On Dual Connectivity in Next-Generation Heterogeneous Wireless Networks

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

by

Pradgnya Kiri Taksande
(Roll No. 123079003)

Under the guidance of:
Prof. Abhay Karandikar
and
Prof. Prasanna Chaporkar

Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai Mumbai 400076
2020

To my parents (Neeta and Devendra Kiri), husband (Niraj),
sister (Madhavi), in-laws (Maya and Gautam Taksande)

Thesis Approval

The thesis titled

On Dual Connectivity in Next-Generation Heterogeneous Wireless Networks

by

Pradgnya Kiri Taksande

(Roll No. 123079003)

is approved for the degree of

Doctor of Philosophy

Prof. Rajesh Hegde (Examiner)

Prof. Saravanan Vijayakumaran (Examiner)

Prof. Abhay Karandikar (Advisor)

Prof. Prasanna Chaporkar (Co-Advisor)

Prof. Indrajit Mukherjee (Chairman)

Date: 10/01/2020

Place: Mumbai

Declaration

I, Pradgnya Kiri Taksande, declare that this written submission represents my ideas in my own words and wherever others' ideas or words are included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/fact/data/source in my submission. I understand that any violation of the above will be cause for disciplinary action by IIT Bombay and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 10|01|2020

Place: Mumbai

Pradnya K.T.

Pradgnya Kiri Taksande

Abstract

In a heterogeneous wireless network, small cells exist alongside wide-coverage macro cells. Due to the difference in the coverage areas of small cells and macro cells, a large number of handovers takes place for a mobile User Equipment (UE) in a heterogeneous network. This leads to an increase in radio link failures and signaling overhead in the network. One of the approaches to overcome these challenges is the Dual Connectivity (DC) technique proposed by 3rd Generation Partnership Project (3GPP), in which a UE is connected to a macro cell and a small cell simultaneously. One of the nodes, which handles the control-plane of the UE is known as the Master Node (MN), and the other, which handles the user-plane of the UE is known as the Secondary Node (SN). As a result, a dual connected UE whose MN is a macro cell undergoes a handover only when it moves out of the coverage of the macro cell; thus, reducing the number of handovers and signaling overhead. SN is connected to MN via a non-ideal backhaul link, and SN may belong to the same Radio Access Technology (RAT) as the MN or to a different RAT.

In a heterogeneous network, dual connectivity can be used to improve the mobility robustness in the network. In DC, a UE can obtain resources from both MN and SN and thus, improve its throughput performance. A body of literature exists on dual connectivity in the areas of mobility robustness and throughput improvement. However, the performance of DC in other areas such as UE-perceived delay has not been studied. A new type of bearer called split bearer is introduced for dual connectivity in 3GPP standards. The backhaul link between MN and SN incurs an additional delay when the split bearer is configured for a dual connected UE. Due to this, different packets belonging to a single split bearer may reach the UE at different points in time. In this thesis, we investigate the issue of splitting traffic in DC to minimize the delay in the system. Further, there exists a trade-off between minimizing delay and blocking traffic, since limiting

traffic in the system leads to lower delays in the system. We address this trade-off by formulating an optimization problem to minimize the average delay in the system subject to a constraint on the blocking probability of arriving traffic using Markov Decision Process (MDP) framework. We obtain the optimal policy by solving this problem using Lagrangian approach and value iteration algorithm. We propose two heuristic policies, which have low computation and storage complexity as compared to the optimal policy. Moreover, the proposed policies achieve near-optimal performance.

The signaling procedures for DC, as defined by 3GPP in which the MN and SN, belong to the same RAT are different from the case when they belong to different RATs. Moreover, the interface between MN and SN is distinctly defined for scenarios where MN and SN belong to specific RATs. Further, a UE is allowed to dual connect when MN and SN belong to specific RATs only. The existing architecture is not favorable for dual connected UEs connected with base stations belonging to different RATs. To overcome these challenges, we propose SMRAN, a Software-Defined Networking (SDN) based multi-RAT Radio Access Network (RAN) architecture. This architecture brings uniformity and simplicity to the network interactions, allows flexibility in the network, and helps the network perform load balancing and mobility management effectively. This architecture seamlessly integrates multiple RATs for dual connectivity and defines a unified interface for communication between RAN and core network. The signaling procedures for dual connectivity between multiple RATs are standardized, and a common interface for communication between the multi-RAT nodes is proposed. Other features such as flexibility in processing the data-plane protocol layers at any node (MN or SN) are possible in SMRAN. Based on this architecture, we propose an algorithm for selecting UEs for dual connectivity. We demonstrate a reduction in the control signaling and an improvement in system performance using SMRAN as compared to the existing 3GPP architecture. Although SMRAN is proposed keeping in mind the challenges in multi-RAT dual connectivity, it is a generic architecture providing simplicity and flexibility.

The extension of DC to multiple connections is known as Multi-Connectivity (MC), in which a UE is connected to multiple base stations simultaneously. It is known that in the case of UEs with multiple connections, Proportional Fair (PF) scheduling maximizes the PF utility defined as the sum of the logarithm of long term throughput of UEs. A

UE, however, has a limited number of interfaces for communication with base stations. Moreover, the signaling overhead to maintain so many connections for a single UE is cumbersome to manage for a service provider. Hence, our objective is to improve the PF utility with the constraint that each UE can maintain a maximum of two connections. We propose PF-DC, a low complexity, modified PF scheduling scheme for dual connectivity based on SMRAN. One of the key challenges that arise in dual connectivity is the selection of two suitable BSs for UE association, taking into consideration the network load and channel conditions. We propose four UE association algorithms for dual connectivity. We demonstrate that the PF-DC scheme improves proportional fairness in the system compared to the standard PF scheduling scheme. We also demonstrate that as compared to single connectivity, dual connectivity provides a remarkable gain in the proportional fairness in the system.

To evaluate our work and its performance, we have developed a Network Simulator-3 (ns-3) based evaluation platform to design specific scenarios pertaining to each problem. We perform extensive simulations in ns-3 to compare the performance of our proposed algorithms with existing algorithms using practical network parameters. We demonstrate through these simulations that our proposed algorithms outperform the existing algorithms.

Contents

Acknowledgments

Undertaking PhD has truly been a life-changing experience for me and it would not have been possible without the support and guidance that I received from several people.

I would like to express my special appreciation and gratitude to my advisor, Prof. Abhay Karandikar, for believing in me and encouraging me to pursue Ph.D. Without his guidance and constant feedback, this PhD would not have been achievable. Since my first day, he believed in me like nobody else and gave me endless support. On an academic level, he taught me the fundamentals of conducting scientific research. Under his supervision, I learned how to define a research problem, find a solution to it, and finally publish the results. On a personal level, he inspired me by his hardworking and passionate attitude. I am honored to be a student under him.

I am grateful to my co-advisor, Prof. Prasanna Chaporkar, for encouraging my research and for allowing me to grow as a research scientist. I am very much thankful to him for accepting me as a student at a critical stage of my Ph.D. Under his guidance, I successfully overcame many difficulties and learned a lot.

I warmly thank Mr. Pranav Jha for his valuable advice, constructive criticism and extensive discussions around my work. I would also like to thank my research progress committee members, Prof. Gaurav Kasbekar and Prof. Saravanan Vijayakumaran, for their timely feedback and comments during my progress report presentations. I take this opportunity to thank all the professors of the Electrical Department from whom I have learned a lot, either through course lectures, seminars, or merely through one or two interactions.

I would like to thank my PhD colleagues at Infonet Lab - Arghyadip, Akshatha, Shashi, Anushree, Meghna, Indu and Sadaf, who have helped me throughout my journey as a doctoral candidate at IIT Bombay. Be it paper reviews or mock presentations or

any general technical discussions; they are the ones who built my thought process and cautioned me for any mistakes.

I appreciate the support and help of Sonal, Margaret, Beena, Sangeeta Ma'am, Aditya Sir and Rajesh Sir in all official works. I also appreciate the prompt response provided by department office staff, in particular, Santosh Sir and Madhu Ma'am.

Relinquishing a lucrative job as a software engineer and joining a doctoral program in India is not easy considering the social pressure. I thank my family for their belief in me and for providing unconditional support all through these years. I thank my parents, Mrs. Neeta Kiri and Mr. Devendra Kiri for giving me the best education and encouraging me throughout my childhood to make education the base of my life. I thank my sister Madhavi for her love and understanding. I very much appreciate my in-laws Mrs. Maya Taksande and Mr. Gautam Taksande, for their patience and support throughout the journey. I owe my deepest gratitude to my husband Niraj for his eternal support and understanding of my goals and aspirations and for being the wonderful person that he is. Without his support and encouragement, I would not have had the courage to embark on this journey in the first place. He was there for me whenever I needed it the most and has been patient with me whenever I did not have time for him. This thesis is dedicated to my wonderful family.

I would like to express my gratitude to my mentor Dr. Daisaku Ikeda, without whose encouragement and guidance I would not be the person that I am today.

January 2020 Pradgnya Kiri Taksande

List of Acronyms

3GPP	Third Generation Partnership Project
5G	Fifth Generation
API	Application Programming Interface
BS	Base Station
CDU	Centralized Data-plane Unit
CMDP	Constrained Markov Decision Process
CN	Core Network
DC	Dual Connectivity
DDU	Distributed Data-plane Unit
DP	Dynamic Programming
DRB	Data Radio Bearer
eNB	evolved NodeB
EPC	Evolved Packet Core
EPS	Evolved Packet System
ETU	Extended Typical Urban
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FIFO	First-In-First-Out
FP	Fixed Pico
FTP	File Transfer Protocol
gNB	next-Generation NodeB
GPF	Global Proportional Fairness
GPRS	General Packet Radio Services
GRE	Generic Routing Encapsulation
GTP	GPRS Tunneling Protocol

IP	Internet Protocol
JFI	Jain's Fairness Index
L1	Layer 1
L2	Layer 2
LM	Lagrange Multiplier
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
LWA	LTE-WLAN Aggregation
LWAAP	LTE-WLAN Aggregation Adaptation Protocol
MAC	Medium Access Control
MC	Multi-Connectivity
MCG	Master Cell Group
MDP	Markov Decision Process
MeNB	Macro eNB
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MN	Master Node
MR-DC	Multi-RAT Dual Connectivity
NE-DC	NR E-UTRA DC
NG-C	Next Generation Control plane
NG-U	Next Generation User plane
NGEN-DC	NGRAN E-UTRA NR-DC
NGRAN	Next Generation Radio Access Network
NR	New Radio
NR-DC	NR-NR DC
OF	OpenFlow
PBCH	Physical Broadcast Channel
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit

PF	Proportional Fairness
PGW	Packet Data Network Gateway
PHY	Physical Layer
QoS	Quality of Service
RAB	Radio Access Bearer
RAN	Radio Access Network
RAT	Radio Access Technology
RLC	Radio Link Control
RP	Random Pico
RRC	Radio Resource Control
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
S1AP	S1 Application Protocol
SA	Stochastic Approximation
SC	Single Connectivity
SCG	Secondary Cell Group
SCTP	Stream Transmission Control Protocol
SDAP	Service Data Adaptation Protocol
SDN	Software-defined Networking
SeNB	Small cell eNB
SGA	Stochastic Gradient Algorithm
SGW	Serving Gateway
SN	Secondary Node
SNR	Signal to Noise Ratio
SRB	Signaling Radio Bearer
SRC	SDN RAN Controller
TCP	Transmission Control Protocol
TMSI	Temporary Mobile Subscriber Identity
UDP	User Datagram Protocol
UE	User Equipment

UMTS	Universal Mobile Telecommunications System
VIA	Value Iteration Algorithm
VoIP	Voice over Internet Protocol
WLAN	Wireless Local Area Network
WT	WLAN Termination
X2-AP	X2 Application Protocol

List of Symbols

α_i	Probability of batch size $G = i$
β	Lagrangian multiplier
B	Queue size expressed in number of packets
B_{max}	Constraint on blocking probability
δ	Weight on blocking probability of background traffic
G	Batch size
$\bar{G}$	Mean batch size
λ_1	Average arrival rate of batches for foreground users
λ_2	Average arrival rate of batches for background users
μ_m	Average service rate of a packet in MeNB subsystem
μ_d	Average service rate of a packet in backhaul subsystem
μ_s	Average service rate of a packet in SeNB subsystem
n_1	Number of resources in MeNB subsystem
n_2	Number of resources in SeNB subsystem
N_1	Capacity of MeNB subsystem expressed in number of packets
N_2	Total capacity of backhaul and SeNB subsystems expressed in number of packets

List of Tables

List of Figures

Chapter 1

Introduction

There has been an unprecedented rise in the demand for mobile data traffic in the recent past, due to proliferation and ease of accessibility of smart handheld devices. According to an estimate, the growth in mobile data traffic has intensified, and it has grown 17-fold in the past five years [**?**]. In the future also, it is expected that mobile data traffic will expand at a rate of 46% annually from 2017 to 2022 and by the year 2022, it is projected to reach 77.5 exabytes per month [**?**]. At the same time, the density of users is not uniform across regions. For instance, there are certain regions with high user density, known as hotspot areas. Further, data traffic is bursty in nature, and traffic demands differ in the requirement of Quality of Service (QoS). This leads to spatio-temporal variations in the data traffic patterns. The primary challenge for the next generation cellular systems lies in meeting these varying demands of mobile data traffic.

The service providers are needed to satisfy diverse traffic demands, along with providing coverage at reasonable costs. However, the spectrum available for mobile communication is limited, and it is an expensive resource. This makes it even more challenging to meet the increasing traffic demands. The need of the hour is to improve the spectral efficiency while providing coverage at reasonable costs. Hence, it is imperative to reuse the available spectrum spatially. For this purpose, low power base stations (small cells) are overlaid on the homogeneous macro cell network. The deployment of small cells overlaid on a homogeneous network is known as a Heterogeneous Network (HetNet) as depicted in Figure **??**. In such a network, macro cells provide coverage, whereas small cells serve hotspot areas and enable enhanced QoS. The macro and small cell nodes can

belong to different Radio Access Technologies (RATs) such as Fourth Generation (4G), Fifth Generation (5G), or Wireless Local Area Network (WLAN).

The 4G and 5G cellular systems have been introduced to provide higher data rates along with high spectral efficiency within the limited available spectrum. 3rd Generation Partnership Project (3GPP) is a standardization body that develops and maintains mobile telephony standards [**?**]. Long Term Evolution (LTE) of Universal Mobile Telecommunications System (UMTS) and LTE-Advanced (LTE-A) standards have been initiated by 3GPP as a part of 4G cellular systems. The standardization of LTE was performed as a part of 3GPP Release 8 and 9. LTE-A is standardized as 3GPP Release 10 onwards. 3GPP Release 15 and 16 are being standardized as 5G. 5G services are required to provide higher throughputs, lower latencies, ultra-high reliability along with low cost and low power consumption. 4G mobile broadband access, along with the New Radio (NR) access technology, will provide 5G wireless access [**?**]. In LTE terminology, a base station is known as an evolved NodeB (eNodeB or eNB). In 5G, a base station is known as a next-Generation NodeB (gNB), and it provides New Radio (NR) access to users on the air interface.WLAN small cells use unlicensed spectrum for their operation. Due to their ease of installation and low hardware cost, WLAN access points are commonly deployed in commercial offices, homes, and hotspots.

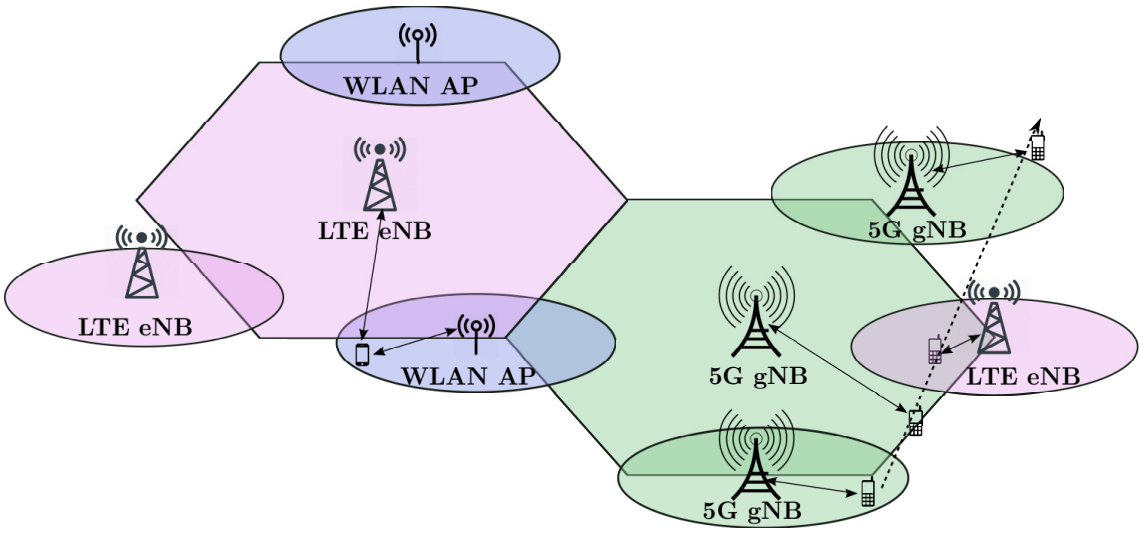

Figure 1.1: Heterogeneous network.

In a HetNet, as illustrated in Figure **??**, consider a mobile User Equipment (UE) following a trajectory shown by the dashed line. The signal strength received from differ-

ent nodes at the UE varies due to its movement. As a result, it undergoes an increased number of handovers due to the presence of small cells in the HetNet than that in a homogeneous network. The increase in the number of handovers leads to increased signaling between the Core Network (CN) and Radio Access Network (RAN). To overcome these mobility challenges in HetNets, the technique of Dual Connectivity (DC) was introduced by 3GPP as a part of LTE-A Release 12 [**?**]. In this technique, a UE is connected to a macro cell as well as a small cell simultaneously. The control-plane functions of the UE are handled by the macro cell, while the user-plane functions can be handled either by the macro cell or small cell or both. In this case, the UE undergoes a handover only when it moves out of the coverage area of the macro cell. This reduces the number of handovers from small cells to macro cells and vice versa, making the number of handovers in HetNets comparable to that in a homogeneous network. Thus, DC provides mobility robustness in the system. A dual connected UE receives user-plane data from both the macro cell and small cell, thus improving its throughput. In this thesis, we explore and analyze the various benefits of dual connectivity.

We consider a scenario, as illustrated in Figure **??** for our analysis. In this chapter, we provide a background on the dual connectivity technique. In Section **??**, we discuss LTE network architecture. We provide an overview of different types of DC and their architectures in Section **??**. Section **??** discusses the challenges in dual connectivity. Software-Defined Networking (SDN) is an emerging paradigm for network management and control. In this thesis, we use the SDN paradigm to design a multi-RAT network. We present an overview of SDN in Section **??**. The motivation behind the thesis is discussed in Section **??**. In Section **??**, we highlight the contributions and organization of the thesis.

1.1 Overview of Long Term Evolution

LTE is the first cellular communication standard, which provides packet-switched data as well as voice services [**?**]. As compared to the previous standards, LTE provides high data rates as well as extensive coverage. It also enables seamless mobility and flexible bandwidth deployments. We outline the network architecture and resource structure of LTE in subsequent subsections.

1.1.1 LTE Network Architecture

LTE network called Evolved Packet System (EPS) comprises two parts: Evolved Packet Core (EPC) or Core Network (CN) and Evolved UMTS Terrestrial Radio Access Network (E-UTRAN). It provides an all-Internet Protocol (IP) end-to-end connectivity for UEs. E-UTRAN deals with radio access to the end-users, and CN provides access to the Internet as well as user control functions. E-UTRAN is made up of essentially one type of node, eNB, while the CN consists mainly of Packet data network GateWay (PGW), Serving GateWay (SGW) and Mobility Management Entity (MME). Each of these network nodes is interconnected with each other through standardized interfaces. Figure **??** illustrates LTE network architecture along with RAN protocol stack. E-UTRAN (or simply RAN) consists of eNBs and the radio interface to transmit/receive signals and data to/from UEs.

PGW acts as a mobility anchor for handovers between 3GPP and non-3GPP technologies. It performs policy and QoS enforcement, packet filtering and charging as well as assigns IP addresses to UEs and connects UEs to the Internet. The inter-working with other 3GPP technologies, e.g., 3G, takes place via the SGW, which acts as a mobility anchor. All the user IP packets are transferred through SGW. MME is the main control node for signaling between the CN and UE. It manages connections and authentication for users as well as their bearers.

The eNBs are connected to the EPC via an S1 interface, particularly to the MME via an S1-MME interface and to the SGW via an S1-U interface, while eNBs are connected among themselves via an X2 interface. Each eNB handles Radio Resource Management (RRM) for the respective UEs associated with it, e.g., transmit power for a UE, resources to be allocated to a UE, frequency of operation, etc. Radio Resource Control (RRC) layer of an eNB performs control signaling towards the MME via S1-MME interface and towards other eNBs via X2 interface (X2-C). The information about load and interference management is shared via the X2 interface for load balancing and interference coordination in the network. This control signaling takes place through X2 Application Protocol (X2-AP), Stream Control Transmission Protocol (SCTP), IP, Layer 2 (L2) and Layer 1 (L1) as depicted in the figure. Similarly, the control signaling with MME takes place over the S1-MME interface using S1 Application Protocol (S1-AP), SCTP, IP, L2, and L1.

RRC layer of eNB is responsible for configuring all the data-plane layers and the

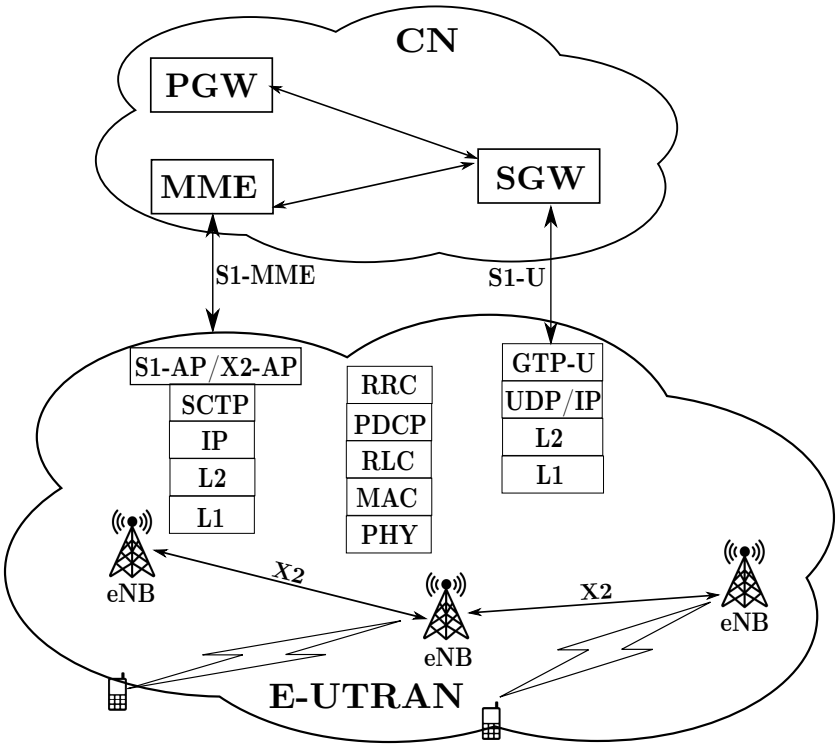

Figure 1.2: LTE network architecture.

establishment of Signaling Radio Bearers (SRBs) [**?**]. It is also responsible for initial security activation, the establishment of Data Radio Bearers (DRBs), and handovers. The establishment, modification, and release of an RRC connection are also the functions of the RRC layer. Further, configuration and activation of measurements by UE as well as an exchange of system information and UE capability information with CN are functions of the RRC layer.

The data transfer from CN to RAN takes place through S1-U interface between SGW and eNBs using the protocol layers GPRS Tunneling Protocol for user-plane (GTP-U), User Datagram Protocol (UDP), IP, L2 and L1 as illustrated in the figure. The data transfer from eNB to UE takes place via the radio interface using Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), Medium Access Control (MAC) and Physical (PHY) protocol stack.

The PDCP layer performs ciphering and IP header compression. It is also responsible for integrity protection and in-sequence delivery of packets for handovers. Concatenation/segmentation, handling of retransmissions, and in-sequence delivery of packets to higher layers are the functions of the RLC layer. MAC layer is responsible for multiplex-

ing of logical channels, uplink and downlink scheduling, and retransmission of packets. The physical layer is responsible for coding/decoding, modulating/demodulating, and transmission of data over the air interface.

In the next section, we take a look at the resource structure in LTE.

1.1.2 Resource Structure in LTE

LTE uses Orthogonal Frequency Division Multiple Access (OFDMA) technique at MAC layer for allocating resources to multiple users in the downlink. OFDMA uses Orthogonal Frequency Division Multiplexing (OFDM) technique at the physical layer. OFDM divides the available bandwidth into multiple orthogonal subcarriers. These subcarriers, either individually or in groups, can be used to carry independent data streams by using different modulation schemes. These multiple orthogonal subcarriers are then allocated to multiple users. Thus, OFDM provides a high degree of robustness against channel frequency selectivity.

In the time domain, LTE transmissions are divided into radio frames. Each radio frame is 10 ms long. A radio frame is subdivided into ten equal-sized subframes of 1 ms duration. Each subframe is composed of two slots of length 0.5 ms with each slot consisting of 7 OFDM symbols. The smallest unit of resource that can be assigned to a UE in LTE is known as a Physical Resource Block (PRB). Each PRB has a bandwidth of 180 kHz, divided into subcarriers of 15 kHz each. Figure **??** illustrates the PRB structure. N_{RB}^{DL} represents the total number of PRBs in a given bandwidth.

Each PRB consists of 12 subcarriers with 7 symbols on each subcarrier. Thus, every PRB has 84 OFDM symbols which can be used to transmit data based on the modulation scheme used. The smallest resource structure in LTE is the resource element, which consists of one 15 kHz subcarrier for a duration of one symbol. Thus, 84 resource elements make up one PRB. Although there are 84 symbols available per PRB for transmission in every radio subframe, all of these symbols cannot be used to transmit user data. There are various types of control channels in LTE such as Physical Downlink Control Channel (PDCCH), Physical Broadcast Channel (PBCH), the control signals of which are transmitted on some of these symbols.

In each subframe, the user data is transmitted over a transport block of dynamic

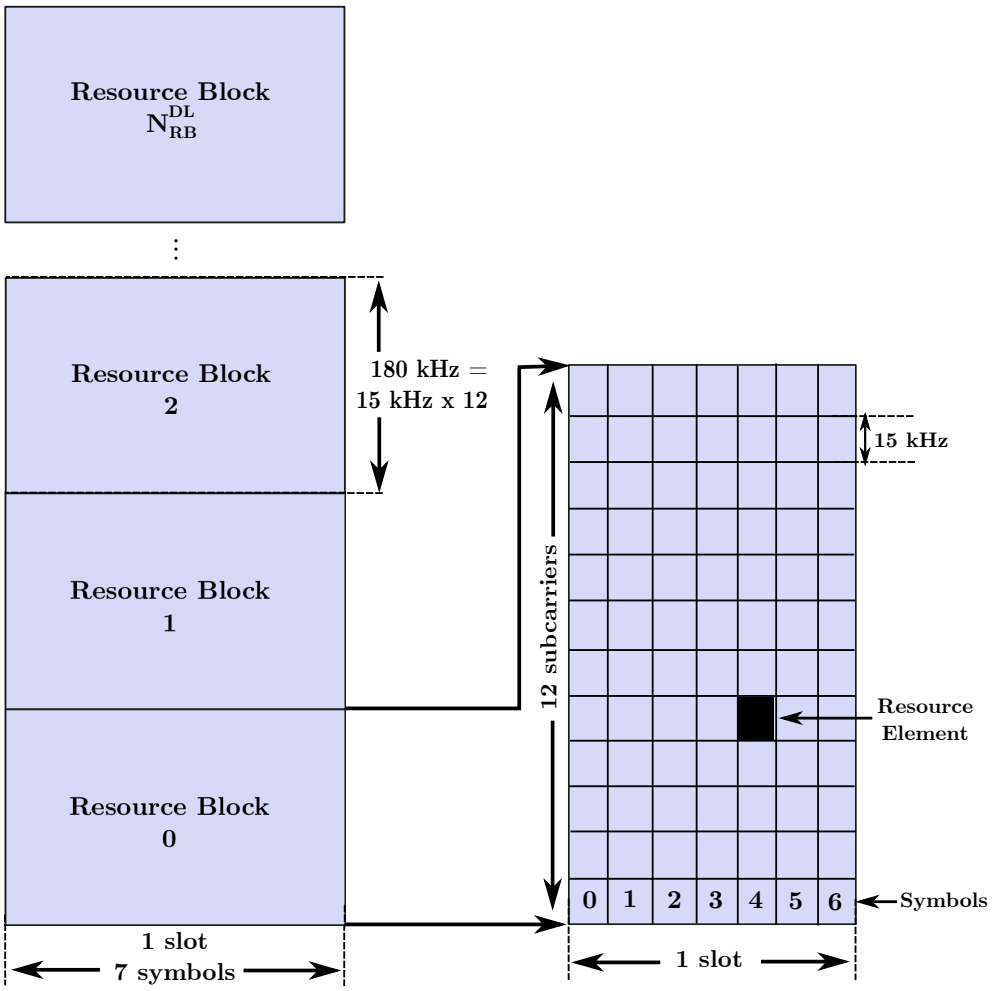

Figure 1.3: Resource structure in LTE.

size on the Physical Downlink Shared Channel (PDSCH). A transport block is a MAC Protocol Data Unit (PDU). Transport blocks are passed down from the MAC layer to the PHY layer once per subframe duration. The number of bits that are to be transmitted per symbol is determined by the modulation order. The adaptive modulation and coding scheme in LTE decides the modulation order. Thus the modulation and the coding scheme used by the physical layer ultimately decides the transport block size a UE can receive in each PRB.

1.2 Introduction to Dual Connectivity

Dual Connectivity is defined as an "operation where a UE uses the radio resources provided by at least two different network points connected with a non-ideal backhaul link

while in RRC_CONNECTED state" [?]. Figure ?? illustrates the dual connectivity functionality. A UE with multiple transceivers using radio resources of two base stations simultaneously is called a dual connected UE. One of the base stations is called Master Node (MN) and the other Secondary Node (SN). SN is connected to the MN via a non-ideal backhaul link. This backhaul link has a finite capacity and incurs a delay of the order of milliseconds during transmission [?]. MN, which is typically a macro base station, handles the control-plane functionality of the UE, while the user-plane functionality can be handled by either MN or SN, or both. MN may be a single cell or a group of cells called Master Cell Group (MCG). Similarly, the group of cells associated with SN is called Secondary Cell Group (SCG). Release 12 UEs are equipped with multiple radio transceivers, and they can be served by two base stations simultaneously.

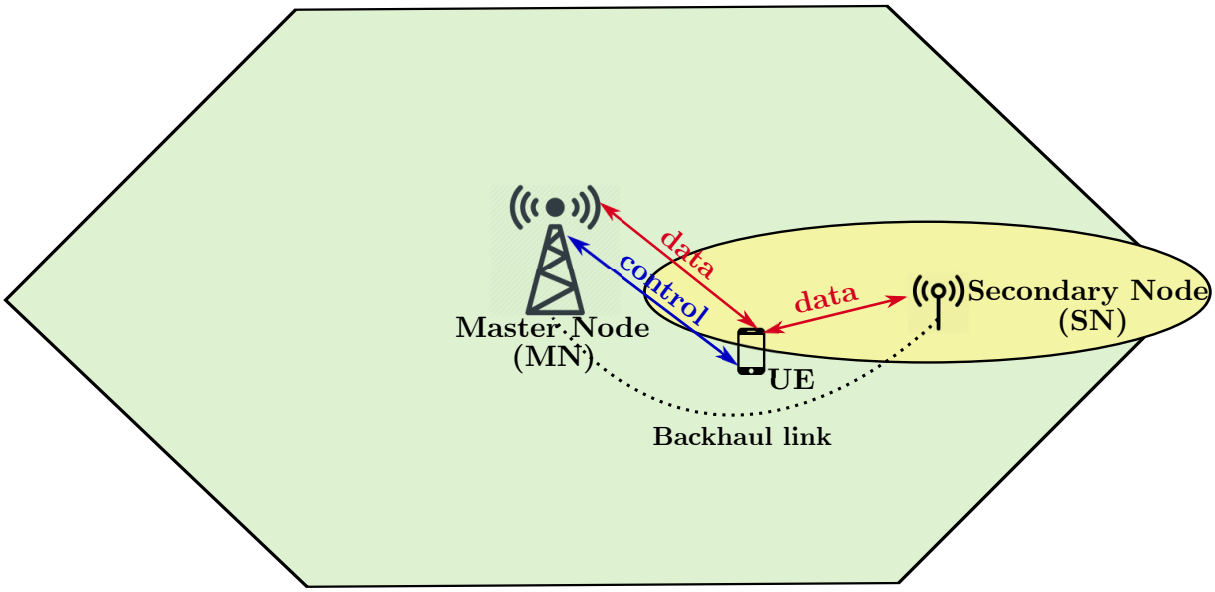

Figure 1.4: Dual connectivity.

The design goals for which 3GPP had introduced dual connectivity are as follows [?]:

- Providing mobility robustness: DC prevents the increase in the number of handovers in a HetNet, thus making the mobility performance comparable to a homogeneous network. It prevents the increase in handover failures and radio link failures in HetNets.

- Preventing increased signaling load due to frequent handovers: A large number of handovers leads to increased signaling load between the RAN and CN. DC prevents

this increase since it limits the handovers to macro-only.

- Improving per-user throughput and system capacity: A dual connected UE uses resources of two base stations simultaneously, thus improving its throughput and QoS.

Varied modes of dual connectivity have been defined by 3GPP based on the radio access technologies of master node and secondary node. Dual connectivity between master node and secondary node belonging to LTE technology is defined in [**?**], dual connectivity between master node belonging to LTE and secondary node belonging to WLAN technology is characterized in [**?**] as LTE-WLAN Aggregation (LWA), and dual connectivity between master node and secondary node belonging to either 5G or LTE technology is known as Multi-RAT Dual Connectivity (MR-DC) [**?**]. We elaborate on each of these in the following subsections.

1.2.1 Dual Connectivity in LTE

In this case, a dual connected UE is connected to a Macro eNB (master node) and a Small cell eNB (secondary node) simultaneously. Macro eNB (MeNB) handles the control-plane, and Small cell eNB (SeNB) and MeNB manage the user-plane for the UE. MeNB and SeNB can operate at same or different carrier frequencies. SeNB is connected to the MeNB via X2 interface, which is a non-ideal backhaul link. The non-ideal link has limited capacity and higher latency as compared to an ideal link [**?**]. The radio protocol architecture is explained next.

In this form of DC, only MeNB communicates with CN and exchanges the control messages with the UE. For signaling information exchange between MeNB and SeNB, the X2 interface is used. There are three types of bearers, viz., MCG, SCG, and split bearers [**?**] as depicted in Figure **??**. For MCG and split bearers, S1-U interface is terminated at the MeNB, whereas for SCG bearers it terminates at the SeNB. MCG bearer uses radio resources of the MeNB only, whereas SCG bearer uses radio resources of the SeNB only. In the case of the split bearer, data units are split at the PDCP layer and routed to the SeNB via X2 interface. Split bearer, thus, uses radio resources of both MeNB and SeNB. Data radio bearers can be configured as any of the three types of bearers. Signaling radio

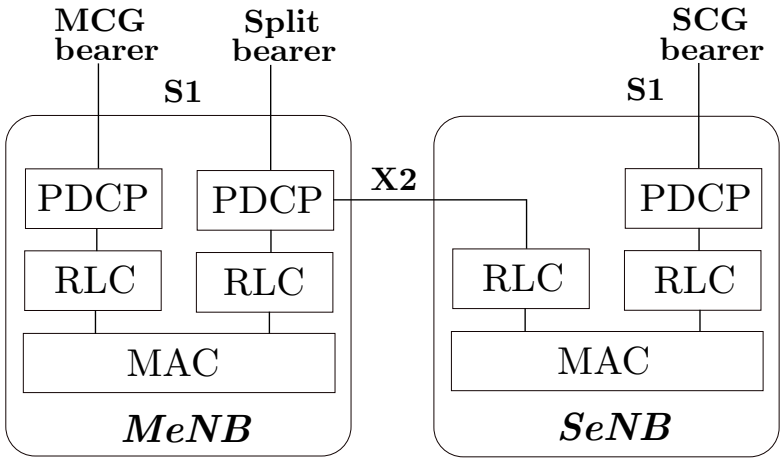

Figure 1.5: Radio protocol architecture for dual connectivity [**?**].

bearers, however, can be configured as MCG bearer type only.

The user-plane connectivity of a UE depends on the type of bearer configured. There are two different user-plane architectures depending upon node at which S1-U is terminated [**?**]. The two architectures depend on which bearer types are configured. These two architectures are explained next.

Architecture 1A

In this architecture, CN segregates the traffic according to the radio bearers and decides to route each radio bearer through the MeNB or the SeNB. Thus, each eNB serves different radio bearers as instructed by the CN. The complexity of splitting the connections lies at the CN. In this architecture, however, SeNB needs to be connected to the CN via a fiber-based backhaul, which is not practically feasible if the number of small cells in the network is significant. Further, CN has to take care of control signaling for both types of eNBs. As shown in Figure **??**, two different radio bearers are routed through MeNB and SeNB, which then simultaneously reach the UE. The SeNB has the same PDCP, RLC and MAC layer functionalities as those of MeNB.

Architecture 3C

In this architecture, the radio bearer is split at the PDCP layer of MeNB. This is also known as split bearer architecture since a single bearer is split into two parts. In this case, only MeNB has a user-plane connection with the CN, and user-plane data exchange

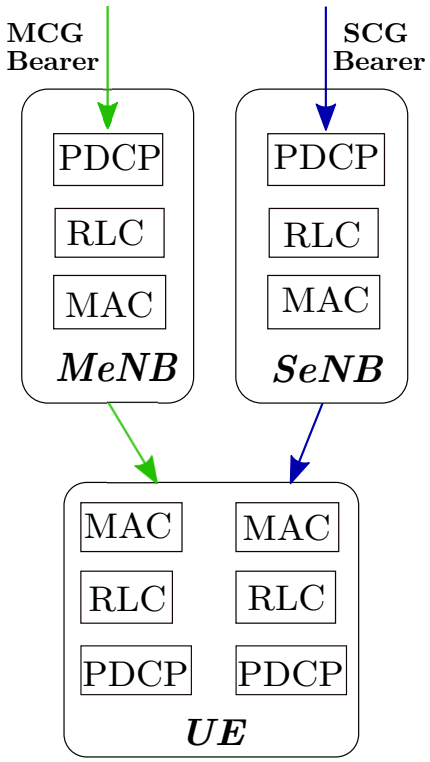

Figure 1.6: Architecture 1A user-plane protocol stack at eNB and UE.

takes place between MeNB and SeNB via the backhaul link. The backhaul link is used by MeNB to share control signals and data traffic with SeNB. SeNB uses the backhaul link to share its status information with the MeNB. MeNB handles the control signaling of dual connected UE with the CN and the complexity of controlling the data traffic and splitting lies at MeNB. Hence, signaling at the CN does not increase as in 1A architecture; however, control information needs to be shared among the two eNBs through the backhaul link. Since the splitting takes place at MeNB, better and faster inter-eNB load balancing can be achieved.

Figure **??** depicts the user-plane protocol stacks at MeNB, SeNB, and UE for the split bearer architecture. MeNB splits the user-plane data at the PDCP layer, part of which is assigned to the RLC layer of SeNB to send to the UE and the remaining part is assigned to the RLC layer of MeNB. MeNB and SeNB perform independent scheduling of the data arriving at their respective MAC layers. At the PDCP layer of UE, reordering of packets received from two independent RLC entities is performed.

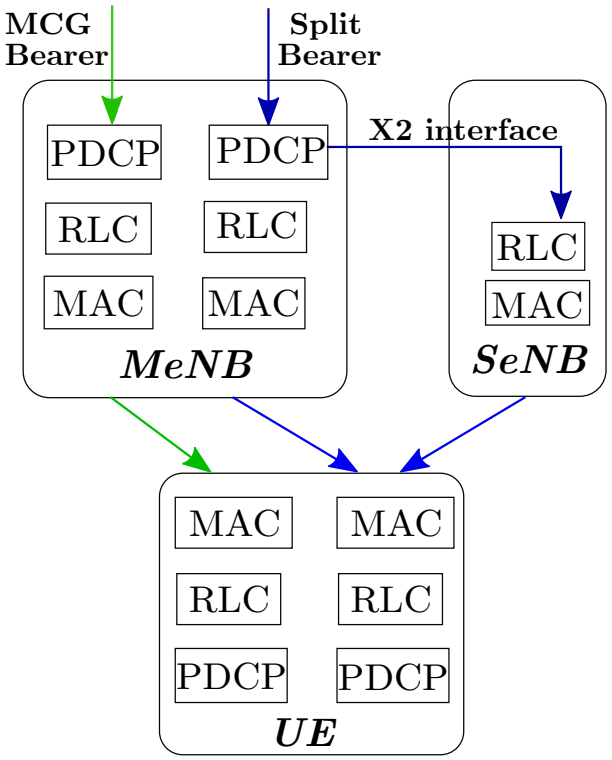

Figure 1.7: Architecture 3C user-plane protocol stacks at eNB and UE.

1.2.2 LTE-WLAN Aggregation (LWA)

Similar to dual connectivity in LTE, in 3GPP Release 13 and Release 14, radio level aggregation of traffic over LTE and WLAN has been introduced as LWA so that the resources provided by the individual systems can be used by a UE [**?**]. Figure **??** illustrates the user-plane and control-plane signaling in LWA. To take the role of SN, 3GPP has specified a logical node called WLAN Termination (WT). WT is connected to the eNB through a specific interface called Xw interface. For UEs configured with LWA, the control-plane connection can take place only through the eNB, whereas the data can be routed via a split LWA bearer using both eNB and WT. There is a switched LWA bearer, which uses radio resources of WT only, but whose radio protocols are located in both WT and eNB.

SRBs can be configured through eNB only (S1-MME interface), and control information is then shared between eNB and WT using Xw-Control plane (Xw-C) interface [**?**]. S1-U terminates in the eNB for both split and switched bearers. For split LWA bearer, data is split at the PDCP layer of eNB and routed to WT via Xw-User plane (Xw-U) interface. LWA Adaptation Protocol (LWAAP) [**?**] is used to transfer PDUs between eNB

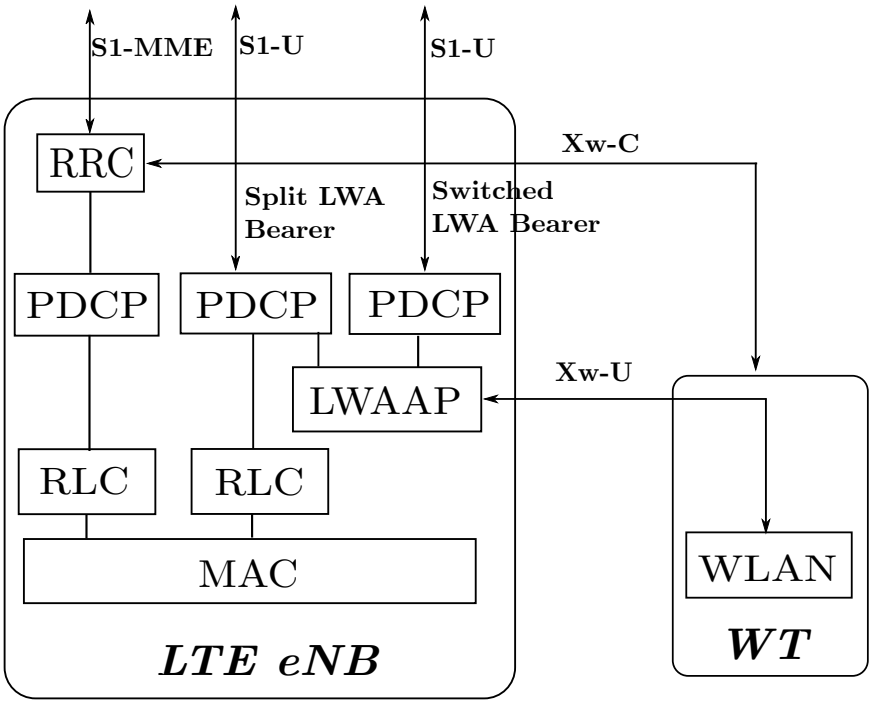

Figure 1.8: LWA architecture.

and WT. The working of LWA is similar to that of Release 12 LTE dual connectivity. LWA plays a valuable part in integrating WLAN into 3GPP standards.

1.2.3 Multi-RAT Dual Connectivity (MR-DC)

Multi-RAT Dual Connectivity was introduced by 3GPP as a part of LTE Release 15 [**?**]. Multi-RAT Dual Connectivity [**?**] is a generalization of dual connectivity explained in Section **??**, where the core network may be either evolved packet core (LTE Core) or 5G Core (5GC). A UE with multiple transceivers may be configured to utilize resources provided by LTE eNB and 5G gNB simultaneously. In this case, the master node and secondary node can be either eNB or gNB. The master and secondary nodes are connected via a network interface Xn and at least the master node is connected to the core network. Based on the type of node acting as master and secondary, MR-DC is classified into three types when the core network belongs to the 5G technology. (i) Next Generation RAN (NGRAN) E-UTRA New Radio dual connectivity (NGEN-DC) is a type of MR-DC where eNB acts as the master node and gNB acts as the secondary node. (ii) When gNB acts as master node and an eNB acts as secondary node, it is called NR E-UTRA

dual connectivity (NE-DC). (iii) NR-NR Dual Connectivity (NR-DC) is a type of MR-DC where both master and secondary nodes are gNBs.

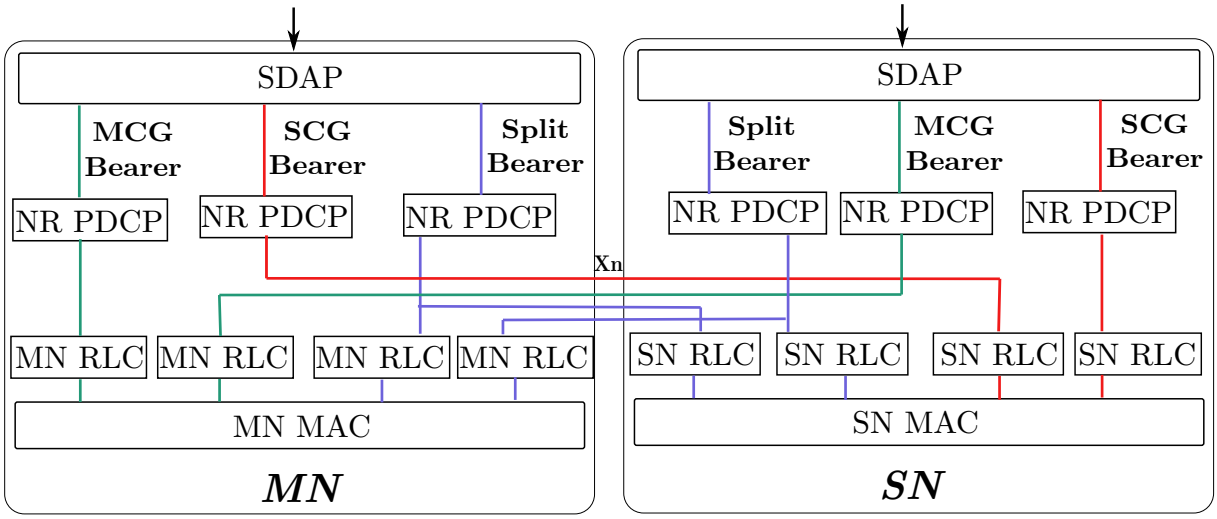

Figure 1.9: MR-DC architecture.

Figure **??** illustrates the bearer set-up in MR-DC (NGEN-DC, NE-DC, and NR-DC). In this architecture, there are three types of bearers, viz., MCG bearer, SCG bearer, and split bearer, as explained in Section **??**. However, in this architecture, each of the three bearer types can be terminated at either MN or SN. Data can be transported using any of these bearer types. SRBs can, however, be configured in many ways. In MR-DC, there is a new signaling radio bearer SRB3, which is established between SN and UE. Besides, there is a split SRB, which is an SRB between the MN and the UE, allowing the duplication of control information via both MN and SN. MN always sends the initial SN RRC configuration via MCG SRB (SRB1) but succeeding reconfigurations may be carried via master or secondary node. Moreover, irrespective of the technology of master and secondary nodes, NR PDCP is used for all bearer types.

1.3 Challenges in Dual Connectivity

In this section, we present the challenges arising in HetNets due to dual connectivity. A HetNet consists of multiple nodes belonging to different RATs and having varying coverages. In such a network, one of the interesting problems is the association of users with single or dual connectivity. The association of users with one or more base stations

depends on several factors such as signal strength received from different base stations, load at different base stations, coverage, etc. User association is an important topic because the performance of the network depends on it. An appropriate procedure should be in place for user association.

In a HetNet, the dual connected users are present alongside users with single connectivity. While scheduling, dual connected users may be allocated more resources as compared to single connected users, and hence, there may be a considerable difference between the throughputs of these two sets of users. The presence of dual connected users may then detrimental to the performance of single connected users. Hence, the fairness of all users needs to be taken into account while scheduling.

In the case of 1A architecture for dual connectivity in LTE, the two bearers take two distinct paths to reach the dual connected UE. However, in the case of 3C architecture, there exists a split bearer, which uses the resources in both MN and SN. The packets belonging to the split bearer can take two paths, either via MN or via SN. An appropriate path can be chosen to achieve a global objective, for instance, to achieve load balancing in the system or for maximizing a system performance metric. One of the interesting problems is how to split the traffic appropriately between the two paths to achieve a certain objective.

As explained in the previous section, the different modes of DC have similar bearers, viz., MCG, SCG, and split bearers. However, the signaling procedures in these modes of DC are inconsistent. In LTE dual connectivity, there is only one signaling bearer, and it is served by the MN only. When SN has to share radio control information with a dual connected UE, it needs to send this information to the MN over X2 interface, which in turn sends it to UE. Further, SN has to get the data bearer related control information from the MN through the X2 interface. This leads to an exchange of a large number of signaling messages between MN and SN for the dual connected UE. Moreover, this exchange takes place over a non-ideal backhaul link (X2 interface) between MN and SN, which has a latency of the order of milliseconds [**?**] leading to additional delay in the signaling exchange. In LWA, there is only one signaling bearer between CN and eNB, and all control information exchange with the dual connected UE takes place via eNB. The control information is exchanged between WT and eNB via the Xw-C interface. In MR-

DC, there are three types of signaling radio bearers, viz., MCG, SCG, and split bearers. The initial signaling, however, can take place via an MCG bearer only. Further, in this case, the control signaling exchange between MN and SN takes place via the Xn interface. Thus, there is no uniformity in the existing architecture for UEs dual connected to nodes belonging to different RATs.

In a HetNet, base stations may belong to different RATs. In such a scenario, UEs can be dual connected to base stations belonging to different RATs. Macro base stations can act as the master nodes for all UEs. The question arises which base station to select as the secondary node for UEs. In the existing architecture, macro base station chooses the small cell from multiple potential candidate cells based on factors such as signal strength. However, the decision taken by macro base stations are based on a local view of the network, and they may not be globally optimal. If there is a centralized entity at which all network information is available, then it will have a global view of the network, and optimal performance can be achieved. Additionally, in the existing architecture, there is no provision for a dual connected UE to connect with a WT and gNB simultaneously. If there is a centralized entity to manage the multi-RAT nodes, then this problem may be solved. Software-defined networking is an emerging paradigm for network control and management in which the network can be centrally controlled. If the SDN framework can be applied to such a HetNet, the problems in exiting architecture can be resolved. We present an overview of SDN in the next section.

1.4 Overview of Software-Defined Networking

Software-Defined Networking refers to an emerging framework for network management, which centralizes control of the network [**?**]. It separates the control (logic) plane from the forwarding (routing) plane. It mainly consists of an application layer, a control layer, and an infrastructure layer, as illustrated in Figure **??**. The infrastructure layer consists of network devices which accept instructions from the controller via an interface southbound from the control layer. This interface is called the control-data plane or southbound interface. Control layer consists of a logically centralized controller which provides an abstract view of the network to the application layer. Application plane

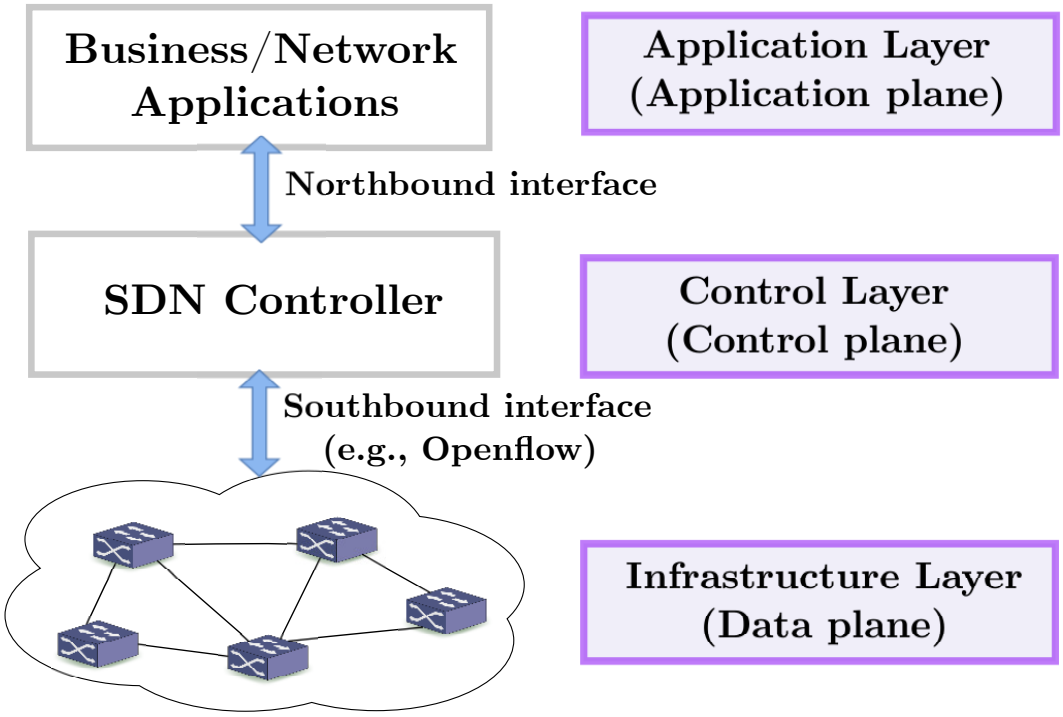

Figure 1.10: Software-defined networking.

comprises business or network applications which convey their requirements and network characteristics to the controller via an interface northbound to the controller.

In SDN, network intelligence (control-plane) is segregated from the network operation (forwarding/data plane), which leads to flexible centralized networks. SDN allows a vendor-independent control over the network from a single controller, thus simplifying the network design. The network devices are also simplified since now there is no need for them to process different protocol standards but merely accept instructions from a central control entity. OpenFlow [**?**] is an open communications standard defined for the southbound interface. Granular control of OpenFlow's flow-based control model allows the network to respond to real-time changes at the device, session, and application levels. In addition, open Application Programming Interfaces (APIs) can be used for implementing the northbound interface. Open APIs can be used between the control and applications layers to enable applications to operate on an abstraction of the network, thus provisioning flexibility and programmability in the network. SDN, thus, provides a unified and logically centralized network design, which is scalable, flexible, and practical.

In the next section, we present the motivation behind the thesis.

1.5 Motivation for the Thesis

In Section **??**, we have introduced challenges arising in different HetNet scenarios using the dual connectivity technique. We now discuss various aspects of DC on which the thesis is based.

The main design goals behind the introduction of DC are mobility robustness, decrease in control signaling, and improvement in per-user throughput, as detailed in Section **??**. There exist different approaches in the literature to improve the mobility robustness and throughput using DC. However, the problem of additional delay in dual connectivity due to the existence of non-ideal backhaul between the two base stations has not been addressed. The backhaul link incurs an extra delay in the transmission when the split bearer is configured for a dual connected UE. Due to this, different packets belonging to a single bearer reach the UE at different points in time. A packet p_1 belonging to bearer b_1 may arrive early at the UE because it is transmitted directly by MN, whereas packet p_2 also belonging to the same bearer b_1 may reach the UE after a substantial delay because this packet was routed by MN through an SN which was overloaded.

The existing literature has addressed the split bearer problem of dual connectivity from the perspective of throughput maximization. However, not much literature exists on the application of dual connectivity for minimizing the delay in the system. Moreover, in a practical scenario, users arrive and depart in a dynamic fashion. This dynamic behavior needs to be incorporated into the model while formulating the problem. To minimize the delay in the system, the incoming users may be blocked, since blocking leads to fewer users in the system and hence lower delay. Hence, there exists a trade-off between minimizing delay and blocking of users. We address this trade-off in our work. We formulate the optimization problem using a Markov Decision Process (MDP) and solve it using a Dynamic Programming (DP) framework.

As discussed in Section **??**, dual connectivity between different types of RATs is defined differently by 3GPP. The control signaling procedures in these different types of DC are not uniform and differ in many aspects. For instance, in LTE DC, control information for a UE can be transferred via MN only while in MR-DC, it can be transferred via MN or MN and SN both. Moreover, the interface between MN and SN differs in these three types. In LTE dual connectivity, MN and SN exchange the control information

for a dual connected UE through X2 interface between them. In the case of LWA, this exchange takes place using Xw interface, and in MR-DC, it takes place via an Xn interface. Because of a single control-plane connection for a UE to the CN via the MN, the volume of control information to be exchanged between MN and SN is significant. This increases the control load on each of the MNs for dual connected UEs. Again, this control load increases with the number of dual connected UEs, and it also leads to additional delay in the system. Due to these issues, the existing RAN architecture is not favorable for the management of dual connected UEs in a multi-RAT network. The concept of SDN introduced in the previous section can be instrumental in achieving integrated control in a multi-RAT network.

In a HetNet, the dual connected users are present alongside users with single connectivity. In such a scenario, the presence of dual connected users may, however, hamper the performance of single connected users. For instance, dual connected users may utilize resources in more than one base stations, while single connected users may be allocated the remaining resources in a single base station. Thus, dual connected users might receive higher throughput as compared to the single connected users. There is a need for a common scheduling algorithm running at all the base stations in order to maintain fairness among all users. Various network or user parameters can be optimized in designing a scheduling algorithm. For instance, network-wide throughput or user fairness can be maximized.

Proportional Fair (PF) is a type of scheduling algorithm which balances throughput maximization along with the provision of fairness to users in the system. It achieves this by assigning a priority to users at each time-slot based on their current achievable rate and the historical average throughput received until that time. When users are connected to multiple base stations, the authors in [**?**] propose a modified PF scheduling algorithm that maximizes proportional fairness in the system. However, in this work, they assume that users are connected to all available base stations. This assumption is not valid in practice since a UE has a limited number of interfaces to connect to more than one base station at the same time. Moreover, the signaling overhead to maintain so many connections for a single UE is cumbersome to manage for a service provider. Maintenance of multiple connections requires resources, for instance, power, load, and processing requirements at

the base station as well as at the CN. A practical scenario needs to be considered, and a reasonable PF scheduling scheme needs to be introduced to resolve these problems.

In this thesis, we try to address these challenges.

1.6 Contributions and Organization of the Thesis

In this section, we highlight the contributions of the thesis. The thesis is organized into six chapters. Chapter **??** presents the related literature and some of the open research challenges. Chapters 3-5 explain our contributions in detail. We develop a simulation setup in Network Simulator-3 (ns-3) [**?**] in order to simulate multiple dual connectivity scenarios as mentioned in each chapter. We compare the performance of proposed solutions with various existing solutions using this ns-3 setup.

Chapter-wise contributions are described next.

- In Chapter **??**, we consider the split bearer architecture of DC (see Section **??**) and model it using the MDP framework. We aim to minimize the average delay in the system subject to the blocking probability of arriving traffic. This is the first work that deals with the trade-off between minimizing the delay and blocking probability. The model captures the dynamic nature of traffic in the system, that is, users' arrivals and departures. Our model captures the user perspective by minimizing the delay experienced by them as well as the service provider perspective of assigning preference to the two types of users considered in the model. We obtain optimal traffic splitting policy to minimize the average delay in the system subject to a constraint on the blocking probability of arriving traffic. We also propose two heuristic policies with lower computation complexity than that of the optimal policy. We compare our policies with a traditional policy and demonstrate that dual connectivity indeed minimizes the delay in the system if the incoming traffic is routed appropriately. We perform ns-3 simulations to compare the performance of these different algorithms. The variation of the average system delay and blocking probability is studied for changes in different system parameters.

- Chapter **??** discusses the issues with the existing architecture in relation to the different types of dual connectivity. We propose an SDN based multi-RAT RAN (SM-

RAN) architecture which simplifies the existing architecture from the perspective of dual connectivity. The architecture integrates 3GPP RATs (LTE and 5G NR) with non-3GPP access technologies (WLAN) and devises a common standard interface for communication across RATs. Other RATs can also be included in the architecture with minor modifications. The different variants of dual connectivity across multiple RATs are seamlessly integrated, and their signaling procedures simplified in this architecture. It provides flexibility in setting up control-plane communication paths for a dual connected UE at any network node. SMRAN brings simplicity to the network interactions, allows flexibility in the network, and helps the network in performing load balancing and mobility management functions effectively. Based on this architecture, we propose an algorithm for selecting users for dual connectivity. We demonstrate a reduction in the control signaling and an improvement in system performance using SMRAN as compared to the legacy architecture.

- Chapter **??** presents the impact of dual connectivity on the proportional fairness in the system. One of the challenges in dual connectivity is the selection of two suitable BSs for UE association. Based on the SMRAN architecture, we propose various user association algorithms for dual connectivity. We propose a low complexity centralized PF scheduling scheme for dual connectivity (PF-DC) based on SMRAN and investigate its performance for the proposed user association algorithms. We demonstrate that PF-DC scheme outperforms the standard PF scheme in terms of PF utility and average user throughput. We perform extensive simulations in ns-3 and observe the performance of the proposed schemes. These simulations are performed under different scenarios with different placement of small cells and varying distribution of users. We demonstrate that dual connectivity, along with the PF-DC scheme, gives remarkable gains in PF utility over single connectivity and performs almost close to the optimal multi-connectivity scheme in cellular networks. Further, the PF-DC scheme has low complexity as well as low signaling overhead.

We summarize and suggest future research directions in Chapter **??**.

Chapter 2

Dual Connectivity: Relevant Literature and Open Research Areas

In this chapter, we present the relevant literature in the area of dual connectivity. In the work presented in this thesis, we focus on dual connectivity in the downlink, i.e., a user is connected to two base stations simultaneously to download data. This chapter focuses on literature for application of dual connectivity in the downlink. As mentioned in the previous chapter, the main goals for the introduction of dual connectivity technique are mobility robustness, improvement in per-user throughput, among others. In this chapter, we have categorized the existing works based on various aspects of dual connectivity that they cater to. The extension of dual connectivity technique to multiple connections, i.e., user can connect to multiple base stations at the same time, is known as Multi-Connectivity (MC). Since dual connectivity is a special case of multi-connectivity, we incorporate the literature of MC wherever applicable in this chapter. In dual connectivity or multi-connectivity, in general, a UE can be connected to nodes belonging to multiple radio access technologies. As mentioned in the previous chapter, SDN provides a centralized control that can programmatically manage and orchestrate a multi-RAT network. In this chapter, we also discuss some of the existing literature on SDN based multi-RAT network control.

2.1 Dual Connectivity in Heterogeneous Networks

As pointed out in Chapter **??**, with the emergence of HetNets, there is an increase in the number of handovers for a UE with high mobility. The control-plane user-plane split [**?**] or phantom cell [**?**] and dual connectivity [**?**] concepts have been introduced to overcome these mobility challenges in HetNets. In the control-plane user-plane split architecture, control-plane is handled by one node, and user-plane is managed by another node. As such, the first node is a conventional base station, but the second node only provides data service to the users. The first node is typically a macro base station so that a UE undergoes a handover only when it moves out of coverage area of the macro cell, and the second node is usually a small cell for capacity improvement. Hence, small cells are called phantom cells since they are only meant to carry user data traffic. [**?**] discusses control-plane user-plane split architecture, where control-plane is handled by one node, and user-plane is managed by another node. It also proposes a signaling technique to differentiate between a master node and a secondary node in dual connectivity. [**?**] discusses the phantom cell concept in detail. An overview of dual connectivity as standardized in 3GPP Release 12 has been given in [**?**]. The authors also demonstrate using simulations that DC can improve user throughput and mobility performance in HetNets. [**?**] discusses potential challenges in dual connectivity in the LTE network.

We classify the existing literature on dual connectivity based on what aspects of DC they deal with in subsequent subsections.

2.1.1 Mobility Robustness

With the introduction of HetNets, the number of handovers for a UE moving between small cells increases and radio link failures for a mobile UE increase. Further, this leads to an increase in the signaling overhead in the network. The technique of dual connectivity resolves these problems by connecting a UE simultaneously with a macro cell and a small cell. In the case of user mobility, DC provides robustness by reducing the number of handovers, handover failures and signaling in a heterogeneous network as compared to a homogeneous network. In this section, we review the literature on mobility robustness using dual connectivity.

A majority of the work in this area addresses new mobility management mechanisms for DC to improve mobility robustness in the system [?, ?, ?, ?]. In [?], the authors simulate a high mobility highway scenario in LTE within dual connectivity framework and evaluate data interruption time during handovers and cell management operations. They claim that with dual connectivity using split bearer architecture, data interruption time is reduced by 4% as compared to the single connectivity case, thus providing mobility robustness. [?] proposes an architecture in which the DC technique is applied among small cells in an ultra-dense network. Under the proposed architecture, anchor small cells are selected for UEs, acting as the master nodes while other small cells act as secondary nodes, which only provide data service to UEs. The authors propose mobility management procedures for dual connectivity between two small cells and evaluate them using different handover parameters. The authors demonstrate that for a low mobility UE, handover failure rate and ping pong rate are lower in the proposed scheme as compared to those in the existing standard LTE scheme. [?] proposes a Mobility Robustness Optimization (MRO) scheme where the handover parameters for a UE in intra-frequency dual connectivity scenario are adjusted based on its speed and handover history. The authors demonstrate using simulations that the MRO scheme gives superior mobility performance as compared to the 3GPP scheme in which handover parameters are constant.

In HetNets, when a UE moves out of the coverage area of a macro cell, an inter-macro handover procedure is performed. In the case of dual connectivity, inter-macro handover includes a macro-macro handover consisting of C-plane transition and a small cell-small cell handover consisting of U-plane transition. Both these handovers should be successfully performed for a smooth inter-macro handover. In this case, the inter-macro handover becomes inefficient since it leads to interruption either in control signaling or data transfer for a UE. This problem is considered in [?], and the authors propose a prevenient handover scheme, a two-level handover mechanism comprising small cell-macro cell and macro cell-small cell handovers. The authors demonstrate that this scheme leads to smooth handovers and shorter average service interruption times as compared to that in existing standard handover schemes with analytical expressions as well as simulations. This scheme, however, leads to an increase in handover signaling as compared to that in standard handover scheme, as two handovers need to be performed.

In [?], the authors develop analytical models to prove the reduction in handover signaling in SDN-based RAN architecture with respect to the conventional approach. In [?], the authors develop a simulation framework as an extension of the ns-3 simulator for dual connectivity between 4G and 5G cells. They also propose fast network handover procedures to improve mobility performance in HetNets as compared to traditional hard handover procedures. They propose dynamic changes in handover parameters and demonstrate that it leads to reduced packet loss, reduced latency, and improved throughput stability. In [?], the authors study the mobility and reliability aspects of dual connectivity between LTE macro cells and 5G/LTE small cells in a dense urban scenario. The authors conclude that the benefits of mobility robustness are moderate, and the instantaneous throughput is improved. However, there is a loss in the session throughput for FTP type traffic.

In [?], the authors consider a Cloud-RAN architecture, where access points with only radio head functionalities, are connected to a central unit via a fiber link. The authors propose a method to select a set of coordinated cells for mobility-related link failures and throughput degradation of cell-edge users. The authors consider a scenario with a homogeneous placement of co-channel small cells. As the number of cells for multi-connectivity increases, radio link failures are reduced to zero, and the throughput of cell-edge users is improved. The authors in [?] consider C/U plane split architecture, where a macro cell handles the control-plane of a UE, and a small cell handles the user-plane of the UE. They propose a scheme that predicts future handover events for small cells and expected handover time depending on signal strengths and UE context information such as location, speed, and handover history. The scheme is performed at the UE, and it predicts the target small cell as well as the expected handover time. It is shown through simulations that the handover latency under the proposed scheme is reduced as compared to standard LTE handover.

In [?], the authors propose a new node called mobility anchor, which helps a macro eNB to handle the handovers of small cells in a dual connectivity environment. The macro eNB offloads the handover signaling related to small cells to the mobility anchor. They demonstrate, through simulations, that in a scenario with a large number of small cells, the number of handovers handled by the macro eNB is reduced. In [?], the authors

evaluate the mobility performance of dual connectivity in three different scenarios; (1) 3GPP specific scenario, (2) Europe city scenario and (3) Tokyo city scenario. Time-of-stay of a UE in a cell is a significant mobility performance indicator since it is inversely proportional to the number of mobility events of UE. They evaluate time-of-stay and demonstrate that the time-of-stay is longer in scenarios 2 and 3 as compared to that in the scenario 1. This is because the user movement is restricted to streets in scenarios 2 and 3, while the users can move freely in scenario 1. They also demonstrate that the percentage of dual connected users is more when the small cells have a spatially uniform distribution. In [**?**], the authors propose a mobility scheme in dual connectivity performed by the UEs autonomously. This mobility scheme gives UE the autonomy of deciding the target small cell, and when to initiate the mobility event, thus, taking small cell control load off the network. This scheme reduces the amount of signaling significantly as compared to the standard dual connectivity.

2.1.2 Throughput Improvement

In this section, we review some of the representative work on throughput improvement using dual connectivity. While some of the literature on dual connectivity aims to maximize the total network throughput in the system [**?**, **?**], rest of the literature addresses the problem of improving the per-user throughput [**?**, **?**]. The authors in [**?**] propose an optimal resource fraction algorithm to split the available resources at the macro cell between dual connected users. They consider a single macro cell and multiple small cells scenario. The problem of splitting the resources of the macro cell among all users is formulated as a sum log throughput maximization problem. The idea is to split the traffic in the ratio of the corresponding rates at the macro and small cells to maximize network-wide proportional fairness.

In [**?**], the authors determine the type of bearer, viz., MCG, SCG or split, to be configured for all users to maximize the sum rate of all users in a two-tier network. The problem is formulated as an integer programming problem, to which the authors propose a sub-optimal algorithm. They demonstrate through simulations that the proposed algorithm provides higher sum throughput as compared to only 1A (MCG, SCG) or 3C (MCG, split) configurations. In [**?**], the authors propose heuristic algorithms to improve

user performance and achieve load balancing in the network using dual connectivity. They propose algorithms for bearer splitting and balancing the load between macro and small cells. The simulations are performed in different heterogeneous scenarios using video-like traffic, which has a minimum rate requirement. It is demonstrated that the proposed algorithms improve user satisfaction as well as achieve load balancing in the network compared to traditional dual connectivity.

In [**?**], an opportunistic cell association algorithm is proposed for configuring dual connected users, which improves the throughput experienced by users. They demonstrate that the downlink per-user throughput is improved with dual connectivity in a realistic urban scenario using system-level simulations.

The problem of traffic splitting between macro cell and small cell in dual connectivity is formulated as a multi-objective optimization problem in [**?**]. Various system parameters such as minimum and maximum data rate requirement of users, capacities of backhaul links between macro and small cells are considered in the problem formulation. Based on the problem formulation, the authors propose two algorithms; one in which user throughput is maximized and the other in which energy consumption in the system is minimized. With simulations, the proposed algorithms are compared with existing algorithms in terms of user throughput and energy savings. In [**?**], the traffic splitting problem in a multi-RAT network with multi-connected users is considered. The authors formulate an optimization problem to maximize a general utility function which includes maximizing sum rate, maximizing minimum rate, and proportional fairness as its special cases and an optimal traffic splitting algorithm is proposed. It is demonstrated through simulations in LTE-WLAN network that the proposed algorithm improves the median and edge throughput of users as compared to single-RAT selection techniques.

2.1.3 Delay Minimization

In this section, we review the approaches to minimize delay using dual connectivity. The split bearer configuration of dual connectivity can be utilized to minimize the delay observed at the user. In [**?**], the traffic splitting problem in a single macro cell and a single small cell scenario is modeled using fluid approximation. Their objective is to minimize expected delay in the system. Individually myopic and global myopic policies are proposed

and demonstrated to provide minimum expected delay as compared to join the shortest queue policy. The idea is to obtain an approximation to the expected delay or cost in the two paths and then choose the path with minimum delay or cost. This work, however, considers a single split bearer and a single user in the system.

A general multi-RAT wireless network, where a user can use resources belonging to multiple RATs, is considered in [**?**, **?**]. In [**?**], each RAT is modeled as an M/G/1 queuing system, and the objective is to determine the traffic splitting among RATs to minimize the maximum delay in the system. A convex optimization problem is formulated, and a traffic splitting algorithm is proposed. It is demonstrated that splitting traffic among two RATs brings down the delay as compared to a single RAT system. In [**?**], a multi-RAT system is modeled as a queuing system, and the objective is to minimize the expected delay in the system. A packet dispatching algorithm is proposed, and it is demonstrated to give lower delay as compared to an algorithm in which the packets are routed to the system with a shorter queue. In [**?**], the authors consider the split bearer scenario in dual connectivity. They propose a cascade controller design to restrict the difference between the travel time along the two paths within a pre-specified interval.

2.1.4 Energy Saving

In this section, we review the existing works on saving energy using dual connectivity. The technique of dual connectivity can be exploited to save energy in the system by enabling only those small cell nodes that are required for traffic offloading [**?**, **?**] or by using energy harvesting to reduce the total on-grid power consumption [**?**, **?**]. A two-tier network consisting of a macro cell and multiple small cells is considered in [**?**]. Small cells operate in either active mode or sleep mode and are activated only when needed by their respective macro cells. When in sleep mode, small cells cannot transmit pilot signals for channel estimation to the users in their coverage areas. Hence, corresponding to each small cell, a database is maintained at the macro cell containing a mapping from geographical locations in the small cell to Signal to Noise Ratios (SNR) of UE-small cell links. When a new UE arrives in the system, based on the geographical coordinates of the UE, the SNR from all small cells is obtained from the database. The small cell with the highest SNR is selected for activation. It is demonstrated through simulations that the

proposed scheme yields energy savings and average user throughput gains as compared to macro-only and all active small-cells scenarios. In [**?**], a two-tier HetNet using dual connectivity framework is considered. Initially, small cells are in sleep mode, and they are activated by their respective macro cells when needed. The objective in this work is to maximize the energy efficiency of the network, which is defined as the ratio of the total capacity of all cells to the total power consumption in the system. The power consumption in the backhaul link is considered in addition to the power consumption of macro and small cells. A small cell activation mechanism is proposed, and it is demonstrated through simulations that the proposed mechanism provides significant power savings with minimal implementation overhead as compared to load-aware and location-based schemes.

Dual connectivity technique can be used to save power at base stations by using a combination of energy harvesting sources to power macro and small cells. [**?**] gives an overview of the architecture, problems and possible solutions in smart energy management using dual connectivity. The benefit of exploiting the DC feature for traffic scheduling and minimizing the total on-grid power consumption has been investigated in [**?**]. In their model, each macro cell and each small cell has a hybrid energy supply, including on-grid power supply from electric power companies and renewable energy through energy harvesting devices (e.g., solar panels). The objective is to minimize the total on-grid power consumption of all macro and small cells subject to QoS constraints of users and power constraints at macro and small cells. The authors in [**?**] propose an algorithm to determine the transmit power and rate allocation for dual connected users. The authors demonstrate through simulations that their proposed algorithms outperform other fixed splitting algorithms in terms of energy savings and success probability of traffic delivery.

In [**?**], the authors consider a multi-tier network with single and dual connected users. They aim to maximize the energy efficiency such that the users' rate requirements are met. They form an optimization problem and propose a condition, which when satisfied, the user should be dual connected. The condition is based on the bandwidth allocated to the base station and the channel gain of UE-base station link. The trade-off between rate and transmit power at the base station is exploited in this work. They demonstrate that the proposed scheme provides higher energy efficiency as compared to fixed splitting and single connectivity.

2.1.5 Proportional Fair Scheduling

In this section, we review some of the existing literature on proportional fair scheduling and its usage alongside dual connectivity. As mentioned in Chapter **??**, proportional fair is a type of scheduling algorithm which maximizes the throughput in the system while at the same time ensures fairness to all users in the system [**?**].

For dual connectivity, the existing literature on PF scheduling is rather limited [**?**, **?**, **?**]. In [**?**], the authors consider LWA scenario, where traffic for each user may be split across an LTE macro cell and a WLAN small cell. The authors formulate a PF utility maximization problem to determine the optimal traffic splitting ratio at the macro cell. An optimal resource allocation algorithm is proposed, which gives enhanced user throughput performance as compared to standard algorithms. The authors in [**?**] consider a system with a single macro cell and multiple small cells. They introduce a matching based small cell selection algorithm for improving PF utility in the system. However, they assume that all users are dual connected with connection to the macro cell and propose a distributed algorithm for selecting the small cell for these users.

In [**?**], the authors consider a HetNet scenario containing multiple macro and small cells. They propose user association algorithms exploiting dual connectivity that maximize the weighted sum rate as well as proportional fairness system utilities subject to per-user rate constraints. They demonstrate that the proposed algorithms outperform standard single connectivity algorithms in terms of user throughput. An ultra-dense network consisting of LTE and 5G cells is considered in [**?**], where multi-connectivity is enabled. The authors propose an improved PF scheduling scheme, where the user first dynamically selects the cooperative base stations based on its received signal strength and then determines its priority based on load balancing, user characteristics, and fairness. The resources are scheduled based on the priority of all users. It is demonstrated that the proposed scheme improves the system throughput as compared to the standard proportional fair scheme.

Thus, dual connections for UEs can be utilized to increase throughput, minimize delay, enhance the energy efficiency, improve mobility robustness as well as proportional fairness in the system. In a HetNet, UEs can be dual connected to nodes belonging to different RATs. Management and control of such a multi-RAT network with single and

dual connected UEs is one of the interesting problems in wireless networks. In the next section, we summarize the existing works on the design of multi-RAT networks.

2.2 Multi-RAT Architectures in Wireless Networks

The next-generation wireless network is an amalgamation of multiple radio access technologies such as LTE, WLAN, and 5G. Each of these RATs have their unique architectures, protocols, and procedures. Dual connectivity is a technique which can be used between different combinations of RATs. Existing works [?, ?, ?] have proposed multi-RAT architectures to manage dual/multiple connections for users. A unified framework to manage and deal with multiple RATs is the need of the hour. Software-defined networking is a networking paradigm to make networks centrally controlled. SDN decouples the control plane from the data-plane and supports a centralized controller to control all the functions. SDN can be instrumental in integrating and providing unified control over a multi-RAT network. Existing works [?, ?, ?, ?, ?, ?, ?] have proposed multi-RAT architectures based on SDN.

SDN framework can be used either to control the core as well as radio access network together [?, ?, ?, ?] or to control the radio access network only [?, ?, ?]. The former case is difficult to deploy in practice since it needs to model all the control functions of core network as well as RAN in the controller, whereas the latter case allows easy incremental deployment, since only RAN component needs to be changed. However, the former approach brings more flexibility and control in network design. [?] gives an overview of the application of SDN in wireless networks with certain use cases and their advantages and shortcomings. It also gives an introduction to some standardization efforts in SDN.

[?] presents a software-defined architecture for LTE system, modifying the core network as well as RAN. The control functions in CN and RAN are moved to the controller to manage the LTE system centrally. Along with RRC, the controller also contains control functions of core network entities, specifically, MME and SGW. The data plane functions of CN entities and eNBs are moved to the data-plane or infrastructure layer. The authors evaluate the proposed architecture using experimental testbed. They propose an algorithm for load-based handover in the proposed architecture and demonstrate that it

leads to an improvement in user throughput and reduction in signaling as compared to standard handover in LTE architecture.

The work in [**?**] considers an ultra-dense network consisting of traditional macro and small cells along with mmWave access points. The authors propose an SDN architecture consisting of a centralized superior SDN controller and a localized subordinate SDN controller. While the superior SDN controller supervises the working of all subordinate SDN controllers to guarantee global load balancing and energy efficiency, each subordinate SDN controller controls a local area containing several cells. The subordinate SDN controller controls user association, load balancing, and resource allocation within its service area. Configuration of service areas and cells of subordinate controllers, dynamic subordinate management, and mobility management of UEs are functions of the superior SDN controller. The functions of CN and RAN are divided among the two layers of SDN controllers. They demonstrate that significant improvement in network load balance and energy efficiency can be achieved in this architecture compared to traditional architecture.

In [**?**], the authors propose a centralized SDN architecture for 5G modifying CN as well as RAN. The control functions of 5G RAN are relocated to the core network. It is demonstrated that this leads to a reduction in signaling between CN and RAN, and an improvement in mobility management. In [**?**], the authors redesign a multi-RAT network using SDN. They propose a centralized SDN based wireless network controller for controlling the multi-RAT nodes as well as gateways in the core network. The control functionality from the core as well as RAN is aggregated at a central network controller, while the data-plane functions remain at the respective nodes in CN and RAN. The main advantage of this architecture is that all the network control is unified at the controller, and the RAN nodes can belong to any technology, e.g., LTE, 5G.

In [**?**], the authors propose SoftRAN, which modifies the RAN in LTE architecture to segregate it into a centralized RAN controller and radio elements. The centralized controller takes decisions that affect the neighboring elements in the network such as radio resource management, handovers and transmit power allocation, while each radio element takes decisions that affect locally such as resource allocation. It is demonstrated using some example scenarios that this architecture brings load balancing and user QoS enhancement. In [**?**], the authors abstract RAN into two parts: a centralized master

controller and several distributed agents situated at eNBs. The distributed agents perform only time-critical applications such as scheduling. The other functions, such as event management and configuration of agents, are performed by the master controller. The authors also propose FlexRAN protocol for communication between master controller and agents.

In [?], the authors propose OpenRAN, where the RAN is divided into three parts: an SDN controller, cloud computing resource pool, and wireless spectrum resource pool. The wireless spectrum resource pool consists of remote radio units for radio access to users. The cloud computing resouce pool consists of cloud computing hardware such as baseband units and base station controllers. These are then controlled centrally by an SDN controller. However, there is no clear separation between the control-plane and data-plane in these works [?, ?, ?]. In [?], the authors give an overview of cloud-RAN, a novel architecture, which centralizes the base stations and provides a cooperative RAN solution. The basic idea behind cloud-RAN is to pool the baseband-processing units from all base stations to a central location for statistical multiplexing gain. This reduces the cost incurred in base station establishments and increases the energy efficiency of the system.

Some of the existing works have proposed architectures on dual/multi-connectivity, but these are not based on SDN [?, ?, ?]. In [?], several options for the protocol architectures in 5G are proposed. For instance, one of the architectures proposes a common Medium Access Control (MAC) layer across all RAT nodes. These architectures are examined individually and evaluated based on aspects such as signaling overhead, inter-RAT connectivity, etc. However, no qualitative analysis of these architectures is provided in this work. Similar approaches about radio protocol architectures for multi-connectivity in 5G networks have been proposed in [?]. They present a set of different potential architectures for 5G radio access network and discuss their functional requirements. In [?], the authors propose some multi-connectivity radio protocol architectures. In one of the architectures, there is a centralized PDCP layer which exists in the cloud, and lower layers (RLC, MAC, PHY) exist in the access points. The authors analyze these architectures, and some of the reliability aspects of 5G are highlighted. However, these works [?, ?, ?] are specific to 5G networks only and do not consider other RATs.

Thus, a significant body of work exists on architecture design in heterogeneous wireless networks, and efforts have also been made to design dual/multi-connectivity wireless network.

2.3 Discussion and Research Topics

In this chapter, we have reviewed existing literature on the basis of different applications of dual connectivity. Note that dual connectivity is used for overcoming the mobility challenges in HetNets and for improving the throughput in the system. Dual connectivity is also useful in improving energy efficiency in the system. A significant body of literature exists on the application of dual connectivity in improving mobility robustness, throughput, energy efficiency in the system. There exist different approaches in the area of PF scheduling using dual connectivity. However, the application of dual connectivity on the delay in the system has not been addressed satisfactorily in the literature. Specifically, the trade-off between minimizing delay and blocking of arriving traffic has not been addressed in the literature. Moreover, in a practical scenario, users arrive and depart in a dynamic fashion. This dynamic behavior has not been considered in the literature before. We consider all these points and formulate an optimization problem in Chapter **??**.

We have reviewed the existing approaches to multi-RAT architectures in this chapter. Various architectures have been proposed to bring centralization in the design of cellular networks. However, these architectures do not apply to dual connectivity. Some works have proposed architectures for multi-connectivity, but these do not include a multi-RAT system, or there is no clear separation of control-plane and data-plane in these works. As mentioned in Chapter **??**, there are different variants of dual connectivity defined in 3GPP standards. The control signaling procedures in these different types of DC are inconsistent with each other. For instance, in case of LTE DC and LWA, the control signaling from core network can only take place via the master node, which is an eNB in these variants. In the case of MR-DC, the control signaling with core network can take place via any of the three types of bearers, viz., MCG, SCG or split bearers with the condition that signaling initiation can take place via MCG bearer only. Further, the control signals are conveyed by MN to SN via the interface between them. Thus, a lot of

signaling exchange takes place between MN and SN. Moreover, the interfaces for signaling exchange between MN and SN vary based on the RAT of MN and SN. Hence, to simplify the existing architecture and bring uniformity among the different variants of DC, we propose an SDN based multi-RAT RAN architecture in Chapter **??**.

The literature on proportional fair scheduling has been presented in this chapter. Specifically, the works on proportional fair scheduling in the context of dual or multi-connectivity have been highlighted. In the case of multiple connections of users, a new PF scheduling scheme has been proposed in [**?**] that maximizes the proportional fairness in the system. However, the authors consider that all users are connected to all available base stations. This would, however, require a UE to have multiple transceivers to communicate with all base stations in the system. In practical scenarios, a UE would have a limited number of transceivers. Moreover, there arises large signaling overhead in the network to maintain multiple connections per UE. Therefore, we consider a practical scenario of UEs with dual transceivers and propose PF-DC - a simple PF scheduling scheme for dual connectivity in Chapter **??** to resolve these problems. This scheduling scheme has low complexity and is easy to implement in practice.

In the next chapter, we formulate an optimization problem to minimize delay in the system in a scenario comprising single and dual connected users.

Chapter 3

Optimal Traffic Splitting Algorithm in a Heterogeneous Network

As mentioned in the previous chapter, few studies exist on the application of dual connectivity to minimize delays in the system. In this chapter, we investigate the impact of dual connectivity on the delay in the system. We model the split bearer architecture (see Section **??**) of dual connectivity using a queuing theoretic system. In the split bearer architecture, two types of bearers may be created: Master Cell Group (MCG) and split bearer. Both types of bearers originate from the core network and pass through Master Node (MN) since only MN has a connection with the core network. For MCG bearer, data reaches a user via MN. However, in the case of the split bearer, data is split at the PDCP layer of MN, and a part of the data is routed via the Secondary Node (SN). The dual connected user simultaneously receives a part of the data of the split bearer from MN and the remaining portion from SN. Thus, data from the split bearer takes two different routes to reach a user. In this chapter, we address the problem of optimal route selection for dual connected users when multiple routes are available to the user. In this chapter, without loss of generality, we assume Macro eNodeB (MeNB) acts as MN, and Small cell eNB (SeNB) acts as SN.

Due to diversity in the paths taken by data to reach a user in the case of a split bearer, the difference between reception times corresponding to the two paths may be significant. Thus, for a dual connected user, the delay experienced in receiving data corresponding to a single transmission may be considerable. Moreover, there exist legacy

User Equipments (UEs) which only connect to a single eNB, i.e., MeNB. For these users (henceforth referred to as background users), data transfer can only take place via MeNB. The data transfer for dual connected users (henceforth referred to as foreground users) can take place via MeNB or SeNB or both. We aim to minimize the delays in data reception corresponding to each transmission. Minimization of delay may, however, lead to the blocking of users because blocking the users decreases the number of packets in the system. Further, MeNB resources are shared between background and foreground users. Hence, a constraint on the blocking probability of both background and foreground users is introduced. Our objective is to determine an optimal splitting policy to minimize average delay in the system subject to a constraint on blocking probability of background and foreground users. We formulate this problem as a Constrained Markov Decision Problem (CMDP).

The authors in [**?**] propose a split bearer algorithm for video traffic to improve the user perceived data rate. In [**?**], the optimal splitting ratio for minimizing the queuing delay in the system is calculated for a single UE by modeling the split bearer architecture using a fluid model. The authors in [**?**] obtain optimal traffic split over multiple Radio Access Technologies (RATs) such that the maximum average delay across different RATs is minimized. These works [**?**, **?**, **?**], however, do not consider user arrival and departure. Further, in [**?**], the authors consider the maximization of expected delays in different RATs as the optimization parameter. However, in our work, we consider the total delay defined as the time between the transmission of a packet and the time it is received at the UE. This delay is then averaged over all the packets transmitted. Thus, we deal with the total delay as the system metric.

The main contributions of our work are enumerated here.

1. We aim to minimize average delay in the system subject to a constraint on the weighted average of the blocking probability of background and foreground users. The application of dual connectivity to minimize delays in the system has not been investigated in the literature.

2. We formulate the optimization problem using the Markov Decision Process (MDP) framework and determine the optimal policy. The model captures the dynamic nature of traffic in the system, that is, users' arrivals and departures.

3. The method of obtaining the optimal policy is, however, computation-intensive. Hence, we propose two heuristic algorithms with low computation and space complexity.

4. We conduct extensive simulations in ns-3 to compare the proposed algorithms with the optimal policy as well as with some traditional algorithms. We demonstrate that the proposed algorithms outperform other algorithms.

5. The variations in average system delay and blocking probability are studied for changes in different system parameters.

6. We demonstrate that with dual connectivity, we can indeed achieve minimum delay in the system if incoming traffic is routed appropriately.

The chapter is structured as follows. Section **??** presents the system model. The problem formulation as a constrained MDP and solution methodology are described in Section **??**. The various heuristic traffic splitting algorithms are explained in Section **??**. The computational complexity and implementation issues of optimal policy and proposed algorithms are introduced in Section **??**. Section **??** discusses numerical results of simulations performed in ns-3 and analyzes the results obtained. Section **??** concludes the chapter.

3.1 System Model

A HetNet may be composed of several macro and small cells. We consider a scenario, as illustrated in Figure **??**, consisting of a single macro cell and a single small cell for simplicity. This model can be easily extended to multiple small cells within the coverage area of a macro cell. It consists of a small cell with its coverage overlapping with that of a macro cell. MeNB and SeNB operate at different carrier frequencies and are connected via a non-ideal backhaul link (X2 interface). SeNB uses this backhaul link to share its status information with the MeNB, and MeNB uses it to share control/data information with the SeNB.

We segregate UEs into two categories. Legacy UEs which are present in the coverage area of the macro cell and can connect to MeNB only are categorized as background UEs.

UEs which are present in the coverage area of the small cell and capable of dual connectivity to MeNB and SeNB are categorized as foreground UEs. Note that MeNB manages the connectivity of both sets of UEs. The user-plane data belonging to foreground users can, however, be managed by MeNB or SeNB, while that belonging to background users can be managed by MeNB only. The data traffic streams for these two sets of UEs are each assumed to constitute two Poisson arrival streams with rates λ_1 and λ_2, respectively. The download data traffic corresponding to the set of UEs that can connect to both types of eNBs is assumed to form a foreground Poisson arrival stream with rate λ_1. The data traffic corresponding to the set of UEs that can connect to only MeNB is assumed to form a background Poisson arrival stream with rate λ_2. We assume that UEs are uniformly distributed in the coverage area and are stationary. However, in Section **??**, we demonstrate through simulations that we obtain similar results for mobile users as well.

Figure **??** depicts the system model. The overall system comprises three subsystems, viz., MeNB subsystem (subsystem 1) and SeNB subsystem (subsystem 2) and backhaul subsystem. We model MeNB, SeNB subsystems, and backhaul link as queuing subsystems. The flow controller is situated at the MeNB. The downlink traffic corresponding to the two traffic streams arrives at the flow controller, which then decides to route each type of traffic appropriately to the two eNB subsystems, based on a global network utility. We formulate this globally optimal control problem as a stochastic dynamic programming problem in this section. Our model is a generic formulation of the split bearer dual connectivity problem to optimize a global utility. The flow controller can route incoming traffic of foreground users to either the MeNB or SeNB subsystem appropriately. It can either allow incoming traffic of background users to join the MeNB subsystem or reject it. MeNB and SeNB subsystems have n_1, n_2 number of servers, respectively, to serve the arrivals, whereas the backhaul subsystem has a single server. Each server on the MeNB subsystem takes an exponentially distributed amount of time with mean $1/\mu_m$ to serve a packet. Each server on the SeNB subsystem takes an exponentially distributed amount of time with mean $1/\mu_s$ to serve a packet. Similarly, the service time of server on the backhaul link is exponentially distributed with mean $1/\mu_d$. The latency of the backhaul link in practical systems is of the order of milliseconds [**?**]. Hence, we assume $\mu_d \gg \mu_s$. All queuing systems are work conserving and follow First-In-First-Out (FIFO) queuing

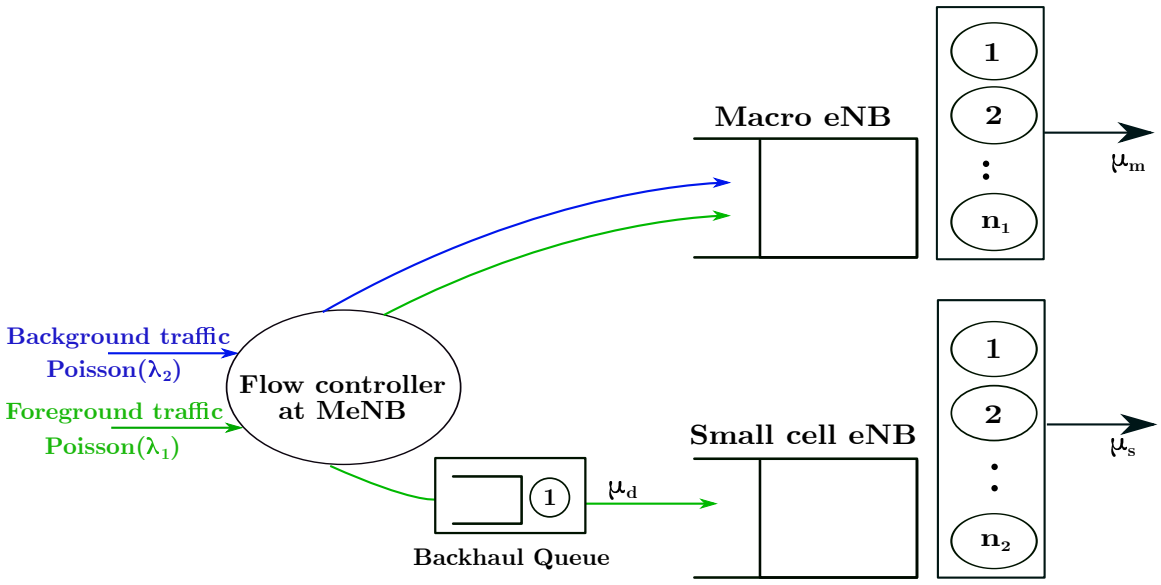

Figure 3.1: System model.

discipline.

The data traffic over the Internet is self-similar [?, ?], which gives it a bursty nature. We adopt a batch arrival process to model its bursty nature. We consider batch arrivals with a random number of packets in a batch. The batch size G follows a discrete probability distribution given by

$$\alpha_i = P(G = i), \quad i = 1, 2, \cdots, \tag{3.1}$$

with mean batch size $\bar{G}$.

As MeNB and SeNB subsystems have a limited number of servers, we assume that the excess traffic is placed in a buffer of size B at each subsystem. The buffer may have a finite size in practice. However, we assume B to be large enough compared to the batch arrival rate to neglect the probability of buffer overflow and packet drops. Thus, a total of $N_1 = B + n_1$ packets can be accommodated in the MeNB subsystem. When the flow controller routes the foreground traffic to the SeNB subsystem, the traffic is first served by the backhaul subsystem. We assume that a buffer of size B at the SeNB subsystem is shared between the SeNB and backhaul subsystems. Thus, a total of $N_2 = B + n_2$ packets can be accommodated in the SeNB and backhaul subsystems. Since $\mu_d \gg \mu_s$, this is a reasonable assumption. We assume that each packet has a fixed size. We assume that each packet requires one server to get served. Once packets join any of the two subsystems, scheduling of packets in both subsystems takes place independent of each

other. We assume that both types of traffic are assigned equal priority while allocating resources.

3.1.1 State Space

Let the number of packets in each of the subsystems be denoted as the state of the system. The continuous-time stochastic process $\{X(t)\}_{t\geqslant 0}$ denotes the state of the system at time instant t, which is defined over the state space $\mathcal{S}$. We define a 4-tuple (s_1, s_2, s_3, e) as the state of the system $s \in \mathcal{S}$. Here, s_1, s_2, and s_3 represent the number of packets in the queue plus the number of packets currently in service in the MeNB, backhaul and SeNB subsystems, respectively. Since a maximum of N_1, N_2 packets can be accommodated in each subsystem, $0 \leqslant s_1 \leqslant N_1, 0 \leqslant s_2 + s_3 \leqslant N_2$. The event that takes place when the system is in state s_1, s_2, s_3 is denoted by e. The possible events in the model are explained next.

Events e_{-2}, e_{-1}, e_0 denote the event of a departure of a packet from MeNB, backhaul, and SeNB subsystems, respectively. The packets departed from MeNB and SeNB subsystems leave the system completely, whereas the packets departing from the backhaul subsystem join the SeNB subsystem. If there is a foreground batch arrival of size $G = 1, 2, ..., n$ then events are denoted as $e_1, e_2, ..., e_n$, respectively. If there is a background batch arrival of size $G = 1, 2, ..., n$ then events are denoted as $e_{n+1}, e_{n+2}, ..., e_{2n}$, respectively. Since the state of the system changes only at arrival or departure instants, the state at other points in time can be deduced from the state at these epochs. For simplicity, we explain the state space and state dynamics for maximum batch size $n = 2$. e_1, e_2 represent foreground traffic arrival of batch size 1 and 2, respectively. e_3, e_4 represent background traffic arrival of batch size $1, 2$, respectively. Consider $n_1 = 5, n_2 = 5$ and queue size $B = 10$. Then, $0 \leqslant s_1 \leqslant 15, 0 \leqslant s_2 + s_3 \leqslant 15$. Thus, state $s = (3, 2, 6, e_2)$ indicates that there are 3 packets in service in the MeNB subsystem, 1 packet in service plus 1 packet in queue in the backhaul subsystem, and 5 packets in service ($n_2 = 5$) plus 1 packet in the queue of the SeNB subsystem when foreground traffic with batch size 2 (e_2) has arrived.

3.1.2 Decision Epochs

The time instants at which a decision needs to be taken are called decision epochs. In our model, the arrival instants of background and foreground traffic are the decision epochs. At the same time, time instants at which departure of a packet takes place are decision epochs. These events at which decisions need to be taken are incorporated in the state of the system to simplify the model.

3.1.3 Action Space

We denote $\mathcal{A}$ as the action space. At each decision epoch, the flow controller needs to take an action $a \in \mathcal{A}$ based on the current state of the system $s \in \mathcal{S}$. For foreground traffic arrival, a decision needs to be taken whether to admit the arrival or to route the arrival to MeNB subsystem or to SeNB subsystem or to split the arriving batch between the two subsystems. For background traffic arrival, the controller needs to take a decision of allowing the arrival into the MeNB subsystem. The action space $\mathcal{A}$ depends on the maximum size of the batch n. For $n = 2$, the action space is as follows:

$$\mathcal{A} = \begin{cases} a_0, & \text{Do nothing / Block,} \\ a_1, & \text{Route all packets to SeNB subsystem,} \\ a_2, & \text{Route to both the subsystems,} \\ a_3, & \text{Route all packets to MeNB subsystem.} \end{cases}$$

A foreground traffic arrival of one packet (event e_1) can be blocked (a_0) or accepted in SeNB subsystem (a_1) or accepted in MeNB subsystem (a_3). A foreground traffic arrival with batch size 2 (event e_2) can be blocked (a_0) or accepted in SeNB subsystem (a_1) or accepted in MeNB subsystem (a_3) or split between the two subsystems (a_2). If both the subsystems are completely occupied, foreground traffic is blocked (a_0) on arrival. A background traffic arrival with one packet in a batch (event e_3) can be either blocked (a_0) or accepted in MeNB subsystem (a_3). A background traffic arrival with batch size 2 (event e_4) can be either blocked (a_0) or accepted in MeNB subsystem (a_3). If MeNB subsystem is completely occupied on a background traffic arrival, then the arrival is blocked (a_0). In the event of a departure of a packet (events e_{-2}, e_{-1}, e_0), the only possible action is to do nothing (a_0).

3.1.4 State Dynamics

At each decision epoch, the controller takes an action $a \in \mathcal{A}$ depending on the state of the system $s \in \mathcal{S}$. Depending on the state and action taken, the system transitions to another state with a certain probability. Let $T_{ss'}(a)$ denote the transition probability from state s to state s' under action a. Denote $\nu(s_1, s_2, s_3)$ as the sum of arrival rates and departure rates, when the current state is $s = (s_1, s_2, s_3, e)$,

$$\nu(s_1, s_2, s_3) = \lambda_1 + \lambda_2 + \min(s_1, n_1)\mu_m + \min(s_2, 1)\mu_d + \min(s_3, n_2)\mu_s. \tag{3.2}$$

Note that $\nu(s_1, s_2, s_3)$ is independent of the event e.

The transition probabilities from state $s = (s_1, s_2, s_3, e)$ with $e = e_{-2}, e_{-1}, e_0, e_1, e_2, e_3, e_4$ to state s' using action a_0 are given by:

$$T_{ss'}(a_0) = \begin{cases} \frac{\min(s_1,n_1)\mu_m}{\nu(s_1,s_2,s_3)}, & s' = (s_1 - 1, s_2, s_3, e_{-2}), \\ \frac{\min(s_2,1)\mu_d}{\nu(s_1,s_2,s_3)}, & s' = (s_1, s_2 - 1, s_3 + 1, e_{-1}), \\ \frac{\min(s_3,n_2)\mu_s}{\nu(s_1,s_2,s_3)}, & s' = (s_1, s_2, s_3 - 1, e_0), \\ \frac{\lambda_1\alpha_1}{\nu(s_1,s_2,s_3)}, & s' = (s_1, s_2, s_3, e_1), \\ \frac{\lambda_1\alpha_2}{\nu(s_1,s_2,s_3)}, & s' = (s_1, s_2, s_3, e_2), \\ \frac{\lambda_2\alpha_1}{\nu(s_1,s_2,s_3)}, & s' = (s_1, s_2, s_3, e_3), \\ \frac{\lambda_2\alpha_2}{\nu(s_1,s_2,s_3)}, & s' = (s_1, s_2, s_3, e_4). \end{cases} \tag{3.3}$$

The transition probabilities for state $s = (s_1, s_2, s_3, e)$ with $e = e_1, e_2$ and action a_1 are as follows:

$$T_{ss'}(a_1) = \begin{cases} \frac{\min(s_1,n_1)\mu_m}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1 - 1, s_2 + G, s_3, e_{-2}), \\ \frac{\min(s_2+G,1)\mu_d}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1, s_2 + G - 1, s_3 + 1, e_{-1}), \\ \frac{\min(s_3,n_2)\mu_s}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1, s_2 + G, s_3 - 1, e_0), \\ \frac{\lambda_1\alpha_1}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1, s_2 + G, s_3, e_1), \\ \frac{\lambda_1\alpha_2}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1, s_2 + G, s_3, e_2), \\ \frac{\lambda_2\alpha_1}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1, s_2 + G, s_3, e_3), \\ \frac{\lambda_2\alpha_2}{\nu(s_1,s_2+G,s_3)}, & s' = (s_1, s_2 + G, s_3, e_4). \end{cases} \tag{3.4}$$

For state $s = (s_1, s_2, s_3, e)$ with $e = e_2$ and action a_2, the transition probabilities are:

$$T_{ss'}(a_2) = \begin{cases} \frac{\min(s_1+1,n_1)\mu_m}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1, s_2+1, s_3, e_{-2}), \\ \frac{\min(s_2+1,1)\mu_d}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1+1, s_2, s_3+1, e_{-1}), \\ \frac{\min(s_3,n_2)\mu_s}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1+1, s_2+1, s_3-1, e_0), \\ \frac{\lambda_1\alpha_1}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1+1, s_2+1, s_3, e_1), \\ \frac{\lambda_1\alpha_2}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1+1, s_2+1, s_3, e_2), \\ \frac{\lambda_2\alpha_1}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1+1, s_2+1, s_3, e_3), \\ \frac{\lambda_2\alpha_2}{\nu(s_1+1,s_2+1,s_3)}, & s' = (s_1+1, s_2+1, s_3, e_4). \end{cases} \tag{3.5}$$

For state $s = (s_1, s_2, s_3, e)$ with $e = e_1, e_2, e_3, e_4$ and action a_3, the transition probabilities are:

$$T_{ss'}(a_3) = \begin{cases} \frac{\min(s_1+G,n_1)\mu_m}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G-1, s_2, s_3, e_{-2}), \\ \frac{\min(s_2,1)\mu_d}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G, s_2-1, s_3+1, e_{-1}), \\ \frac{\min(s_3,n_2)\mu_s}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G, s_2, s_3-1, e_0), \\ \frac{\lambda_1\alpha_1}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G, s_2, s_3, e_1), \\ \frac{\lambda_1\alpha_2}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G, s_2, s_3, e_2), \\ \frac{\lambda_2\alpha_1}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G, s_2, s_3, e_3), \\ \frac{\lambda_2\alpha_2}{\nu(s_1+G,s_2,s_3)}, & s' = (s_1+G, s_2, s_3, e_4). \end{cases} \tag{3.6}$$

3.1.5 Cost Function

Let $c(s, a)$ denote the cost incurred when the system is in state $s = (s_1, s_2, s_3, e)$ and action $a \in \mathcal{A}$ is taken. We define this cost as the expected delay encountered by the arriving batch of packets. Determining the expected delay for a batch of packets in a multi-server queuing system given the number of packets in the system is computationally intensive. Hence, we use a different approach to determine the expected delay for the arriving batch. Since the system is modeled as a queuing system, we can apply Little's Law. According to Little's Law, the expected delay in the system is proportional to the expected number of packets in the system [**?**]. Hence, we define the cost rate as the number of packets in the system. Hence, the cost function is defined as the number of packets multiplied by the expected time until the next decision epoch. The cost function is a measure of the

service received by users (user perspective). The cost for different states and actions is listed in Table **??**.

Table 3.1: Cost table.

(e, a)	$c(s, a)$
$(\{e_{-2}, e_{-1}, e_0\}, a_0)$	$(s_1 + s_2 + s_3)\ \nu(s_1, s_2, s_3)$
$(\{e_1, e_2, e_3, e_4\}, a_0)$	$(s_1 + s_2 + s_3)\ \nu(s_1, s_2, s_3)$
(e_1,a_1)	$(s_1 + s_2 + 1 + s_3)\ \nu(s_1, s_2 + 1, s_3)$
(e_2,a_1)	$(s_1 + s_2 + 2 + s_3)\ \nu(s_1, s_2 + 2, s_3)$
(e_2,a_2)	$(s_1 + 1 + s_2 + 1 + s_3)\ \nu(s_1 + 1, s_2 + 1, s_3)$
$(\{e_1, e_3\},a_3)$	$(s_1 + 1 + s_2 + s_3)\ \nu(s_1 + 1, s_2, s_3)$
$(\{e_2, e_4\},a_3)$	$(s_1 + 2 + s_2 + s_3)\ \nu(s_1 + 2, s_2, s_3)$

Minimization of average delay in the system may, however, lead to the blocking of both background and foreground traffic. The functions $b_b(s, a)$ and $b_f(s, a)$ are defined as the blocking cost functions of background and foreground traffic, respectively.

$$b_b(s,a) = \begin{cases} 1, & \text{if state } s = (s_1, s_2, s_3, e_3) \text{ and action } a = a_0, \\ 2, & \text{if state } s = (s_1, s_2, s_3, e_4) \text{ and action } a = a_0, \\ 0, & \text{else.} \end{cases} \tag{3.7}$$

$$b_f(s,a) = \begin{cases} 1, & \text{if state } s = (s_1, s_2, s_3, e_1) \text{ and action } a = a_0, \\ 2, & \text{if state } s = (s_1, s_2, s_3, e_2) \text{ and action } a = a_0, \\ 0, & \text{else.} \end{cases} \tag{3.8}$$

A fixed cost is incurred whenever foreground or background traffic is blocked. We define the total blocking cost $b(s, a)$ as the weighted sum of background and foreground traffic blocking cost.

$$b(s,a) = \delta b_b(s,a) + (1-\delta) b_f(s,a). \tag{3.9}$$

The parameter δ, $0 \leqslant \delta \leqslant 1$, decides the fraction of weight to be assigned to the blocking probability of each traffic type. Depending on the requirement of the service provider, parameter δ can be chosen. For instance, if the service provider wants to charge extra for

providing dual connectivity service, he/she can give a higher weight $(1-\delta)$ to foreground traffic which comprises users with multiple transceivers capable of dual connectivity and low weight (δ) to legacy users (background users). Thus, our model captures the perspective of the service provider as well.

We explain the problem formulation in the next section.

3.2 Problem Formulation and Solution Methodology

As batches arrive at random time instants, a sequential decision-making procedure is required to accept/reject the arriving traffic or to route it to a suitable subsystem. Due to the dynamic and probabilistic nature of the system model, the state transitions need to be controlled by a sequence of appropriate decisions. We aim to minimize the average delay in the system subject to a constraint on the total blocking probability. We consider the long-run average cost per time unit as the performance criterion of the MDP. We model this using the continuous-time CMDP framework.

An MDP is a stochastic control problem defined by a 5-tuple $\{\mathcal{S}, D, \mathcal{A}, T_{st}(a), c(s,a)\}$ with the following components:

- $\mathcal{S}$ is a finite set of states.
- D is a set of decision epochs.
- $\mathcal{A}$ is a finite set of actions, $\mathcal{A}(s)$ is the set of actions for a particular state s.
- $T_{st}(a)$ is the transition probability to go from current state s to next state t if action $a \in \mathcal{A}(s)$ is chosen.
- $c(s,a)$ is the cost if the current state is s and the action $a \in \mathcal{A}(s)$ is chosen.

These components of the MDP problem are defined in Section **??**. Thus, our system model fits within the MDP framework.

3.2.1 Formulation as Constrained Markov Decision Process

A decision rule $d_t: \mathcal{S} \rightarrow \mathcal{A}$ is a mapping from state space to action space at decision epoch t. A policy $\{d_1, d_2, ..., d_t, ...\}$ for controlling the system specifies the decision rule

to be used at each decision epoch. If decision rules are independent of time point t, then the policy is called stationary; it has the form $\{d, d, ...\}$. It assigns to each state s a fixed action $a = a_s$ and always uses this action whenever the system is in state s. We are interested in determining a control policy, which is optimal for the problem under consideration.

The state of the system (s_1, s_2, s_3) consists of the number of packets in each of the three subsystems, viz., MeNB, backhaul, and SeNB subsystems, which forms a continuous-time Markov chain. Since the service rate of each of these subsystems is positive, the departure times in the system are positive. Arrival rates in the system are also positive, making arrival times positive. Hence, the state $(0, 0, 0)$ is reachable from any state in the system. This is true for any policy chosen for the system. There cannot be two closed communicating classes for the embedded Markov chain, as the state $(0, 0, 0)$ has to be a member of the communicating class. Hence, the embedded Markov chain associated with this model contains a single recurrent class plus, perhaps, some transient states making it a unichain.

Since the state space and the action space of our model are finite, and the embedded chains are unichain, the optimal policy exists, and it is a stationary policy [**?**]. Denote Π as the set of stationary policies. The long-run time-averaged delay and blocking can be expressed as,

$$\bar{C} = \lim_{W\to\infty} \frac{1}{W} \ \mathbb{E}_\pi\Big\{ \sum_{n=0}^{W-1} c(s(n), a(n))\Big\}, \tag{3.10}$$

and,

$$\bar{B} = \lim_{W\to\infty} \frac{1}{W} \ \mathbb{E}_\pi\Big\{ \sum_{n=0}^{W-1} b(s(n), a(n))\Big\}, \tag{3.11}$$

respectively, where W is the number of decisions taken until time t, $s(n)$ and $a(n)$ denote the state and action at decision epoch n. $\mathbb{E}_\pi$ is the expectation operator under policy $\pi \in \Pi$. Note that the limits in (**??**) and (**??**) exist since we are considering stationary policies [**?**]. The objective of our model is to obtain a policy that minimizes $\bar{C}$ subject to a constraint (say, B_{max}) on $\bar{B}$.

$$\text{Minimize } \bar{C} \text{ Subject to } \bar{B} \leq B_{max}. \tag{3.12}$$

The problem (**??**) is a CMDP problem with average cost and finite state and action spaces. It is widely known that a stationary randomized optimal policy [**?**] exists.

3.2.2 Solution Methodologies

In this section, we solve the problem formulated in the previous section.

Uniformization

To solve the continuous-time MDP problem (**??**), we first convert it into a discrete-time MDP problem using the uniformization technique [**?**]. In this technique, delay cost function, blocking cost function, and event probabilities are transformed using a number τ such that $0 < \tau < \min_{s,a} \tau(s,a)$. Here, $\tau(s,a)$ is the expected time until the next decision epoch, if action a is chosen in the state $s = (s_1, s_2, s_3)$. Note that $\tau(s,a)$ is independent of the event taking place when the system is in state $s = (s_1, s_2, s_3)$. Consider the discrete-time Markov decision model, with same state space and action space as the continuous-time MDP model in Section **??** and delay cost, blocking cost, and transition probabilities modified as:

$$\hat{c}(s,a) = c(s,a)/\tau(s,a), \tag{3.13}$$

$$\hat{b}(s,a) = b(s,a)/\tau(s,a), \tag{3.14}$$

and

$$\hat{T}_{ss'}(a) = \begin{cases} (\tau/\tau(s,a))T_{ss'}(a), & s \neq s', \\ (\tau/\tau(s,a))T_{ss'}(a) + [1 - (\tau/\tau(s,a))], & s = s'. \end{cases} \tag{3.15}$$

This discrete-time Markov decision model has the same class of stationary policies as the original continuous-time MDP model. For each stationary policy $\pi \in \Pi$, the long-run average cost per time unit in the discrete-time model is equal to that in the continuous-time MDP model.

Lagrangian Approach

The Lagrangian approach [**?**] is now used to convert the CMDP problem into an unconstrained one. Let $\beta \geq 0$ denote the Lagrangian Multiplier. The total cost function is then given by,

$$\hat{h}(s,a;\beta) = \hat{c}(s,a) + \beta\hat{b}(s,a). \tag{3.16}$$

The dynamic programming equation below provides the optimal value function $V(s)$ for a given state s, thus yielding the optimal policy.

$$V(s) = \min_{a} \left\{ \hat{h}(s, a; \beta) + \sum_{s' \in \mathcal{S}} \hat{T}_{ss'}(a) V(s') \right\}. \tag{3.17}$$

For a particular value of β, this dynamic programming equation can be solved using the Value Iteration Algorithm (VIA) [**?**].

Gradient Descent Algorithm

At a particular value of $\beta = \beta^*$, minimum cost is obtained for the constrained problem. This value of β^* is found using the gradient descent algorithm. The value of β is approached using the equation,

$$\beta_{n+1} = \beta_n + \frac{1}{n}(\bar{B}_n - B_{max}), \tag{3.18}$$

where $\bar{B}_n$ is the blocking probability obtained using the policy π_{β_n} at iteration n. For this value of β^*, the optimal policy is a mixture of two stationary policies, which can be determined by a small deviation ϵ of β^* in both directions. This results in two policies $\pi_{\beta^*-\epsilon}$ and $\pi_{\beta^*+\epsilon}$ with associated average blocking probability $\bar{B}_{\beta^*-\epsilon}$ and $\bar{B}_{\beta^*+\epsilon}$, respectively. Define a parameter q such that $q\bar{B}_{\beta^*-\epsilon} + (1-q)\bar{B}_{\beta^*+\epsilon} = B_{max}$. The optimal policy π^* of the CMDP is a randomized mixture of the two stationary policies ($\pi_{\beta^*-\epsilon}$ and $\pi_{\beta^*+\epsilon}$), such that at each decision epoch, the first policy is chosen with probability q and the second policy is chosen with probability $(1-q)$. Thus, the optimal policy is given by,

$$\pi^* = q\pi_{\beta^*-\epsilon} + (1-q)\pi_{\beta^*+\epsilon}. \tag{3.19}$$

3.3 Heuristic Traffic Splitting Algorithms

In this section, we introduce some traffic splitting algorithms. We introduce Join the Shortest Queue (JSQ) scheme, which is a common scheme used in queuing systems. We use JSQ as a benchmark scheme for comparison of our algorithms. This is followed by two heuristic traffic splitting algorithms that we propose.

3.3.1 Join the Shortest Queue (JSQ) Scheme

In this scheme, the arrivals join the subsystem with the shortest queue. The action is chosen as below:

$$a = \begin{cases} \text{Accept arrival in SeNB subsystem if } s_1 > s_2, \\ \text{Accept arrival in MeNB subsystem if } s_1 \leq s_2, \\ \text{Block, if both subsystems are full.} \end{cases} \tag{3.20}$$

This scheme routes the arriving traffic to the subsystem, which is less occupied. In this scheme, if both the subsystems have exhausted their resources, then incoming traffic is blocked. Note that the above scheme is followed for foreground traffic only. Background traffic is routed to the MeNB subsystem whenever resources exist in the subsystem; else it is blocked.

3.3.2 Proposed Heuristic Traffic Splitting Algorithms

In this section, we propose two heuristic traffic splitting algorithms based on the MDP system model discussed in Section **??**.

Join the Shortest Queue with Proactive Blocking Algorithm (JSQPB)

Join the Shortest Queue is a policy where the traffic is routed to the subsystem with fewer number of packets. JSQPB combines proactive blocking along with join the shortest queue policy. This algorithm chooses the action with minimum cost, which is blocking, whenever the blocking probability is strictly within the threshold B_{max}, else it follows join the shortest queue policy.

JSQPB algorithm (see Algorithm **??**) starts by initializing blocking probability for the foreground (BP_f) and background traffic (BP_b) to zero (Line 2). Whenever an arrival occurs in the system (Line 4), the fraction of foreground arrivals blocked is determined (Line 5) by dividing the number of foreground arrivals blocked by the total number of foreground arrivals until that time. Similarly, Line 6 determines the fraction of background arrivals blocked. Line 7 calculates the total blocking probability as a weighted sum of the two blocking probabilities determined in Lines 5 and 6. If the value of blocking probability obtained is less than allowed maximum value B_{max} minus an infinitesimal value ϵ (Line

Algorithm 1 Join the Shortest Queue with Proactive Blocking Algorithm (JSQPB)

1: Input: $\lambda_1, \lambda_2, \mu_m, \mu_s, \delta, c(\cdot), B_{max}$.
2: Initialization : $BP_f \leftarrow 0, BP_b \leftarrow 0$.
3: **while** True **do**
4: **if** Arrival **then**
5: $BP_f \leftarrow$ Fraction of foreground arrivals blocked.
6: $BP_b \leftarrow$ Fraction of background arrivals blocked.
7: Determine $BP = \delta BP_f + (1-\delta)BP_b$.
8: **if** $BP < B_{max} - \epsilon$ **then**
9: $a^* \leftarrow a_0$ - Block
10: **else**
11: **if** $s_1 < s_2 + s_3$ **then**
12: $a^* \leftarrow a_3$ - Route to MeNB
13: **else**
14: $a^* \leftarrow a_1$ - Route to SeNB
15: **end if**
16: **end if**
17: **end if**
18: **end while**

8), then it chooses the action with the minimum cost from the set of feasible actions for that state (Line 9). This action is a_0 (block) for all states since the action of blocking provides a minimum delay in each state. If the condition in Line 8 is not satisfied, then it checks which subsystem has a fewer number of packets (Line 11). This scheme considers backhaul subsystem and SeNB subsystem as one subsystem. If the MeNB subsystem has lower number of packets as compared to the SeNB subsystem, then it routes the arrival to the MeNB subsystem, else it routes it to the SeNB subsystem. This algorithm chooses the action with minimum cost and proactively blocks the arrivals, whenever the allowed limit of blocking probability ($B_{max} - \epsilon$) is available.

Algorithm 2 Modified Join the Shortest Queue with Proactive Blocking Algorithm (MJSQPB)

Input: $\lambda_1, \lambda_2, \mu_m, \mu_s, \delta, c(\cdot), B_{max}$.
Initialization : $BP_f \leftarrow 0, BP_b \leftarrow 0$.
while True **do**
  if Arrival **then**
    $BP_f \leftarrow$ Fraction of foreground arrivals blocked.
    $BP_b \leftarrow$ Fraction of background arrivals blocked.
    Determine $BP = \delta BP_f + (1-\delta)BP_b$.
    if $BP < B_{max}$ **then**
      $a^* \leftarrow a_0$ - Block
    else
      for Each packet in arrival **do**
        if $s_1 < s_2 + s_3$ **then**
          $a^* \leftarrow a_3$ - Route the packet to MeNB
        else
          $a^* \leftarrow a_1$ - Route the packet to SeNB
        end if
      end for
    end if
  end if
end while

Modified Join the Shortest Queue with Proactive Blocking Algorithm (MJSQPB)

We propose Modified Join the Shortest Queue with Limited Blocking (MJSQLB) Algorithm, which is similar to JSQPB algorithm with one modification. In this algorithm, join the shortest queue policy is followed for each packet, instead of a batch. For instance, if a batch of two packets arrives, then the first packet is routed to the subsystem with a fewer number of packets. For the second packet, we again determine the subsystem with a fewer number of packets and then route the second packet to it.

MJSQPB algorithm (see Algorithm **??**) determines the fraction of background and

foreground arrivals blocked (Lines 5 and 6). The total blocking probability (BP) is then determined in Line 7. If the value of blocking probability obtained is less than the allowable limit $B_{max} - \epsilon$, then choose the action from the set of feasible actions (in that state) with minimum cost, which is a_0 (block) (Line 9). Else, for each packet in the arrival, the algorithm determines the subsystem with a fewer number of packets (Line 12) and routes the packet to that subsystem. This algorithm chooses proactively blocks the arrivals, whenever the allowed limit of blocking probability ($B_{max} - \epsilon$) is available, else it opportunistically routes each packet to the subsystem with a fewer number of packets.

3.4 Computational Complexity and Implementation Analysis

In this section, we analyze the computation complexity of the optimal policy and that of the proposed traffic splitting algorithms. Let $|\mathcal{S}|$, $|\mathcal{A}|$ denote the cardinality of the state space and action space, respectively. The policy iteration algorithm [**?**] to obtain the optimal policy determines the optimal policy by traversing the entire policy space. Hence, the computation complexity of the policy iteration algorithm is $O(|\mathcal{A}|^{|\mathcal{S}|})$, since there are at most $|\mathcal{A}|^{|\mathcal{S}|}$ distinct policies. This algorithm, therefore, terminates in an exponential number of iterations.

The computation complexity of the proposed algorithms (JSQPB, MJSQPB) is the same since these algorithms are simple and only depend on the state of the system. The algorithms are executed online when arrivals occur, and the computation complexity of proposed algorithms is $O(1)$ since we do not need any computation for these algorithms. As compared to the computation complexity of traditional policy iteration algorithm ($O(|\mathcal{A}|^{|\mathcal{S}|})$), the complexity of the proposed algorithms is reduced from exponential to a constant.

The optimal policy is obtained offline and stored in the format of a table. The space complexity for the optimal policy is $O(|\mathcal{S}|)$. The proposed algorithms are computed online and their space complexity is $O(1)$. The proposed heuristic algorithms are thus considerably efficient than the classical optimal algorithm.

The optimal policy requires the knowledge of arrival rates, cost function, Lagrangian

Multiplier, in addition to other parameters. The optimal policy chooses to block the arrivals for which the total delay would have been large. The knowledge of the system enables the optimal policy to provide the minimum delay among all policies. The proposed algorithms proactively block the traffic when the blocking probability is less than the allowable limit, and at other times they allow the packets in the system. In these proposed algorithms, the blocking of traffic does not depend on any system parameter. The advantage of the proposed policies is that they are simple to implement in a general system, without the need to know any parameters of the system. Thus, the algorithms have low computation as well as space complexity, and they can be easily implemented without the knowledge of traffic arrival rates.

3.5 Numerical Results and Discussions

In this section, we compare the performance of the optimal policy with the proposed algorithms and the JSQ scheme. We outline the simulation model next.

3.5.1 Simulation Model

The simulation scenario consists of an LTE macro node and an overlapping LTE small cell node situated at a distance of 250m from the macro node. The small cell node is connected to the macro node via a wired link with a latency which is exponentially distributed with parameter μ_d secs. Background users are situated in the coverage area of the small cell. Foreground users are situated in the non-overlapping area of the macro cell. The network parameters used for simulation are based on 3GPP standards [**?**] and are enlisted in Table **??**. The algorithms are implemented in ns-3 [**?**], a discrete-event simulator using 3GPP standardized parameters and extensive simulations are performed to average over a large number of randomized scenarios. A special downlink application is developed in ns-3 to simulate the batch type of traffic. The traffic destined for users is generated at a remote server. On reaching the gateway, the traffic is routed appropriately based on the traffic splitting algorithms. The total delay encountered by each packet is noted and averaged over a fixed number of packets.

The parameters of the system model used for simulations are enumerated in Table

Table 3.2: Network parameters.

Parameter	**Value (Macro node, Small cell node)**
Macro ISD, Pico radius	1000 m, 200 m
Transmit power	46 dBm, 30 dBm
Frequency	2, 2.6 GHz
Antenna	Omnidirectional, Omnidirectional
Antenna height	32 m, 12 m
Path loss (d_0 in km)	128.1 + 37.6 log(d_0), 140.7 + 36.7 log(d) dB
Fading	3GPP Extended Typical Urban (ETU) model

??. The system parameter values have been chosen to keep the system computationally tractable. We assume that $\mu_s \geqslant \mu_m$. This is because the achievable rate in the coverage area of a small cell is typically higher than that in the whole macro cell [**?**]. Although for the computation purpose, we assume a maximum batch size of two, the analysis presented in this chapter holds for any general batch size. The performance metrics are the average delay in the system and the blocking probability of arriving traffic.

Table 3.3: System model parameters.

Parameter	**Value**
Batch size Distribution (α_1, α_2)	0.5, 0.5
Number of Resources (n_1,n_2)	6, 6
Queue size (B)	5
Service rates (μ_m, μ_d, μ_s)	3, 20, 3.2 packets/s
Blocking probability weight parameter (δ)	0.5

We consider two scenarios wherein the users are stationary and mobile, respectively, and present their analysis in subsequent subsections.

3.5.2 Stationary User Scenario

In this section, we present the comparison of different algorithms in a scenario where users are stationary. We evaluate the performance of the algorithms in terms of average delay and blocking probability of arriving traffic in the system.

Foreground arrival rate variation

Figures **??** and **??** illustrate the plots of average system delay and the fraction of traffic blocked for different algorithms under varying foreground traffic arrival rate λ_1 and background traffic arrival rate fixed at $\lambda_2 = 1$ batches/s. The optimal policy gives minimum delay as compared to other algorithms. The background traffic in the system is less and uses the resources in the MeNB subsystem. Initially, the amount of foreground traffic in the system is low, and it uses the resources in the SeNB subsystem without queuing. As foreground arrival rate increases, foreground traffic uses resources of both MeNB and SeNB subsystem opportunistically, and background traffic uses resources of the MeNB subsystem. As λ_1 increases from 2 to 16 batches/s, the foreground traffic in the system increases utilizing the SeNB subsystem completely and the MeNB subsystem partially. The queuing delay in the system rises, leading to an increase in average delay in the system.

The blocking probability constraint B_{max} has been set to 0.04. JSQPB and MJSQPB algorithms proactively block the incoming traffic whenever the margin of blocking probability B_{max} is available. So their blocking probability is almost constant at $B_{max} - \epsilon$. For JSQ, traffic is only blocked when the system is full. Hence, JSQ gives low blocking probability, and it increases with increase in λ_1. When λ_1 is low ($\lambda_1 \leq 8$), both MeNB and SeNB subsystems occupy a fewer number of packets and rarely reach their capacities. The optimal policy blocks more traffic as the margin of B_{max} is available. As λ_1 increases ($\lambda_1 > 8$), the optimal policy has to accommodate a large amount of traffic, and both MeNB and SeNB subsystems are completely utilized. Hence, the optimal policy blocks less amount of traffic. JSQ is a simple algorithm which routes the arriving traffic to the subsystem with the lowest state. However, it provides a higher average delay as compared to other algorithms. JSQPB and specifically, MJSQPB perform quite close to the optimal policy.

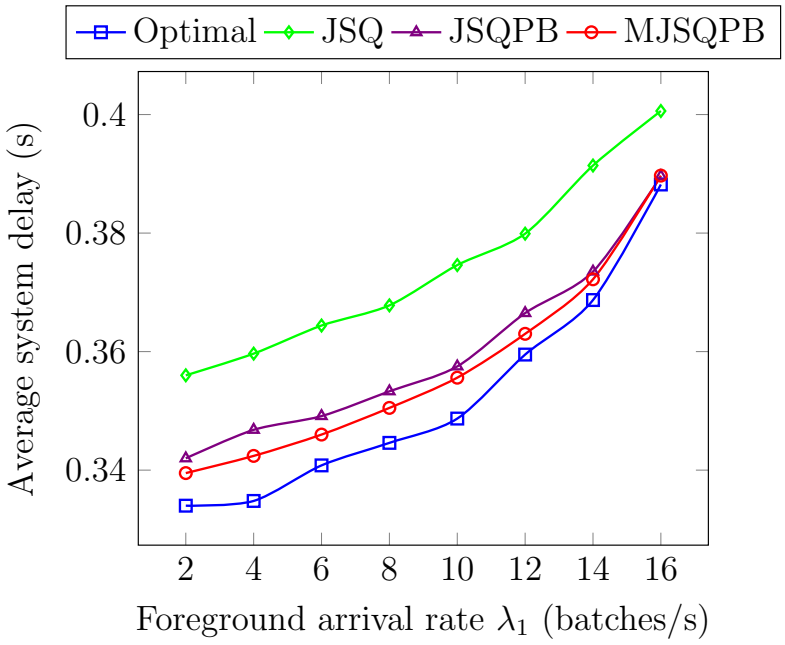

Figure 3.2: Plot of average system delay vs foreground arrival rate (λ_1) for different algorithms.

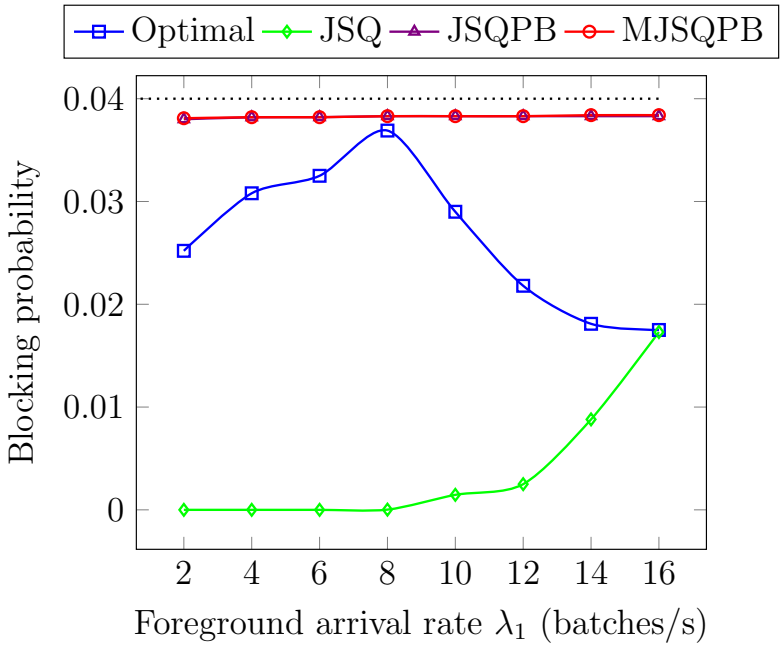

Figure 3.3: Plot of blocking probability vs foreground arrival rate (λ_1) for different algorithms.

Background arrival rate variation

Figures **??** and **??** depict the average system delay and the fraction of traffic blocked for various algorithms under varying λ_2 and foreground traffic arrival rate fixed at $\lambda_1 = 1$

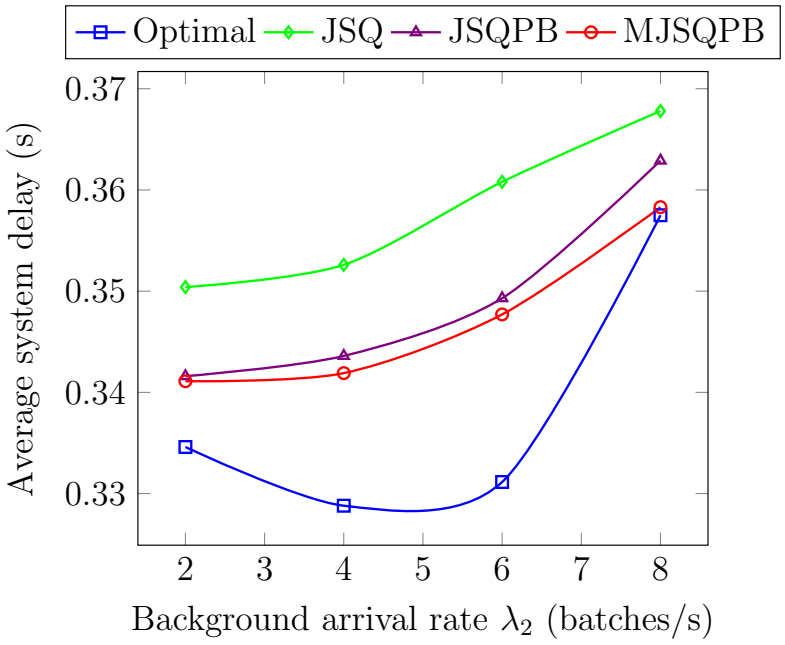

Figure 3.4: Plot of average system delay vs background arrival rate (λ_2) for different algorithms.

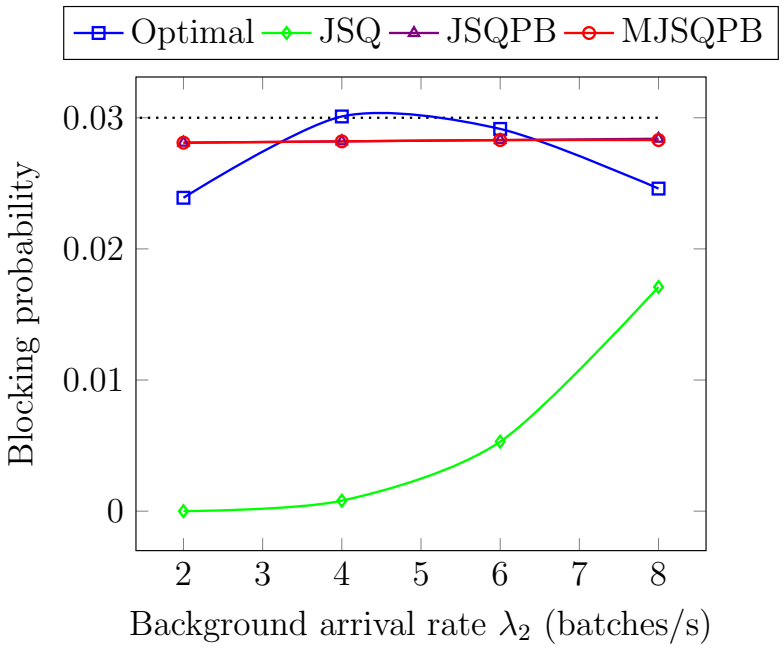

Figure 3.5: Plot of blocking probability vs background arrival rate (λ_2) for different algorithms.

batches/s. The optimal policy gives minimum average delay in the system. Initially, $\lambda_1 = 1$, and $\lambda_2 = 2$. The amount of traffic in the system is low, and traffic is also blocked, as shown by the blocking probability plot of optimal policy. Hence, the average delay in

the system is reduced. As λ_2 increases from 4 to 6 batches/s, background traffic in the system increases whereas foreground traffic remains the same. Foreground traffic is routed to the SeNB subsystem, and background traffic uses resources of the MeNB subsystem. A significant amount of traffic is blocked as well at this time. Hence, the average delay in the system is low. As λ_2 increases from 6 to 8 batches/s, both MeNB and SeNB subsystems are completely occupied, and background traffic packets start queuing at the MeNB. Therefore, queuing delay in the system increases, thereby raising the average delay in the system steeply.

After the optimal policy, MJSQPB algorithm gives minimum delay followed by JSQPB algorithm. MJSQPB algorithm proactively blocks the arrivals till the margin $B_{max} - \epsilon$ is reached, else it routes each packet to the subsystem with a fewer number of packets. JSQPB algorithm proactively blocks the arrivals till the margin $B_{max} - \epsilon$ is reached, else it routes the arrival to the subsystem with a fewer number of packets. So their blocking probability is almost constant at $B_{max} - \epsilon$. As λ_2 increases, blocking probability of these two algorithms remains constant, and the average delay in the system increases as the number of packets in the system increase. JSQ algorithm routes the traffic to the subsystem with a fewer number of packets and only blocks the traffic when the system is full. Hence, its blocking probability rises steadily as λ_2 increases.

3.5.3 Mobile User Scenario

In this section, we present the comparison of different algorithms in a scenario where users are mobile. The users move with a random speed in a random direction until they reach a bounding box, at which point, they rebound with a reflexive angle and speed. Their speed is uniformly distributed between 0 and 36 kmph. We evaluate the performance of the algorithms in terms of average delay and blocking probability of arriving traffic in the system.

Foreground arrival rate variation

Figures **??** and **??** illustrate the variation in average system delay, and the fraction of traffic blocked for different algorithms under varying foreground arrival rate λ_1, when the users are mobile. λ_1 is varied from 2 to 16 batches/s, and other parameters such as

$\lambda_2 = 1$ batches/s are kept constant. We observe that the average system delay steadily increases as λ_1 increases from 2 to 16 batches/s. The variation in average system delay and blocking probability is similar to the case when the users are stationary. The optimal policy gives minimum delay, followed by MJSQPB algorithm. JSQPB algorithm also gives almost similar performance as the MJSQPB algorithm, since it is very similar to MJSQPB algorithm. In comparison to other algorithms, JSQ provides a higher average delay.

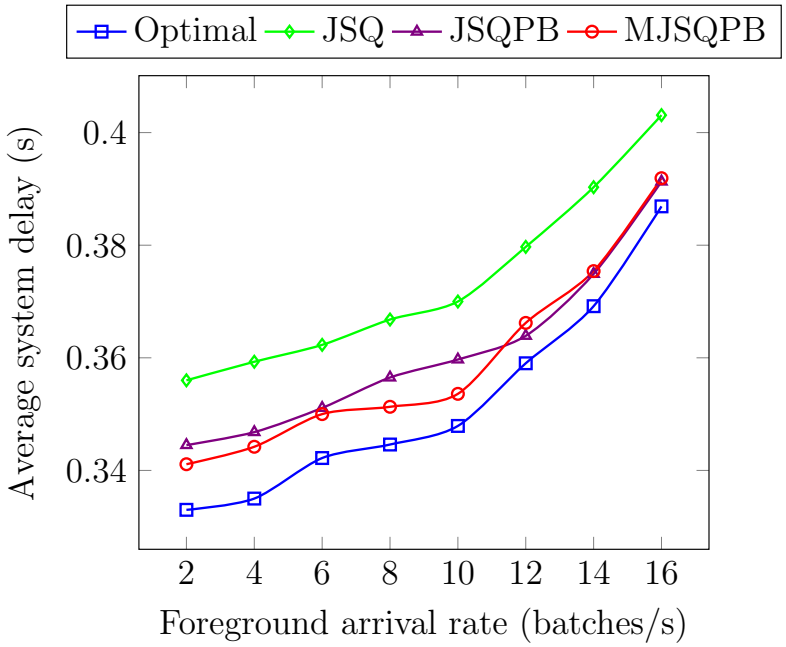

Figure 3.6: Plot of average system delay vs foreground arrival rate (λ_1) for different algorithms when the users are mobile.

Background arrival rate variation

Figures **??** and **??** depict the variation in average system delay, and the fraction of traffic blocked for different algorithms under varying background arrival rate λ_2, when the users are mobile. λ_2 is varied from 2 to 8 batches/s, and the foreground traffic arrival rate is kept constant at $\lambda_1 = 1$ batches/s. The optimal policy provides a minimum average delay in the system as compared to other policies. The variation in the average system delay and blocking probability is similar to the case when the users are stationary. MJSQPB and JSQPB algorithms give average delay close to that of the optimal policy. JSQ provides a higher average delay as compared to other algorithms.

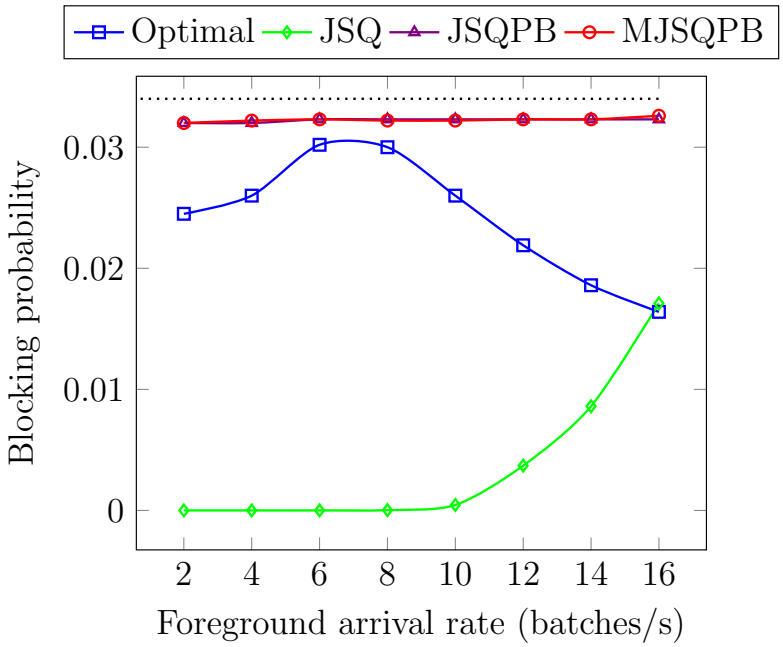

Figure 3.7: Plot of blocking probability vs foreground arrival rate (λ_1) for different algorithms when the users are mobile.

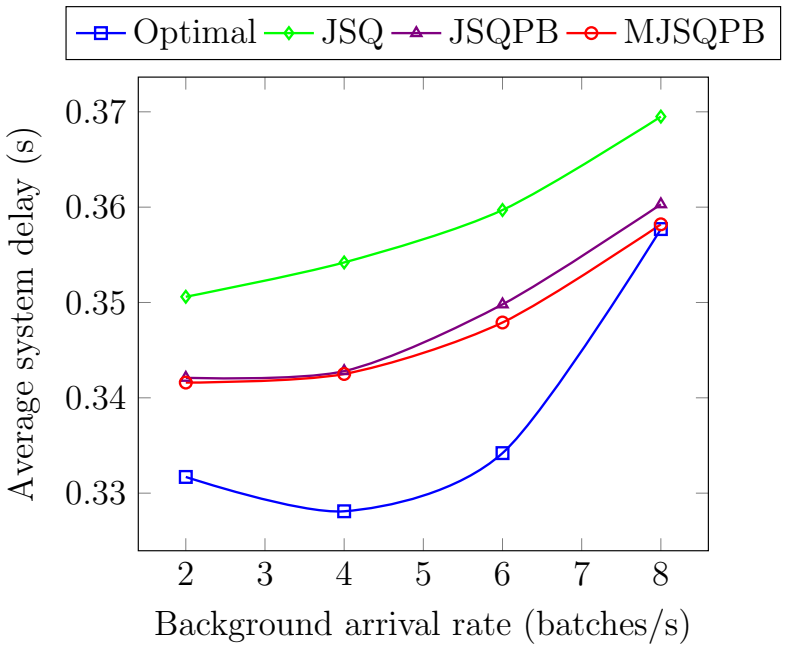

Figure 3.8: Plot of average system delay vs background arrival rate (λ_2) for different algorithms when the users are mobile.

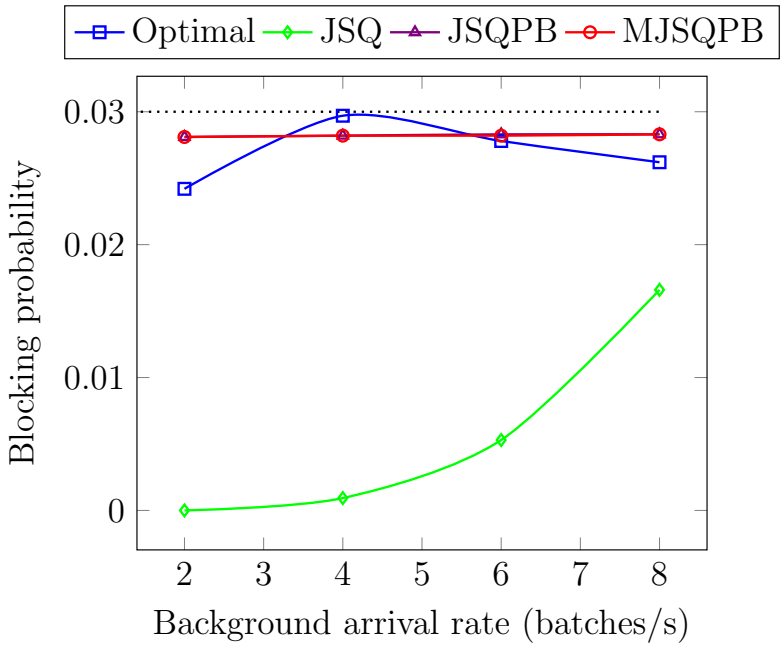

Figure 3.9: Plot of blocking probability vs background arrival rate (λ_2) for different algorithms when the users are mobile.

3.6 Conclusions

In this chapter, we have formulated the optimal traffic splitting problem for foreground traffic to minimize the average delay in the system subject to a constraint on the blocking probability of background as well as foreground traffic. CMDP framework is used to model the system, and the optimal policy is obtained using uniformization, Lagrangian approach, and value iteration algorithm. We have proposed two heuristic algorithms, viz., JSQPB, and MJSQPB, which, although suboptimal, are simple to implement in practice with low computational and space complexity as compared to the optimal policy. We have implemented the optimal policy, JSQPB, MJSQPB, and JSQ algorithms in ns-3 simulator in an LTE specific scenario and performed extensive simulations. We have presented the simulation results, which depict the variation of the system delay and blocking probability for these algorithms under varying arrival rates of foreground and background traffic.

We demonstrate that the MJSQPB algorithm performs better than other algorithms. Further, in MJSQPB and JSQPB algorithms, average delay values are close to that obtained in the optimal policy in case of both stationary and mobile users scenario. JSQ and JSQPB algorithms only route traffic to a single subsystem. They do not split the

incoming traffic, whereas the optimal policy and MJSQPB algorithm use traffic splitting and route each packet to an appropriate subsystem, and the delays provided by the optimal policy and MJSQPB algorithm are lowest compared to JSQ and JSQPB. Thus, we demonstrate that splitting traffic appropriately between the two subsystems can be used to achieve minimum delay in the system. Thus, dual connectivity indeed minimizes the delay in the system by splitting arriving traffic appropriately between the two subsystems. Although we have demonstrated the results in an LTE specific system, the framework is generic and can be extended to other RATs as well.

Chapter 4

SDN Framework for Multi-RAT Dual Connectivity

In a HetNet, nodes belonging to multiple RATs coexist with each other. Varied modes of dual connectivity (see Section **??**) are feasible in such a network. The UEs should be able to make use of and change into these varied modes of dual connectivity with ease and flexibility. However, the varied modes of dual connectivity are not consistent with each other. As discussed in Chapter **??**, dual connectivity between LTE Master Node (MN) and LTE Secondary Node (SN) is defined in [**?**], DC between LTE MN and WLAN SN is characterized in [**?**] as LTE WLAN Aggregation (LWA), and DC between LTE/5G MN and LTE/5G SN is known as Multi-RAT Dual Connectivity (MR-DC) [**?**]. Control signaling procedures in these different types of DC are not uniform. Moreover, interfaces between MN and SN are defined differently in each of these variants of DC. Further, the volume of control information to be exchanged between MN and SN for a dual connected UE is significant because of a single control plane connection to the Core Network (CN) via MN. This increases the control load on each of the MNs, which again increases with the number of dual connected UEs. An increased control load leads to additional delays in the system since the backhaul link between MN and SN is non-ideal. Due to these issues, the existing RAN architecture is not suitable for the management of dual connected UEs in a multi-RAT network. One of the approaches to resolve these issues is to use the Software-Defined Networking (SDN) paradigm. SDN is an architectural approach to network management. It facilitates dynamic and efficient, centrally programmed network

configuration. As detailed in Chapter **??**, SDN decouples control plane from the data plane, thus bringing flexibility and efficiency in network control. In this chapter, we introduce SMRAN - an SDN based Multi-RAT RAN architecture, which resolves the problems in existing architecture.

SDN framework is exploited in existing works [**?**], [**?**] to bring agility in network management. In [**?**], the authors propose a software-defined RAN consisting of a centralized RAN controller with individual radio elements. The centralized controller takes decisions that affect the global state of the network. Each radio element takes local control decisions which do not affect the neighboring elements. However, this work deals with LTE networks only. As opposed to this, we propose a fully centralized control plane architecture for multiple RATs. In [**?**], the authors propose a centralized SDN based wireless network controller for controlling multi-RAT nodes as well as gateways in CN. However, in our work, we focus on SDN based control of interworking among multiple RATs at the RAN level from the perspective of dual connectivity.

In [**?**], the authors propose a multi-RAT RAN architecture for multi-connectivity. They propose a common control-plane and user-plane for the joint operation of different multi-RAT technologies such as LTE, 5G, and WLAN. The control-plane functionalities of 5G, LTE, and WLAN are controlled by a common control-plane layer. The common data-plane functionalities of these RATs are controlled by a common data-plane layer. The common control-plane and data-plane reside in a RAN cloud to control the radio entities from a centralized location. However, this architecture is not based on SDN, and although there is centralized control over the RAN, there is no mention of the design of the common data-plane protocol. Moreover, the interfaces between the radio entities are the same as in legacy architectures. As opposed to this, our architecture has a clear separation of data-plane and control-plane.Moreover, a common interface for communication between data-plane entities has been defined to maintain uniformity in the architecture.

The main aspects of SMRAN are as follows:

1. It is an SDN based multi-RAT RAN architecture for DC, wherein the control plane and data plane entities are separated.

2. The architecture integrates 3GPP RATs (LTE and 5G NR) with non-3GPP access technologies (WLAN) and devises a common standard interface for communication

across RATs.

3. SMRAN integrates the following variants of DC across different RATs, and simplifies their signaling procedures. i) LTE DC, ii) LWA, and iii) MR-DC.

4. The architecture allows integration of RAN with both 3GPP 4G Core as well as 5G Core. It uses a unified interface between RAN and CN irrespective of the RAT (LTE, 5G NR, WLAN) being used.

5. It uses a common interface (Xd) for exchange of data between RAN data plane entities.

6. It provides flexibility in setting up control plane communication paths over a radio interface (signaling radio bearers) for a dual connected UE at any network node.

7. It provides flexibility in processing the data plane protocol layers for a dual connected UE across different network nodes.

8. SMRAN is structured to reduce the control signaling overhead for a dual connected UEs.

9. It improves the load and mobility management in the network.

10. It enhances the performance of the system, i.e., in terms of system throughput.

The rest of the chapter has been organized as follows. In Section **??**, we outline the motivation behind the architecture. An in-depth explanation of SMRAN is provided in Section **??**. In Section **??**, we show that there is a significant reduction in control signaling exchange in SMRAN. In Section **??**, we propose a centralized scheme for selecting dual connected UEs, which outperforms a general distributed scheme. The novel features of SMRAN are detailed in Section **??**. We present the conclusions in Section **??**.

4.1 Motivation

In this section, we discuss the limitations of existing 3GPP architectures. In LTE dual connectivity, as explained in Section **??**, there is a single RRC connection, and the Signaling Radio Bearer (SRB1 and SRB2) is served by the master node only. When secondary

node has to share radio control information with UE, it needs to send this information to the MN over the X2 interface, which in turn sends it to UE. Further, SN has to get the data bearer related control information from the MN through the X2 interface. This leads to an exchange of a large number of signaling messages between MN and SN for a dual connected UE. The amount of control information exchange increases with an increase in the number of dual connected UEs in the network. Moreover, this exchange takes place over the non-ideal backhaul link between MN and SN, which has a latency of the order of milliseconds [**?**] leading to additional delays in the signaling exchange.

As described in Section **??**, in LWA, there is a single UE specific signaling connection between CN and eNB. Further, all control information exchange with UE takes place via eNB, i.e., both SRB1 and SRB2 are carried by eNB. The control information is exchanged between WLAN Termination (WT) and eNB via the Xw-C interface. In MR-DC, as mentioned in Section **??**, there are three types of Signaling radio bearers: SRB1 and SRB2 can be split across both MN and SN, and SRB3 is through SN. The initial signaling, however, can take place through MN only.

Thus, the SRB establishment procedure is not uniform across different DC architectures. A significant amount of control information exchange takes place between MN and SN via the non-ideal backhaul link causing additional delay. In addition, there are different interfaces defined for communication between different types of RAT nodes. For example, the X2 interface is defined for communication between eNBs, the Xw interface for communication between eNB, and WT and the Xn interface for communication between eNB/gNB and gNB.

In the existing architecture, the control-plane functions are distributed across the network nodes in RAN. It is challenging to get load information of all the cells at a central location. Hence, the traditional method of connecting a UE to a node having the best signal strength fails to take into account the possibly differing load levels across cells in a heterogeneous environment. Note that load and interference management information is shared via X2/Xw/Xn interfaces between two network nodes for load balancing and interference coordination between them. This sharing, however, typically takes place between a small number of neighboring nodes only. Hence, the decisions taken for a dual connected UE association locally by a network node may not be optimal.

To simplify the complexity of the existing architecture and to achieve better load balancing in the network, we propose SMRAN - an SDN based multi-RAT RAN architecture, where the control and data planes are separated. As a result, signaling and data flow management are simplified. Further, data plane nodes are controlled by a centralized controller which has a global view of the load and interference conditions in the network. Thus, the problem of selection of appropriate MN and SN for a dual connected UE can be addressed efficiently.

4.2 Description of SMRAN

In this section, we describe SMRAN in detail. Figure **??** illustrates the SMRAN architecture, along with the protocol stacks. In this architecture, there is no change in the CN, whereas RAN is divided into two parts:

- An SDN RAN Controller (SRC) with the combined control plane functionality of LTE eNBs, 5G gNBs, and WLAN WT points.
- Diverse data plane entities, also called d-eNB (data-eNB), d-gNB (data-gNB), and d-WT (data-WT) - created out of LTE eNB, 5G gNB, and WT, respectively, by removing their control functions.

The CN can comprise LTE Core or 5G Core (5GC). The data plane entities can further be grouped into Centralized Data-plane Units (CDUs) facing the CN and Distributed Data-plane Units (DDUs) facing the UEs. CDUs can act as RAN data plane aggregation units, which combine the data from a group of data plane entities and send it to the CN. The data plane entities which act as DDUs have radio access to the UEs.

In legacy LTE architecture, each eNB handles RRM and control functions for the respective UEs associated with them. However, in SMRAN, RRM, control plane, and the associated protocol layers of LTE eNB (RRC, RRM, and S1-AP) have been transferred to the SRC. Similarly, the control functions of 5G gNB (RRC, RRM, and Next Generation Application Protocol (NG-AP)) has been shifted to the SRC. SRC implements the control functions of WT (WT-C). Control functions of all the nodes are thus, centralized at SRC. There is no control information exchange between the data plane entities. This results in

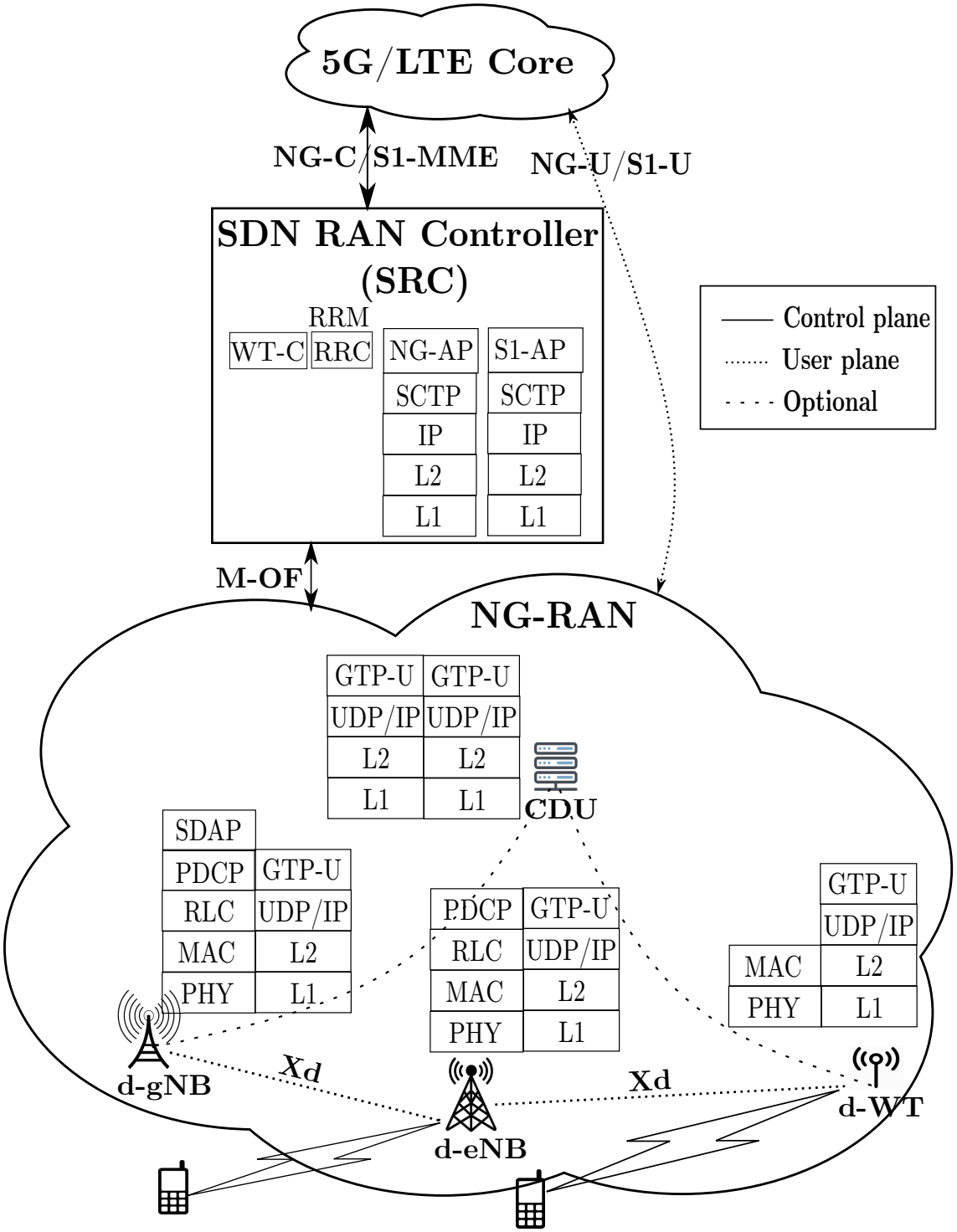

Figure 4.1: SMRAN: an SDN based Multi-RAT RAN architecture for dual connectivity.

a reduction in the amount of signaling required in DC operations such as cell addition, cell change, and cell release, as demonstrated in the next section.

The data plane entities consist of data plane functions as well as the associated protocol layers for communicating with a UE over the air interface. For instance, the protocol layers of Service Data Adaptation Protocol (SDAP), PDCP, RLC, MAC, PHY of gNB and PDCP, RLC, MAC, PHY of eNB exist in d-gNB and d-eNB, respectively to exchange data with a UE over the air interface. Similarly, d-WT hosts the protocol stack for communicating data with a UE over the WLAN interface. We define a new interface Xd for the exchange of data between the RAN nodes. Xd is a common interface for communication between data plane nodes belonging to different RATs. Xd interface comprises GTP-U, UDP/IP, L2, L1 protocol stack. Data is exchanged between the CN and data plane entities through the Next Generation User-plane (NG-U) or S1-U user-plane interface using GTP-U, UDP/IP, L2, L1 protocol stack.

SRC uses the Next Generation Control plane (NG-C) or S1-MME interface for the exchange of control information with the CN depending on whether it is communicating with 5G or LTE CN. Thus, there is a unified interface between SRC and CN irrespective of the RAT being used.

SRC has a global view of all data plane entities to take effective decisions. For instance, based on the radio channel, memory, and load information for d-gNB, d-eNB, and d-WT at the SRC, an appropriate path for a bearer can be chosen.

SRC controls the data plane entities with a southbound interface using a Modified OpenFlow (M-OF) protocol. The data plane entities such as d-eNBs act as OpenFlow switches and the SRC acts as the OpenFlow Controller in the proposed architecture. The main function of M-OF protocol is the establishment of a path between UE and radio data plane entities. For example, one of the functions of M-OF protocol is to establish radio bearers between d-eNB and UE. Once radio bearers are established between UEs and d-eNBs, control signaling information (through SRBs) and data (through DRBs) can be exchanged between the SRC/CN and UEs. M-OF protocol is based on OpenFlow (OF) [**?**] and OF-Config [**?**] protocols with some modifications.

An OpenFlow physical port corresponds to a hardware interface of an OpenFlow switch [**?**]. An OpenFlow logical port is a higher-level abstraction that is defined on

top of OpenFlow physical ports. The OF logical port is similar to the OF physical port but may have a tunnel-id as an extra meta-data field associated with it [**?**]. These OF concepts need to be mapped to wireless networks, especially cellular networks, after some modifications to define the M-OF protocol. As explained below the concepts of physical ports and logical ports or tunnels in OpenFlow are adapted for the M-OF protocol.

We explain M-OF protocol using an example. Consider the data communication between CN and UE via a d-eNB. The radio link between UE and d-eNB can be treated as an OF physical port, and each bearer to a UE can be treated as a logical port. The Temporary Mobile Subscriber Identity (TMSI), the UE id available at the base station, along with the bearer id (tunnel-id) is used to identify the logical port. The GTP tunnels corresponding to individual radio bearers for a UE (both DRBs/SRBs) towards the CN are directly mapped to the concept of OF tunnels (logical ports). The IP address of d-eNB along with the UDP port number and the GTP tunnel-id can be treated as the logical port on the d-eNB side on the S1-U interface.

OF-Config protocol is used to configure the OF physical ports (radio links for UEs) and OF logical ports (radio bearers). The radio bearer (logical port) configuration requires LTE PDCP/RLC layer configuration parameters to be supplied by the SRC to d-eNB, while radio link (physical port) configuration requires LTE MAC and PHY layer parameters to be sent by SRC to d-eNB. The existing OF physical port and logical port configuration parameters in OF-Config do not use these PDCP/RLC/MAC/PHY layer parameters. In M-OF protocol, we add these parameters to the existing parameters of OF physical and logical ports to adapt it for the wireless network. OpenFlow protocol supports tunneling mechanisms like Generic Routing Encapsulation (GRE) but does not support GTP tunneling at present. Therefore, OF protocol is modified (M-OF) to support GTP tunneling.

Analogous to OF interface, M-OF interface uses a flow table with (port id, action) tuples. The port id represents the ingress port where the packets belonging to the flow (bearer) enter. For instance, in the case of uplink, it consists of UE id and bearer-id. The action consists of the 'output' action, where the flow gets routed to a specific output port. The output port can be the IP address of d-eNB along with the UDP port number and the GTP tunnel-id. Each GTP tunnel on the S1-U interface of the d-eNB is mapped

to a DRB (tunnel) on the radio interface of d-eNB through this flow table configuration, which is provided by the SRC. Thus, OF-Config is used to configure the ports, and OF is used to route flows through d-eNB based on the port configuration.

4.3 Signaling Reduction in SMRAN

In this section, we describe the signaling procedure for dual connectivity mobility events of small cell addition, small cell change, and small cell release in case of legacy as well as SMRAN architectures. We then demonstrate the reduction in the control signaling on account of SMRAN through simulations.

4.3.1 Control Signaling Procedure in Dual Connectivity

In the case of DC, SeNB addition is initiated by MeNB and is used to establish UE context at the SeNB to provide additional resources from the SeNB to UE. Figure **??** describes the procedure to add an SeNB in legacy LTE [**?**]. The procedure is as follows. MeNB sends addition request to SeNB to allocate radio resources for a Radio Access Bearer (RAB) indicating RAB parameters. In SCG-ConfigInfo, MeNB provides information about MCG configuration, UE capabilities as well as SeNB measurements. In the second message, SeNB allocates radio resources for the UE and sends necessary information in SCG-Config message. MeNB then sends RRCConnectionReconfiguration message to the UE which includes new radio resource configuration of SCG. UE applies the new configuration and replies with RRCConnectionReconfiguration Complete message (Message 4). MeNB, then informs the SeNB that UE has completed the reconfiguration process (Message 5). UE then synchronizes with the SeNB using Random Access Procedure (Message 6). For the transfer of data from MeNB to SeNB, MeNB sends packet Sequence Number status to SeNB (Message 7). The data received by MeNB from SGW is forwarded to SeNB (Message 8). E-UTRAN RAB update procedure takes place, where the path from PGW to MeNB is changed to PGW to SeNB (Messages 9 through 12).

In SMRAN, there exists a centralized RAN controller, which possess all measurement data, interference data, load information as well as channel information for all eNBs and UEs. The procedure for the addition of Small cell d-eNB (Sd-eNB) in SMRAN is presented

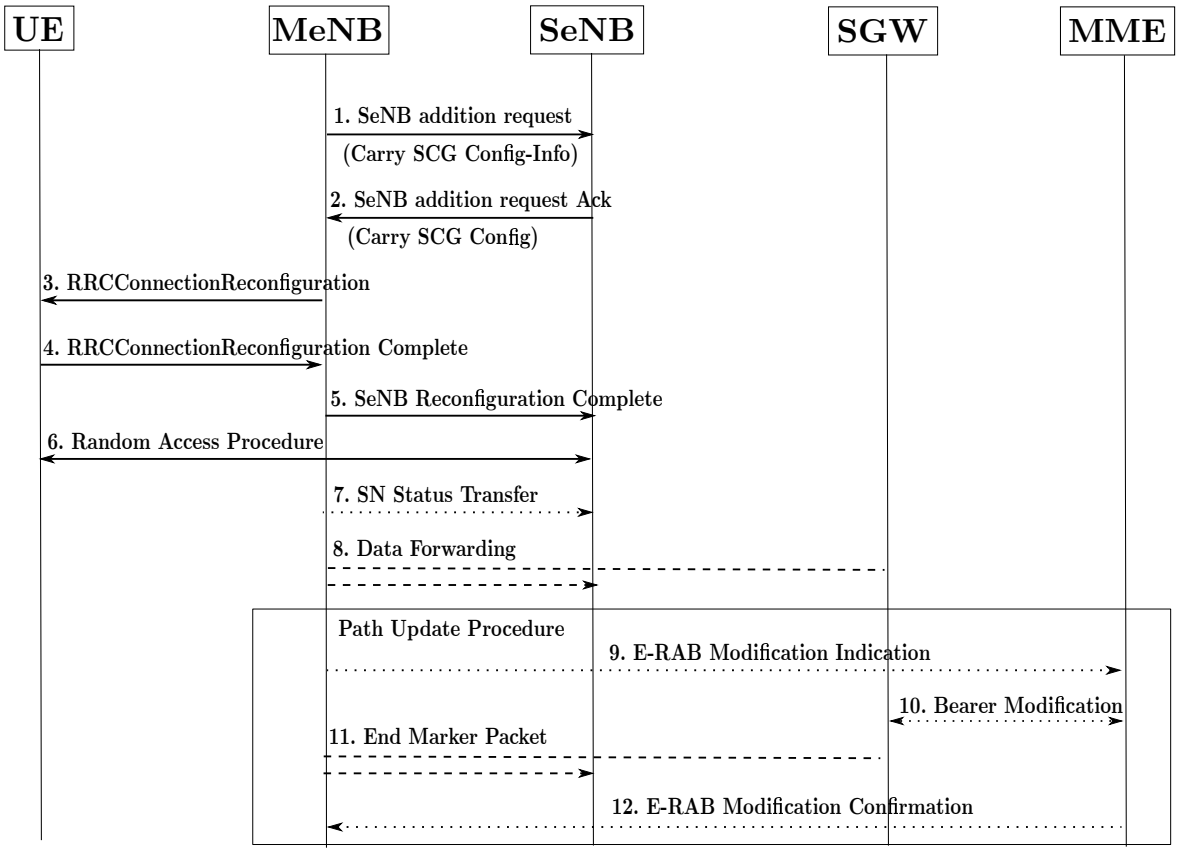

Figure 4.2: SeNB addition procedure in LTE (Adapted from [**?**]).

in Figure **??**. When an Sd-eNB has to be added to a UE for DC, the controller sends Sd-eNB Flow addition command to the Sd-eNB, which contains the bearer characteristics as well as data layers (PDCP, RLC, MAC, PHY) configuration data. The SRC then sends RRCConnectionReconfiguration to UE via the Macro d-eNB (Md-eNB). UE applies the connection configuration and acknowledges it (Message 3). The rest of the messages are the same as that in legacy LTE.

Since RRC (control plane) is segregated from PDCP, RLC, MAC, PHY (data plane) and RRC layer is present in the SRC; SRC can take the decision of Sd-eNB addition and does not need an acknowledgment from Sd-eNB. Thus, we observe that the number of signaling messages is reduced. Earlier 12 messages were exchanged for the addition of SeNB. It has now reduced to 10 with the introduction of SRC. The flow graphs for various DC mobility events such as SeNB addition, SeNB change, and SeNB release in the legacy LTE architecture are presented in Section 10.1.2.8 of [**?**].

The procedures for DC mobility events in SMRAN are displayed in Figure **??**. Ini-

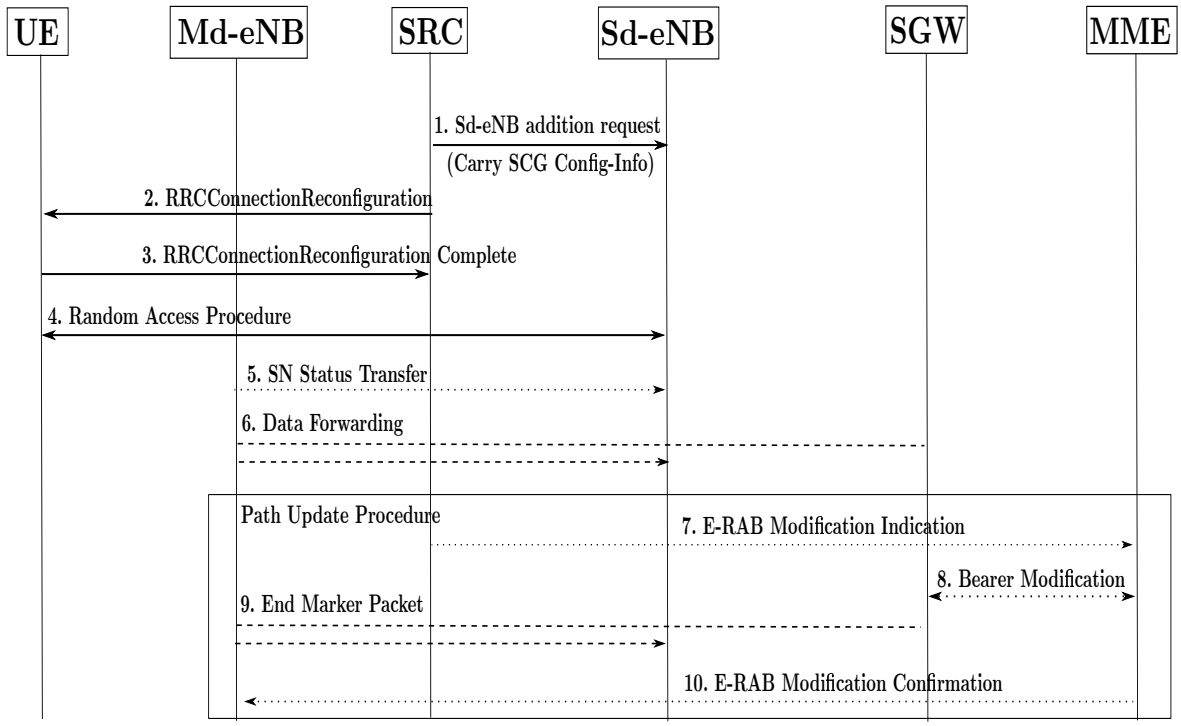

Figure 4.3: Sd-eNB addition procedure in SMRAN.

tially, the first link for the UE is served by Md-eNB, and there is no second link. SRC decides to add Sd-eNB as a second link to the UE. Sd-eNB addition procedure, as explained above, follows. In the legacy LTE architecture, the number of control signaling exchanges for the addition of SeNB between MeNB and SeNB is 3 (SeNB addition request, SeNB addition request acknowledge and SeNB reconfiguration complete [**?**]), which has come down to 1 (Sd-eNB Flow add command) with the introduction of SRC in SMRAN. Now, UE has two links served by Md-eNB and Sd-eNB. SRC decides to change the small cell from Sd-eNB to Target Sd-eNB (T-Sd-eNB). It sends Flow Add command to T-Sd-eNB and Flow Release command to Sd-eNB. In SeNB modification procedure, the number of control signaling messages in LTE was 5 (SeNB Addition request, SeNB Addition request Acknowledge, SeNB Release request, SeNB Reconfiguration Complete, UE Context release) which has reduced to 2 (T-Sd-eNB Flow addition command and Sd-eNB Flow release command) in SMRAN. UE has two links served by Md-eNB and T-Sd-eNB. SRC decides to release Sd-eNB and sends Flow Release command to T-Sd-eNB. In the legacy architecture, the SeNB release procedure consists of 2 control signaling messages, viz., SeNB Release request, and UE Context release. In SMRAN, this control signaling procedure needs only one control message, viz., Flow release command.

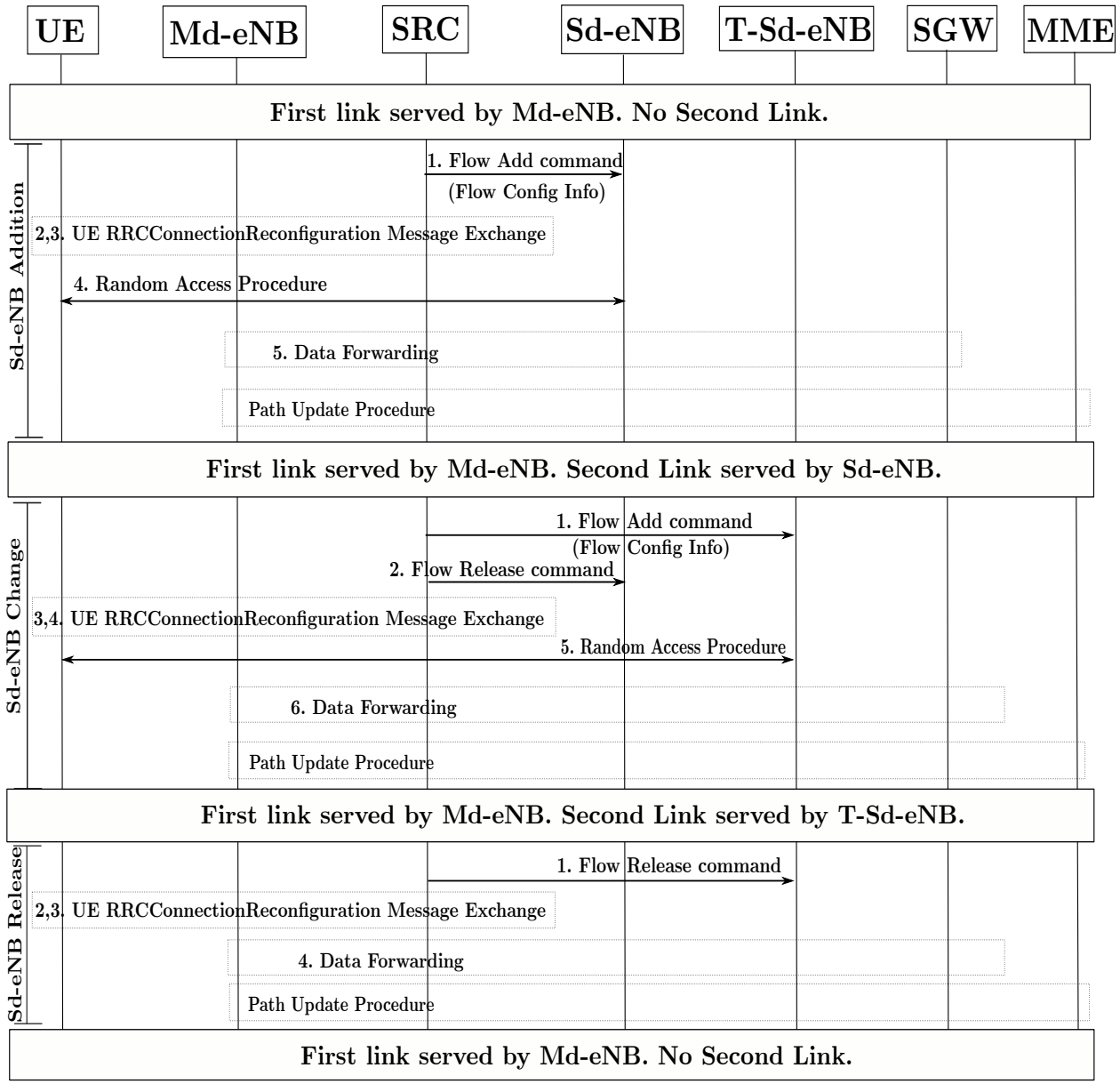

Figure 4.4: Signal flow graph of Sd-eNB addition, Sd-eNB change and Sd-eNB release in SMRAN.

The number of signaling exchanges in the DC mobility events of SeNB addition, SeNB change, and SeNB release in legacy LTE architecture and SMRAN are specified in Table **??**. It presents the count of control signaling message exchanges between the RAN nodes in legacy LTE and those between the SRC and data plane entities in SMRAN.

Table 4.1: Number of signaling messages exchanged between RAN nodes in DC mobility events.

Architecture	**Small cell addition**	**Small cell change**	**Small cell release**
Legacy LTE	3	5	2
SMRAN	1	2	1

4.3.2 Simulation Results

We demonstrate the effect of control signaling reduction in a heterogeneous network scenario. We consider a 7-cell system with inter-site distance (ISD) of 500m. Each macro cell is divided into three hexagonal sectors and four pico cells per sector with overlapping coverage deployed in the central cell. We assume radio spectrum with a bandwidth of 10 MHz deployed in each macro and pico cell. The UEs are uniformly distributed in the center cell. We simulate different scenarios where the UEs move with speeds of 3 kmph, 30 kmph, and 60 kmph, respectively. The UEs move within the sector with a fixed speed in a randomly chosen direction till a boundary is reached. Once they reach this boundary, they bounce back in a random direction such that they remain within a bounding box. Various network parameters are enumerated in Table **??**. We determine the number of DC mobility events, viz., secondary cell additions, secondary cell changes, and secondary cell releases, experienced by a mobile UE in a duration of 1000 secs. Events A2 and A4 [**?**] using Reference Signal Received Quality (RSRQ) measurements are used as triggering events for these DC mobility events. The simulations are performed in ns-3 simulator [**?**].

The simulation results obtained are illustrated in Table **??** for different UE speeds. These values, for instance, indicate that a UE traveling at a speed of 30 kmph experiences an average of 0.02229 SeNB additions, 0.0289 SeNB changes, and 0.0217 SeNB releases per second. The average number of message exchanges between eNBs in the legacy architecture and that between the SRC and d-eNBs in SMRAN are then calculated using Table **??**.

Figures **??**, **??**, and **??** illustrate the average number of signaling message exchanges between the network nodes in the two architectures for UEs moving with speeds 3 kmph,

Table 4.2: Network parameters.

Parameter	**Value (Macro, Pico)**
Macro ISD, Pico radius	500 m, 50 m
Transmit power	46 dBm, 30 dBm
Antenna	Sectored, Omnidirectional
Antenna height	32 m, 10 m
Path loss (d in km)	128.1 + 37.6 log(d), 140.7 + 36.7 log(d) dB

Table 4.3: Average number of DC mobility events per UE per second.

UE speed	**3 kmph**	**30 kmph**	**60 kmph**
SeNB additions	0.00291	0.02229	0.04282
SeNB changes	0.00289	0.0289	0.05843
SeNB releases	0.00233	0.0217	0.0423

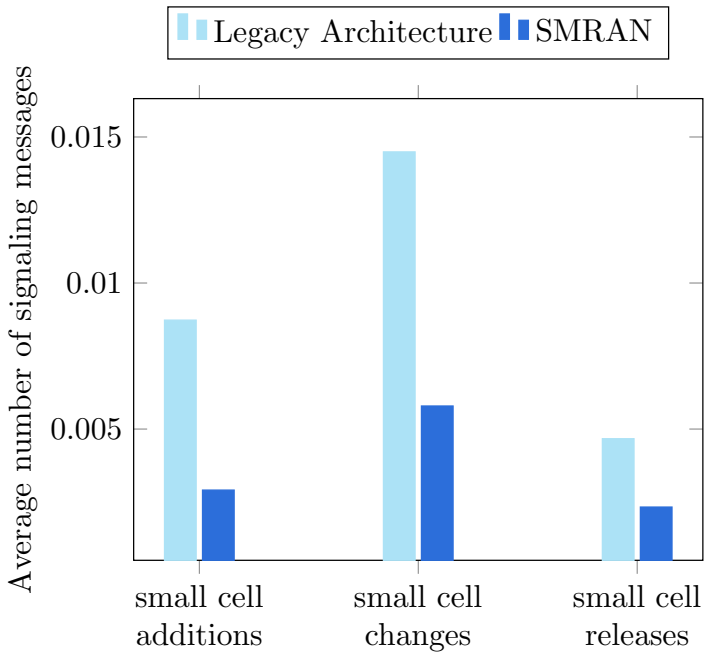

Figure 4.5: Amount of signaling messages exchanged in DC mobility events when UEs are moving at a speed of 3 kmph.

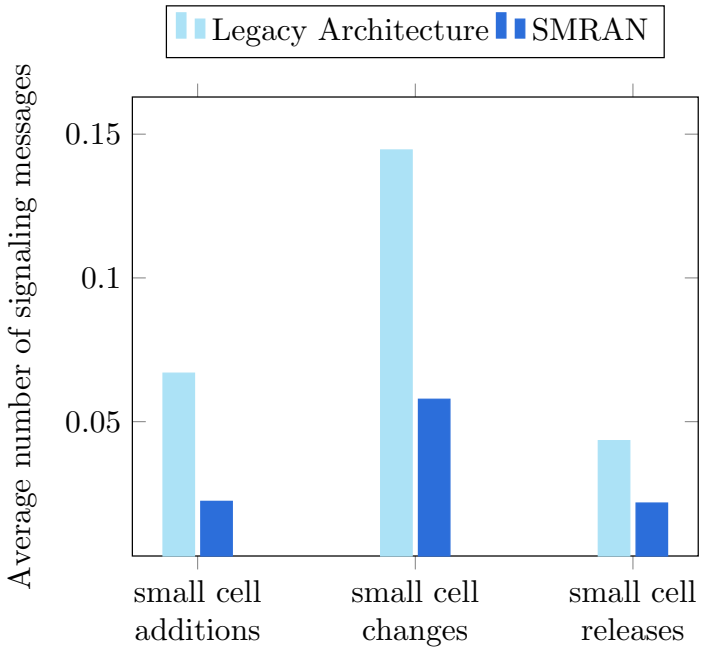

Figure 4.6: Amount of signaling messages exchanged in DC mobility events when UEs are moving at a speed of 30 kmph.

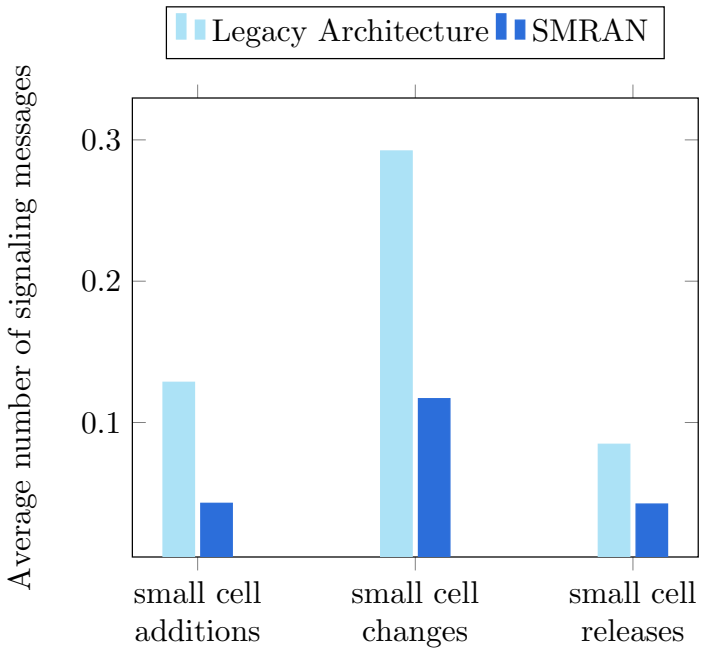

Figure 4.7: Amount of signaling messages exchanged in DC mobility events when UEs are moving at a speed of 60 kmph.

30 kmph, and 60 kmph, respectively. We infer that there is a 66% reduction in signaling for small cell additions in SMRAN. Similarly, there is a 60% and 50% reduction in signaling for

small cell changes and small cell releases, respectively, in SMRAN. Due to the reduction in control signaling messages between SRC and data plane nodes, the delay in initial connection setup for dual connected UEs is also reduced in SMRAN. Thus, SMRAN leads to a reduction in control signaling overhead in the network, thereby improving the response time of the system.

4.4 Evaluation of System Performance

In this section, we demonstrate that decisions taken in SMRAN lead to an improvement in the throughput of the system as compared to those taken in the legacy architecture.

4.4.1 Proposed Algorithm

In SMRAN, SRC implements the RRM function of all data plane entities. Therefore, all signal strength measurements are received at the SRC, which can then take a centralized decision to select the dual connected UEs. To utilize the available capacity and to achieve maximum throughput in the system, we need to select appropriate UEs to be dually connected. We propose a centralized algorithm for selection of dual connected UEs in this section. In our centralized model, there is no difference between MN and SN for a dual connected UE; hence, for our algorithm, we do not differentiate between macro and small cell nodes.

Centralized Dual Connectivity Algorithm

The key idea behind this algorithm is to connect those UEs dually whose secondary radio links are better than some of the existing primary radio links in a particular cell (see Algorithm **??**). Primary link refers to the first radio link to which a UE is associated, and secondary link refers to the second auxiliary link of the UE. As already mentioned, there is no difference between different types of cells in our algorithms. The algorithm is executed periodically to take into account new UE arrivals and departures in the system.

In Algorithm **??**, Reference Signal Received Power (RSRP) of UE a from data plane network node b is denoted by $RSRP(a, b)$. For each UE u, we determine node i^* which

Algorithm 3 Centralized Dual Connectivity Algorithm (CDA).

1: $\mathcal{B} \leftarrow$ Set of all data plane nodes.
2: $\mathcal{U} \leftarrow$ Set of UEs.
3: $count \leftarrow 0$.
4: **for all** $u \in \mathcal{U}$ **do**
5: $\quad i^* \leftarrow \arg\max_{j \in \mathcal{B}}(RSRP(u, j))$
6: $\quad A_1(u) \leftarrow i^*$
7: $\quad$ Append $RSRP(u, i^*)$ to $RSRP_B(i^*)$
8: **end for**
9: $\mathcal{U} \leftarrow$ UEs sorted according to their second best link
10: **for all** $u \in \mathcal{U}$ **do**
11: $\quad$ **while** $count < N$ **do**
12: $\quad\quad j^* \leftarrow \arg\max_{j \in \mathcal{B} \setminus A_1(u)}(RSRP(u, j))$
13: $\quad\quad$ **if** $RSRP(u, j^*)$ is better than at least n existing links in j^* **then**
14: $\quad\quad\quad A_2(u) \leftarrow j^*$
15: $\quad\quad\quad count \leftarrow count + 1$
16: $\quad\quad\quad$ Append $RSRP(u, j^*)$ to $RSRP_B(j^*)$
17: $\quad\quad$ **end if**
18: $\quad$ **end while**
19: **end for**

provides the maximum RSRP (in Line 6) and associate the UE with it. We store this association in vector $A_1(\cdot)$ (Line 7). $RSRP_B(b)$ denotes a vector of RSRP values of UEs which are associated with node b. In Line 8, we append the RSRP value of the associated UE to $RSRP_B(i^*)$. UEs are now associated with the nodes which provide them with the best radio links. In Line 10, we sort all UEs in the order of the strength of their second best link. In the second iteration, we determine the node with the second best link for UE u (Line 13). Only those UEs, whose second best link is better than at least n existing links for that node, where n is a configurable number, are selected for DC (Line 14). The number of UEs that can be dual connected is limited to a fixed positive integer, N (Line 12).

4.4.2 Simulation Results

We consider a 7-cell system with ISD of 500m. Each macro cell is divided into three sectors. Four pico cells per sector with fixed locations are deployed in the central cell. The macro and pico cells operate at different carrier frequencies with a bandwidth of 10 MHz each. UEs are stationary and are uniformly distributed in the central cell. We assume full buffer traffic for all UEs. The other parameters are listed in Table **??**. We compare the proposed algorithm with a distributed threshold-based algorithm (see Algorithm **??**).

The threshold-based algorithm is commonly used for DC [**?**], [**?**], [**?**]. In this algorithm, $\mathcal{B}_M$ and $\mathcal{B}_S$ denote the set of Macro BSs and Small cell BSs, respectively. The primary link is selected from the set of macro cells (Line 2), while the secondary link is selected from the set of pico cells (Line 7). The secondary link, however, is chosen only if its signal strength is greater than a threshold $RSRP_{th}$ (Line 8). We set the value of $RSRP_{th}$ to -78 dBm.

Algorithm 4 THreshold-based Algorithm (THA)

1: **for all** $u \in \mathcal{U}$ **do**
2: $i^* \leftarrow \arg\max_{j \in \mathcal{B}_M}(RSRP(u, j))$
3: $A_1(u) \leftarrow i^*$
4: Append $RSRP(i^*, u)$ to $RSRP_B(i^*)$
5: **end for**
6: **for all** $u \in \mathcal{U}$ **do**
7: $j^* \leftarrow \arg\max_{j \in \mathcal{B}_S}(RSRP(u, j))$
8: **if** $RSRP(j^*, u) > RSRP_{th}$ **then**
9: $A_2(u) \leftarrow j^*$
10: Append $RSRP(j^*, u)$ to $RSRP_B(j^*)$
11: **end if**
12: **end for**

We consider system throughput and UE throughput as the performance indicators for our analysis. The system throughput is defined as the sum of throughput of all users deployed in the system. User throughput in bits per second is defined as the ratio of the number of bits received at the UE and the time between reception of first and

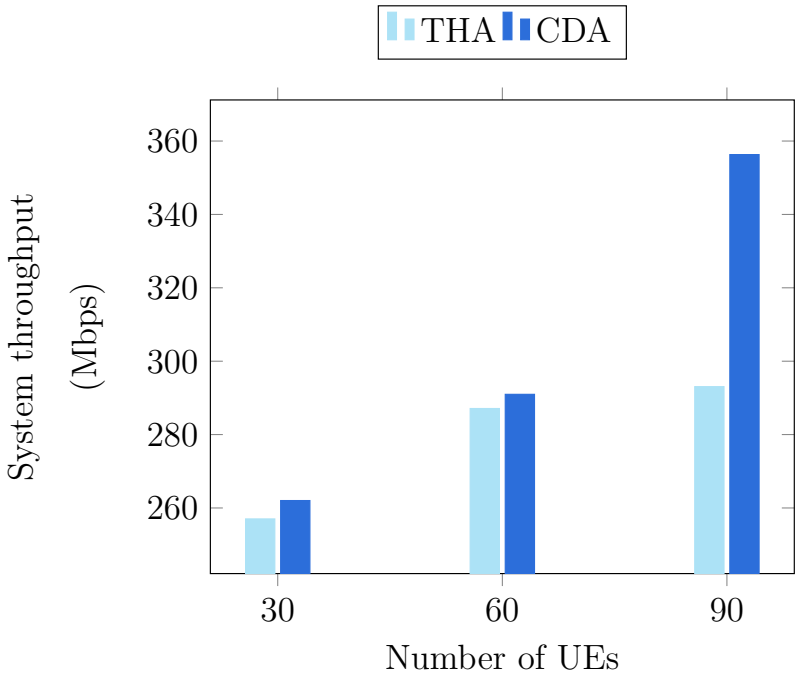

Figure 4.8: System throughput performance of different algorithms.

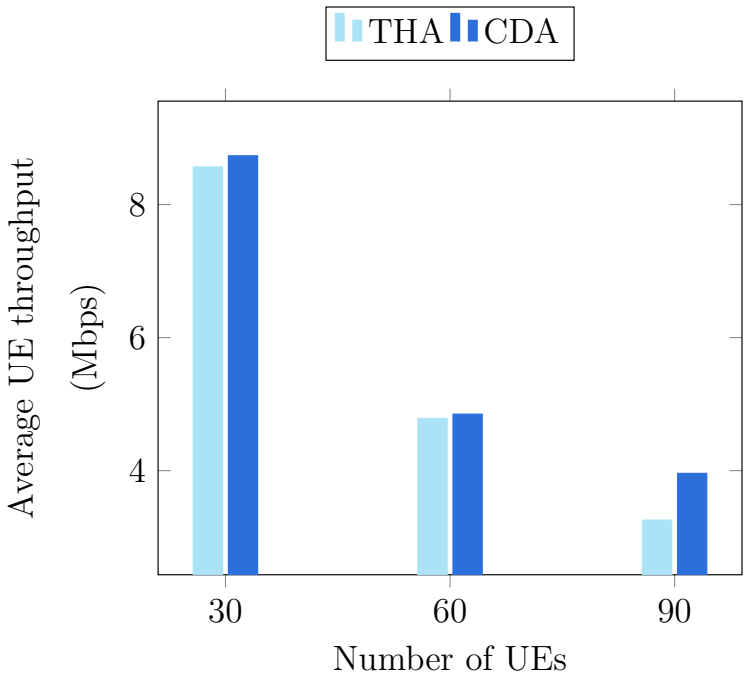

Figure 4.9: Average UE throughput performance of different algorithms.

last packet. The values of system throughput and UE throughput are averaged over ten different random simulations. Figure **??** illustrates the average system throughput and average UE throughput obtained by simulating the distributed algorithm and the proposed centralized algorithm for different UE drops. In the proposed algorithm, SRC

uses the knowledge of UE measurements information gained through a global view of the network. In the CDA algorithm, the values of N and n are considered to be the total number of users in the system and one, respectively. The second link is added to the UE, only if its radio condition is better than at least one existing primary link in that cell. The simulation results demonstrate that the system, as well as the average UE performance, improves in SMRAN. Since secondary links of dual connected UEs are added only when they are better than existing primary links in that cell, the overall throughput of the system is improved.

4.5 Salient Features of SMRAN

In this section, we present the salient features of SMRAN.

4.5.1 Integration of Multiple RATs

The architecture enables the unification of multiple RATs, viz., 3GPP access technologies (LTE, 5G NR) and non-3GPP access technologies (WLAN) under a centralized controller and a common CN. The CN can belong to LTE or 5G Core. It provides a common Xd interface for communication between the network nodes belonging to different RATs.

4.5.2 Integration of DC Variants

The different variants of DC, viz., LTE DC, LWA, and MR-DC (Section **??**) have been integrated in this architecture, and these follow common control signaling procedures. There is no specific requirement that only LTE eNB can carry the control signaling, as in the legacy architecture. Any network node can carry the SRBs.

4.5.3 Unified Interface to the CN

The RAN may contain network nodes belonging to different RATs (LTE, 5G NR, WLAN), but irrespective of their RAT, a unified interface exists from the CN to the SRC. The RAN is tied together by a single interface from the SRC to the CN. This interface may be S1-MME if the CN belongs to LTE or it may be NG-C if the CN belongs to 5G network.

4.5.4 Common Interface for RAN Communication

The RAN may contain network nodes belonging to different RATs (LTE, 5G NR, WLAN). For communication between these network nodes, a common interface called Xd interface is defined. With the introduction of this interface, a uniform way of communicating between network nodes is established.

4.5.5 Centralized Control Logic

The control functions of all nodes in the RAN reside at the SRC. As a result, the overall management of RAN can be improved. Depending on the QoS requirement of traffic flow, the data path can be set up through an appropriate data plane node. For instance, for best-effort type traffic, a data path through d-WT can be set up. For a traffic flow with real-time constraints, e.g., for a voice call, a data path through d-eNB can be established. Similarly, for a low latency application, the data path through a d-gNB can be set up. The control functions at the SRC convey their respective configurations to d-WT, d-eNB, and d-gNB individually. Additionally, the architecture helps in administering the entire RAN with the help of simple configuration commands from the controller using open APIs. Centralized control over RAN allows effective flow control and routing for dual connected UEs.

4.5.6 Flexibility in Setting Up an SRB

In legacy LTE architecture, the SRB is usually served by macro eNB. However, in SM-RAN, RRC functions of all data plane nodes are implemented at the SRC. Thus, there are no specific requirements as to which node should serve the SRB. The SRB can be served by any cell, irrespective of its coverage based on UE's radio conditions. For instance, d-WT can serve the SRB for a UE, which is not possible in existing architecture. Such type of changes may be challenging in the existing architecture but can be easily achieved in SMRAN. This brings adaptability and agility in network flow control.

4.5.7 Flexibility in Processing the Data Plane Protocols

SMRAN can bring flexibility in the processing of data plane protocols across the data plane entities in the RAN. SRC is aware of the load, resources, capabilities, and buffer status of the data plane entities. If a UE connected to a macro cell requests for DC with a small cell, the SRC might decide to perform RLC layer processing at the data plane node (d-eNB/d-gNB) of either the macro cell or small cell based on the resources and capabilities available at the respective cells. Since the processing of data plane protocol layers can be controlled dynamically, it brings flexibility in network control.

4.5.8 Reduction in Control Signaling

SMRAN leads to a reduction in control signaling overhead in the network, thereby improving the performance of the system. Section **??** demonstrates this reduction in control signaling.

4.5.9 Load Management

SRC has a global view of the load and interference conditions of the different multi-RAT nodes in the network. It is responsible for the control functions of RAN. The data transfer between the network nodes in RAN takes place according to these control functions. These control functions allow the SRC to associate UEs with those nodes which are less loaded, especially for the secondary link in case of DC. This leads to higher throughput through better distribution of load across the nodes. We have demonstrated this in Section **??**.

4.5.10 Mobility Management

During a handover in legacy RAN, the change in the path from a source node to a target node at the CN follows path switch procedure. In SMRAN, SRC may decide not to perform path switch procedure in certain scenarios, and allow the source data plane node to act as the anchor node for the UE. For example, if the current session of the UE is expected to be a short one, e.g., a Voice over IP (VoIP) session, such a decision can be taken by the SRC. The data proceeds from the source node to the target node and then to the UE, as illustrated in Figure **??**. A similar action can be taken in case the source

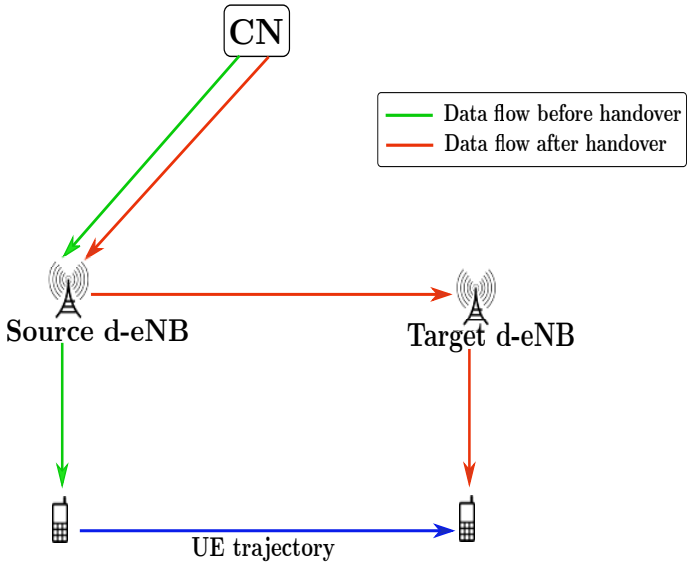

Figure 4.10: Data flow in case of a handover.

node has data stored in its cache that the UE wants to download, as depicted in Figure ??. In this situation, updating the path of the UE to the target node may lead to a longer delay in delivering the data to the UE. This flexibility in mobility management procedures by the centralized controller according to session requirement leads to improved network behavior.

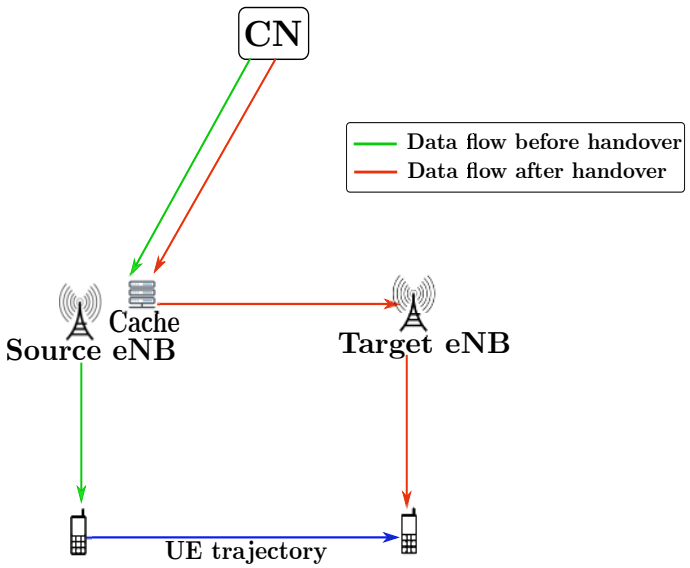

Figure 4.11: Data flow from cache server in case of a handover.

4.6 Conclusions

Due to diverse signaling radio bearer establishment procedures across different variants of dual connectivity, control signaling in current dual connectivity architectures is inherently complex in nature. Moreover, signaling between MN and SN takes place via different interfaces in different types of DC. Again, due to the nonideal nature of this interface, it leads to a delay in the signaling exchange. Further, for a dual connected UE, SN is selected based on a local view of the UE and the base station. To surmount these challenges, we have proposed SMRAN - an SDN based multi-RAT RAN architecture for dual connectivity. Without any change in the CN, RAN is implemented into two parts; a multi-RAT RAN controller and the data plane entities with their respective user-plane protocol stacks. This architecture provides unified control over the functionality of RAN as well as a reduction in the control signaling in the network. SMRAN integrates the 3GPP RATs (LTE and 5G) with non-3GPP access technologies (WLAN) and designs a common interface (Xd) for communication across nodes. The different variants of DC, viz., LTE DC, LWA, and MR-DC, are seamlessly integrated into this architecture. In SMRAN, there is no difference between MN and SN. Any RAN node can function as MN or SN for a dual connected UE. Based on SMRAN, a centralized algorithm for DC is presented, which improves the performance of the system. The architecture enhances the load and mobility management in the network. Thus, SMRAN brings simplicity, flexibility, and uniformity in network interactions and procedures.

Chapter 5

Proportional Fair Scheduling for Dual Connectivity

In the previous chapter, we introduced a centralized multi-RAT RAN architecture (SM-RAN) to improve the performance of dual connectivity. Based on this architecture, in this chapter, we propose a low complexity Proportional Fair (PF) scheduling scheme for dual connectivity (PF-DC). We also propose four UE association schemes for dual connectivity in this chapter.

A scheduling algorithm can meet diverse goals, such as maximizing throughput or improving fairness. The opportunistic scheduling algorithm maximizes the throughput in the system by greedily allocating resources to User Equipments (UEs) with the best channel conditions. However, UEs with poor channel conditions may not be allocated any resources in this algorithm. The round-robin scheduling algorithm maximizes fairness in the system by allocating an equal amount of resources to all UEs. The total throughput, however, can be low in this case. The proportional fair scheduling algorithm maximizes the total system throughput while at the same time ensuring a minimum level of service to each UE. In this way, it seeks to balance throughput maximization with UE fairness in the system.

Proportional Fair (PF) scheduling was first introduced in [**?**] and proved to maximize the PF utility defined as the sum of logarithmic user throughput in a system. The extension of proportional fair scheduling for multiple transmission channels has been proposed in [**?**]. In [**?**], the general PF scheduling scheme is extended to Multiple Input Multiple

Output (MIMO) case. With the emergence of next generation cellular systems such as LTE-A, new mechanisms were proposed. One such mechanism was where a cell could use multiple carriers or bandwidths to enhance the data capacity in the network. In such a scenario, it is challenging to maximize system throughput over multiple carriers. Using independent PF on each carrier will not achieve the maximum PF utility. The extension of PF to multiple carriers in such a scenario is demonstrated in [**?**].

PF scheduling maximizes the PF utility given by $\sum_{u \in \mathcal{U}} \log(\gamma_u)$, where $\mathcal{U}$ denotes the set of UEs in the system, and γ_u is the long-term average throughput that UE u experiences when connected to a Base Station (BS). In a multi-connectivity case, i.e., when a UE is connected to multiple BSs, γ_u is determined by the long-term average throughput experienced by UE u from all BSs to which it is connected. The PF utility obtained in this case is known to be maximized by a Global PF (GPF) scheduling scheme as proposed in [**?**]. The GPF scheduling scheme is a distributed scheme based on the instantaneous data rates and historical throughput of UEs. The basic idea of this scheme is to connect all UEs to every BS and then schedule a UE opportunistically at the BS providing the best channel to the UE. This scheme, however, requires periodic sharing of per-UE throughput information among the BSs. This leads to significant control signaling exchange between BSs. Moreover, a UE may have a limited number of interfaces, typically either one or two, and the assumption of all UEs connected to every BS is not practical. Further, the maintenance of multiple connections requires resources, for instance, power, radio, and processing resources at the network. As a result, the overhead to maintain multiple connections for a single UE may be significantly high.

In this work, we consider a viable scenario where each UE has two interfaces and therefore can connect to a maximum of two BSs at the same time. By limiting the number of connections for a UE to two, the control overhead for each UE is not significant as compared to multiple connections for a UE. We consider a network architecture with a centralized controller (SMRAN) communicating with all BSs and propose a semi-persistent scheduling scheme (PF-DC). In this scheme, the controller collects the per-UE throughput data periodically from BSs, processes it, and distributes the processed data back to the BSs. The data collection at the controller takes place regularly with a period of T secs.

One of the key challenges that arise in dual connectivity is the selection of two suitable BSs for UE association. To deal with it, we propose four association algorithms for dual connected UEs and evaluate the performance of PF-DC in conjunction with these algorithms. We denote the procedure of all UEs connected to every BS and GPF scheduling performed at each BS as All Connectivity Procedure (ACP). The scheme of a UE connected to the BS offering it with the best signal strength and standard PF scheduling performed at each BS is denoted as Single Connectivity Procedure (SCP). The procedure where a UE is dual connected to two suitable BSs using proposed association algorithms and standard PF scheduling performed independently at each BS is denoted as Dual Connectivity Standard Procedure (DCSP). The scheme where a UE is dual connected to two suitable BSs using proposed association algorithms and PF-DC scheduling performed at each BS is denoted as Dual Connectivity Procedure (DCP). We investigate the performance of ACP, SCP, DCSP, and DCP in different HetNet scenarios. We compare the performance of these algorithms based on indicators such as PF utility, total system throughput, average UE throughput, and Jain's Fairness Index (JFI). We demonstrate that DCP achieves a significant increase in PF utility as compared to SCP. The PF utility obtained in DCP is quite close to that of ACP. We also demonstrate that DCP achieves higher PF utility and higher average UE throughput as compared to DCSP.

The major contributions of this work are as follows:

1. We propose PF-DC - a low complexity and easily implementable scheduling scheme for dual connectivity based on SMRAN architecture.

2. We propose four UE association algorithms for dual connectivity.

3. We implement the PF-DC scheme in ns-3 using a centralized architecture. We modified multiple modules and source code of the ns-3 simulator to achieve the same. We conducted extensive simulations to evaluate its performance in diverse HetNet scenarios.

4. We compare SCP, ACP, and DCP and demonstrate through simulations in diverse scenarios that DCP leads to a significant improvement in PF utility as compared to SCP. We demonstrate that the improvement in PF utility from DCP to ACP is,

however, not that pronounced and the computation complexity of the ACP scheme is high as compared to that of the DCP scheme.

5. Simulations are conducted for different values of the period of information exchange (T) of the PF-DC scheme and variation in PF utility is studied for diverse HetNet scenarios.

6. We demonstrate that with dual connectivity, we can improve the proportional fairness in the system.

The chapter is organized as follows. The system description, as well as PF-DC - the proposed scheduling scheme, is presented in Section **??**. The UE association algorithms are proposed in Section **??**. Section **??** discusses the simulation results. The complexity analysis of the PF-DC scheme is detailed in Section **??**, followed by conclusions in Section **??**.

5.1 System Description

We consider SMRAN architecture as introduced in Chapter **??**. The control plane functionality is handled by a unified and logically centralized controller, and the data plane is managed individually at each BS. The SRC supports radio resource control and management functions such as admission control, mobility management, and load balancing. Access to the global view of the system enables SRC to take appropriate decisions such as UE association. On the other hand, scheduling decisions are handled at the individual BSs. Further, the architecture supports different RATs and treats each RAT in the same way.

Consider a RAN consisting of B BSs and U UEs, as illustrated in Figure **??**. Let $\mathcal{B} = \{1, 2, ..., N\}$ denote the set of BSs and $\mathcal{U} = \{1, 2, ..., M\}$ denote the set of UEs. Let the total number of BSs and the total number of UEs be denoted as $|\mathcal{B}| = N$ and $|\mathcal{U}| = M$, respectively, where $|\mathcal{X}|$ denotes the cardinality of set $\mathcal{X}$. The BSs can be macro or pico. The UEs do not differentiate between different types of BSs and connect to any two of them simultaneously. Each BS is assigned a fixed bandwidth and transmits at a fixed power. The system functions in a time-slotted manner with fixed duration time-slot.

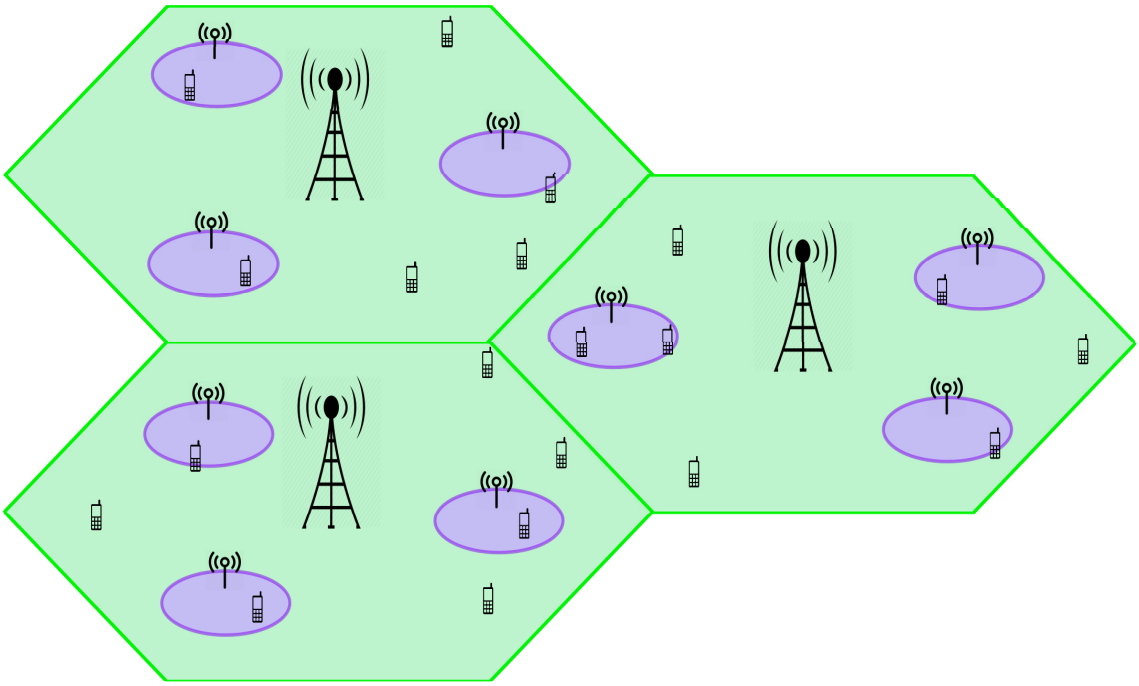

Figure 5.1: System model.

5.1.1 Standard PF Scheduling Scheme

In this section, we discuss the standard PF scheduling scheme. PF scheduling is performed individually at all the BSs. Each BS takes scheduling decisions at the beginning of a time slot. At each BS, the PF metric for all UEs is determined based on the instantaneous rate that can be received by that UE, and the average throughput which the UE has achieved until that time. For instance, PF metric for UE j at BS k is determined by $r_j^k(t)/\bar{r}_j^k(t)$, where $r_j^k(t)$ is the achievable rate by UE j at time t, and $\bar{r}_j^k(t)$ represents the average throughput received by UE j until time t. The UE with the maximum PF metric u_k^* is then selected to be scheduled in slot t by BS k.

$$u_k^* = \arg\max_{j \in \mathcal{U}} \frac{r_j^k(t)}{\bar{r}_j^k(t)}. \tag{5.1}$$

In each time-slot, the average throughput of UE j $\bar{r}_j^k(t)$ is updated using weighted moving average,

$$\bar{r}_j^k(t+1) = (1-\gamma)\,\bar{r}_j^k(t) + \gamma\, r_j^k(t)\, \mathbb{1}\{j = u_k^*\}, \tag{5.2}$$

where, $\mathbb{1}\{A\}$ denotes an indicator function of the event A, and $0 < \gamma < 1$ is a constant. Typically, the value of γ is closer to 1, which ensures that higher weight is given to the historical rate received by the UEs as compared to the current achievable rate. This

enables the PF scheduling algorithm to balance UE fairness and maximization of total throughput.

5.1.2 PF-DC: Proposed PF Scheduling scheme

In this section, we present PF-DC, the proposed PF scheduling scheme for dual connected UEs. This scheme is similar to the standard PF scheduling scheme with two modifications. First, the PF metric is determined using the sum of the average throughput received by the UE from the two BSs to which it is connected. Each BS k chooses the UE with maximum PF metric (u_k^*) to schedule in each time-slot.

$$u_k^* = \arg\max_{j \in \mathcal{U}} \frac{r_j^k(t)}{\sum_{k \in \mathcal{B}} \bar{r}_j^k(t)}. \tag{5.3}$$

The denominator is determined by the summation of the average throughput at the two BSs to which the UE is connected. Second, the calculation of the total average throughput of all UEs is performed at the centralized controller (SRC).

BSs share the average throughput information regarding their respective dual connected UEs with the SRC. In response, the SRC determines the total average throughput of each UE using the shared information. SRC then shares the total average throughput values of UEs with their respective BSs. This throughput information is used individually by the BSs for PF scheduling. The exchange of throughput information between the SRC and BSs takes place at regular intervals with period T. We demonstrate through simulations in Section **??** that there is no significant degradation in the PF utility when this information is shared periodically. The parameter T can be tuned as per the requirements of the service provider. Even if there is a slight delay in sharing of throughput information between SRC and BSs, it does not affect the entire system, since scheduling is performed independently at individual BSs based on the shared information. Thus, synchronization is not required between BSs and SRC.

Algorithm 5 UE Initiated Greedy (UIG)

1: **for all** $u \in \mathcal{U}$ **do**
2: $i^* \leftarrow \arg\max_{j \in \mathcal{B}}(RSRP(j,u))$
3: $A_1(u) \leftarrow i^*$
4: **end for**
5: **for all** $u \in \mathcal{U}$ **do**
6: $j^* \leftarrow \arg\max_{j \in \mathcal{B} \setminus \{A_1(u)\}}(RSRP(j,u))$
7: $A_2(u) \leftarrow j^*$
8: **end for**

5.2 UE Association Algorithms for Dual Connectivity

In this section, we propose different schemes for the selection of appropriate BSs for dual connected UEs. In our model, all the UEs are dual connected to two distinct BSs. We propose four different association schemes for dual connectivity. The schemes are based on Reference Signal Received Power (RSRP) of UEs.

5.2.1 UE Initiated Greedy (UIG) Scheme

As the name suggests, this scheme is initiated at the UE. Each UE selects two distinct BSs that provide the highest value of RSRP for the UE (see Algorithm **??**). $RSRP(a,b)$ represents the RSRP received by UE b from BS a. $A_1(b)$ and $A_2(b)$ denote two BSs to which UE b is associated. For each UE, the BS with maximum RSRP is selected for the first association (Line 2,3). From among the remaining BSs, the UE chooses the BS with maximum RSRP for the second association (Line 6,7).

5.2.2 UE Initiated Greedy with Offloading (UIGO) scheme

In the UIG scheme, the UEs select the BSs without considering the load on the chosen BSs. This leads to an imbalance in the distribution of load across BSs. We modify the UIG scheme and offload the UE to a small cell if the BS for the second association is a macro cell (see Algorithm **??**). The first association is selected as in the UIG scheme. For

the second association, we check if the BS is a macro BS (Line 7). If it is a macro BS, we check if the next best BS is a small cell for the purpose of offloading. $\mathcal{B}_M$ denotes the set of macro BSs (Line 8). If the difference between RSRP of the second and third BS is within a threshold H_1 (Line 9), then we select the third BS, else we select the second BS.

The UE selects the first and second BSs and notifies the SRC. SRC then checks if the second BS is a macro cell to enable offloading. It acquires the third BS details from the UE and then takes a decision about the second association of the UE. This algorithm enables offloading some users from macro cells to small cells if signal strength from a small cell is comparable to that of the macro cell. The holistic view of the network allows SRC to take appropriate decisions for the second association of UEs.

Algorithm 6 UE Initiated Greedy with Offloading (UIGO)

1: **for all** $u \in \mathcal{U}$ **do**
2: $b^* \leftarrow \arg\max_{b \in \mathcal{B}}(RSRP(b,u))$
3: $A_1(u) \leftarrow b^*$
4: **end for**
5: **for all** $u \in \mathcal{U}$ **do**
6: $c^* \leftarrow \arg\max_{b \in \mathcal{B} \setminus \{A_1(u)\}}(RSRP(b,u))$
7: **if** c^* is a macro BS **then**
8: $d^* \leftarrow \arg\max_{b \in \mathcal{B} \setminus \{\mathcal{B}_M \cup c^*\}}(RSRP(b,u))$
9: **if** $(RSRP(c^*,u) - RSRP(d^*,u)) < H_1$ **then**
10: $A_2(u) \leftarrow d^*$
11: **else**
12: $A_2(u) \leftarrow c^*$
13: **end if**
14: **else**
15: $A_2(u) \leftarrow c^*$
16: **end if**
17: **end for**

Algorithm 7 BS Initiated Greedy with UE feedback (BIGU)

1: $U_{state}(\cdot) \leftarrow 2$.
2: **while** $U_{state}(u) > 0$ for all $u \in \mathcal{U}$ **do**
3: **for all** $b \in \mathcal{B}$ **do**
4: $u^* \leftarrow \arg\max_{j \in \mathcal{U}}(RSRP(b, j))$
5: **for all** $c \in \mathcal{B}$ **do**
6: $c^* \leftarrow \arg\max_{c \in \mathcal{B}}(RSRP(c, u^*))$
7: **end for**
8: **if** $RSRP(c^*, u^*) - RSRP(b, u^*) < H_2$ **then**
9: **if** $U_{state}(u^*) = 2$ **then**
10: $A_1(u^*) \leftarrow b$
11: $U_{state}(u^*) \leftarrow U_{state}(u^*) - 1$
12: **else if** $U_{state}(u^*) = 1$ **then**
13: $A_2(u^*) \leftarrow b$
14: $U_{state}(u^*) \leftarrow U_{state}(u^*) - 1$
15: **end if**
16: $RSRP(b, u^*) \leftarrow -\infty$
17: **end if**
18: **end for**
19: **end while**

5.2.3 BS Initiated Greedy with UE feedback (BIGU) scheme

As the name suggests, in this scheme, a BS decides which UE will be associated with it. This scheme is based on a team selection process where a BS has to select a team of UEs. Each BS gets a chance to choose a UE in each round. The BSs are sequenced in a round-robin fashion. Each UE can be selected twice by two distinct BSs, as we allow two connections per UE. The selection of a UE is based on the RSRP at the UE from the corresponding BS. The scheme is detailed in Algorithm **??**. After selection of a UE (u^*) by a BS (Line 4), the UE checks if the signal strength from this BS is within a threshold (H_2) as compared to the signal strength from its best BS (c^*) (Lines 6-8). If not, it rejects the current BS and waits for a better offer from another BS from which it can obtain a

higher signal strength. The process continues till two distinct BSs are selected for each UE. In this scheme, the SRC coordinates between BSs and UEs to select the appropriate BS for UEs.

5.2.4 Stable Matching (SM) scheme

The Stable Matching (SM) scheme is a centralized scheme running at the SRC. In this scheme, the set of UEs is ranked by the set of BSs based on their received signal strengths. Similarly, all BSs are ranked by all UEs based on the received signal strength from respective BSs. These preferences are aggregated at the SRC. Let M, N denote the number of UEs and BSs, respectively. The size of these two sets is made equal by repeating each BS $q = M/N + c$ times, where c is a constant. The constant c is selected such that the load for all BSs is almost equal, thus balancing the load in the system. This problem then converts to a stable matching problem. The problem is solved using the Gale-Shapley algorithm [**?**]. After the first association, the UE preferences are updated by giving the BS of the first association last preference. Similarly, BS preferences are updated. The scheme is repeated once more after changing the preferences of UEs and BSs for the second association. The scheme is detailed in Algorithm **??**. The algorithm terminates when all UEs are associated with exactly two BSs.

Algorithm 8 Stable Matching (SM)

$P_1 \leftarrow$ Each UE u sets its preferences for all BSs in $\mathcal{B}$
$P_2 \leftarrow$ Each BS b sets its preferences for all UEs in $\mathcal{U}$
Repeat each BS $q = |\mathcal{U}|/|\mathcal{B}| + c$ times for load balancing
while all UEs not allocated to some BS **do**
    $A_1(u) \leftarrow$ Solve stable matching problem using Gale-Shapley algorithm (P_1, P_2)
end while
$P_1 \leftarrow$ UE u sets its preferences for all BSs in $\mathcal{B}$ giving $A_1(u)$ as last preference
$P_2 \leftarrow$ BS b sets its preferences for all UEs in $\mathcal{U}$ giving all UEs already associated to it as last preference
while all UEs not allocated to some BS **do**
    $A_2(u) \leftarrow$ Solve stable matching problem using Gale-Shapley algorithm (P_1, P_2)
end while

5.3 Motivation for the Proposed Schemes

In this section, we motivate the usage and application scenarios of the schemes proposed in the previous section.

Consider a scenario, as illustrated in Figure **??**. In the case of UIG algorithm, UE_1 selects BS_1 and BS_2, since they provide with maximum signal strength. Similarly, UE_2 selects BS_1 and BS_4, and UE_3 selects BS_1 and BS_4 since users are greedy in selecting the base stations. In this case, BS_3 remains unused. In this case, the UIGO algorithm would be useful. In the UIGO algorithm, UE_1, UE_2, and UE_3 select BS_1 for their first connection. For the second connection, UE_1 selects BS_2, UE_2 selects BS_4, and UE_3 selects BS_4. SDN-based RAN Controller (SRC) then checks for the third link of UE_2 and UE_3. BS_3 will be selected for UE_3 if its signal strength from BS_3 is within a threshold H_1 from its signal strength from BS_4. Thus, the UIGO algorithm offloads traffic to small cells under certain conditions.

In BIGU algorithm, each BS gets a chance to choose a UE (one at a time) in a round-robin fashion. This algorithm was initially proposed as Base Station Initiated Greedy (BIG) algorithm. BIG algorithm balances the load in the system by distributing the users almost equally between the base stations. Consider the scenario illustrated in Figure **??**. BIG algorithm: Round 1: BS_1 selects UE_1, BS_2 selects UE_1, BS_3 selects UE_4, BS_4 selects UE_3; Round 2: BS_1 selects UE_3, BS_2 selects UE_2, BS_4 selects UE_2; Round 3: BS_1 selects UE_4. In this case, even though UE_2 may receive better signal strength from BS_1 or BS_4, it is forced to connect with BS_2, and it may not receive a satisfactory experience. BIGU algorithm: Round 1: BS_1 selects UE_1, BS_2 selects UE_1, BS_3 selects

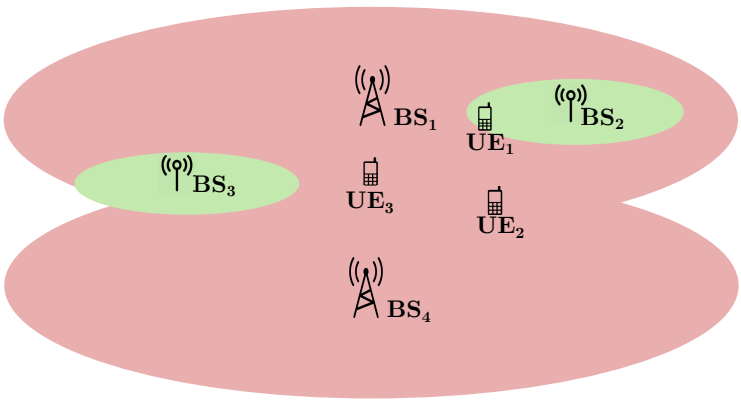

Figure 5.2: Use Case 1.

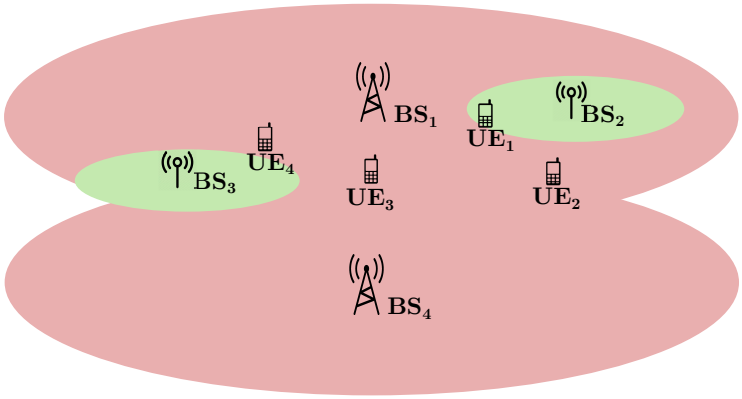

Figure 5.3: Use Case 2.

UE_4, BS_4 selects UE_3; Round 2: BS_1 selects UE_3, BS_2 selects UE_2, But UE_2 can reject BS_2 since it is possible that the next BS can offer it with a better link. BS_4 selects UE_2; Round 3: BS_1 selects UE_3. Thus, the user experience is not compromised in the BIGU algorithm.

5.4 Numerical Results and Analysis

We consider a two-tier HetNet scenario with three macro cells and three pico cells deployed in each macro cell. The operating frequencies of macro and pico BSs are different. The pico cells are deployed using two deployment scenarios. (i) Fixed pico deployment: The locations of pico cells are fixed, and their coverages are non-overlapping. (E.g., see Figure **??**). (ii) Random pico deployment: Pico cells are randomly deployed within the macro cell coverage and may overlap with other pico cells. The UEs are dropped according to two different deployment scenarios, viz., hotspot deployment and uniform deployment. (i) Hotspot deployment: Two-third UEs are deployed uniformly in the coverage area of pico cells, and one-third UEs are deployed uniformly outside the coverage of pico cells but within the macro cell. (ii) Uniform deployment: UEs are deployed uniformly in the coverage area of the macro cell without any consideration of the location of pico cells. Using the combination of pico cell deployments and UE deployments, we have four different deployment scenarios as follows:

- **Scenario 1:** Pico cells are deployed at fixed locations with non-overlapping coverage, and UEs are dropped using hotspot deployment.

- **Scenario 2:** Pico cells are deployed at fixed locations with non-overlapping coverage, and UEs are dropped using uniform deployment.
- **Scenario 3:** Pico cells are deployed randomly within macro cells, and UEs are dropped using hotspot deployment.
- **Scenario 4:** Pico cells are deployed randomly within macro cells, and UEs are dropped using uniform deployment.

The simulations are performed using ns-3 simulator [**?**]. Path loss and fading are considered in the simulation scenario. We assume infinitely backlogged traffic model for all UEs in the system. The characteristics of the network simulated are enumerated in Table **??**. The simulations are performed for a duration of 10000 time slots.

Table 5.1: Network parameters.

Parameter	**Value (Macro, Pico)**
Macro ISD, Pico radius	500 m, 80 m
Transmit power	46 dBm, 30 dBm
Antenna	Omnidirectional, Omnidirectional
Bandwidth	5 MHz, 5 MHz
Antenna height	32 m, 10 m
Path loss (d in km)	128.1 + 37.6 log(d), 140.7 + 36.7 log(d) dB

5.4.1 Comparison of PF-DC with Standard PF Scheduling

We compare the DCP and DCSP procedures in this section. That is, we compare the PF-DC scheme (see Section **??**) with the standard PF scheduling scheme (see Section **??**). The dual connectivity algorithms proposed in Section **??** are used for UE association in DCP and DCSP procedures. We use PF utility and average UE throughput as system metrics to compare these two procedures. As already mentioned, PF utility is a metric which represents a balance between total throughput and fairness in the system. A high value of PF utility indicates a balance between total throughput and fairness in the system.

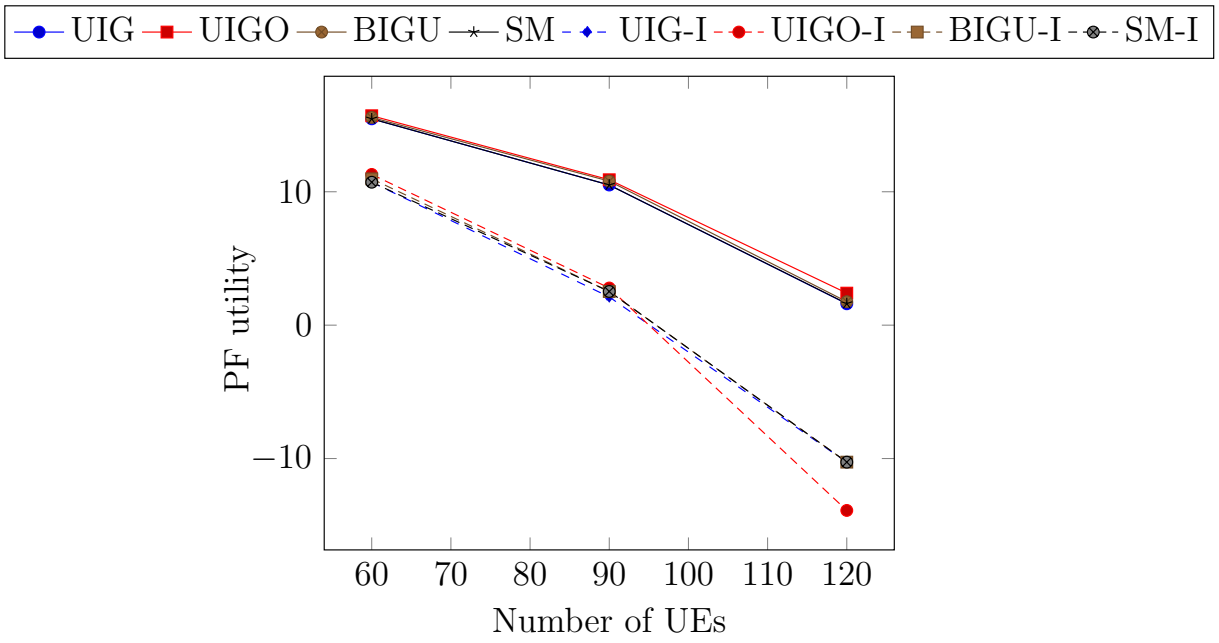

Figure 5.4: PF utility vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 3.

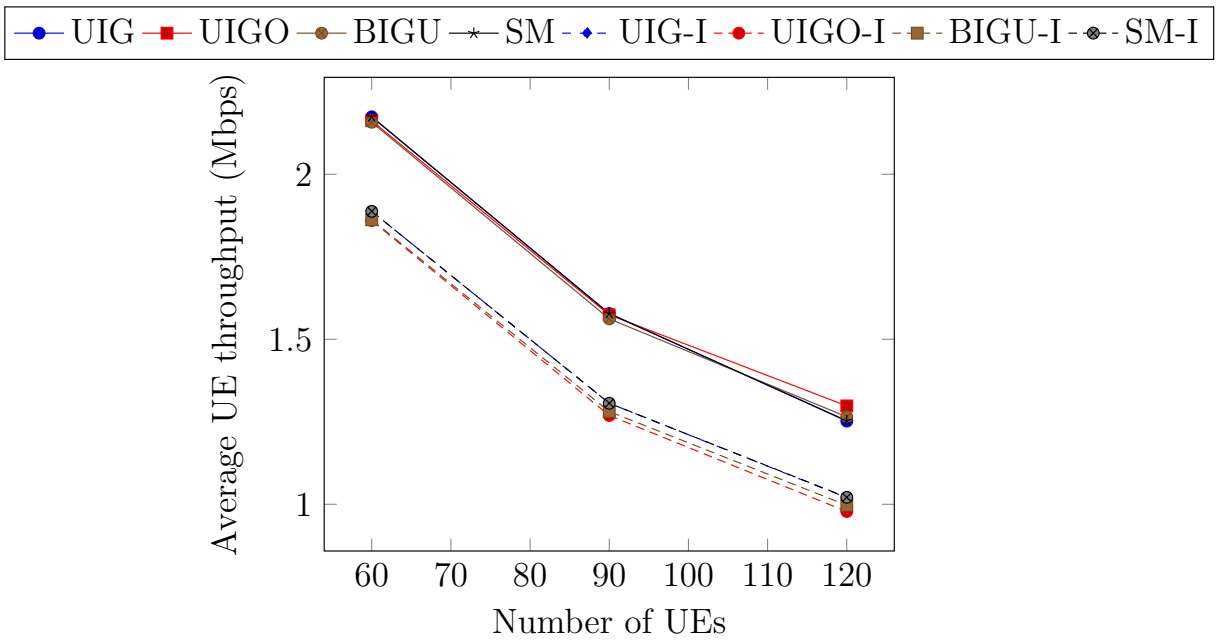

Figure 5.5: Average UE throughput vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 3.

Figures **??** and **??** illustrate the variation of PF utility and average per-UE throughput metrics as a function of the number of UEs in the system in the case of Scenario 3. The plot UIG depicts the UE association chosen according to UIG algorithm and using

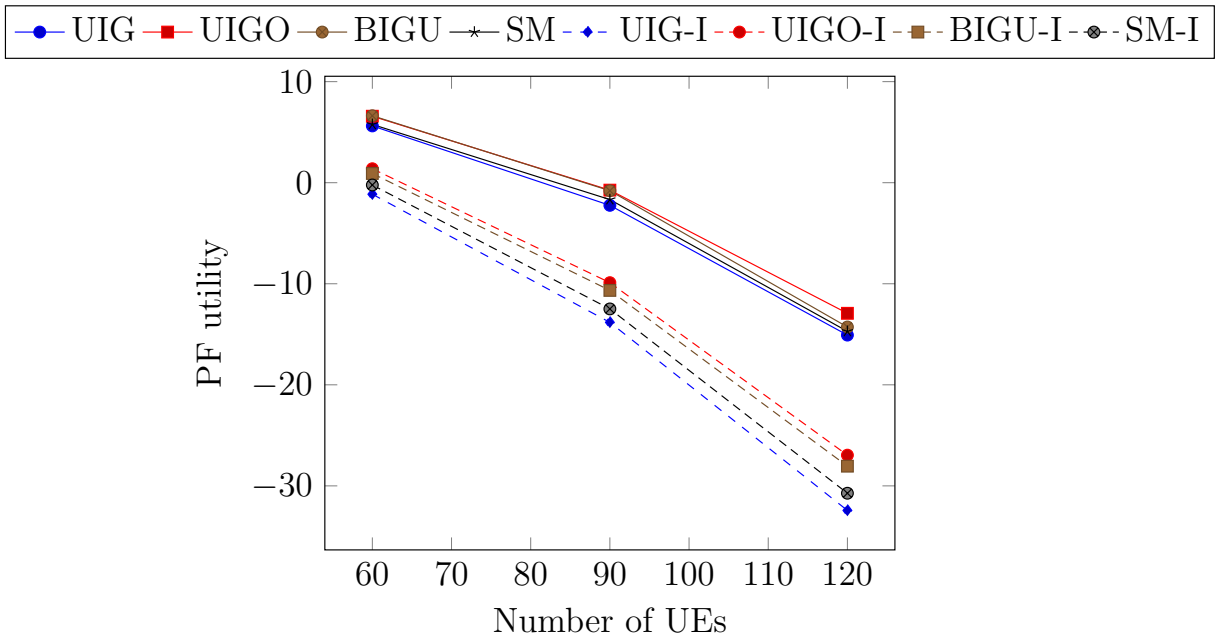

Figure 5.6: PF utility vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 4.

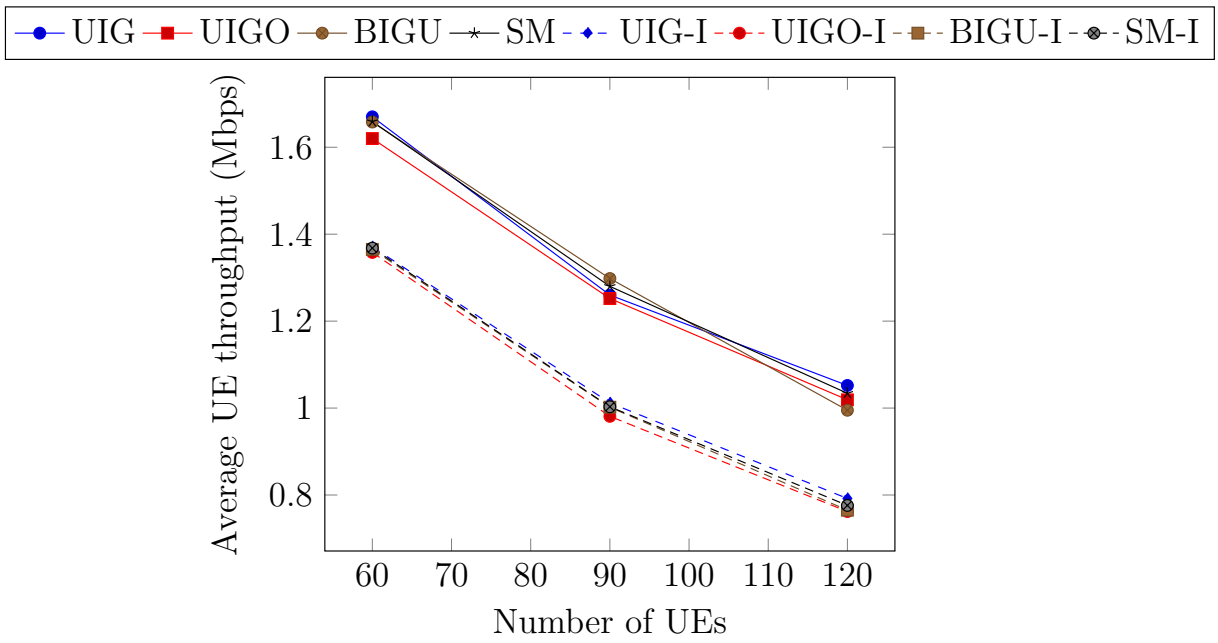

Figure 5.7: Average UE throughput vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 4.

the PF-DC scheme for scheduling. The plot UIG-I depicts the UE association chosen according to UIG algorithm and using standard PF scheduling independently at each BS. Similarly, the plots UIGO, BIGU, and SM denote the UE association chosen according to

UIGO, BIGU, and SM algorithms and using the PF-DC scheme for scheduling. The plots UIGO-I, BIGU-I, and SM-I, represent the UE association chosen according to UIGO, BIGU, and SM algorithms and using standard PF scheduling independently at each BS.

Figures **??** and **??** illustrate the variation of PF utility and average per-UE throughput metrics as a function of the number of UEs in the system in Scenario 4. In Scenarios 3 and 4, we observe similar behavior of PF utility and average UE throughput metrics. As the number of UEs in the system increases, the average per UE throughput decreases as total available capacity is divided among more UEs. We observe a slight improvement in the total throughput with an increase in the number of UEs in the system. However, the fairness in the system decreases with an increase in the number of UEs. UEs in different regions within the coverage area of cells receive a varying amount of throughput, and fairness in the system drops. Therefore, there is a drop in the PF utility of the system. PF-DC scheduling scheme outperforms standard PF scheduling scheme not only in terms of the PF utility but in terms of the average UE throughput as well.

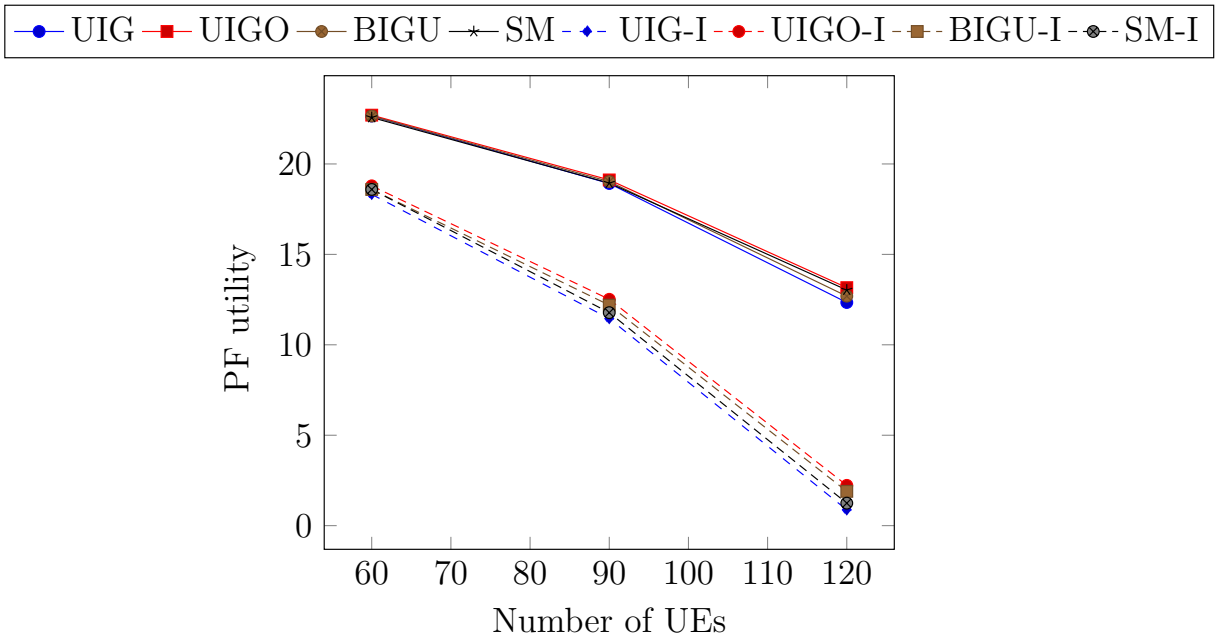

Figure 5.8: PF utility vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 1.

Figures **??** and **??** depict the variation of PF utility and average per-UE throughput in the system as a function of the number of UEs in the system in Scenario 1. Similarly, figures **??** and **??** illustrate the variation of PF utility and average per-UE throughput

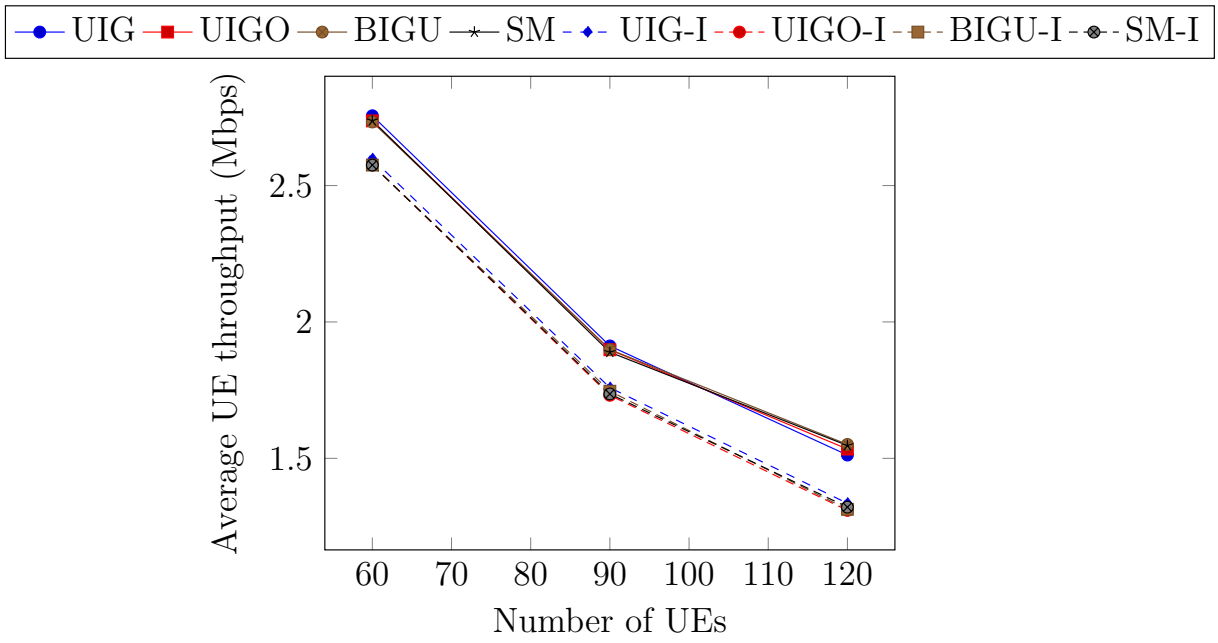

Figure 5.9: Average UE throughput vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 1.

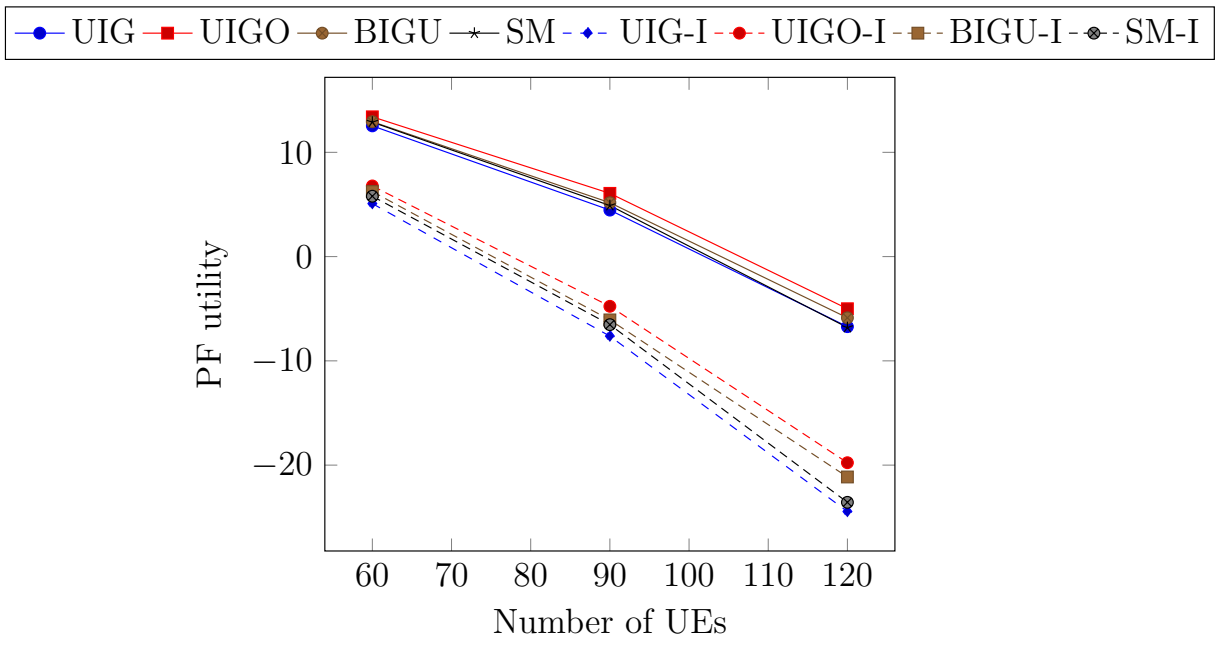

Figure 5.10: PF utility vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 2.

in the system as a function of the number of UEs in the system in Scenario 2. As the number of UEs increases in the system, PF utility as well as average UE throughput in the system decrease. The results for the fixed deployment of pico cells are almost similar

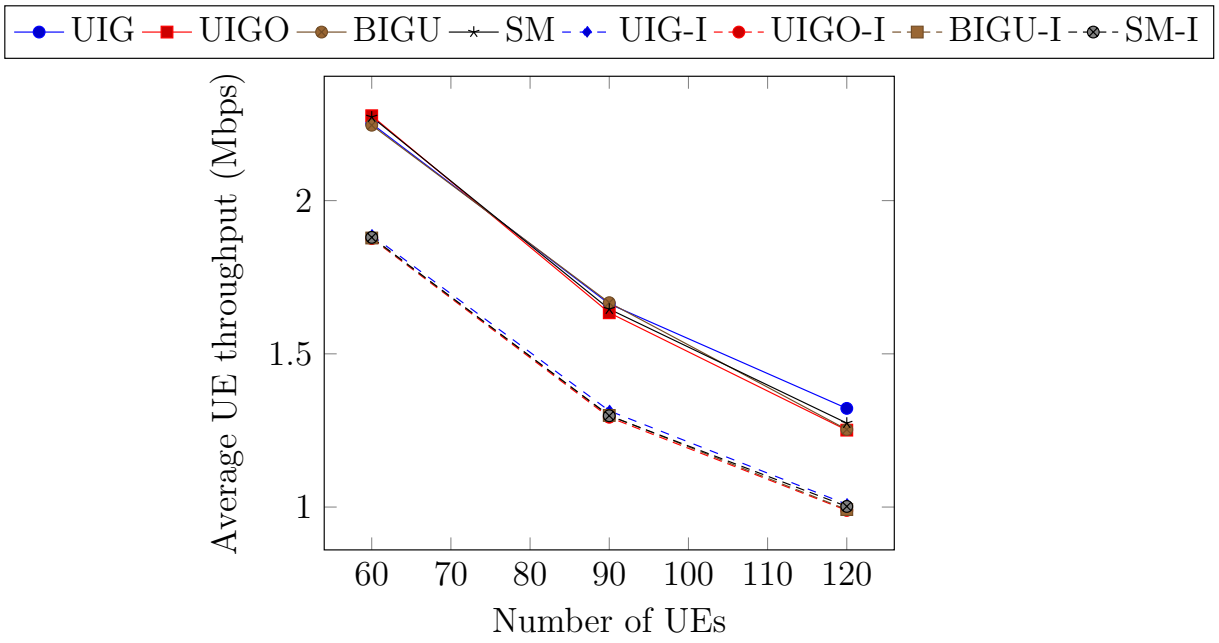

Figure 5.11: Average UE throughput vs number of UEs for PF-DC and standard PF scheduling algorithms in Scenario 2.

to those obtained for the random deployment of pico cells. For hotspot deployment of UEs, there is not much variation in PF utility for different UE association algorithms. However, for the uniform deployment of UEs, UIGO algorithm gives a higher PF utility as compared to other UE association algorithms. One of the interesting findings is that the UE association algorithms give a comparable value of PF utility for all four scenarios. This implies that no matter what UE association scheme we choose, the improvement in PF utility depends only upon the scheduling scheme used. Therefore, the PF-DC scheme improves the average per-UE throughput as well as fairness in the system.

5.4.2 Comparison of Dual Connectivity Procedure (DCP) with All Connectivity (ACP) and Single Connectivity Procedures (SCP)

In this section, we compare the procedures DCP, ACP, and SCP. In SCP, a UE is connected to the BS providing it with maximum RSRP and standard PF scheduling (Section **??**) is performed at every BS. In DCP, the dual connectivity UE association algorithms proposed in Section **??** are used, and PF-DC scheduling (Section **??**) is performed. In ACP, all

UEs are connected to every BS, and GPF scheduling is performed at each BS. The period of information exchange T of PF-DC scheme is considered to be 25 slots. We compare the performance of these procedures based on PF utility, system throughput, and jain's fairness index. JFI denotes the fairness in the throughput values obtained by all UEs in the system. It is given by $\{\sum_{j=1}^{M} x_j\}^2/M\sum_{j=1}^{M} {x_j}^2$, where x_j denotes the long term throughput of UE j. The value of JFI is one if all UEs get equal throughput. Thus, JFI closer to one implies a high level of fairness between the UEs. We demonstrate the performance of SCP, DCP, and ACP in four different scenarios mentioned earlier.

Fixed locations of Pico cells

In this section, we discuss the simulation results of Scenarios 1 and 2. Figures **??**, **??**, and **??** illustrate the variation in PF utility, system throughput, and JFI of ACP, DCP, and SCP as a function of the number of UEs in the system in Scenario 1. As the number of UEs in the system increases, system throughput increases due to an increase in multi-user diversity in the system. However, the rate of increase in system throughput is low. Fairness in the system decreases as the number of UEs increases. Hence, the PF utility decreases with an increase in the number of UEs. We observe that ACP provides high PF utility, i.e., it provides a balance between total throughput and fairness in the system as compared to DCP and SCP. In Scenario 1, the coverage of pico cells is strictly disjoint, and the deployment of UEs is hotspot type. SCP provides the least PF utility since it neither utilizes the entire capacity in the network nor gives fairness to UEs. SCP offers a wide range of throughput to the UEs, which is clear from Figure **??**. The dual connectivity association algorithms (UIG, UIGO, BIGU, and SM) in conjunction with the PF-DC scheduling algorithm (DCP), provide a substantial improvement in PF utility as compared to SCP, but it is slightly less than ACP. Though ACP provides maximum fairness, the system throughput values for DCP and ACP are almost comparable. UIGO algorithm gives better PF utility as compared to the other dual connectivity algorithms in this scenario.

Figures **??**, **??**, and **??** depict the variation of PF utility, system throughput, and JFI of ACP, DCP, and SCP as a function of the number of UEs in the system in Scenario 2. This scenario contains fixed pico cells deployed with disjoint coverage and uniform UE

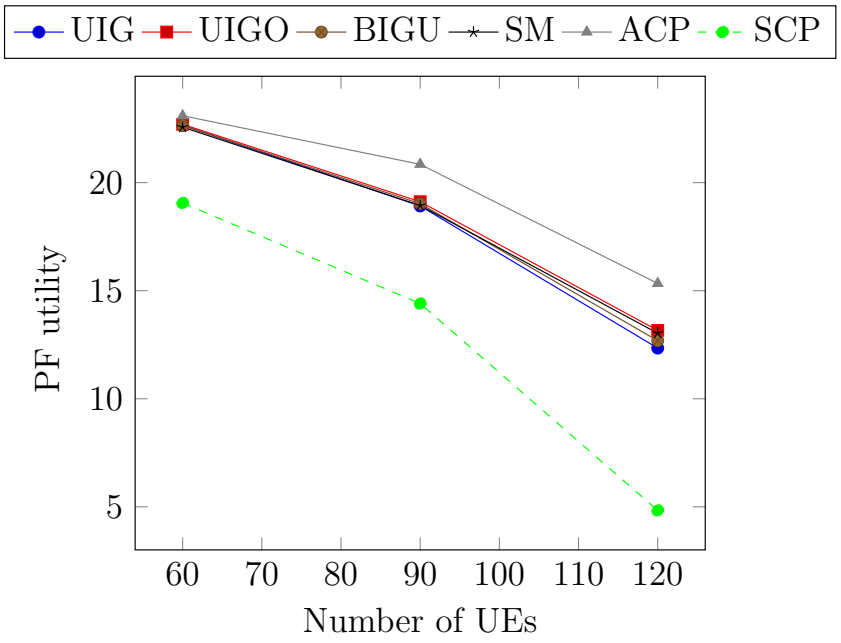

Figure 5.12: Comparison of PF utility for various algorithms in Scenario 1.

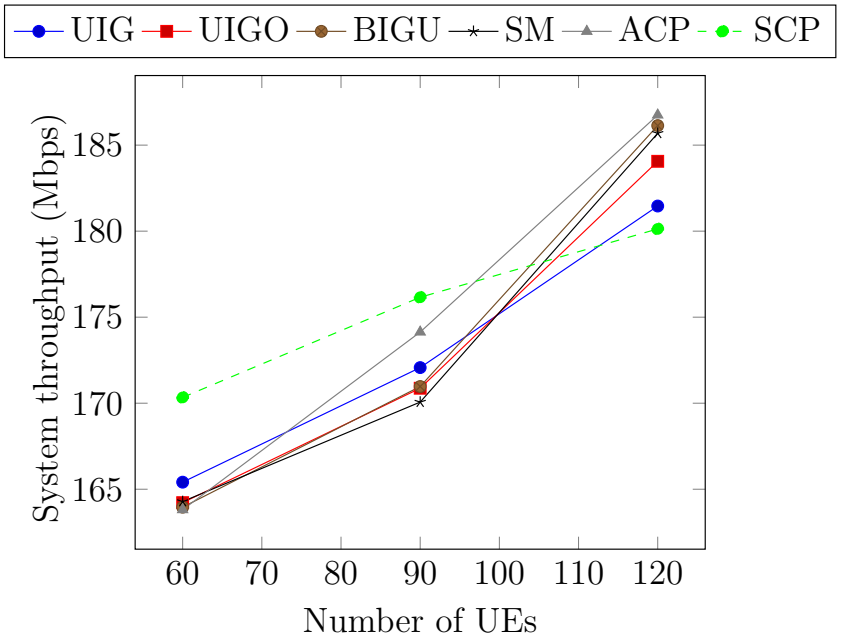

Figure 5.13: Comparison of system throughput for various algorithms in Scenario 1.

deployment. In this scenario, the UEs may not be deployed inside pico coverage. Hence, SCP is performing worse as compared to its performance in Scenario 1. In this case, SCP gives minimum system throughput and minimum PF utility. ACP provides maximum PF utility as it gives maximum fairness to UEs as well as maximum total throughput as compared to other algorithms. In uniform UE deployment, ACP has an advantage

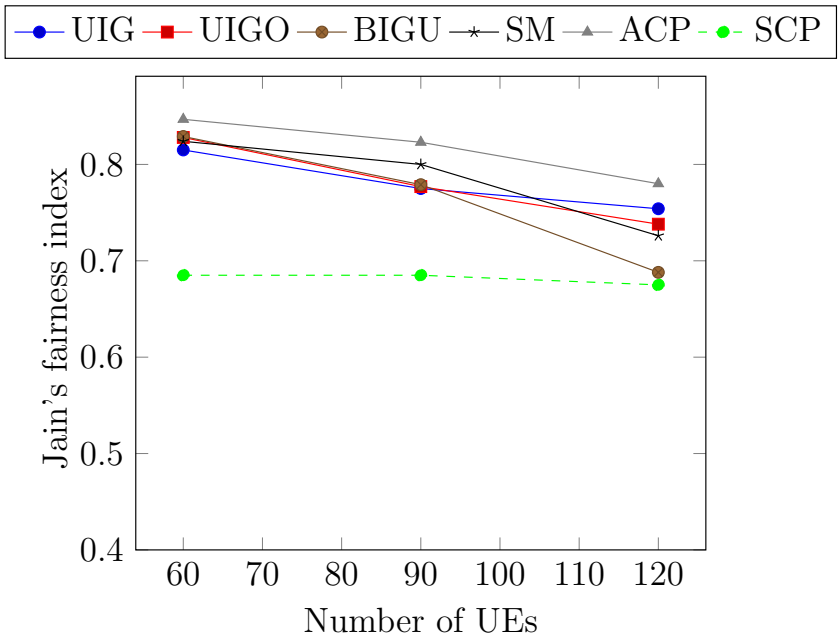

Figure 5.14: Comparison of JFI for various algorithms in Scenario 1.

as it balances the load among the BSs by opportunistically choosing the UE with peak rates at each BS. Dual connectivity algorithms perform slightly worse than ACP due to the limitation in the number of connections, but the total throughput and JFI values are almost comparable for DCP and ACP. UIGO algorithm performs better among the dual connectivity association algorithms in this scenario.

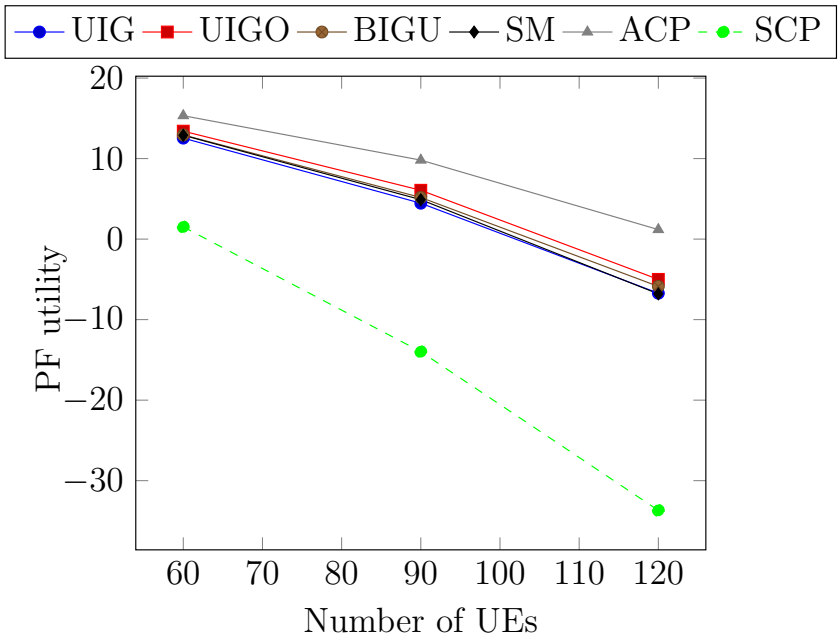

Figure 5.15: Comparison of PF utility for various algorithms in Scenario 2.

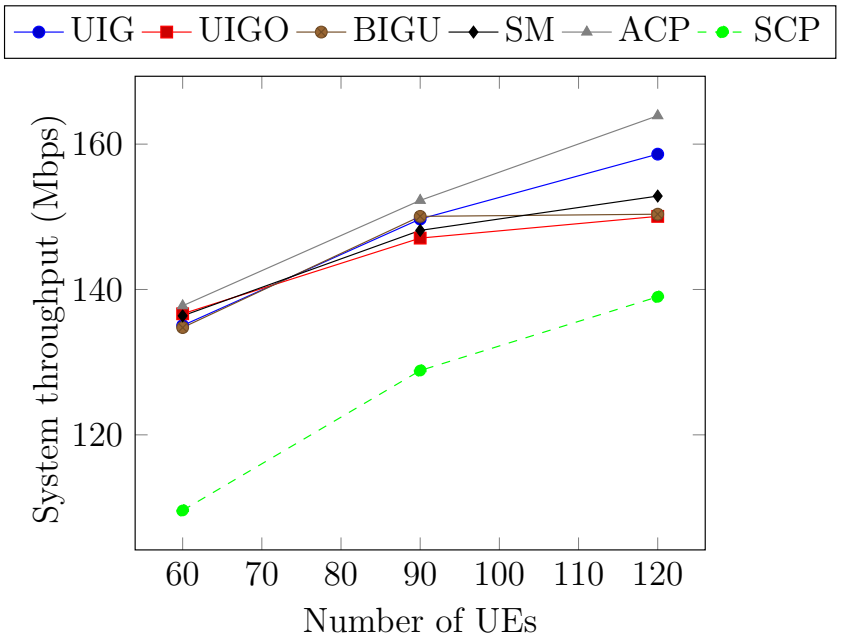

Figure 5.16: Comparison of system throughput for various algorithms in Scenario 2.

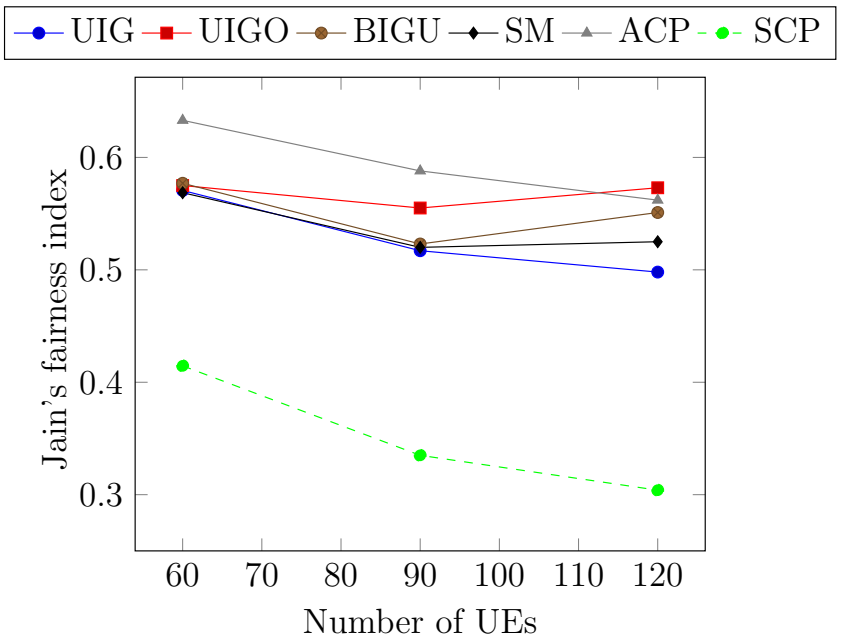

Figure 5.17: Comparison of JFI for various algorithms in Scenario 2.

Random locations of Pico cells

In this section, we present the simulation results of Scenarios 3 and 4. Figures **??**, **??**, and **??** illustrate the variation in PF utility, system throughput, and JFI of ACP, DCP,

and SCP as a function of the number of UEs in the system in Scenario 3. In this case, ACP gives maximum PF utility along with maximum system throughput. The benefit of multiple connections is accentuated by the overlapping coverage of pico cells as well as hotspot deployment of UEs in this scenario. Overlapping coverage works as an advantage for ACP and DCP. However, in SCP, UE connects to the BS from which it receives maximum signal strength. Consider a UE situated in the overlapping coverage area of two cells. In SCP, the UE is connected to one of these two BSs, whereas in DCP and ACP, the UE utilizes the resources of both cells. In SCP, the load among the BSs may not be distributed, and there may be overloaded cells. Hence, as compared to DCP and ACP, SCP gives the worst PF utility in this scenario. In DCP, a substantial increase in PF utility is observed as compared to that of SCP, but it is lower than that of ACP. UIGO algorithm contributes high PF utility as compared to other proposed algorithms.

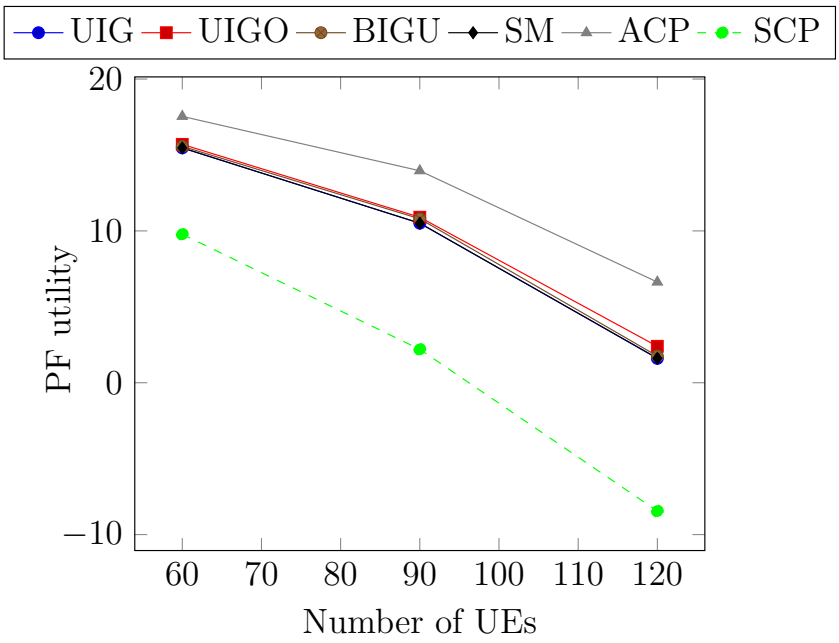

Figure 5.18: Comparison of PF utility for various algorithms in Scenario 3.

The plots for Scenario 4 are depicted in Figures **??**, **??**, and **??**. ACP maintains maximum PF utility by providing high system throughput and high fairness among the UEs. The multiple connections offer benefit to the UEs in overlapping pico coverage areas. SCP provides the lowest PF utility due to weak fairness among UEs as well as low total throughput. UEs are opportunistically connected to their best BS, thus leading to an imbalance in the system. Dual connectivity algorithms provide PF utility lower than

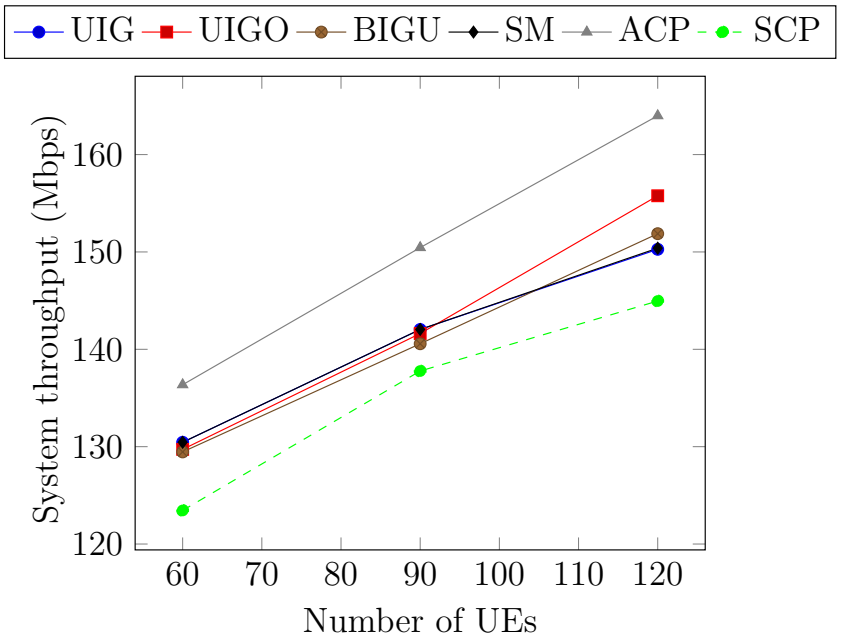

Figure 5.19: Comparison of system throughput for various algorithms in Scenario 3.

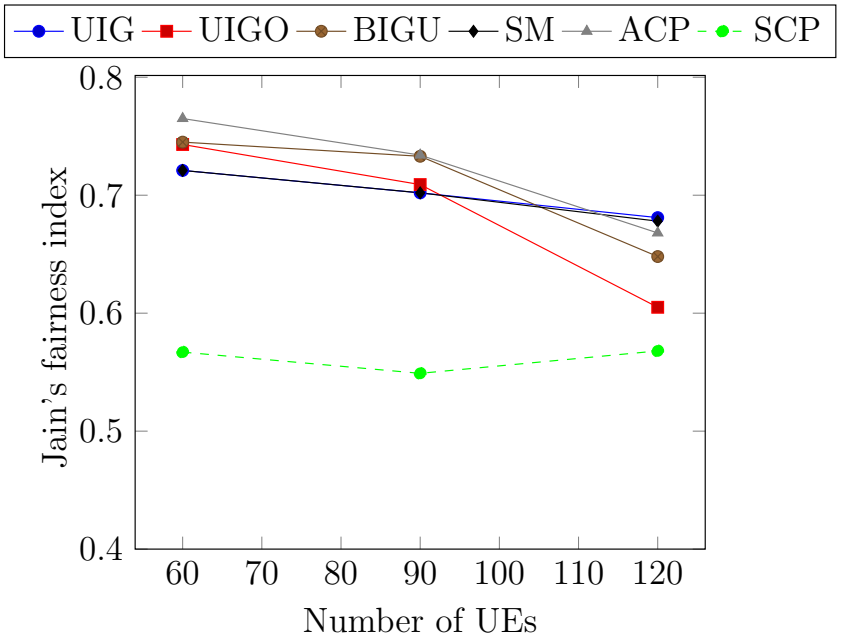

Figure 5.20: Comparison of JFI for various algorithms in Scenario 3.

ACP but much higher than SCP by creating a balance between total throughput and fairness among UEs. UIGO algorithm contributes high PF utility as compared to other proposed algorithms in this scenario.

We present some general inferences from the results obtained. The values for system throughput and PF utility are higher in general for the hotspot deployment of UEs since

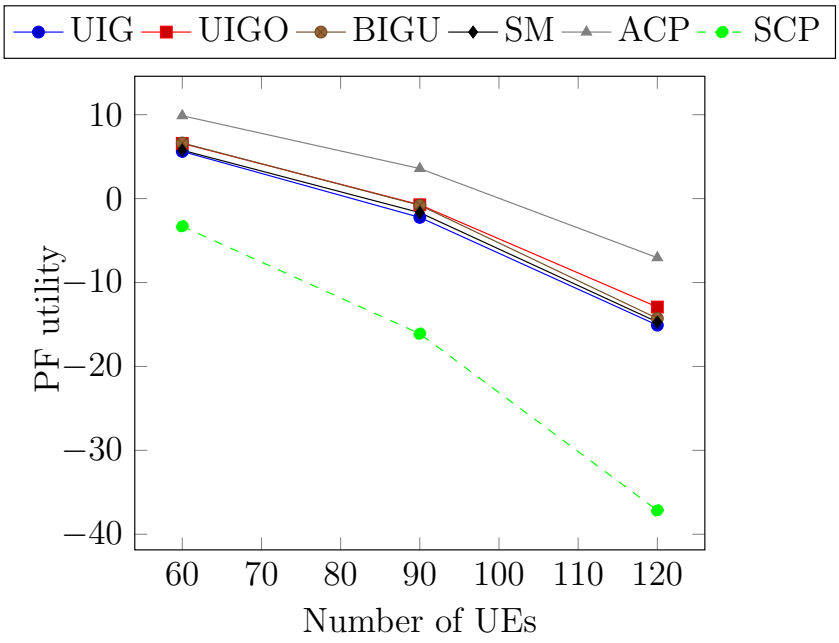

Figure 5.21: Comparison of PF utility for various algorithms in Scenario 4.

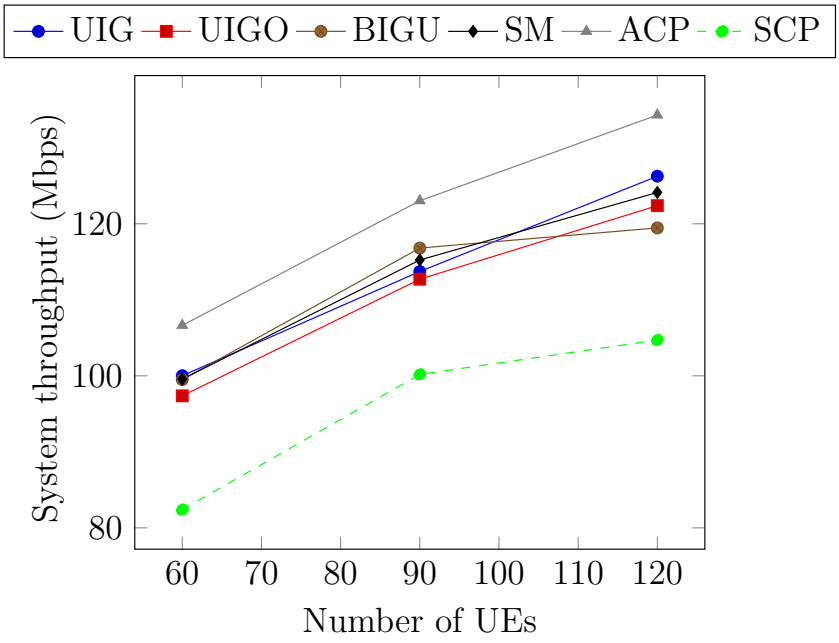

Figure 5.22: Comparison of system throughput for various algorithms in Scenario 4.

more UEs are located in areas where the deployed BSs provide coverage. The performance of proposed association algorithms does not exhibit much variation in terms of the PF utility system metric. This implies that the association of dual connected UEs does not play a major role in the PF utility of the system, but it matters how PF scheduling is performed. There is a remarkable improvement from SCP to DCP in all four scenarios,

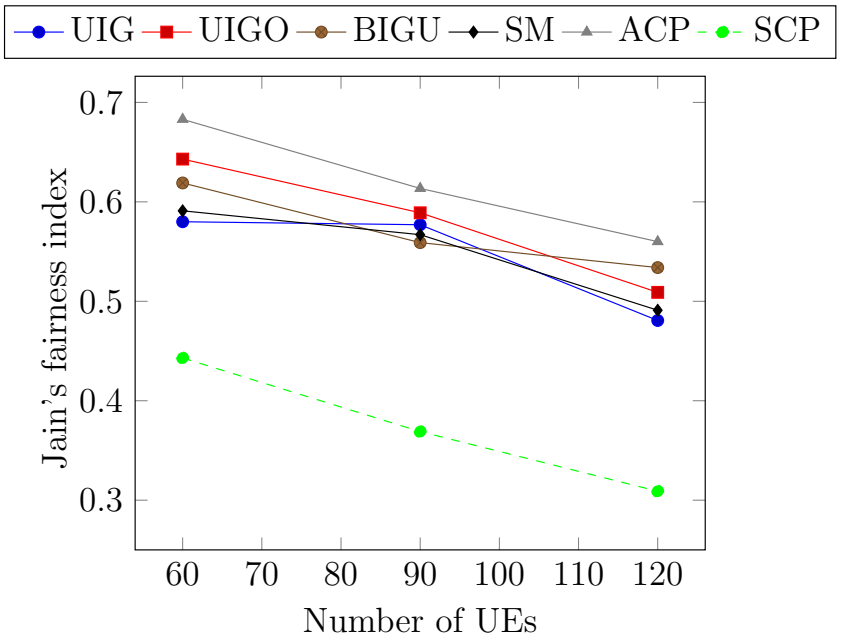

Figure 5.23: Comparison of JFI for various algorithms in Scenario 4.

but the improvement from DCP to ACP is not as much. This improvement from DCP to ACP is even less in Scenario 2. In this scenario, there is no overlap between pico cells, and UEs are distributed uniformly in the macro coverage. The advantage of multiple connections is available for only a few UEs, and hence, the performance of ACP and DCP is almost comparable. We demonstrate in Section **??** that the improvement for DCP to ACP comes at an extra computational cost.

5.4.3 Variation in PF utility with variation in period T

In this section, we demonstrate the variation in PF utility of the system with a change in the period of information exchange T of the PF-DC scheme. We vary the period T and keep other parameters of the network constant. We perform simulations for a total of 90 UEs. The simulation parameters of the network are enumerated in Table **??**. Figures **??** and **??** illustrate the PF utility in the system against variation in period T for Scenarios 1 and 2, respectively. There is not much variation in PF utility in case of Scenario 2 as the period T increases. In the case of Scenario 1, the PF utility decreases as the period T increases from 1 slot to 50 slots. This is because as T increases, the throughput information is shared less frequently with the controller, and as a result, the total throughput in the system is decreased. The values of PF utility are higher for

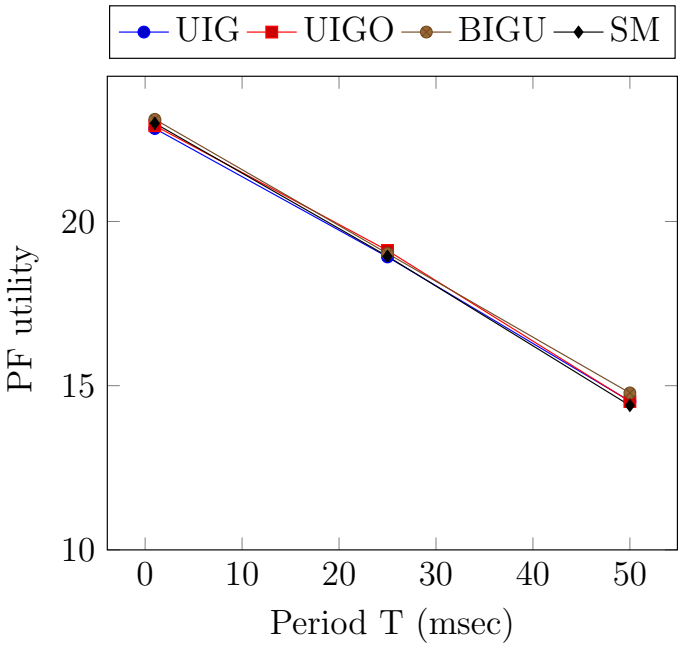

Figure 5.24: PF utility vs period T for proposed algorithms in Scenario 1.

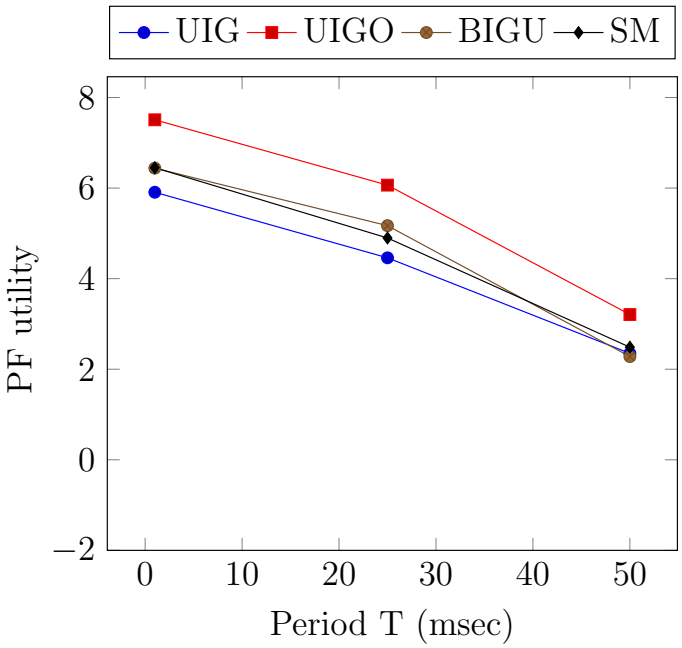

Figure 5.25: PF utility vs period T for proposed algorithms in Scenario 2.

Scenario 1 since more UEs are placed in pico cell coverage areas, and these UEs achieve higher throughput bringing up the total throughput in the system.

Figures **??** and **??** depict the PF utility in the system with variation in period T in Scenarios 3 and 4, respectively. There is not much variation in PF utility in case of

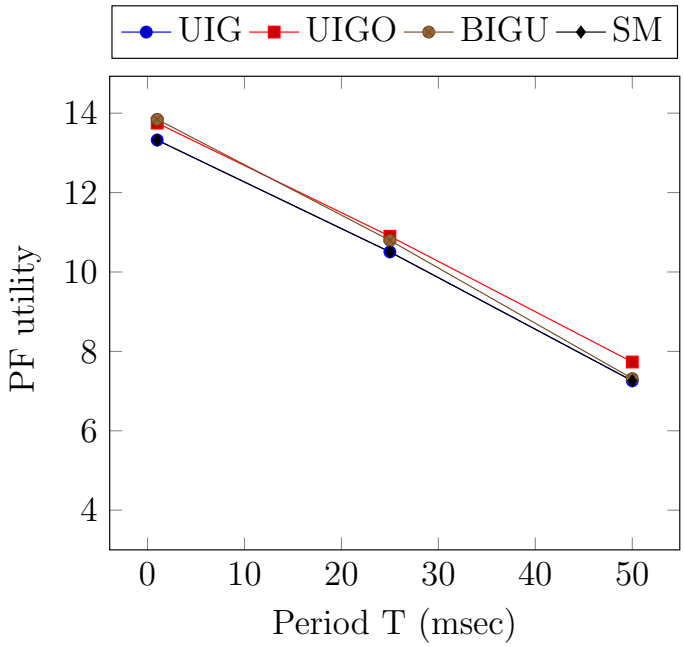

Figure 5.26: PF utility vs period T for proposed algorithms in Scenario 3.

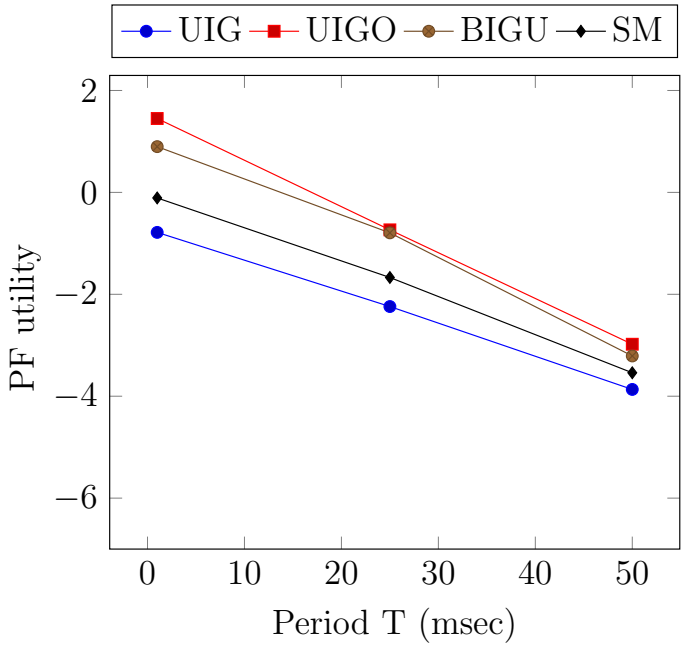

Figure 5.27: PF utility vs period T for proposed algorithms in Scenario 4.

Scenario 4 as the period T increases. In the case of Scenario 3, the PF utility decreases slightly as the period T increases. For instance, the value of PF utility for UIGO algorithm decreases from 14 to 8 as the period T increases from 1 slot to 50 slots. This is because increasing T leads to lower total throughput in the system. Thus, there is not much

variation in PF utility in the system as T is varied from 1 to 50 slots. The value of the period T can be chosen according to the requirements of the service provider.

5.5 Complexity and Implementation Analysis of PF-DC

In this section, we compare the communication and computation complexity of the PF-DC scheme with that of the GPF scheme [**?**]. We consider only signaling information exchange for our analysis. In the case of ACP, all UEs are connected to all available BSs. Hence, in every T time slots, the information is being exchanged between each pair of BSs for each UE. Thus, the communication complexity for ACP comes out to be $\frac{1}{T}MN(N-1)$. In our network architecture, SRC has access to all BSs in the RAN. Each UE is connected to two BSs, and each BS shares its UE information with SRC. SRC processes this information and sends the total throughput information back to the BSs every T time slots. Hence, the communication complexity is $\frac{1}{T} \times 2 \times 2 \times M$.

Let us compare the computational complexity of ACP and DCP. The basic computations required by the PF scheduling algorithm are comparisons, additions, and multiplications. For choosing the UE to be scheduled in a scheduling interval, we need to determine the UE with the maximum PF metric. For determining the maximum from among the UEs connected to a BS, ACP requires $M-1$ comparisons per BS, i.e., a total of $N(M-1)$ comparisons. In the case of DCP, let each BS i has x_i connections. Then the total number of comparisons is $\sum_{i=1}^{N}(x_i - 1)$. This comes out to be $\sum_{i=1}^{N} x_i - N$. The total number of connections for M UEs is $2M$, since we allow each UE to have two connections. Thus, the number of comparisons is $2M - N$. The number of additions and multiplications for PF scheduling in case of ACP are $NM + \frac{1}{T}(N-1)M$ and $3NM$, respectively [**?**]. In the case of DCP, this comes out to be $2M + \frac{1}{T}M$ and $6M$ respectively. Thus, we see that the communication and computation complexity of PF in our network architecture with DCP is less as compared to the PF scheme proposed in [**?**].

5.6 Conclusions

In this chapter, we have proposed a low complexity centralized PF scheduling scheme for dual connectivity (PF-DC), based on SMRAN architecture of Chapter **??**. We have also proposed several UE association algorithms for dual connectivity. We compared the PF-DC scheme with standard PF scheduling independently performed at each BS in conjunction with the proposed algorithms. We demonstrated that dual connectivity, along with the PF-DC scheme outperforms dual connectivity with the standard PF scheduling scheme at each BS. The PF-DC scheme improves the per-UE throughput as well as the fairness in the system. We also compared dual connectivity in conjunction with PF-DC scheduling, with single connectivity (SCP) and all connectivity (ACP). The analysis suggests that the addition of the second radio link along with the proposed PF-DC scheduling algorithm gives remarkable gains in PF utility over SCP. It also demonstrates that further gain in PF utility with additional radio links (ACP) is marginal and comes at a significant additional cost of maintenance of a large number of connections for a UE. We showed that the communication complexity, as well as the computation complexity of the PF-DC scheme, is low as compared to that of the ACP scheme.

From the simulation results, we observed that the proposed UE association algorithms give similar performance as far as the PF utility is concerned. This implies that no matter what UE association scheme we choose, the improvement in PF utility depends only upon the scheduling scheme used. However, the UIGO algorithm gives slightly better PF utility than other proposed algorithms for all four scenarios considered in this work. We also demonstrated the performance of the PF-DC scheme with the variation in the period of information exchange T and observed that the variation is not substantial. The period T can be adjusted to suit the requirements of the service provider. Even though we proposed to use a centralized RAN controller for the execution of the PF-DC algorithm, the proposed PF-DC algorithm can easily be employed in existing HetNets as well where one of the BSs can act as the centralized controller for a subset of UEs.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we have addressed various problems in dual connectivity pertaining to different scenarios and proposed solutions to resolve these problems. As pointed out in Chapter **??**, a majority of the literature addresses improving mobility robustness and maximizing throughput using dual connectivity in a heterogeneous environment. Some existing works have also focused on the energy-saving aspect of dual connectivity. However, few studies exist which consider minimization of average system delay using the traffic splitting capability of dual connectivity. Moreover, this literature does not consider dynamic arrival and departure of users. Further, the trade-off between delay minimization and blocking of users has not been addressed in existing studies. In this thesis, we have considered all these aspects and formulated an optimization problem. To the best of our knowledge, this is the first work which considers dynamic user arrival and departure in delay minimization using dual connectivity, and we conclude that dual connectivity minimizes the average delay in the system subject to a constraint on the blocking probability of arriving traffic.

The second problem that we consider is the non-uniformity in the different architectures proposed in the standards for dual connectivity. Moreover, the signaling overhead is high in the existing architectures for dual connectivity leading to additional delays in inter-node communications. We have addressed these issues and proposes SMRAN - an SDN based Multi-RAT RAN architecture for all variants of dual connectivity. We also define detailed protocols at every node and interface.

We consider the problem of proportional fair scheduling in dual connectivity and propose PF-DC, a simplified, low complexity modified PF scheduling algorithm based on SMRAN. The proposed PF-DC scheme outperforms standard PF scheme in terms of PF utility as well as average user throughput in the system. We also propose heuristic user association algorithms and evaluate their performance in various scenarios. We now discuss our contributions in further detail.

In Chapter **??**, we consider a system comprising a single macro cell and a single small cell and model it using a queuing system. The backhaul link between small cell and macro cell is also modeled as a queuing system. We formulate an optimization problem to minimize the average delay in the system subject to a constraint on the blocking probability of arriving traffic. We frame the problem in an MDP framework along with a constraint. The model captures the dynamic nature of user arrival and departure observed in practice. We consider the arrival of traffic in the form of a batch, since generally in the Internet, traffic occurs in bursts. The cost function in our model is the expected time the batch takes in the system, which is proportional to the expected number of packets in the system. Our model captures the user performance in terms of the delay experienced by it as well as the network perspective in terms of the preference given to background and foreground users.

The constrained MDP problem is solved using gradient descent and value iteration algorithms, and the optimal policy is determined. A special application is implemented in ns-3 to simulate batch type traffic. Extensive simulations are performed in ns-3 to compare the performance of the optimal policy with a traditional policy, known as join the shortest queue policy. The join the shortest queue policy does not have the traffic splitting capability, and hence, the optimal policy is found to perform better than this policy. However, the determination of optimal policy induces a computation complexity of exponential order. We propose two heuristic policies whose computational complexity is constant. Along with reducing the computation complexity, the proposed policies also yield a near-optimal performance. For simulations, we have considered fading in a typical urban environment as well as the mobility of users.

In Chapter **??**, we propose SMRAN - an SDN based Multi-RAT RAN architecture for dual connectivity, which consists of a centralized RAN controller with control-plane

functionality to control multi-RAT RAN nodes. SMRAN integrates 3GPP RATs (LTE and 5G) with non-3GPP access technologies (WLAN) and designs a common interface (Xd) for communication across nodes belonging to these RATs. The different variants of dual connectivity, viz., LTE dual connectivity, LWA, and MR-DC, are seamlessly integrated into this architecture. The architecture is flexible and can easily be extended to include other RATs as well. SMRAN provides integrated control over the functionality of RAN with a common unified interface to the core. SMRAN is a simplified architecture with details of protocols at each of the nodes as well as interfaces.

Due to the separation of control and data planes in SMRAN, the load, mobility, and interference in RAN are better managed. The control signaling in case of dual connectivity is reduced in this architecture, and the system performance is improved with the design of efficient algorithms. We demonstrate this using extensive simulations in practical scenarios in ns-3. SMRAN makes the processing of data plane protocol layers for dual connected users flexible at any of the RAN nodes. Moreover, the signaling connection for UEs can take place at any node, which is not possible in the existing architecture. Therefore, using this architecture for dual connectivity, there is no special requirement of a master node and secondary node. Any node can act as a master or secondary node since the control lies with the controller. Although this architecture is proposed keeping the requirements of dual connectivity into perspective, this is a generic architecture which can be used without dual connectivity as well.

In Chapter **??**, we discuss proportional fair scheduling in dual connectivity. Based on SMRAN, we propose PF-DC - a low complexity PF scheduling mechanism for dual connected users. We compare the performance of PF-DC scheme with the standard PF scheduling algorithm in dual connectivity scenario and observe that PF-DC gives a remarkable improvement in PF utility as well as average per-user throughput. We also propose various user association algorithms for dual connectivity. These algorithms select two appropriate base stations for dual connected users based on different criteria of signal strength. We implement PF-DC scheme in ns-3 using a centralized architecture along with the proposed algorithms.

Extensive simulations have been performed in ns-3 for four different types of scenarios, including fixed and random placement of pico cells. We compare the performance

of the proposed dual connectivity association algorithms using PF-DC scheduling scheme and single connectivity using standard PF scheduling. We observe that there is a significant improvement in PF utility from single connectivity to dual connectivity. On comparison of dual connectivity in conjunction with PF-DC scheduling with all connectivity in conjunction with global PF scheduling, we observe that the improvement in PF utility from dual connectivity to all connectivity is not that remarkable. Moreover, all connectivity leads to an additional cost of increased complexity and processing requirements in the network as compared to dual connectivity. PF-DC scheme has been simulated for different values of the period of information exchange T, and it is observed that the variation in PF utility is not substantial. The period T can be tuned as per the requirements of the service provider. One of the interesting findings of this work is that no matter what dual connectivity algorithms are used for user association, the PF utility depends only on the PF scheduling scheme used. It can be concluded that dual connectivity improves proportional fairness in the system as compared to single connectivity.

In summary, we have designed computationally efficient algorithms for traffic splitting in dual connectivity that minimize the average delay in the system subject to a constraint on the blocking probability of users. We have devised an SDN-based multi-RAT RAN architecture that seamlessly integrates different variants of dual connectivity and provides a simplified network control. We have developed simple heuristic algorithms based on this architecture to select the dual connected users as well as for user association. We have also designed a low complexity modified PF scheduling algorithm based on this architecture. Though specific scenarios were considered for proposing the algorithms in this thesis, these can be generalized and easily adopted for implementation in practical systems such as LTE, 5G.

While some of the problems in dual connectivity have been addressed in this thesis, others still need to be examined. In the next section, we discuss some problems open for future work.

6.2 Future Work

In Chapter **??**, we cast the traffic splitting problem as a Markov Decision Process and then determine the optimal traffic splitting policy for batch type traffic. However, using MDP requires complete knowledge of the system model, including arrival rates, service rates, and transition probabilities. In practice, it is difficult to obtain these exact parameters for modeling the system. Online algorithms or learning-based algorithms can be proposed, which do not require the knowledge of the system model. Another extension could be to include handoff traffic arrival in the model to incorporate mobility. We have minimized expected delay of the arriving traffic in our work. We could also minimize the variance of delay between the two paths instead. The number of resources in MeNB and SeNB subsystems (n_1 and n_2) is constant. Another possible extension could be to make the total number of resources ($n_1 + n_2$) constant and adjust the number of resources allocated to each system (n_1, n_2) periodically, based on the state of the system.

In a general HetNet, single connected users coexist with dual connected users. In such a scenario, dual connected users consume resources from two base stations at the same time. This may leave fewer resources for the use of single connectivity users. Thus, the presence of dual connected users may hamper the performance of SC users. In such a scenario, joint scheduling of all users at a controller situated above the base stations may be beneficial. This could be one of the future directions.

In a multi-tier network with macro and small cells, single connected users are present alongside users with dual connectivity. In such a scenario, the users can bid for dual connectivity access, and the base stations can auction the available resources to these two types of users at a cost. The service providers can earn additional revenue by providing the service of dual connectivity. This problem can be formulated as an optimization problem to determine the set of dual connected users.

Different services require varying quality of service. For instance, VoIP traffic has tight delay requirements but does not require high bit rates and can tolerate rather high packet losses. In contrast, best-effort traffic benefits from higher bit rates but is less delay sensitive as compared to VoIP traffic. Taking such QoS requirements into account, design of user association algorithms for dual connectivity can be studied. Based on the requirements of the user, dual connectivity should be established or released. The energy

perspective of a dual connected user should also be taken into account. For instance, since dual connectivity uses two interfaces of the UE simultaneously, it may lead to more energy consumption at the UE. If the UE is in power save mode or the battery is not sufficiently charged, then dual connectivity can be deactivated for such a UE. Such practical problems in dual connectivity can be studied in the future.

Another possible problem for future work is to analytically analyze the chances of obtaining a secondary link with good signal strength. Out of all the users present, the percentage of users which can have a good secondary link can be determined.

List of Publications

Patents

P1 A. Karandikar, P. Jha, P. K. Taksande, "A Software Defined Network based Multi-Radio Access Technology architecture for Dual connectivity", Patent filed, June 2019 (India) (201921025411).

International Conference Publications

C1 P. K. Taksande, A. Roy, and A. Karandikar, "Optimal traffic splitting policy in LTE-based heterogeneous network," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, 2018.

C2 P. K. Taksande, P. Jha, and A. Karandikar, "Dual connectivity support in 5G networks: An SDN based approach," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, 2019.

C3 P. K. Taksande, P. Chaporkar, P. Jha, and A. Karandikar, "Proportional Fairness through Dual Connectivity in Heterogeneous Networks", in *2020 IEEE Wireless Communications and Networking Conference (WCNC).*

C4 P. K. Taksande, P. Jha, A. Karandikar, and P. Chaporkar, "Open5G: A Software-Defined Networking Protocol for 5G Multi-RAT Wireless Networks", in *2020 IEEE Wireless Communications and Networking Conference (WCNC) - Workshop.*

Submitted/Under Preparation

J1 P. K. Taksande, P. Chaporkar, and A. Karandikar, "Optimal traffic splitting algorithm in heterogeneous network", Under Preparation (Journal).

J2 P. K. Taksande, P. Chaporkar, P. Jha, and A. Karandikar, "Traffic splitting for Proportional Fairness in Heterogeneous Networks", Under Preparation (Journal).