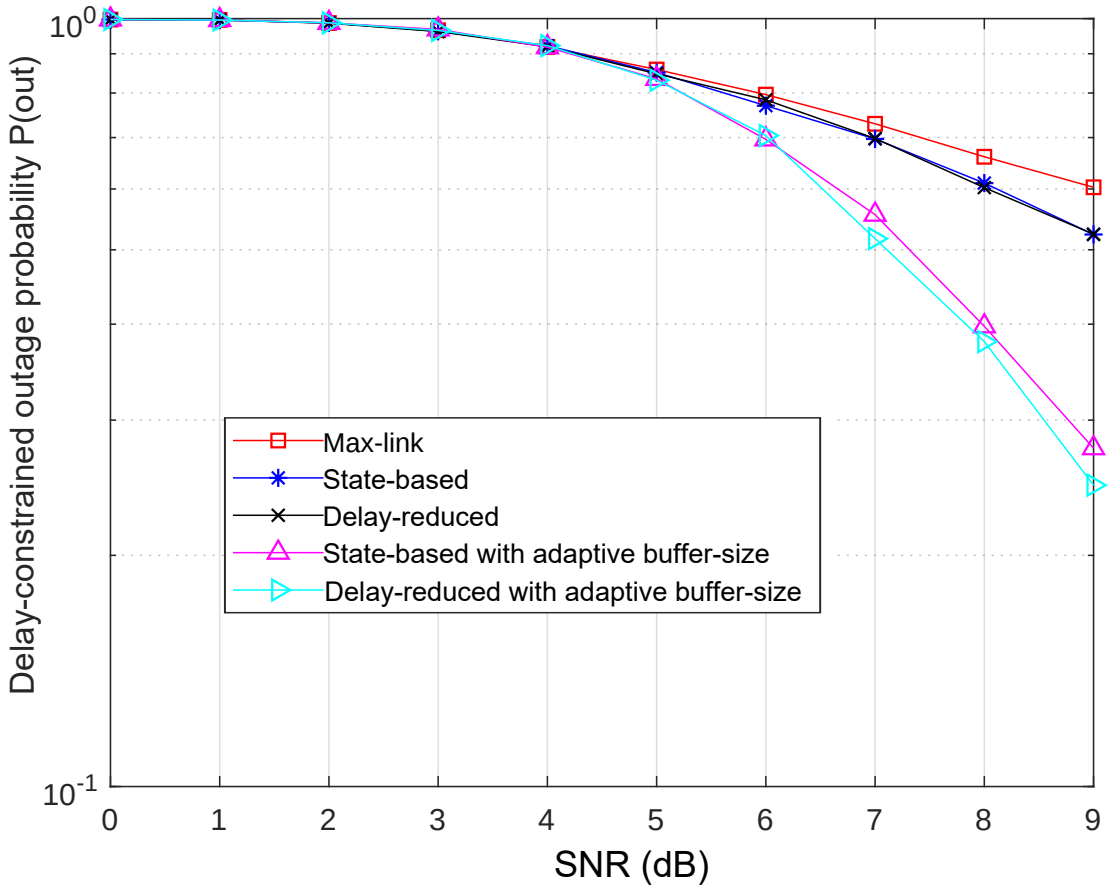

Figure 5.10 Delay-constrained outage probabilities $P(out)$ for the delay-reduced and state-based link selection schemes with and without adaptive buffer-size, where the average channel gains $\Omega_1 = 1$ and $\Omega_2 = 0.5$.

Fig. 5.10 also consider the i.n.i.d. channels except that the $S \to R$ link is stronger than the $R \to D$ link that $\Omega_1 = 1$ and $\Omega_2 = 0.5$. We have the following observations:

- Without the adaptive buffer-size, both the state-based and delay-reduced schemes perform better than the max-link scheme. This is because that, with stronger $S \to R$ link, the buffer is more likely to be saturated than empty. Thus the optimum target buffer length shall be set close to zero, which well matches the state-based or the delay-reduced scheme.

- With the adaptive buffer-size, both the state-based and delay-reduced schemes perform better than their fixed buffer-size counterpart. Particularly, the performance improvement due to adaptive buffer-size is more obvious than those in Fig. 5.7 and 5.9. This is because that, with stronger $S \rightarrow D$ link, the buffer lengths are more likely to grow than shrink, making it have higher impact for the delay overtime on the delay-constrained outage probability.

5.7 Summary

This chapter studies the impact of constraining the delay to a specific target delay on the buffer-aided relay performance. This is done through introducing the delay-constrained outage probability, which combines outages caused by channels and by delay exceeds the target delay. The delay-constrained outage probability is studied for three of the most effective buffer-aided relay selection schemes: the max-link, the state-based and the delay-reduced. Result shows that the delay-constrained outage probability has worse values than the channel outages in the three schemes because of the delay-constraints. Therefore, a novel adaptive buffer-size is proposed to mitigate the delay-constraints challenge. Result show improvement in the delay-constrained outage probability when adaptive buffer-size is applied.

Chapter 6

Conclusions and future work

This chapter summarizes the main contributions of this thesis and the main conclusions that can be drawn from them. In addition, this chapter includes a discussion about possible directions for future research.

6.1 Conclusions

Traditional cooperative networks are known for their ability to achieve higher throughput and diversity gain. On the other hand, giving the cooperative relay networks the buffering capabilities has improved their performance tremendously. However, this improvement in the throughput and the diversity gain can be on the price of longer data queues in the buffers, and ultra-low latency is one of the main requirements in the 5G applications. This thesis has focused on how the buffer-aided relays can be applied in the 5G cooperative networks more effectively. In particular, this thesis has proposed techniques to improve the buffer-aided cooperative relay networks main performance metrics: the throughput, the diversity gain and the average packet delay.

In Chapter 3, we proved the effectiveness of the proposed prioritization-based buffer-aided relay selection scheme through analytical expressions and numerical simulations. The proposed scheme combines the NOMA and the OMA transmissions in the buffer-aided relay network. The proposed scheme uses the OMA transmission when the NOMA is not possible rather than being in outage like the case in [140].

This is why the proposed scheme achieves higher throughput and diversity gain than [140], especially at low SNR ranges. The superiority of the proposed scheme compared to the threshold-based switching scheme is also proved. While all the available schemes for buffer-aided relays in cooperative NOMA has considered a single relay, the proposed scheme has shown its excellence for the multiple relays scenario. And the performance improvement in the proposed scheme is proportional to the number of relays. Chapter 3 also stresses the importance of choosing the proper value of the target length. Thus, in delay-unconstrained applications, selecting the target length is based only on avoiding empty and full buffers.

Chapter 4 studies the impact of the source delay in cooperative relay networks. The result shows that the buffer-aided relays have lower source delay than the conventional (non-buffer-aided relays). In a single relay network, considering the source delay was enough to make the end-to-end delay of the buffer-aided relay lower than in the conventional relay at low SNR values. However, this is not true with multiple relays. Because the delay-reduced [48], which is the best available scheme to minimize the delay, is proportional to the number of relays at the low SNR range. Therefore, two delay reduction techniques are used to tackle the long delays caused by multiple relays. Broadcasting and small buffer size are able to reduce the end-to-end delay of the buffer-aided multiple relay network and make it competitive to the conventional network. In addition, the proposed adaptive target length scheme has succeeded in reducing the buffer-aided relay network delay. The proposed scheme has a better than the delay-reduced even with applying the broadcasting and small buffer size on the delay-reduced.

In Chapter 5, the buffer-aided relay performance degradation caused by constraining the delay to a target delay is analysed by introducing the delay-constrained outage probability. The delay-constrained outage probability is applied on three relay selection schemes: the max-link, the state-based and the delay-reduced. Results show the huge degradation on outage performance caused by exceeding the target delay. The delay-reduced has the best delay profile, but on the price of channel outage. While the state-based has the best delay-constrained outage probability

because it has the best channel outage and better delay performance than the max-link. A novel adaptive buffer-size algorithm is proposed to reduce the delay. When comparing adaptive buffer-size delay-reduced and state-based with their fixed buffer-size counterparts, the ones with adaptive buffer-size have experienced significant delay reduction, which is also, led to better delay-constrained outage performance. Interestingly, the importance of choosing the right target length is proved again. Specifically, when the $R \rightarrow D$ channel is stronger than $S \rightarrow R$, the scheme with the larger target length has better performance, and vice versa.

6.2 Future work

The research presented in this thesis could be extended in several directions. In general, we assumed flat fading Rayleigh channels through the whole thesis. Because it gives a fair comparison with the available relay selection studies since the majority of them applied flat fading Rayleigh in their assumption. It is intuitive to check the performance of the presented systems under different channel models, such as Nakagami-m distribution. Nakagami-m distribution is suitable for special cases like the indoor mobile multi-path environment as mentioned in [122]. Another assumption through this thesis is the availability of the CSI; it is useful to study the impact of the outdated CSI or the statistical CSI on the proposed systems performance because the availability of the CSI is less common in practical scenarios. Besides, study the proposed solutions with AF relays is a possible future research path.

In Chapter 3, studying the impact of the proposed scheme on average packet delay might be necessary to gain a comprehensive performance assessment from all perspectives. And finding the appropriate target length to achieve the optimal trade-off between the diversity and the delay is still open for research. Another interesting direction is to apply the adaptive transmission rate to the proposed scheme to calculate the maximum achievable rate and compare it to the previous schemes. This was the path in previous studies, but it still valid for the proposed scheme.

In Chapter 4, enhancements other than the broadcasting and the small buffer size can be studied while considering the source delay. For example, authors in [144] suggest the NOMA as a delay reduction technique. So, applying the NOMA in the $S \rightarrow R_k$ links rather than the broadcasting may lead to better results. In addition, the optimal values for the threshold and the target lengths of the proposed relay selection scheme is still an open problem.

Extending the work in Chapter 5, to a multiple relay scenario is a natural future direction. Also, applying the machine learning (ML) methods on the proposed schemes to find the optimal target length or buffer size is an attractive future direction as ML methods have proved its ability to tackle very complicated problems. Finally, the door is still open for proposing new selection rules which can achieve better trade-off between the performance metrics.

References

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[2] L. Ahlin, "Jens zander principles of wireless communications," *Studentlitteratur,*, 2000.

[3] N. Ahmed, M. A. Khojestapour, and R. G. Baraniuk, "Delay-limited throughput maximization for fading channels using rate and power control," in *IEEE Global Telecommunications Conference, 2004. GLOBECOM '04.*, vol. 6, Nov 2004, pp. 3459–3463 Vol.6.

[4] A. Al-Dulaimi, X. Wang, and C. I, *Waveform design for 5G and beyond*. IEEE, 2018, pp. 51–76. [Online]. Available: https://ieeexplore.ieee.org/document/8496433

[5] M. S. Ali, E. Hossain, and D. I. Kim, "Coordinated multipoint transmission in downlink multi-cell NOMA systems: Models and spectral efficiency performance," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 24–31, April 2018.

[6] M. Alkhawatrah, Y. Gong, O. Aldabbas, and M. Hammoudeh, "Buffer-aided 5g cooperative networks: Considering the source delay," *In Proceedings of the 3rd International Conference on Future Networks and Distributed Systems*, p. 13, July 2019.

[7] M. Alkhawatrah, Y. Gong, and G. Chen, "Delay-constrained adaptive link selection in buffer-aided relay networks," *IEEE Internet of Things Journal*, In Submission.

[8] M. Alkhawatrah, Y. Gong, G. Chen, S. Lambotharan, and J. Chamber, "Buffer-aided relay selection for cooperative noma in the internet of things, year=2019, volume=6, number=3, pages=5722-5731, keywords=, doi=, issn=, month=June„" *IEEE Internet of Things Journal*.

[9] M. Alkhawatrah, Y. Gong, and S. Lambotharan, "The impact of source delay on end-to-end average packet delay in buffer-aided relay networks," *IEEE Open Journal of the Communications Society*, In Submission.

[10] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[11] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[12] I. Azam, M. B. Shahab, and S. Y. Shin, "User pairing and power allocation for capacity maximization in uplink NOMA," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, July 2019, pp. 690–694.

[13] K. Azarian, H. El Gamal, and P. Schniter, "Achievable diversity-vs-multiplexing tradeoffs in half-duplex cooperative channels," in *Information Theory Workshop*, Oct 2004, pp. 292–297.

[14] R. Baldemair, E. Dahlman, G. Fodor, G. Mildh, S. Parkvall, Y. Selen, H. Tullberg, and K. Balachandran, "Evolving wireless communications: Addressing the challenges and expectations of the future," *IEEE Vehicular Technology Magazine*, vol. 8, no. 1, pp. 24–30, 2013.

[15] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "NOMA/OMA mode selection-based cell-free massive MIMO," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–6.

[16] A. Benjebbour, Y. Kishiyama, Y. Okumura, C. Hwang, and I. Fu, "Outdoor experimental trials of advanced downlink NOMA using smartphone-sized devices," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, June 2018, pp. 1–6.

[17] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *2015 IEEE Conference on Standards for Communications and Networking (CSCN)*, Oct 2015, pp. 18–23.

[18] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept, performance evaluation and experimental trials," in *2015 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct 2015, pp. 1–6.

[19] A. Benjebbour, K. Saito, A. Li, Y. Kishiyama, and T. Nakamura, "Non-orthogonal multiple access (NOMA): Concept and design," *in Signal Processing for G*, vol. 5, 2016.

[20] A. Berman and R. J. Plemmons, *Nonnegative matrices in the mathematical sciences*. Siam, 1994, vol. 9.

[21] D. Bertsekas and J. Tsitsiklis, *Introduction to Probability*, ser. Athena Scientific optimization and computation series. Athena Scientific, 2008. [Online]. Available: https://books.google.jo/books?id=yAy-PQAACAAJ

[22] Bin Zhao and M. C. Valenti, "Practical relay networks: a generalization of hybrid-ARQ," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 1, pp. 7–18, Jan 2005.

[23] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 659–672, March 2006.

[24] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, Mar. 2006.

[25] A. Bletsas, H. Shin, and M. Z. Win, "Cooperative communications with outage-optimal opportunistic relaying," *IEEE Transactions on Wireless Communications*, vol. 6, no. 9, pp. 3450–3460, Sep. 2007.

[26] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[27] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, 2008.

[28] I. Chatzigeorgiou, I. J. Wassell, and R. Carrasco, "On the frame error rate of transmission schemes on quasi-static fading channels," in *2008 42nd Annual Conference on Information Sciences and Systems*, 2008, pp. 577–581.

[29] G. Chen, J. P. Coon, A. Mondal, B. Allen, and J. A. Chambers, "Performance analysis for multi-hop full-duplex iot networks subject to poisson distributed interferers," *IEEE Internet of Things Journal*, pp. 1–1, 2018.

[30] G. Chen, J. Tang, and J. P. Coon, "Optimal routing for multihop social-based D2D communications in the internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1880–1889, June, 2018.

[31] G. Chen, Z. Tian, Y. Gong, and J. Chambers, "Decode-and-forward buffer-aided relay selection in cognitive relay networks," *IEEE Trans. Veh. Tech.*, vol. 63, no. 9, pp. 4723–4728, Nov. 2014.

[32] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "A mathematical proof of the superiority of NOMA compared to conventional OMA," *arXiv preprint arXiv:1612.01069*, 2016.

[33] L. Chettri and R. Bera, "A comprehensive survey on internet of things (IoT) toward 5G wireless systems," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 16–32, 2020.

[34] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294–2323, thirdquarter 2018.

[35] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5G on the horizon: Key challenges for the radio-access network," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 47–53, Sep. 2013.

[36] D. Deng, L. Fan, X. Lei, W. Tan, and D. Xie, "Joint user and relay selection for cooperative NOMA networks," *IEEE Access*, vol. 5, pp. 20 220–20 227, Sep. 2017.

[37] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan 2016.

[38] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 416–419, Aug 2016.

[39] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5g nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, Aug 2016.

[40] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5g systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, 2015.

[41] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun 2016.

[42] Y. Fan, C. Wang, J. Thompson, and H. V. Poor, "Recovering multiplexing loss through successive relaying using repetition coding," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4484–4493, December 2007.

[43] F. R. Farrokhi, A. Lozano, G. J. Foschini, and R. A. Valenzuela, "Spectral efficiency of FDMA/TDMA wireless systems with transmit and receive antenna arrays," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 591–599, 2002.

[44] G. P. Fettweis, "The tactile internet: Applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.

[45] R. G. Gallager, *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[46] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[47] F. Gomez-Cuba, R. Asorey-Cacheda, and F. J. Gonzalez-Castano, "A survey on cooperative diversity for wireless networks," *IEEE Communications Surveys Tutorials*, vol. 14, no. 3, pp. 822–835, 2012.

[48] Y. Gong, G. Chen, and T. Xie, "Using buffers in trust-aware relay selection networks with spatially random relays," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5818–5826, Sep. 2018.

[49] Y. Gong, G. Chen, and T. Xie, "Using buffers in trust aware relay selection networks with spatially random relays," *to appear in IEEE Trans. Wire. Commun.*, pp. 1–1, 2018.

[50] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective non-orthogonal multiple access scheme," *IEEE Trans. Veh. Technol*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[51] D. Gunduz and E. Erkip, "Source and channel coding for quasi-static fading channels," in *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, 2005, pp. 18–22.

[52] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[53] G. Hampel, C. Li, and J. Li, "5G ultra-reliable low-latency communications in factory automation leveraging licensed and unlicensed bands," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 117–123, May 2019.

[54] S. Han, I. Chih-Lin, Z. Xu, and Q. Sun, "Energy efficiency and spectrum efficiency co-design: From NOMA to network NOMA," *IEEE COMSOC MMTC E-Letter*, vol. 9, no. 5, 2014.

[55] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, 2015.

[56] A. Host-Madsen and Junshan Zhang, "Capacity bounds and power allocation for wireless relay channels," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 2020–2040, June 2005.

[57] R. A. Howard, *Dynamic probabilistic systems: Markov models*. Courier Corporation, 2012, vol. 1.

[58] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3gpp lte: challenges and solutions," *IEEE Communications Magazine*, vol. 50, no. 2, pp. 156–162, 2012.

[59] H. Huang, J. Xiong, J. Yang, G. Gui, and H. Sari, "Jrate region analysis in a full-duplex-aided cooperative nonorthogonal multiple-access system," *IEEE Access*, vol. 5, pp. 17 869–17 880, Aug. 2017.

[60] T. E. Hunter and A. Nosratinia, "Cooperation diversity through coding," in *Proceedings IEEE International Symposium on Information Theory,*, June 2002, pp. 220–.

[61] A. Ikhlef, D. S. Michalopoulos, and R. Schober, "Buffers improve the performance of relay selection," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, Dec 2011, pp. 1–6.

[62] A. Ikhlef, D. S. Michalopoulos, and R. Schober, "Max-max relay selection for relays with buffers," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1124–1135, March 2012.

[63] V. Jamali, N. Zlatanov, and R. Schober, "Bidirectional buffer-aided relay networks with fixed rate transmission—part II: Delay-constrained case," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1339–1355, March 2015.

[64] Y. Jing and H. Jafarkhani, "Single and multiple relay selection schemes and their achievable diversity orders," *IEEE Transactions on Wireless Communications*, vol. 8, no. 3, pp. 1414–1423, March 2009.

[65] M. F. Kader and S. Y. Shin, "Coordinated direct and relay transmission using uplink NOMA," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 400–403, June 2018.

[66] J. Kim, J. Lee, J. Kim, and J. Yun, "M2M service platforms: Survey, issues, and enabling technologies," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 61–76, 2014.

[67] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, Sep. 2005.

[68] I. Krikidis, T. Charalambous, and J. S. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, May 2012.

[69] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, Dec 2004.

[70] J. N. Laneman, G. W. Wornell, and D. N. C. Tse, "An efficient protocol for realizing cooperative diversity in wireless networks," in *Proceedings. 2001 IEEE International Symposium on Information Theory (IEEE Cat. No.01CH37252)*, June 2001, pp. 294–.

[71] Y. Liang, V. V. Veeravalli, and H. Vincent Poor, "Resource allocation for wireless fading relay channels: Max-min solution," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3432–3453, Oct 2007.

[72] J. D. Little and S. C. Graves, "Little's law," in *Building intuition*. Springer, 2008, pp. 81–100.

[73] J. Liu, Y. Xu, and X. Jiang, "End-to-end delay in two hop relay MANETs with limited buffer," in *2014 Second International Symposium on Computing and Networking*, Dec 2014, pp. 151–156.

[74] M. Liu, T. Song, and G. Gui, "Deep cognitive perspective: Resource allocation for NOMA based heterogeneous IoT with imperfect SIC," *IEEE Internet of Things Journal*, pp. 1–1, 2018.

[75] Y. Liu, G. Pan, H. Zhang, and M. Song, "On the capacity comparison between MIMO-NOMA and MIMO-OMA," *IEEE Access*, vol. 4, pp. 2123–2129, 2016.

[76] A. Lozano and A. M. Tulino, "Capacity of multiple-transmit multiple-receive antenna architectures," *IEEE Transactions on Information Theory*, vol. 48, no. 12, pp. 3117–3128, 2002.

[77] K. Lu, Z. Wu, and X. Shao, "A survey of non-orthogonal multiple access for 5G," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Sep. 2017, pp. 1–5.

[78] X. Lu and R. C. d. Lamare, "Buffer-aided relay selection for physical-layer security in wireless networks," in *WSA 2015; 19th International ITG Workshop on Smart Antennas*, March 2015, pp. 1–5.

[79] S. Luo and K. C. Teh, "Buffer state based relay selection for buffer-aided cooperative relaying systems," *IEEE Trans. Wire. Commun.*, vol. 14, no. 10, pp. 5430–5439, Nov. 2015.

[80] S. Luo and K. C. Teh, "Adaptive transmission for cooperative NOMA system with buffer-aided relaying," *IEEE Communications Letters*, vol. 21, no. 4, pp. 937–940, April 2017.

[81] L. Lv, J. Chen, and Q. Ni, "Cooperative non-orthogonal multiple access in cognitive radio," *IEEE Communications Letters*, vol. 20, no. 10, pp. 2059–2062, Oct 2016.

[82] T. Lv, Y. Ma, J. Zeng, and P. T. Mathiopoulos, "Millimeter-Wave NOMA transmission in cellular M2M communications for internet of things," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1989–2000, June, 2018.

[83] B. Ma, H. Shah-Mansouri, and V. W. S. Wong, "Full-duplex relaying for d2d communication in millimeter wave-based 5g networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4417–4431, 2018.

[84] N. Marchenko, T. Andre, G. Brandner, W. Masood, and C. Bettstetter, "An experimental study of selective cooperative relaying in industrial wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 3, pp. 1806–1816, 2014.

[85] M. Marcus and B. Pattan, "Millimeter wave propagation: spectrum management implications," *IEEE Microwave Magazine*, vol. 6, no. 2, pp. 54–62, 2005.

[86] M. Medard and D. N. C. Tse, "Spreading in block-fading channels," in *Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems and Computers (Cat. No.00CH37154)*, vol. 2, 2000, pp. 1598–1602 vol.2.

[87] D. S. Michalopoulos and G. K. Karagiannidis, "Performance analysis of single relay selection in Rayleigh fading," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3718–3724, October 2008.

[88] P. K. Mishra, S. Pandey, and S. K. Biswash, "A device-centric scheme for relay selection in a dynamic network scenario for 5g communication," *IEEE Access*, vol. 4, pp. 3757–3768, 2016.

[89] E. M. Mohamed, B. M. Elhalawany, H. S. Khallaf, M. Zareei, A. Zeb, and M. A. Abdelghany, "Relay probing for millimeter wave multi-hop d2d networks," *IEEE Access*, vol. 8, pp. 30 560–30 574, 2020.

[90] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012, vol. 34.

[91] P. Monsen, "channel reuse in orthogonal multiple-access systems," in *2002 IEEE International Conference on Communications. Conference Proceedings. ICC 2002 (Cat. No. 02CH37333)*, vol. 1. IEEE, 2002, pp. 237–241.

[92] R. U. Nabar, H. Bolcskei, and F. W. Kneubuhler, "Fading relay channels: performance limits and space-time signal design," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 1099–1109, Aug 2004.

[93] R. Narasimhan, "Throughput-delay performance of half-duplex hybrid-ARQ relay channels," in *2008 IEEE International Conference on Communications*, May 2008, pp. 986–990.

[94] N. Nomikos, T. Charalambous, I. Krikidis, D. N. Skoutas, D. Vouyioukas, M. Johansson, and C. Skianis, "A survey on buffer-aided relay selection," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1073–1097, Secondquarter 2016.

[95] N. Nomikos, T. Charalambous, D. Vouyioukas, and G. K. Karagiannidis, "Low-complexity buffer-aided link selection with outdated CSI and feedback errors," *IEEE Transactions on Communications*, vol. 66, no. 8, pp. 3694–3706, Aug 2018.

[96] N. Nomikos, D. N. Skoutas, and P. Makris, "Relay selection in 5g networks," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2014, pp. 821–826.

[97] J. R. Norris, *Markov chains*. Cambridge University Press, 1998, no. 2.

[98] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 74–80, Oct 2004.

[99] M. Oiwa and S. Sugiura, "Reduced-packet-delay generalized buffer-aided relaying protocol: Simultaneous activation of multiple source-to-relay links," *IEEE Access*, vol. 4, pp. 3632–3646, 2016.

[100] M. Oiwa, C. Tosa, and S. Sugiura, "Theoretical analysis of hybrid buffer-aided cooperative protocol based on max-max and max-link relay selections," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 9236–9246, Nov 2016.

[101] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26–35, May 2014.

[102] L. P. Qian, A. Feng, Y. Huang, Y. Wu, B. Ji, and Z. Shi, "Optimal SIC ordering and computation resource allocation in MEC-aware NOMA NB-IOT networks," *IEEE Internet of Things Journal*, pp. 1–1, 2018.

[103] H. Ramazanali, A. Mesodiakaki, A. Vinel, and C. Verikoukis, "Survey of user association in 5g hetnets," in *2016 8th IEEE Latin-American Conference on Communications (LATINCOM)*, 2016, pp. 1–6.

[104] B. Rankov and A. Wittneben, "Spectral efficient protocols for half-duplex fading relay channels," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 379–389, February 2007.

[105] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile

communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[106] T. Riihonen, S. Werner, and R. Wichman, "Hybrid full-duplex/half-duplex relaying with transmit power adaptation," *IEEE Transactions on Wireless Communications*, vol. 10, no. 9, pp. 3074–3085, Sep. 2011.

[107] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of loopback self-interference in full-duplex MIMO relays," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5983–5993, Dec 2011.

[108] K. Saito, A. Benjebbour, A. Harada, Y. Kishiyama, and T. Nakamura, "Link-level performance of downlink NOMA with SIC receiver considering error vector magnitude," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[109] K. Saito, A. Benjebbour, Y. Kishiyama, Y. Okumura, and T. Nakamura, "Performance and design of SIC receiver for downlink NOMA with open-loop SU-MIMO," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 1161–1165.

[110] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2013, pp. 611–615.

[111] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance of downlink non-orthogonal multiple access (NOMA) under various environments," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[112] K. R. Santhi, V. K. Srivastava, G. SenthilKumaran, and A. Butare, "Goals of true broad band's wireless next wave (4g-5g)," in *2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No.03CH37484)*, vol. 4, 2003, pp. 2317–2321 Vol.4.

[113] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. part I. system description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, 2003.

[114] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.

[115] M. Shafi, J. Zhang, H. Tataria, A. F. Molisch, S. Sun, T. S. Rappaport, F. Tufvesson, S. Wu, and K. Kitao, "Microwave vs. Millimeter-Wave propagation channels: Key differences and impact on 5G cellular systems," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 14–20, December 2018.

[116] S. K. Sharma, I. Woungang, A. Anpalagan, and S. Chatzinotas, "Toward tactile internet in beyond 5g era: Recent advances, current issues, and future directions," *IEEE Access*, vol. 8, pp. 56 948–56 991, 2020.

[117] A. A. M. Siddig and M. F. M. Salleh, "Buffer-aided relay selection for cooperative relay networks with certain information rates and delay bounds," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 499–10 514, 2017.

[118] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.

[119] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, 2014.

[120] Z. Tian, G. Chen, Y. Gong, Z. Chen, and J. A. Chambers, "Buffer-aided max-link relay selection in amplify-and-forward cooperative networks," *IEEE Trans. Veh. Tech.*, vol. 64, no. 2, pp. 553–565, Feb. 2015.

[121] Z. Tian, Y. Gong, G. Chen, and J. A. Chambers, "Buffer-aided relay selection with reduced packet delay in cooperative networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 3, pp. 2567–2575, Mar. 2017.

[122] Z. Tian, "Buffer-aided cooperative networks," Ph.D. dissertation, Loughborough University, 2015.

[123] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[124] R. Urgaonkar and M. J. Neely, "Delay-limited cooperative communication with reliability constraints in wireless networks," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1869–1882, Mar. 2014.

[125] M. Vaezi, G. A. Aruma Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 900–919, Dec 2019.

[126] E. C. van der Meulen, "Three-terminal communication channels," 1971.

[127] B. Wang, K. Wang, Z. Lu, T. Xie, and J. Quan, "Comparison study of non-orthogonal multiple access schemes for 5G," in *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, June 2015, pp. 1–5.

[128] W. Wicke, N. Zlatanov, V. Jamali, and R. Schober, "Buffer-aided relaying with discrete transmission rates for the two-hop half-duplex relay network," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 967–981, Feb 2017.

[129] B. Xia, Y. Fan, J. Thompson, and H. V. Poor, "Buffering in a three-node relay network," *IEEE Transactions on Wireless Communications*, vol. 7, no. 11, pp. 4492–4496, November 2008.

[130] M. Xu, F. Ji, M. Wen, and W. Duan, "Novel receiver design for the cooperative relaying system with non-orthogonal multiple access," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1679–1682, Aug 2016.

[131] P. Xu, Z. Ding, I. Krikidis, and X. Dai, "Achieving optimal diversity gain in buffer-aided relay networks with small buffer size," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8788–8794, Oct 2016.

[132] P. Xu, J. Quan, Z. Yang, G. Chen, and Z. Ding, "Performance analysis of buffer-aided hybrid NOMA/OMA in cooperative uplink system," *IEEE Access*, vol. 7, pp. 168 759–168 773, 2019.

[133] C. Yang, J. Li, M. Guizani, A. Anpalagan, and M. Elkashlan, "Advanced spectrum sharing in 5G cognitive heterogeneous networks," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 94–101, 2016.

[134] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[135] Z. Yang, Z. Ding, Y. Wu, and P. Fan, "Novel relay selection strategies for cooperative NOMA," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 114–10 123, Nov 2017.

[136] A. Zafar, M. Shaqfeh, M. Alouini, and H. Alnuweiri, "Resource allocation for two source-destination pairs sharing a single relay with a buffer," *IEEE Transactions on Communications*, vol. 62, no. 5, pp. 1444–1457, May 2014.

[137] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, March 2016.

[138] P. Zhang, X. Yang, J. Chen, and Y. Huang, "A survey of testing for 5G: Solutions, opportunities, and challenges," *China Communications*, vol. 16, no. 1, pp. 69–85, Jan 2019.

[139] Q. Zhang, J. Jia, and J. Zhang, "Cooperative relay to improve diversity in cognitive radio networks," *IEEE Communications Magazine*, vol. 47, no. 2, pp. 111–117, February 2009.

[140] Q. Zhang, Z. Liang, Q. Li, and J. Qin, "Buffer-aided non-orthogonal multiple access relaying systems in rayleigh fading channels," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 95–106, Jan 2017.

[141] S. Zhang and V. K. N. Lau, "Multi-relay selection design and analysis for multi-stream cooperative communications," *IEEE Transactions on Wireless Communications*, vol. 10, no. 4, pp. 1082–1089, April 2011.

[142] J. Zhao, Z. Ding, P. Fan, Z. Yang, and G. K. Karagiannidis, "Dual relay selection for cooperative NOMA with distributed space time coding," *IEEE Access*, vol. 6, pp. 20 440–20 450, 2018.

[143] J. Zhao, Z. Ding, P. Fan, Z. Yang, and G. K. Karagiannidis, "Dual relay selection for cooperative NOMA with distributed space time coding," *IEEE Access*, vol. 6, pp. 20 440–20 450, April 2018.

[144] X. Zhao and W. Chen, "Non-orthogonal multiple access for delay-sensitive communications: A cross-layer approach," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 5053–5068, July 2019.

[145] B. Zhou, Y. Cui, and M. Tao, "Stochastic throughput optimization for two-hop systems with finite relay buffers," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5546–5560, Oct 2015.

[146] F. Zhou, Y. Wu, Y. Liang, Z. Li, Y. Wang, and K. Wong, "State of the art, taxonomy, and open issues on cognitive radio networks with NOMA," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 100–108, April 2018.

[147] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2744–2757, Dec 2017.

[148] N. Zlatanov, R. Schober, and P. Popovski, "Throughput and diversity gain of buffer-aided relaying," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, Dec 2011, pp. 1–6.

[149] N. Zlatanov, R. Schober, and P. Popovski, "Buffer-aided relaying with adaptive link selection," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1530–1542, Aug. 2013.

[150] N. Zlatanov, R. Schober, and P. Popovski, "Buffer-aided relaying with adaptive link selection-fixed and mixed rate transmission," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2816–2840, May. 2013.

[151] M. Zorzi and R. R. Rao, "Geographic random forwarding (GeRaF) for ad hoc and sensor networks: multihop performance," *IEEE Transactions on Mobile Computing*, vol. 2, no. 4, pp. 337–348, Oct 2003.