UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Enhanced mobility management mechanisms for 5G Networks

Akshay Jain

PhD program in Network Engineering

Enhanced Mobility Management Mechanisms for 5G Networks

Doctoral thesis by:
Akshay Jain

Thesis Advisors:
Dr. Elena López-Aguilera
Dr. Ilker Demirkol

Department of Network Engineering
Barcelona, Spain
July 2020

Acknowledgments

This thesis is not just an embodiment of the work that has been done over the course of last three and half years, but it is also a symbol of how many people come together to make something, that at many stages seemed outright impossible, possible.

At the very outset, I would like to express my profound gratitude to my doctoral thesis advisors Dr. Elena López-Aguilera and Dr. Ilker Demirkol. They have been instrumental in helping me develop not just my technical skills but also as a human being during this entire process. For me they have been like a family away from my family, and during the numerous coffees and work related trips I was able to learn extensively about the art of doing research. In a more formal setting, they have been instrumental in guiding me through some of the most difficult moments during my research by providing not just technical guidance but also moral support. For that I am eternally grateful to them. With these experiences, it is my hope that someday I will be able to emulate not just their wisdom but their kindness and empathetic nature as well, whilst forging my career in this increasingly competitive world.

All of this would of course have not been possible without the unwavering support of my mother and father. They have forever been my bedrocks, and I truly believe that it is the values that they instilled in me which have allowed me to conquer many of the challenges that I met along this journey. Moreover, it is their sheer strength and patience which has inspired me to forge ahead at many junctures of my short career. For this I am indebted to them for life. I would also like to thank the rest of the family here, for being understanding and supportive whenever I needed them to be. I would like to specially mention Mr. Amit Jain for inspiring me to take up wireless communications and Ms. Pranjali Jain for being the elder sibling I never had and providing all the moral support.

During the period of this thesis, I met many researchers who became colleagues, and ultimately best friends and in some cases even brothers/sisters. I would specially like to mention Mr. Rakibul Islam Rony, Mr. Girma M. Yilma, Mr. Mikel Irazabal, Dr. Francesco Devoti, Mr. Victor Baños-Gonzalez, Dr. Jorge E. Gaitán Pitre, Mr. Matteo Vincenzi, Mr. Nikolaos Giatsoglou, Mr. Lanfranco Zanzi, Dr. Jian Song, Dr. Alejandro S. Gonzalez, Dr. Adriana Fernández-Fernández, Dr. Leonardo Ochoa-Aday, Ms. Irian Leyva-Pupo, Mr.

Alejandro, Mr. Carlos P., Mr. Ahmed, Mr. Asif Habibi, Dr. Khalid, Dr. Xavier Costa, Dr. Vincenzo Sciancalepore, Dr. Marco Di Renzo, Dr. Birkan H. Yilmaz, Dr. Sergi Abadal, Dr. Albert Cabellos-Aparicio, Dr. Eduard Alarcón, Dr. Ali Sadeghian, Dr. Lim Deoksu and Mr. Haazy Haastrup for playing a significant role towards the completion of my thesis. For the various conversations (technical/personal) and memorable moments, I am eternally grateful.

Outside of work, my life in Barcelona has been enriched by interactions with people from various walks of life. They have helped me settle down in this vibrant city and have been my support throughout. I would like to specially mention Ms. Esther Xalabarder, Mr. Marti Rodrigo, Ms. Marisol, Ms. Laura Vargas, Mr. Carlos, Ms. Jacqueline, Mr. Flavio Fernandez, Ms. Silvina Sanchez, Mr. C Anand Iyer, Ms. Debarati Shome, Mr. Dhaval Gadariya, Ms. Judith Murray, Mr. Alan Urquhart and Ms. Boglarka Nagy for their unwavering support and helping me through this process. I will forever be grateful to you.

My journey to this point has its roots to my time back in the USA. It is here where I learnt my art under the most challenging circumstances. But again I was blessed by the grace of god, and I met some of the best friends and colleagues over here. I would like to specially mention Ms. Stuti Joshi, Dr. Akanksha Pandey, Mr. Rahul Bhatia, Dr. Abhilash Paneri, Mr. Subodh Chaturvedi, Mr. Ankit Gupta, Mr. Chris Blower and Mr. Akash Dhruv for their unwavering support, belief and unmatchable acts of kindness towards me. For their friendship and contributions towards my development as a person and as a researcher, I am extremely grateful.

I would also like to take this opportunity to mention the people in India who have been instrumental in my development as a person, since the time of my Bachelors. Specifically, I would like to thank Mr. Victor Roy, Mr. Siddhant Dash, Mr. Amit Kumar, Ms. Sadhvi Aggarwal, Mr. Animesh Kumar, Mr. Bhavik Gattani, Mr. Vinayak Iyer, Ms. Shruti Mahajan, Ms. Kanika Patoria, Mr. Nimish Shah, Mr. Nishant Tilokani and Mr. Keshav Mathur. I am eternally grateful and indebted to you for your extremely vital and significant contributions towards my life as a professional and as a person.

Last but not the least, I would like to thank UPC, the doctoral school, the entire Network Engineering department, the various cafeteria and administrative staff members, the many anonymous journal/conference reviewers as well as the thesis reviewers for their unflinching support, guidance and consideration towards my goal of completing this thesis in the best possible manner.

I would like to state here that, I have tried my best to mention everyone who has been a part of this journey. However, if in any case somebody's name doesn't appear, then I sincerely and whole-heartedly request your pardon.

Preface

Wireless standards such as 2G, 3G, 4G and currently 5G, promise incremental improvements in the Quality of Experience (QoE) and Quality of Service (QoS) when compared to their predecessor technologies (e.g., 5G promises better QoE and QoS than 4G/3G/2G). The quantitative measures of QoE and QoS relate to improved throughput, reliability, etc., from the perspective of the user. These measures of QoE and QoS are tightly coupled with the type of applications (e.g., Emergency services will require low latency and high reliability, while broadband services will require high bandwidth with less stringent latency and reliability measures as compared to emergency services). Further, up until 4G, industry and to some extent academia, through 3GPP, IETF and ETSI, defined methods that served the networks infallibly. However, with the industry facing a significant downturn in their revenues, the prospect of integrating other business verticals as well as moving towards a more softwarized (and thus economical) network deployment approach has led to the advent of 5G and hence, a revolution.

However, such revolutions, as we may know from our knowledge in history, involves significant transformations in each section of the community. Similarly, many mechanisms that served the legacy telecommunication networks for months, years or decades are now being identified as being non-usable or at best sub-optimal. The reason being, increased heterogeneity, complexity and density within the 5G networks as compared to any other legacy system. And so, one such class of mechanisms, which are also extremely critical for any wireless standard, are the Mobility Management (MM) mechanisms. Mobility Management mechanisms ensure the seamless connectivity and continuity of service for a user when it moves away from the geographic location where it initially got attached to the network. But, and as we have already indicated, the 5G network characteristics render the legacy MM approaches as being either non-usable or inefficient.

Hence, in this thesis, we firstly explore the various mechanisms that have been employed or conceived to perform Mobility Management in legacy (2G/3G/4G) as well as 5G networks. Further, based on the 5G requirements as well as the initial discussions on Beyond 5G networks, we provision a novel qualitative gap analysis. We also define the persistent

challenges that exist with regards to MM mechanisms for 5G and beyond networks. Based on these challenges, we define the potential solutions and a novel framework for the 5G and beyond MM mechanisms. This novel framework specifies a complete stack of MM mechanisms at the access network, core network and at the extreme edge network (users/devices) level, that will help satisfy the requirements for the 5G and beyond MM mechanisms.

Following this, and as part of the defined novel MM framework, we present a novel on-demand MM service strategy. This on-demand feature provisions the necessary reliability, scalability and flexibility to the MM mechanisms. These three characteristics, as we elaborate in more detail in the thesis, will be the pillars of future MM mechanisms. It is important to state here that such an on-demand framework will ensure that appropriate resources and mobility contexts are defined for users who will have heterogeneous mobility profiles, i.e., pedestrian, vehicular, high speed traffic, etc., along side applications with versatile QoS requirements in a network with multiple Radio Access Technologies, such as LTE, Wi-Fi, 5G, etc.

Next, based on the novel MM framework for 5G and beyond mechanisms that we have defined in this thesis, we tackle the problem of core network signaling that occurs during MM in 5G/4G networks. A novel handover signaling mechanism has been developed, which eliminates unnecessary handshakes during the handover preparation phase as well as preserves the legacy data structures. This not only allows for ease of transition to future softwarized network architectures but also simultaneously leads to significant reduction in latency, processing cost and transmission cost of handover signaling. Note that, to perform our analysis we utilized data from Greek and Japanese network operators as well as from telecom vendors such as Cisco. A further enhancement of the aforementioned proposed handover signaling mechanism has also been provided, wherein a premonition of a handover failure is utilized to design the handover preparation phase signaling. This consequently results in additional performance gains as observed through our quantitative evaluation. We then perform a comparative analysis of the proposed strategy and the 3GPP handover signaling strategy on a network wide deployment scenario, wherein the performance gains through our proposed strategy are further highlighted.

Lastly, a novel user association and resource allocation methodology, namely AURA-5G, has been proposed. The developed methodology addresses scenarios wherein applications with heterogeneous requirements, i.e., enhanced Mobile Broadband (eMBB) and massive Machine Type Communications (mMTC), are present simultaneously. Consequently, a first approach in literature, wherein a joint optimization process for performing the user association and resource allocation while being cognizant of heterogeneous application requirements, has been performed. Concretely, the methodology aims at not only assigning an AP to a

user, but also aims at reserving the appropriate resources, i.e., bandwidth at the chosen APs, given the heterogeneous application requirements and other prevailing network constraints. As mentioned, we capture the peculiarities of this important mobility management process through the various constraints, such as backhaul requirements, dual connectivity options, available access resources, minimum rate requirements, etc., that we have imposed on a Mixed Integer Linear Program (MILP). The objective function of this established MILP problem is to maximize the total network throughput of the eMBB users, while satisfying the minimum requirements of the mMTC and eMBB users defined in a given scenario. Through numerical evaluations we show that our approach outperforms the baseline user association scenario in terms of achievable system throughput for all possible constraint combinations. The baseline scenario being, to attach the users to an AP with the best Signal to noise ratio towards them and then dividing the access resources equitably amongst all competing users at a given AP. Moreover, to ensure the applicability of the devised methodology, we have presented a system fairness analysis, as well as a novel fidelity and complexity analysis for the same. Notably, for the fidelity analysis, we analyze how well the system satisfies the latency and backhaul utilization constraints. Further, for the complexity analysis, we observe the time to converge to an optimal solution as well as the number of Monte Carlo trials in which our framework is able to determine such an optimal solution, i.e., the solvability of the MILP problem. An extension of this work has then been briefly summarized in the future works section, wherein we comment about the possibility of integrating the Ultra-reliable low latency communication (URLLC) services into the AURA-5G framework via a multi-objective optimization methodology.

Given the aforementioned efforts, we believe that this thesis has significantly advanced the area of Mobility Management for 5G and beyond networks by provisioning methods and system concepts that address many of the important persistent challenges. This has been reinforced by the broad acceptance of our work into multiple globally recognized conferences, reputed journals as well as a patent application.

List of Publications

[J3] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "User Association and Resource Allocation in 5G (AURA-5G): A Joint Optimization Framework", Submitted to Elsevier Computer Networks, pp. 1–35, 2020. (Area: Computer Science; Quartile: Q1 (13/53); IF: 3.03 (2018))

[PT1] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Handover Method and System for 5G Networks", WO 2019/229219 A1 (WIPO PCT), pp. 1-98, Dec. 2019. (Positive International Search Report)

[J2] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Are Mobility Management Solutions Ready for 5G and Beyond?", Accepted in Elsevier Computer Communications, pp. 1–36, 2020. (Area: Telecommunications; Quartile: Q2 (37/88); IF: 2.816 (2018))

[J1] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Evolutionary 4G/5G Network Architecture Assisted Efficient Handover Signaling", IEEE Access, vol. 7, pp. 256–283, Dec. 2018. (Area: Telecommunications; Quartile: Q1 (19/88); IF: 4.098 (2018))

[C4] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Improved Handover Signaling for 5G Networks", IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2018, pp. 164–170, Sept. 2018.

[C3] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Enhanced Handover Signaling through Integrated MME-SDN Controller Solution", IEEE 87th Vehicular Technology Conference VTC Spring 2018, pp. 1–7, Jun. 2018.

[C2] R. I. Rony, **A. Jain**, E. Lopez-Aguilera, E. Garcia-Villegas, and I. Demirkol, "Joint access-backhaul perspective on mobility management in 5G networks", IEEE Conference on Standards for Communications and Networking, CSCN 2017, pp. 115–120, Sept. 2017.

[C1] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Mobility Management as a Service for 5G Networks", IEEE ISWCS 2017, pp. 1–6, Jun. 2017.

Index

List of Figures

List of Tables

List of Abbreviations

2G	Second generation
3G	Third generation
3GPP	Third Generation Partnership Project
4G	Fourth generation
5G	Fifth generation
5G NORMA	5G Novel Radio Multiservice adaptive network Architecture
5G NR	5G New Radio
5GPPP	The Fifth Generation infrastructure Public Private Partnership
ANDSF	Access Network Discovery and Selection Function
AMF	Access and Mobility management Function
AR	Access Router
AuR	Augmented Reality
ASN.1	Abstract Syntax Notation number one
BBU	Baseband Unit
BCE	Binding Cache Entry
BH	Backhaul
BS	Base Station
CA	Carrier Aggregation
CAPEX	Capital Expenditure
CDN	Content Delivery Network
CN	Core Network
CoA	Care-of-Address
CoMP	Coordinated Multi-Point transmission
CP	Control Plane
CrN	Correspondent Node
CRAN	Centralized Radio Access Network
CSI	Channel State Information
D2D	Device-to-Device

DL	Downlink
DP	Data Plane
E2E	End-to-end
EDGE	Enhanced Data Rate for GSM Evolution
eNB	Evolved Node-B
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
FA	Forwarding Agent
F-AP	Fully equipped Access Points
FBAck	Fast Binding Acknowledgement
FBU	Fast Binding Update
FDD	Frequency Division Duplex
FH	Fronthaul
FHO	Frequent Handover
FMIPv6	Fast MIPv6
FNA	Fast Neighbor Acknowledgement
FP7	7th Framework Programme for Research and Technological Development
GA	Genetic Algorithm
gNB	next generation NodeB
GPRS	General Packet Radio Services
GSM	Global System for Mobile Communications
HA	Home Address
HARQ	Hybrid Automatic Repeat Request
HetNet	Heterogeneous Networks
HMIPv6	Hierarchical MIPv6
HO	Handover
HSPA	High Speed Packet Access
IEEE	Institute of Electrical and Electronics Engineers
IE	Information Element
ITU-R	International Telecommunication Union-Radio Communication Section
KPI	Key Performance Indicator
LCoA	Local CoA
LIPA	Local IP Access
LMA	Local Mobility Anchor
LOS	Line of Sight

LTE	Long Term Evolution
LTE-A	Long Term Evolution Advanced
LWA	LTE-WLAN Aggregation
MAC	Medium Access Control
MADM	Multi-Attribute Decision Making
MAG	Mobility Access Gateway
MAP	Mobility Anchor Point
MBS	Macro-Base Station
MC	Macro-cell
MIH	Media Independent Handover
MIMO	Multiple Input and Multiple Output
MIPv4	Mobile IPv4
MIPv6	Mobile IPv6
MME	Mobility Management Entity
MMT	Multi-Mode Terminal
mmWave	Millimetre Wave
MN	Mobile Node
MPTCP	Multipath Transmission Control Protocol
MTC	Machine Type Communication
MU-MIMO	Multi-User MIMO
NBI	Northbound Interface
NC	Network Controller
NFV	Network Function Virtualization
NFVO	Network Function Virtualization Orchestrator
NGC	Next Generation Core
NGFI	Next Generation Fronthaul Interface
NGPON	Next Generation Passive Optical Network
NG-RAN	Next Generation Radio Access Network
NLoS	Non-Line of Sight
OF	OpenFlow
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operating Expenditure
PCRF	Policy and Charging Rules Function
PDCP	Packet Data Convergence Protocol
P-GW	Packet Gateway

PHY	Physical
PMIPv6	Proxy MIPv6
PoC	Proof of Concepts
QoE	Quality of Experience
QoS	Quality of Service
RACH	Random Access Channel
RAN	Radio Access Network
RANaaS	Radio Access Network as a Service
RAT	Radio Access Technology
RCoA	Regional CoA
RNC	Radio Network Controller
RRC	Radio Resource Control
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Remote Radio Unit
RSRP	Reference Signal Received Power
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
SAE-GW	System Architecture Evolution-Gateway
SBI	Southbound Interface
SBS	Small-cell Base Stations
SC	Small-cell
SCF	Small Cell Forum
SCTP	Stream Control Transmission Protocol
SDN	Software Defined Networking
SDN-C	SDN-Controller
SeMMu	SDN enabled Mobility Management unit
SeNB	Source eNB
SGSN	Serving Gateway Support Node
S-GW	Serving Gateway
SINR	Signal to Interference and Noise Ratio
SIPTO	Selected IP Traffic Offload
SLA	Service Level Agreement
SMF	Session Management Function
SMS	Short Message Service
TA	Tracking Area

TAU	Tracking Area Update
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TEID	Tunnel Endpoint Identifiers
TeNB	Target eNB
T-SeMMu	Target SeMMu
UDN	Ultra-Dense Network
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications Systems
UPF	User Plane Function
VLAN	Virtual LAN
VNF	Virtualized Network Function
VR	Virtual Reality
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access

Chapter 1

Introduction

General Packet Radio Service (GPRS), introduced as part of the 3GPP Release 97 (R'97), kick-started the age of mobile data networks. Ubiquitous connectivity and availability of mobile data have now become quintessential aspects of human life, which pioneers in the area of telecommunications had foreseen at the time of R'97. This sparked a cycle of innovation so as to cater to the ever increasing demands for data and services coupled with the exponential growth in mobile broadband subscriptions. As a consequence, technologies such as Enhanced Data Rate for GSM Evolution (EDGE), 3rd generation networks (3G), High Speed Packet Access (HSPA), and the current 4G-Long term Evolution (LTE) were introduced, and have since become mainstays of the telecommunication infrastructure.

However, the aforesaid technological evolution has not lead to a similar growth in the revenues for the operators. Through Figure 1.1 we observe that the growth of Average Revenue per User (ARPU) has been declining as compared to the corresponding enhancements in the data rates [1, 2]. Contributing to this has been, low cost of data plans, high land acquisition and spectrum licensing cost, exorbitant hardware prices and inflexibility to switch hardware at will due to incompatibility issues. Moreover, and as mentioned above, multiple reports such as [3–6] project that the amount of data consumed by mobile broadband users will increase exponentially by 10-15 folds from 3.2 Exabytes (EB) in 2014 to 52 EB by 2021. In addition, the number of mobile broadband subscriptions are also expected to follow a similar trend and reach 9 billion by 2021.

These pressing issues have prompted the industry and consequently the academic community to explore new verticals apart from mobile broadband, such as emergency services, Internet of Things (IoT), etc. Further, the exploration of new avenues for enhancing data rates while reducing the incurred latency have also become critical work items in the ongoing 5G standardization process, given the diverse Quality of Service (QoS) requirements that

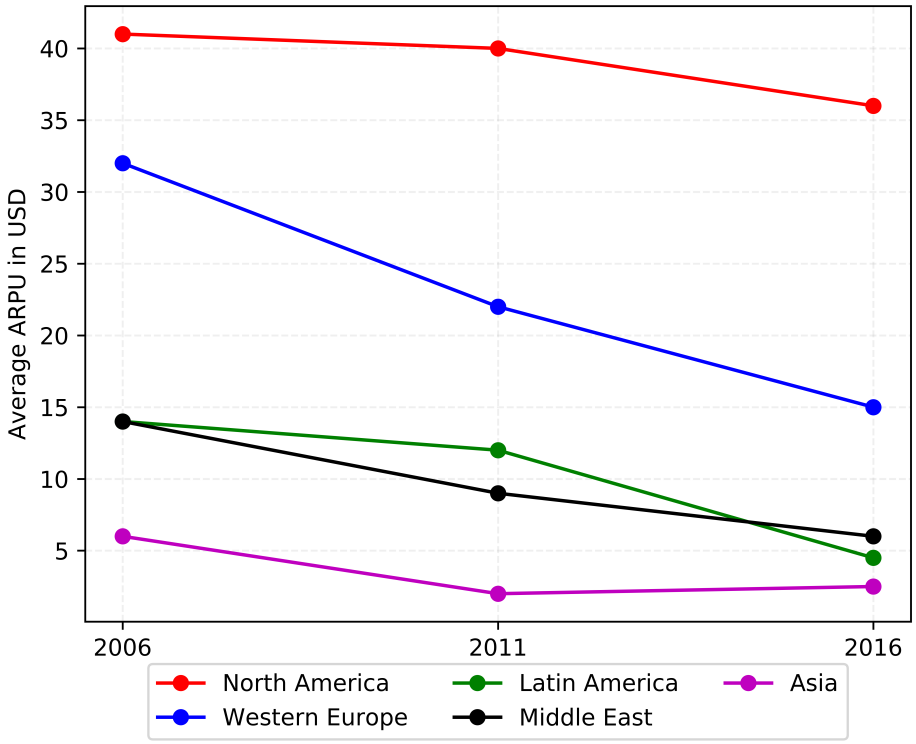


Figure 1.1: Average ARPU across the major geographic Regions

5G will aim to satisfy. Concretely, the 5G standard promises to enhance the user experience through 100x increase in data rates over those possible in LTE [7, 8]. Additionally, it also aims to provide less than 1 ms latency alongside ensuring the requested reliability for delay critical services, such as emergency services, Augmented Reality (AuR)/ Virtual Reality (VR), etc, [7, 8].

Moreover, the ITU has clearly outlined the requirements that the 5G networks will have to satisfy [9]. Based on the study carried out in [9, 10], these have been enlisted in Table 1.1. However, these requirements, that act as guiding principles for the development of 5G, represent significant challenges as well. One of those significant challenges, and as has been documented with every other wireless standard released, is to ensure seamless mobility whilst ensuring the demanded QoS and Quality of Experience (QoE). Concretely, managing the mobility of users whilst still provisioning data rates in excess of 1 Gbps, with heterogeneous application and mobility profiles, and extreme reliability will be one of the most challenging aspects for 5G networks. To better understand this challenge, we briefly discuss the principle behind mobility management and the challenges it faces in 5G in the text that follows.

Table 1.1: Expectations from 5G Networks

Parameter	Support
Data Rates	×10-100 times more than the current 4G-LTE framework
Mobility	Support for high data rate services even at 500 km/h
Heterogeneous Networks	Support for mobility amongst heterogeneous Radio Access Technologies, such as 4G, 5G, Wi-Fi, etc., as well as ability to perform multi-connectivity
CAPEX/OPEX of the networks	Sustainable
New deployment capabilities	Easy
Wireless device density	Support for massive deployment of wireless devices: ×10-100
End-to-End latency	<1ms
QoE	Consistent, but according to the user profile, i.e., user mobility, application in use, etc.
Energy efficiency	High

1.1 Mobility Management

1.1.1 Concept and Criticality

The ability of the modern day wireless networks to allow seamless mobility and maintain continuity of service even at great distances from the initial point of attachment to the network is what makes them extremely popular. This aforesaid critical functionality is broadly defined as mobility management, which also determines the ubiquity of any wireless technology.

Consequently, over the years mobility management has been included in every wireless standard as an enabler and as a non-negotiable component. For example, the GSM (2G) standard adopted a hard handover (HO) approach with an Signal to Noise Ratio (SNR)/Received Signal Strength Indicator (RSSI) based base station (BS) selection mechanism [11], whilst the CDMA (2G) and WCDMA (3G) standard adopted a soft HO approach [11] by also determining the suitability of an BS using RSSI/SNR. Next, 4G technologies such as LTE have specific methods to handle mobility scenarios, e.g., the X2 and S1 handovers. Essentially, the X2 method allows for the possibility of a fast HO if possible. This is so because, the X2 HO method does not necessitate any core network (CN) signaling, as compared to the S1 handover. Additionally, the BS suitability in 4G-LTE is determined based on the RSSI reported by the user [12].

In addition to the cellular technologies, IEEE 802.11 suite of technologies also allow for intra-domain mobility, wherein a change in BS is permitted through the utilization of CAP-

WAP [13]. Concretely, CAPWAP permits an access controller to manage multiple BSs, thus allowing for a global view and hence, the capability of managing intra-domain handovers. Moreover, inter-domain HO was up until now a significant challenge. However, ongoing research efforts in the area of Licensed Assited Access (LAA), which aims to allow LTE networks to contend for resources on the unlicensed bands alongside the Wi-Fi BSs [14], and LTE Wi-Fi Aggregation (LWA), which aims to integrate Wi-Fi services to the 3GPP core network at the PDCP layer [15], have been gradually attempting to alleviate this shortcoming. Certain other technologies such as *Bluetooth*, which are close range communication technologies, do not incorporate mobility management suites within them. Consequently, this is one of the major reasons with regards to their adoption as mainstream wireless standard in many mobility applications, being deterred.

Given this extremely vital nature of mobility management in any wireless technology, 5G networks are expected to incorporate MM approaches that will enable it to provision users/devices with uninterrupted connectivity, low latency, faster data rates and support at higher ground speeds, while satisfying Table 1.1.

1.1.2 The 5G Aspect

5G networks are being designed and developed such that they can support upto 10 Gbps data rates, provide ultra-low latency services, i.e., approximately 1 ms, as well as service nearly 10^6 users per km^2 [9]. To achieve these goals, multiple enablers for 5G, such as Software Defined Networking (SDN), Network Function Virtualization (NFV), Distributed Mobility Management (DMM), Mobile Edge Computing (MEC), Device-to-Device Communications (D2D), Ultra Dense Networks (UDN), Multi-Radio Access Technology (M-RAT), i.e., ability to utilize multiple RATs in conjunction simultaneously, Joint Access and Backhaul design methods (also known as Integrated access backhauling in 5G), millimeter Wave (mmWave) and Cloud Radio Access Network (CRAN) have been proposed and discussed at length in the research community [16].

Concretely, SDN and NFV aim to provision a softwarized network framework, which will be able to grant the network operators with enhanced flexibility in-terms of deployment of hardware as well as services [7, 16]. It will also provision a global view to the network which can then be utilized for optimizing the functionality of existing network functions, such as MM [17]. Next, the DMM paradigm aims at de-centralizing the MM operation and hence, making it more flexible and resilient [18]. This will consequently assist the 5G network operators in provisioning seamless mobility for the users in a highly dynamic network environment. Another enabler, which is aimed at improvising the QoS for the users via

reduction in service times as well as provisioning of compute facilities near the network edge, is the MEC paradigm [19]. Furthermore, the D2D communications will enable information sharing as well as extended connectivity near the network edge [7]. This will facilitate 5G networks in provisioning better scenario specific, i.e., context-based, services.

Additionally, the UDN, M-RAT, mmWave and Joint access and backhaul strategies, aim at provisioning extreme flexibility on the radio side in terms of resource availability as well as resource sharing. Specifically, UDN and mmWave techniques aim at increasing the spectral efficiency of the network by bringing the BSs closer to the users and opening up the higher frequency bands, respectively [7]. Moreover, via the M-RAT technique, it will be possible for any given user to be able to connect to multiple BSs belonging to different RATs. Through Release-15, 3GPP has already standardized the concept and functional characteristics of dual connectivity [20]. And whilst the aforementioned strategies primarily increase the resource availability, the Joint access and backhaul design mechanism aims to address the issue of on-demand and context based resource sharing. Multiple works, such as [C2], have already envisioned how the joint design mechanism can enhance the performance of 5G networks. Lastly, C-RAN aims at provisioning a flexible RAN deployment procedure, and hence, a flexible RAN split. This essentially will assist operators in deploying lower cost RRHs and centralizing the processing of the data, which will eventually lead to significant processing gains [21].

However, and also according to our contribution [J2], these aforesaid enablers do not instill the required reliability, *flexibility* and scalability necessary to ensure the seamless mobility aspect of 5G networks. This is so because, while the SDN and NFV paradigms give a global view of the network, the signaling required to gather such information for mobile users can quickly drown the entire network with control messages [J2]. The DMM paradigm on the other hand, whilst handling the mobility without a central controller and solving the core network signaling and latency issue [22–24], can be quite detrimental at the access network level. The reason being that it requires control signaling amongst the routers to exchange the context of the migrating user. Any disruption in the link or an abruptly large number of migrations can present significant challenges to the DMM strategy. Additionally, techniques such as M-RAT, UDN, C-RAN and mmWave, complicate the development of an effective MM strategy because they increase the dimensionality of the search space as well as alter the behavior of the physical channel, as compared to the sub-6 GHz based 2G/3G/4G standards.

Lastly, the MEC paradigm through its close proximity to the access network can help alleviate issues regarding latency as well as core network signaling. However, when users are mobile their services will also need to be replicated/migrated. So far research efforts such

as [25–27] have not been able to provision methods which meet the latency and efficient compute capacity utilization requirements in the event that service replication/migration is required.

1.2 Motivation

From our discussions so far, it is evident that in 5G networks the design and development of MM solutions will be faced with multitude of challenges. It is these existent challenges that have contributed significantly to the motivation for the work that has been presented in this thesis. Hence, we firstly consolidate these challenges as follows:

- The ultra high density characteristic of the 5G networks, wherein there will be an exponential increase in the number of devices it serves, will be an important challenge [7, 9]. Moreover, a similar increase in the number of BSs, which will cater to these devices, is also expected. Thus, to manage the mobility contexts as well as the signaling involved in such a dense scenario will present a significant challenge.

- Extreme heterogeneity in the network, wherein the users have different services with different QoS requirements along side the heterogeneous RATs provided by network operators, will pose significant challenges towards the design and development of future MM solutions. The reason being that, currently a *one size fits all* approach is being utilized. However, given the aforesaid heterogeneity, it will be important that the future MM strategies consider each service type's requirement individually. Note that, here by different services we mean the Ultra-reliable low latency (URLLC), enhanced Mobile Broadband (eMBB) and massive Machine Type Communications (mMTC) services, as defined in the 5G standards [7]. Concretely, the URLLC services will require extreme reliability, not only in terms of the bit error rate performance but also in terms of link outage probability, as well as low latency, approximately 1ms or less. Furthermore, the eMBB services will necessitate extremely high data rates, i.e., upto 10 Gbps, to support applications such as VR/AuR, etc. Additionally, the mMTC services will require that the network supports extremely large deployment of devices, i.e., of the order of 10^6 per km^2 [7, 9].

- Given the heterogeneity in service types, RAT types, as well as the broader range of support, in terms of speeds (upto 500 km/hr), that the 5G network aims to support, the users will consequently have a broader variety of mobility profiles. Furthermore, the increased density of network and the choice of RATs will translate to an extremely high

dimensional solution space to determine appropriate user-BS associations and resource allocation schemes. Thus, to be able to determine these appropriate resources and associations, future MM schemes will have to traverse through the aforesaid extremely high dimensional solution space to find an optimal solution. This will consequently perpetuate the challenging nature of 5G scenarios for MM.

- The current methods as well as legacy mechanisms do not provision a unified and complete MM framework. Concretely, multiple studies highlighting individual MM mechanisms, such as DMM [22–24], LTE handover [12], Dual Connectivity [20], etc., exist. However, as stated above, a comprehensive suite of MM strategies at the access, core and extreme edge (users/devices) level, that will help satisfy the 5G MM requirements, is not available yet.

Given these broad challenges, the motivation for the work done in this thesis is to build on the decades of experience gained by the research community in developing effective MM strategies, and advance it such that the newly developed MM strategies can cater to the requirements of the 5G networks and even beyond. Such an approach should also compulsorily take into account the 5G enablers mentioned in Section 1.1.2.

Further, the current 5G standards by 3GPP [28,29] provision mobility management methods through the handover signaling sequences as specified in [29] as well as procedures for communicating with non-3GPP techniques through LAA and LWA. Additionally, the academic community has presented multiple solutions for 5G MM. However, specific *applicable and implementable* MM procedures that will be able to handle the complexity, which the 5G framework introduces, continue to elude. Consequently, another motivating factor for the work done in this thesis has been to also develop solutions that are tangible, adaptable and employable by both industry and academia.

And so, in the following section we highlight the *objectives* of this thesis, followed by the *contributions* of the work embodied in this thesis to achieve the aforementioned objectives.

1.3 Objectives and Contributions

While the main objective of this thesis is to investigate and develop **Enhanced Mobility Management mechanisms for 5G Networks**, we broke it down further into several relatively smaller yet significantly important and critical objectives. We discuss them as follows:

- *Suitability of existing MM algorithms:* While we seek to innovate and develop new methods to handle the complex network scenario that 5G will present, it is always

prudent to understand and analyze if already existing mechanisms can assist in meeting these requirements in part or in full. Henceforth, we developed a novel qualitative analysis whereby we investigated some of the most prevalent MM approaches from the legacy mechanisms as well the MM mechanisms being developed for 5G networks by current academic and industrial efforts. Additionally, we have also taken cognizance of the emerging discussions related to Beyond 5G (B5G) networks and their corresponding enablers. As a consequence, we have extended this particular study to B5G scenarios as well. And so, through the aforementioned qualitative analysis we have concretized the persistent challenges that continue to exist, potential solutions to these challenges and a framework for future MM mechanisms, thus also paving the way for our subsequent research efforts. An embodiment of this work, i.e., reference [J2], titled "Are mobility management mechanisms ready for 5G and Beyond?" has been submitted to *Elsevier Computer Communications* Journal and is currently under review.

- *On-demand Mobility Management:* An important aspect for the MM mechanisms to cater the extremely dense and heterogeneous 5G networks will be to provision a flexible and on-demand strategy. This is so because, the current day networks provision a *one size fits all* approach. Such a strategy will be counter-productive given the 5G network characteristics. Hence through our work, titled "Mobility Management as a Service for 5G Networks" [C1], which was presented at the *2017 IEEE ISWCS conference*, we have provisioned an on-demand MM framework which also explores the multiple avenues of introducing flexibility in provisioning MM service.

- *Handover mechanisms for 5G networks:* One of the most basic elements of managing mobility of users is to be able to allow them to seamlessly transition from one BS to another. This seamless transition can be either in the same tracking area/Central entity domain/PLMN or they might be transitioning to an BS of another tracking area/Central entity domain/PLMN. By central entity here we mean SGSN/GGSN for 2G/3G networks, MME in 4G-LTE and SMF in the 5G network. In any of the aforementioned scenarios, continuity of service with the required QoS needs to be maintained. This burdens the current handover strategies because 5G networks will be extremely dense and heterogeneous. Hence, we developed a novel handover method and system that enhances the current handover mechanisms by upto 50% in terms of latency, processing cost and transmission cost for the various handover scenarios defined by 3GPP [29]. Note, the evaluations were performed by utilizing real network data from Greek and Japanese operators as well as from vendors such as Cisco. Further, an SDN based method and system has been proposed which not only enhances 5G

handover signaling, but it also enhances the inter-RAT handover between 5G/4G and 4G/3G-2G networks.

Consequently, some of the initial work that focused on the 4G Intra-RAT and 4G/3G-2G Inter-RAT handover method was published as "Enhanced handover signaling through integrated MME-SDN controller solution" in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)* [C3]. Further, a subset of the 5G/4G Inter-RAT Handover signaling method wherein the N26 interface does not exist was then presented as "Improved handover signaling for 5G networks" in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) Conference* [C4]. This work was then followed by a publication titled "Evolutionary 4G/5G network architecture assisted efficient handover signaling" in *IEEE Access Journal* [J1]. The work focused on other 5G Inter- and Intra-RAT handovers as well as a novel handover preparation aware handover rejection methodology. Further, a novel network wide analysis, utilizing data from a Greek and Japanese network operator, was also provided. Lastly, based on the work done with regards to the HO signaling along side a more detailed development of the SDN based system, a patent application titled "Handover Systems and Method for 5G Networks" was filed with the *European Patent Office* [PT1]. Currently, we have obtained a positive international search report (ISR) for our *PCT application*.

- *Application aware User Association Methods:* In 5G, the heterogeneity will arise not only from the different type of BSs but also from the different application types that will need to be served. Thus, in order to ensure that they receive the requested QoS, it will be equally critical to determine the best application to BS association. In our work we consider only single application per user, and hence, it can be treated as a traditional user association paradigm. However, note that the work presented in this thesis can be easily extended to multiple applications, with different QoS requirements, per user.

  Henceforth, we developed a novel Mixed Integer Linear Programming (MILP) based optimization framework, known as AURA-5G, that evaluates the topology and finds the most optimal application (user) association given the multiple real network constraints. Further, we have evaluated scenarios wherein only eMBB services exist, and where both eMBB and mMTC services exist together. Through this work we have also provisioned a working tool for the community so that it can be utilized for research as well as implementation. Given, implementation being one of the intended aspects, we have additionally performed extensive fidelity, performance and complexity analysis. As

a consequence, an embodiment of this work, titled "User Association and Resource Allocation in 5G (AURA-5G): A Joint Optimization Framework", is currently under review with the *Elsevier Computer Networks* journal [J3].

1.4 Thesis Outline

With the motivation and contributions of our work now highlighted, we specify the organization of this thesis in the text that follows.

In Chapter 2, we discuss the state of the art of mobility management strategies as well as the various avenues where it can be employed. We also take cognizance of the emerging studies related to B5G networks and its enablers from the perspective of mobility management. Next in Chapter 3, a novel qualitative gap analysis, wherein we have evaluated the legacy as well as the currently proposed MM mechanisms, has been provided. Following this, we have presented a novel discussion on the persistent challenges, potential solutions to these challenges and a framework for 5G and beyond MM mechanisms. This consequently lays down the foundation for our subsequent work and also chapters.

Thus, in Chapter 4, we present a novel on-demand MM paradigm. We detail its concept and methodology as well as the various benefits it presents for 5G MM mechanisms. Further in Chapter 5, we present an extensive discussion on handover signaling and the current 5G standards for the same. We then highlight the shortcomings and present our approach. Following this discussion, we present our analytical approach and the resultant comparative analysis for the myriad scenarios that 3GPP specifies. Additionally, a novel network wide analysis has also been presented to concretize the benefits that our approach provisions for 5G networks over the current standards. Then, in Chapter 6, we explore another dimension of MM methods wherein a new novel user association strategy has been proposed. We also propose a novel framework, namely AURA-5G, which can be utilized/implemented by industry and academia. Henceforth, we also present an analysis for the framework highlighting its fidelity and complexity.

Lastly, we provide conclusions for the work done in this thesis in Chapter 7 and then discuss our future work proposals in Chapter 8. The thesis is concluded with an Appendix that consists of additional figures that have not been illustrated in the main text.

Chapter 2

State of the Art in Mobility Management

Overview

In this chapter we present a detailed background with regards to the main mobility management mechanisms conceived and developed for the wireless networks. We firstly highlight the various functional requirements from future MM procedures, which is then followed by a detailed discussion on the various mechanisms/strategies in mobility management. These mechanisms/strategies are categorized as legacy and current state-of-the-art mechanisms/strategies. Note that, these discussions are carried out while being cognizant of the ongoing discussions with regards to B5G networks and their enablers. Additionally, we provision a novel 5G Service based architecture diagram along side a unique classification of the current state-of-the-art mechanisms. Note that, in Chapter 3, we build upon this state of the art and present a novel qualitative analysis with regards to the suitability of the myriad MM mechanisms, discussed in this chapter, for 5G and beyond MM mechanisms.

Contributions

[J2] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Are Mobility Management Solutions Ready for 5G and Beyond?", Accepted in Elsevier Computer Communications, pp. 1–36, 2020. (Quartile: Q2; IF: 2.816 (2019))

[C2] R. I. Rony, **A. Jain**, E. Lopez-Aguilera, E. Garcia-Villegas, and I. Demirkol, "Joint access-backhaul perspective on mobility management in 5G networks", IEEE Conference on Standards for Communications and Networking, CSCN 2017, pp. 115–120

Future wireless networks define a very challenging environment for mobility management (MM) solutions, due to the significant increase in density (in terms of both users and de-

ployed base stations), in heterogeneity (given the various radio access technologies (RATs) supported), as well as in programmability (the network as well as the environment can be programmable). To achieve an ubiquitous network service in such challenging environments, it is critical to devise effective MM strategies that facilitate seamless mobility by allowing users to traverse through the network without losing connectivity and service continuity.

One of the traditional approaches for allowing applications to serve a user in mobile scenarios has been to maintain network connectivity through handovers based on criteria such as Radio Signal Strength Indicator (RSSI), Signal to Interference and Noise Ratio (SINR), Reference Signal Received Quality (RSRQ), Reference Signal Received Power (RSRP), etc. However, in addition to the signal quality parameter centric handovers, modern day applications necessitate that other parameters such as available core network bandwidth, End-to-End (E2E) latency, backhaul bandwidth and backhaul reliability [30] are also taken into consideration. Moreover, maintaining Quality of Service (QoS), e.g., provisioning service continuity, link continuity, required bit-rate and latency, during mobility scenarios has been one of the primary objectives for novel MM mechanisms. Multiple strategies to satisfy such QoS criteria such as service migration [31], service replication [26], path reconfiguration [24], etc., have been proposed by the research community. MM solutions for 5G and beyond networks are also expected to ensure E2E connectivity and session continuity through the maintenance/preservation of IP address of the user towards the core network entity that provisions the service for the corresponding user.

To motivate further, we consider an illustrative example of the future mobility scenario is presented in Figure 2.1, which shows the extraordinary nature of complexity that the future networks will present for MM. As shown in the Figure 2.1(a), a mobile user equipment (UE) is connected to multiple RATs (5G BS/ Long Term Evolution (LTE) evolved NodeB (eNB)/visible light communications (VLC) and Light Fidelity (LiFi) Small-cells [32–34], etc.), while having a delay tolerant and a delay sensitive application datastream (flows) with distinct QoS profiles. Also, the BS through which the delay tolerant flow is being served to the user has a good wireless link with a meta-surface in the vicinity. Note that, meta-surfaces are thin, but electrically significant, surfaces that enable the possibility of engineering the channel through the manipulation of phase, amplitude and polarization of the incident wave [35–37]. In addition to the meta-surfaces, future networks will also consist of mobile BSs such as drones, as shown in Figure 2.1(a). Note that, the density of meta-surfaces and drone BSs will also be extremely high in future networks. Further, in the scenario illustrated, we consider the use case wherein the drone BS is servicing a D2D cluster, and connecting it to the core network through one of the ground based BSs. The D2D cluster over the course of its existence does not generate packets as frequently as the

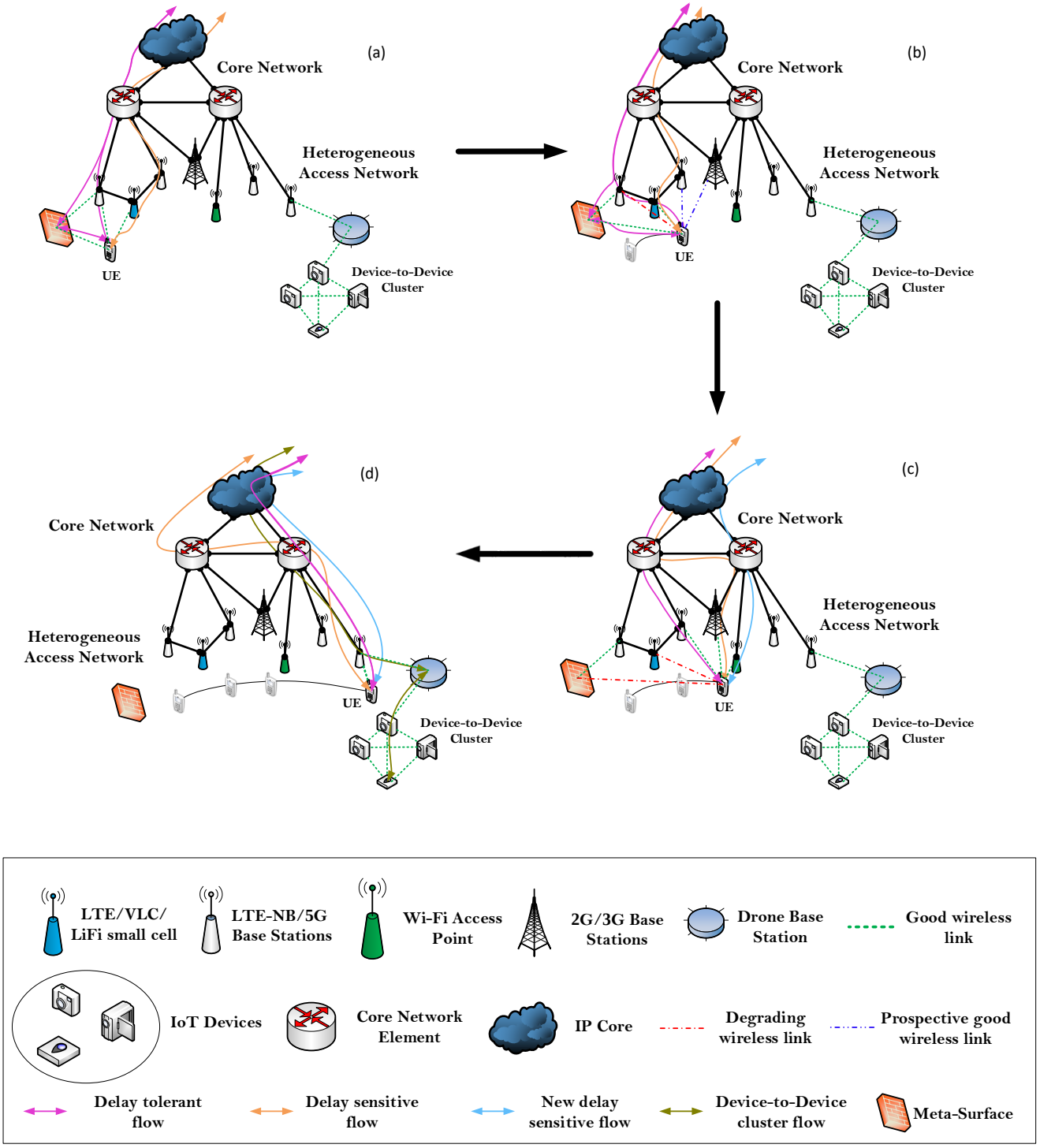


Figure 2.1: An illustrative 5G and beyond network mobility scenario.

other users, since the cluster devices mainly host Internet of Things (IoT) applications.

Next, in Figure 2.1(b), as the user moves, it starts to register wireless links with better signal quality from other BSs as compared to those it is already associated to. It is imperative to state here that, the BSs can be from the same or different network operators. Henceforth, a careful and efficient RAT and BS selection for each flow will be necessary as part of the future MM mechanisms. It is interesting to observe that while the BS used for serving the delay tolerant flow in Figure 2.1(a) no longer has a good link quality, through the meta-surfaces and their programmable nature it still has a good wireless link to the user and hence is able to serve it.

Following the new RAT/BS association, flows pertaining to the user are redirected through the most optimal path. Novel MM mechanisms that aim to service the 5G and B5G networks will require efficient route optimization methods to perform the same. Additionally, the MM mechanisms will also need to implement IP forwarding so as to ensure E2E link continuity. In Figure 2.1(c) we then observe that as the user moves further, the RAT/BS selection and optimal routing methods are continually implemented. Further, when a new application request is generated, as seen in Figure 2.1(c), an appropriate RAT and BS for the given flow is selected alongside the route that satisfies the requested QoS. Lastly, in Figure 2.1(d), it can be seen that alongside the user's flows, the D2D cluster's flows are also being serviced by network. However, the D2D cluster is firstly serviced by a drone BS, which then relays information to/from the ground based BSs. These ground based BSs assist in serving the data flows generated from the devices in the D2D cluster by relaying the data to the relevant servers in the core network.

Given the complexity of the scenario presented in Figure 2.1, it is evident that no single MM mechanism will form the solution to all the possible situations and scenarios that will be prevalent. And, although current MM mechanisms propose methods for careful RAT and BS selection, IP packet forwarding, route optimization, and session management, a more than 10-fold increase in user density coupled with the heterogeneity in flow types and network will extremely limit their capabilities. New user applications such as Augmented Reality, Virtual Reality, Vehicle-to-Everything (V2X), etc., will present very restrictive delay requirements, exceptionally high reliability and bandwidth requirements [38], that will consequently severely challenge the capabilities of current MM strategies. Further, the RAN technologies themselves are expected to undergo important transformation in the future networks given the significant interest in VLC, Li-Fi, etc., [32, 33]. Whilst both Li-Fi and VLC, being TeraHertz (THz) bandwidth technologies, enable near Terabits per second (Tbps) speeds, they are significantly impaired by the environment. This consequently has significantly more detrimental effects on the user QoS during mobility scenarios, which we will discuss in further

detail in the later sections.

Also, owing to the telecom operators' desire to serve more industry verticals, a new set of mobility patterns will emerge. For example, a platoon of vehicles moving coherently together, vehicles disbanding from one platoon to join another, ultra-fast moving users (in excess of 500 km/h), moving base stations (such as those on drones [39]), etc., thus introducing another dimension to the MM problem. Henceforth, the ability to serve devices with mobility patterns that will be more diverse and challenging as compared to current day network scenarios, will be a significant challenge towards the design, development and deployment of 5G and beyond MM mechanisms. An additional yet significant challenge will be to manage and potentially reduce the control plane (CP) signaling load due to mobility events, as expressed in our contribution [J1].

Thus, a fresh perspective, wherein MM solutions are decentralized and flexible, can support multiple use cases simultaneously and account for the various other radical changes in 5G and B5G networks with reliability, is required. Note that, decentralization will permit MM mechanisms to service the exponentially increasing number of users coupled with different mobility profiles (e.g., static IoT devices and users in high-speed trains). On the other hand, flexibility will allow them to adapt to the user, network and/or environment context (e.g., QoS, user mobility profile, network load, flow types, meta-surfaces, etc.). Additionally, reliability will aid in provisioning seamless mobility as well as in satisfying the ultra-reliable criterion for future wireless network applications.

References [40] and [41] aim to provide new MM strategies via Software Defined Networking (SDN) based MM and multi-RAT mobility. However, they do not elaborate on the myriad challenges that future MM mechanisms will encounter, such as time complexity, signaling overhead, etc. Similarly, while in [42] MM strategies, such as advanced cell association, group handovers, etc., have been discussed to address the heterogeneity in the mobility patterns and profiles that will arise in 5G, they fall short in addressing the challenges such as core network signaling, complexity, etc., that 5G and beyond MM solutions will face. Further, surveys such as [43] and [44] are restricted to the current network architecture, and hence, fail to provide a MM perspective for 5G and beyond networks. In addition, while [34] aims to provide insights into the requirements, architecture and key technologies for B5G networks, it does not address the critical issue of MM in B5G networks.

And so in this chapter, through our contributions [J2] and [C2], we firstly present a novel discussion on the functional requirements and design criteria for 5G and Beyond MM mechanisms in Section 2.1. Next, in Section 2.2 and 2.3 we provision a detailed discussion on the existing/legacy and current state-of-the-art MM mechanisms to pave the way to qualitatively analyze their suitability for 5G and beyond MM mechanisms in the next chapter, i.e.,

Chapter 3. We also provide a novel classification of the current state-of-the-art mechanisms based on where they are implemented or create an impact within the network, i.e., CN, access network (AN) and extreme edge network. Additionally we also provision a mapping of these classifications onto the 5G service based architecture (SBA) defined by 3GPP [45], which subsequently in Chapter 3 will assist in explicitly indicating the gaps that still exist.

2.1 Future MM Strategies: Functional Requirements and Design Criteria

Future wireless networks, in addition to being dense, heterogeneous and extensively programmable, will serve multiple industry verticals as well as accommodate multiple tenants on the same network infrastructure [36, 46]. These transformations, some of which are being discussed by the research community [35, 47], represent a paradigm shift from the current network architecture design. As a consequence, MM mechanisms need to be re-evaluated and/or re-designed. For this, we first present the functional requirements of MM mechanisms for future wireless networks in Table 2.1, based on the characteristics we derive from the current and future network scenarios.

From Table 2.1, it can be observed that the MM solutions for 5G and beyond networks will have to adapt and evolve, so as to be able to serve the future wireless networks efficiently. Further, MM solutions will need to be redesigned so that they are flexible, scalable and reliable to ensure the requested QoS and seamless mobility. Apart from these requirements, there are certain criteria that will impact the design and development of future MM solutions. Consequently, in the following text we present an insight into these myriad design criteria and their impact on 5G and beyond MM.

2.1.1 Centralized vs. Hierarchical vs. Distributed Solution

While a centralized solution might offer optimality given its global view, a distributed approach can offer more reliability by eliminating the Single Point of Failure (SPoF) problem as well as avoiding congestion at a specific network node. Instead, a hierarchical approach can incorporate the benefits of both aforesaid techniques. For example, in LTE, MME is the mobility management entity with the Serving Gateway (S-GW) being the mobility anchor, and hence, it is a centralized solution. However, Distributed Mobility Management (DMM) [53] assists in decentralization of the traditional MM mechanisms, wherein instead of having a single MM anchor for all the flows on a UE, the anchors are now distributed. By distribution of MM anchors here we mean that, when a flow is initiated to/from a UE, the

Table 2.1: Functional Requirements from 5G and beyond MM

Req #	Current Scenario	5G and Beyond Scenario	Resulting MM Functional Requirement
R1	Single RAT connectivity	Multi-RAT connectivity	Support for multi-RAT MM as well as efficient RAT selection methods.
R2	Support for mobile broadband applications	Support for eMBB, mMTC and URLLC applications. These applications will have different QoS requirements [48]	Context based support
R3	UE density of $10^5 devices/km^2$ [9]	UE density of $\geq 10^6 devices/km^2$ [9]	Ability to support the increasing user density
R4	Network is vendor driven [16]	Network is softwarized [16]	Ability to utilize the SDN, NFV, MEC, etc.
R5	ground based network with static radio towers	BSs and relay stations may be carried on drones in 5G and beyond networks [49, 50]	Support for mobility of both UEs and BSs
R6	MM protocols are standardized for 2G-4G networks and devices	5G and beyond networks and devices will be gradually rolled out. They are fundamentally different from 4G, 3G and 2G networks	Backwards compatibility
R7	Frequency band: Sub 6 GHz	Sub 6 GHz, millimeter Wave (mmWave) [47], Terahertz communication [32, 33]	Increased robustness, given the significant impact on VLC and mmWave and, thus challenging seamless mobility
R8	Finest granularity of tracking and localization is $< 50m$ [51]	Finest granularity of tracking and localization is a beam ($< 10cm$) [51]	Ability to utilize the advanced level of granularity to provision better mobility and tracking performance in dense urban or high speed scenarios
R9	The complexity is driven mainly via user requirements in a homogeneous network	The complexity is a combination of different user types, different QoS requirements, heterogeneous RAT scenarios, heterogeneous backhaul scenarios [52] and ultra dense nature of the network	Adequate flexibility (they should accommodate for the increased heterogeneity) and tractable solutions (fast and low computational complexity) with well managed power consumption for the increased network complexity
R10	Requested services and data is always hosted in the IP Multimedia Subsystem (IMS) core	Requested services and data in 5G and beyond networks can now be hosted at the network edge, through MECs [16]	MM mechanisms should provide adequate Service Migration [31]/ Service Replication [26] support to ensure the required QoS from the applications
R11	Support for mobility up to 350 km/h	Support for mobility up to and beyond 500 km/h proposed [9]	Avoiding the *one size fits all* approach, i.e., ensuring flexibility to accommodate multiple demanding mobility profiles

anchor may be chosen dependent on the flow requirements. For example, given a new flow originating to/from a UE, a MM anchor is chosen which might be very close to the UE to assist in network offloading purposes, whereas pre-existing flows might still be served from the MM anchors to which they were first assigned, so as to avoid service disruptions. Hence, it would provide more reliability. The hierarchical method on the other hand, will combine the centralized and distributed approaches to offer the reliability of the distributed approach (through decentralization of mobility anchors) and the optimality of the centralized approach (e.g. through master and slave network management entities). An example of such a distributed/hierarchical approach can be found in the upcoming 5G networks, wherein through SDN and NFV there is a separation between the CP, i.e., Access and Mobility Function (AMF)- Session Management Function (SMF) for mobility management, and the data plane (DP), i.e., OpenFlow (OF) switches, etc., [24,54]. A more detailed explanation with regards to the AMF and SMF functioning has been provided in Section 2.3.

2.1.2 Computational Resources

The computational resource locations and their corresponding computational power will determine the degree to which the mobility management mechanisms can be distributed. For example, edge clouds can aid not only in MM related computation (e.g., RAT and BS selection) but can also enable faster access to content through caching. In addition, for 5G and beyond networks, it will also be critical for the MM mechanisms to determine whether services need to be migrated or replicated [25,26,55], so as to maintain service continuity and hence ensure the required QoS. Note that, by service replication we mean that the services being requested by a user undergoing a mobility event are replicated to other edge servers. Further, by service migration [25, 31] we imply that the services being accessed by a user undergoing a mobility event are migrated to the next edge cloud server where the user is expected to move to.

2.1.3 Backhaul Considerations

Network densification and the prohibitively expensive nature of installing optical fibre as backhaul [52] will render the backhaul scenario in 5G and B5G wireless networks to be extremely heterogeneous, i.e., they will be composed of both wired and wireless links. Further, the backhaul wireless links will consist of multiple radio access techniques such as microwave, mmWave, VLC or LiFi, co-existing together [33]. These transformatory trends will need to be taken into consideration while developing future MM mechanisms, as:

- Congestion or multiple-hops in the backhaul can impact the E2E latency, and consequently, the perceived QoS [C2].
- Backhaul reliability will be critical given the relatively poor penetration capability of mmWave [56] and additionally, strong atmospheric absorption features for VLC [32]. Thus, during mobility, attaching to an BS with a poor backhaul link quality can correspondingly lead to degradation in QoS since, there can be increased packet loss or even an outage altogether.

2.1.4 Context

A multitude of parameters, such as user mobility profiles, type of flows, network and user policies, BS signal quality, network load, backhaul-fronthaul options, etc., constitute the context. Additionally, MM mechanisms for 5G and B5G networks will have to service users with different mobility profiles, accessing different services. Hence, the available contextual information will be valuable for any future MM mechanism. For example, in [41], network load aware MM methods present an improvement of 75% in throughput at the cell edge as compared to the context agnostic methods, thus reinforcing the aforesaid criteria.

2.1.5 Granularity of Service

Granularity in MM services (e.g., based on flow, subscriber or mobility profile) will be an important component for MM methods to provision optimal solutions for 5G and B5G networks. Further, the type of granularity offered, i.e., per-flow based, mobility based, etc., will depend on the user context as well as the network conditions. Hence, innovative mechanisms like the Mobility Management-as-a-Service (MMaaS) paradigm, as developed in our contribution [C1], will be required. In MMaaS, on-demand MM solutions can be employed by or assigned to UEs. For example, if a device is moving at a high speed and there is another device, say an IoT device, that is stationary, then a mobility based granularity of service can be adopted. Based on this service granularity provision, the high mobility device can be allocated resources on Macro-cells whilst the stationary device can be served by Small-cells. Another important example being that of network slices. Network slicing, the concept, typically refers to a resource based logical slicing of the existing network infrastructure to support multiple verticals and corresponding operators that serve them [57]. In such scenarios, on-demand MM will be necessitated by the network slices, as they will cater to services with differing mobility requirements and patterns, such as the URLLC and eMBB services.

2.1.6 D2D Service Availability

The availability of D2D services will determine how the mobility management mechanism is executed, as D2D can assist in providing seamless mobility through CP information and/or DP data forwarding. This will be specially relevant in scenarios involving V2X [58], wherein for example, the vehicles, that are outside the coverage area of the infrastructure network (IN) or are experiencing a deep fade with the IN, can exchange data with it by relaying their information through other vehicles, over the PC5 interface [58], that might be nearby and within the coverage area of the IN or are experiencing better channel conditions with it.

2.1.7 Physical Layer Considerations

The introduction of massive MIMO and mmWave technology will certainly impact current MM methods. Concretely, in urban environments the mmWave links will face extensive blockage alongside their limited range due to the propagation characteristics. Hence, this will require densification, which introduces the possibilities of frequent handovers (FHOs). Here by FHOs, we refer to the fact that in a dense network environment, such as those in 5G, the users will be subjected to handover scenarios more frequently as compared to that in the current networks. On the other hand, beamforming through massive MIMO antennas can be utilized to track moving users and hence, provide them with high QoS through higher throughput and better localization services.

Further, for B5G networks, VLC and meta-surfaces have emerged as the main enablers. Note that, VLC will be challenged extremely by the existing environment. This is so because, it operates in the Terahertz range of frequencies, thus making most objects in the environment as blocking agents. Also, meta-surfaces will lead to programmable environments, which will create the issue of dimensionality for an optimal solution.

Henceforth, the physical (PHY) layer techniques require consideration in any MM mechanism development for 5G and beyond networks.

2.1.8 Control Plane Signaling

An important target of future MM mechanisms will be to reduce the CP signaling induced during handovers. Studies, such as our contribution [C3], have proposed enhanced handover signaling mechanisms for an SDN-based core network architecture, such that the transmission and processing cost as well as the overall latency during a handover process is reduced whilst ensuring the Capital Expenditure (CAPEX) does not rise significantly. Such a procedure will enhance the QoS for the user while switching base stations and hence, will be

critical to the future MM suite.

Given the requirements in Table 2.1 and the aforesaid design considerations, a complete overhaul of MM mechanisms for future wireless networks might result in optimal MM solutions. However, the time to develop and market them will be correspondingly longer. Hence, it is prudent to explore and evaluate the myriad legacy as well as current state-of-the-art mechanisms and standardization efforts, and evaluate their suitability as *enablers for MM* in 5G and beyond wireless networks. But, before we perform such an analysis, a detailed background into these mechanisms has been provided in Sections 2.2 and 2.3. Concretely, we have divided the existing mechanisms into two categories, i.e., Legacy mechanisms (2G/3G/4G defined by 3GPP and also non-3GPP solutions such as Wi-Fi) and Current mechanisms (5G as defined by 3GPP and other relevant solutions proposed by academic and standards bodies).

2.2 Mobility Management: Legacy Mechanisms

Mobility Management, as has already been stated, permits a user to stay connected even when it moves beyond the geographic boundaries of the network to where it first attached to. This property, as a result, also determines the ubiquity of a given wireless standard. Further, given the softwarized characteristic of the 5G and beyond networks, and the centralized nature of current and legacy mobility management strategies, it becomes even more critical to explore the various avenues/aspects of future MM strategies. Henceforth, as part of our detailed study, we first reflect back on some of the most significant MM strategies that have served the wireless networks well up until now. Concretely, in the following subsections we present a discussion on the MM strategies for pre-5G networks.

2.2.1 3GPP based MM techniques

In this subsection, the various mobility management techniques developed and deployed by 3GPP, as part of the LTE framework, have been discussed in detail.

2.2.1.1 LTE Handover Mechanisms

3GPP based LTE [59] standard is the most widely accepted and subscribed wireless standard today. With the global subscription reaching 1,100 million [60] for LTE in 2015 and expected to increase by four times by 2021, it is imperative to understand the mobility management mechanisms as prescribed by 3GPP for LTE.

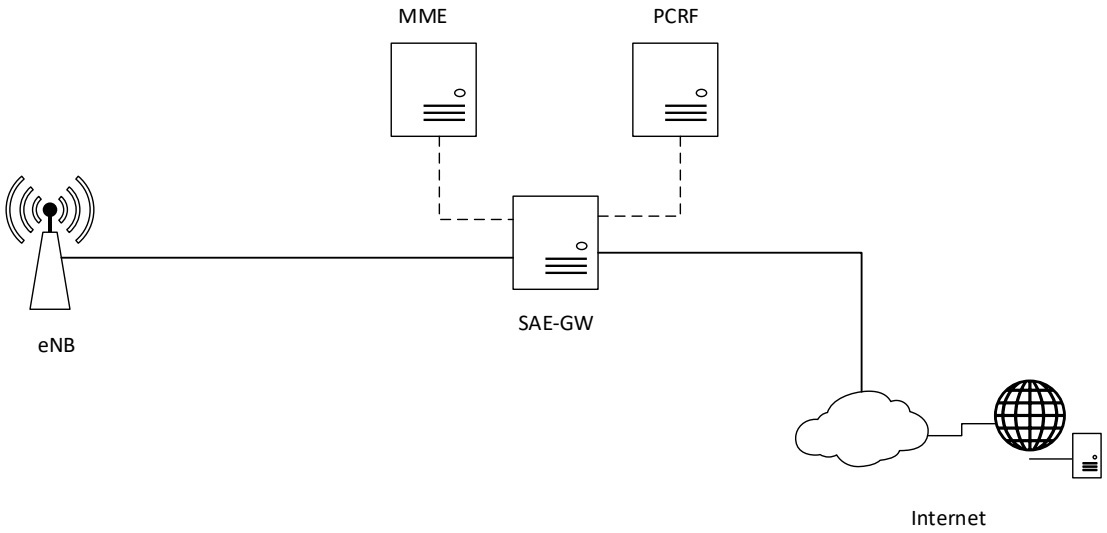


Figure 2.2: Basic LTE Architecture.

In order to understand mobility management in LTE, it is essential to first understand the LTE core structure and the entities that perform mobility management. Figure 2.2 provides a diagrammatic representation of the LTE architecture. From Figure 2.2, the Mobility Management Entity (MME), as the name suggests, is responsible for the handover and policy management functions. It not only performs the negotiations for resources in the core and access networks (in case of S1 handover) but it is also responsible for communicating with the Policy and Charging Rules Function (PCRF) and the System Architecture Evolution-Gateway (SAE-GW). Note that, the SAE-GW is a single entity representing both Serving Gateway (S-GW) and Packet Gateway (P-GW) together as one unit. And so, in addition to the MME, P-GW is another entity that is connected with the mobility management of the UE. It is known that the LTE network is an all-IP network, and hence, every UE that accesses the network receives an IP address. This IP address is assigned by the P-GW and it acts as the anchor for the same until the UE stays in its domain. Note that, layer 3 mobility is an important component in ensuring continuity of the service whilst allowing mobility. This is so because, many application services are not designed to handle a change in IP address without service interruption. And hence, techniques to allow for seamless mobility in such scenarios is of great interest when studying mobility management.

It is important to state here that mobility management in LTE involves: 1) Handover (when the UE is in active state) and 2) Cell re-selection and Tracking Area Update (TAU) (when the UE is in idle state).

Handovers: In LTE, handover may either be within the same Tracking Area (TA) (or

a TA registered in the MME) or it might be to a TA that is not associated with the serving MME, and hence may entail the extra step of tracking area registration. Further, LTE provides two types of handovers, i.e., X2 HO and S1 HO [61].

Whilst within the same tracking area and performing a HO, the UE first sends its measurement updates to the Source eNB (SeNB). The SeNB is the eNB to which the UE is currently attached to. And so, if the measurement report results in the decision to HO, the SeNB then checks for the presence of the X2 interface to the target eNB (TeNB). TeNB is the eNB that is chosen to which the UE has to be handed over. Thus, in case it is absent, S1 HO is initiated. For S1 HO (illustrated in Figure 2.3), the SeNB informs the MME about its decision, which in turn informs the TeNB about the HO request. After negotiating for the resources, the TeNB formulates an indirect route to the SAE-GW and informs the MME about it. An indirect route refers to a route that allows the SeNB to tunnel the packets to the TeNB via the SAE-GW. Hence, the MME informs the SAE-GW about the same, and also asks the SeNB to form an indirect route with the SAE-GW. This allows for the SeNB to tunnel the downlink packets to the TeNB whilst the HO is being executed. Further, the SeNB also informs the TeNB about the sequence number of the downlink and uplink packets. Next, the SeNB signals the UE to perform the HO, and starts tunnelling its DL packets to the TeNB where they are buffered. After the HO is completed, the UE indicates to the TeNB that it has completed the HO. This allows the TeNB to start transmitting the buffered DL packets and accept the UL flow. Further, the TeNB informs the MME about the same, which then informs the SAE-GW and SeNB. The SAE-GW then switches the path to the most optimized one, i.e., it now invalidates the indirect route, and at the same time SeNB releases the UE context. In this way the LTE S1 HO is executed and complete. An illustrative description of the S1 HO process is presented in Figure 2.3.

However, and as mentioned above, in case the SeNB has an X2 interface to the TeNB, the HO negotiation is performed through the X2 interface. The benefit of the X2 interface is that it does not entail any signaling with the elements in the Evolved Packet Core (EPC), and hence it reduces the signaling load in the core network as well as reduces the interruption time. Consequently, the QoS for the users is also improved. The drawback to this method is that to have an X2 interface eNBs need to be connected to each other either through fibre or microwave links, which increases the CAPEX and OPEX for the service provider. When X2 HO is the selected method, SeNB passes the HO request message to the TeNB. After resource negotiation, the TeNB sends back a HO acknowledgement message to the SeNB. Further, the SeNB now informs the UE to start the HO, and tunnels the DL packets to the TeNB. As soon as the HO is confirmed, and TeNB receives this message from the UE, it informs the MME of the event. The MME now requests the SAE-GW to switch the

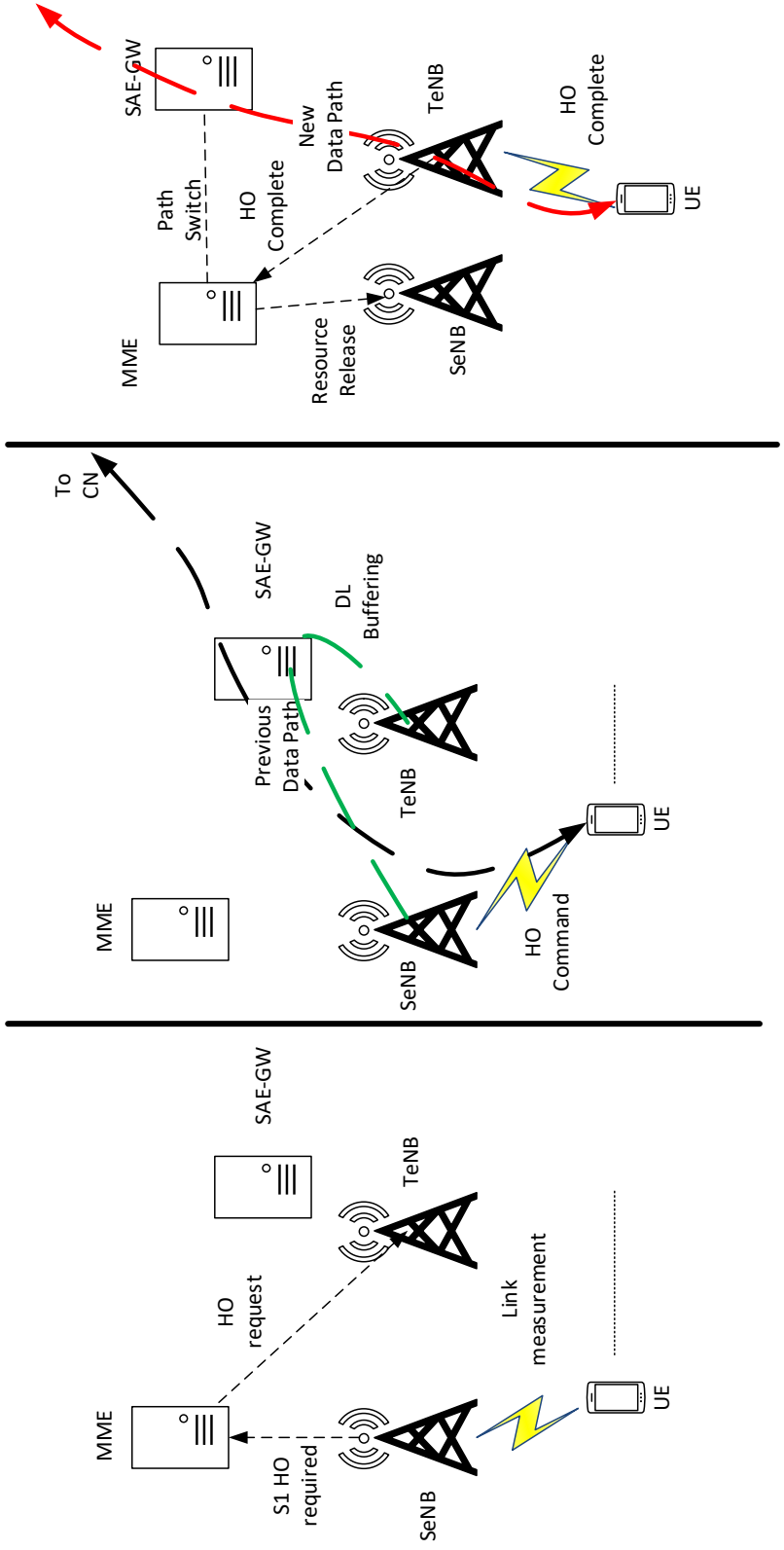


Figure 2.3: 3GPP S1 Handover

paths, and after the path has been switched for the DL, a release resource message is sent to the SeNB. This causes the SeNB to release all the resources reserved for the UE that just completed HO from its domain to that of the TeNB. And hence, in this way the LTE X2 HO is executed. Similar to Figure 2.3, in Figure 2.4 an illustrative description for the X2 HO is provided.

Cell re-selection and Tracking area update: Cell re-selection and TAU happen when the UE is idle, i.e., it does not have any active sessions running. The process of cell re-selection is fairly simple process (here we initially assume that the UE moves within the TAs that are already registered at the MME). Firstly, the UE while camping on a particular cell performs neighbour cell and serving cell measurements. After ranking the cells based on the RSSI, if the serving cell RSSI is greater than the threshold then no re-selection is performed. However, in the case when the neighbouring cell RSSI is greater than the threshold as well as that of the serving cell, cell re-selection is performed and the UE now camps on the new cell. Further, in the event that this new cell is in a TA that is different from those registered at the MME, a TAU message is sent to the MME. The MME then registers the new TA and also sends back a list of TAs to the UE. This list allows the UE to traverse in the TAs mentioned without performing any TA update.

In the event, that the UE is not idle and it performs a HO, a HO with TAU is performed. In this case, the HO procedure is the same as mentioned before, the only additional signaling being that at the end of the HO a TAU, as discussed above, is performed.

2.2.1.2 3GPP Dual Connectivity, LTE-WLAN Aggregation and LWIP

The Dual Connectivity (DC) concept allows a user to camp on two BSs simultaneously. Concretely, a UE can be connected to a Small-cell (SC) and a Macro-cell (MC) at the same time, wherein the MC and SC are connected to each other via the X2 interface. According to 3GPP, all control plane communications, including resource allocation on SC, are performed via the corresponding MC, to which the UE is associated to. Note that, DC was introduced by 3GPP for LTE during Release-12. But, it is in Release-13 that this concept matured, wherein multiple usage scenarios, architecture and the operational characteristics were defined. A detailed description of the same has been presented in [62]. Furthermore, during Release-13, the concept of LTE-WLAN aggregation (LWA) was standardized [15]. Through LWA, a UE can simultaneously receive packets over both the LTE and the Wi-Fi interfaces, wherein the aggregation (and splitting in the eNB) of these two physically distinct data streams takes place at the Packet Data Convergence Protocol (PDCP) layer in the protocol stack. However, note that the LWA functionality is defined only for the downlink [63].

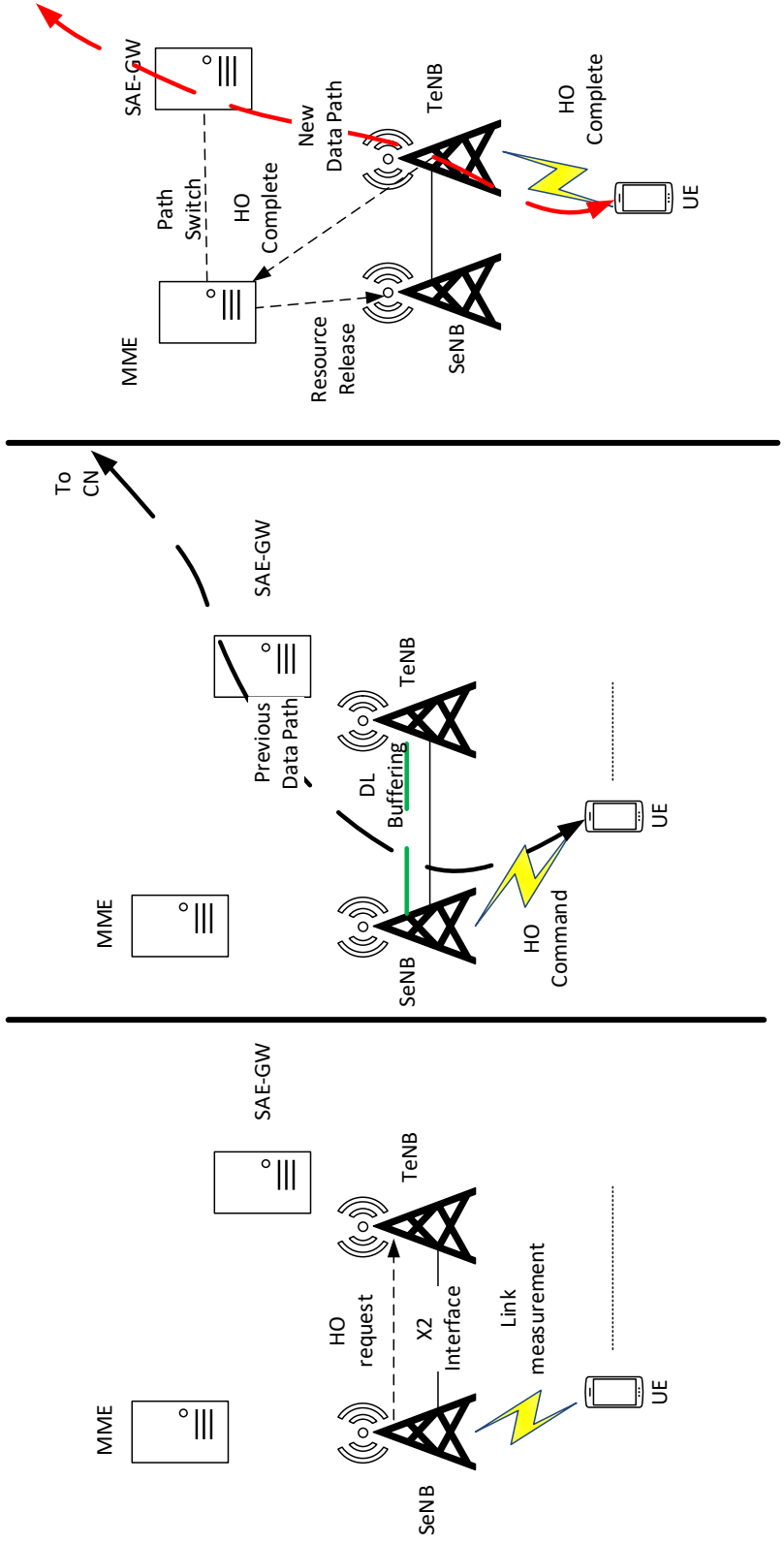


Figure 2.4: 3GPP X2 Handover

Another technique, similar to LWA, is the LTE WLAN integration using IP security tunnel (LWIP) [64]. In this technique, while the objective is similar to that of LWA, i.e., to integrate the LTE and WLAN technologies, it is done at the network layer in LWIP. However in LWA, the integration is performed at the PDCP layer. Further, LWIP, unlike LWA, can be implemented for both uplink and downlink.

2.2.1.3 3GPP Traffic Offloading

Traffic offloading essentially helps the operator to reduce the amount of traffic in its core network [65] by either offloading it to another domain in its network or to a completely different network such as Wi-Fi, the latter of which will be explored in more detail later. And so, specifically with this objective 3GPP introduced the Local IP Access (LIPA) and Selected IP Traffic Offload (SIPTO) frameworks [53, 65, 66]. The main reason to analyze the traffic offload frameworks is that they implicitly involve mobility from one domain to the other. And hence, the requirements of traffic offloading also need to be taken into consideration whilst designing the mobility management scheme and policies.

3GPP through the LIPA and SIPTO framework made provisions for traffic offloading approach in 3GPP networks. LIPA allows the network to offload the traffic locally if both the mobile node (MN) and the Correspondent Node (CrN) are in the same domain. By same domain we refer to the fact that the MN and CrN are associated with the same Home eNB (HeNB). Figure 2.5 presents the scenario involving LIPA traffic offload. Additionally, LIPA also allows for a local breakout to the Internet/CrN domain, hence bypassing the core network during the flow of data. An important challenge of LIPA with regards to MM is that, session continuity for LIPA connections during mobility events is not supported.

Further, the SIPTO framework enables the network to offload traffic to a geographically proximal gateway, hence, allowing the network to reduce traffic load on a particular gateway. Figure 2.6 provides an illustrative description as to how SIPTO framework operates. As can be seen from Figure 2.6, the traffic to the UE is offloaded to a set of gateways that are geographically close to the UE's point of attachment to the access network, which here are P-GW2 and S-GW [53]. Next, during 3GPP Release-10, the concept of IP Flow Mobility (IFOM) was also introduced. IFOM allows a UE to offload, if possible, data sessions to the Wi-Fi network from the 3GPP network. Consequently, through IFOM, a UE can maintain data flows belonging to the same packet data network (PDN) connection simultaneously on both the 3GPP and the Wi-Fi network [66]. However, while in LWIP the connection eventually passes through the LTE core network, in IFOM the offloaded connections pass through the WLAN network, and onto the IMS core.

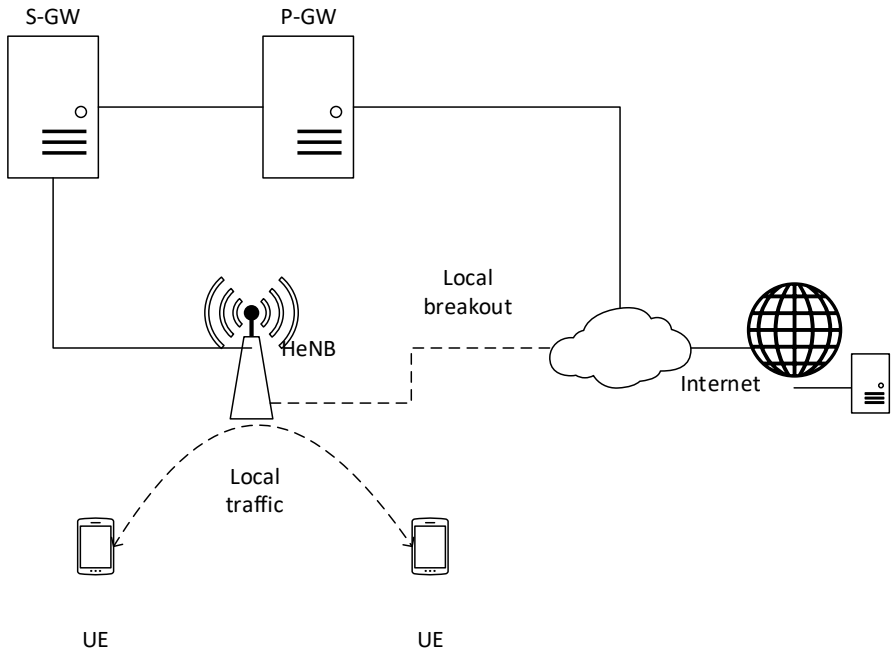


Figure 2.5: Local IP Access (LIPA)

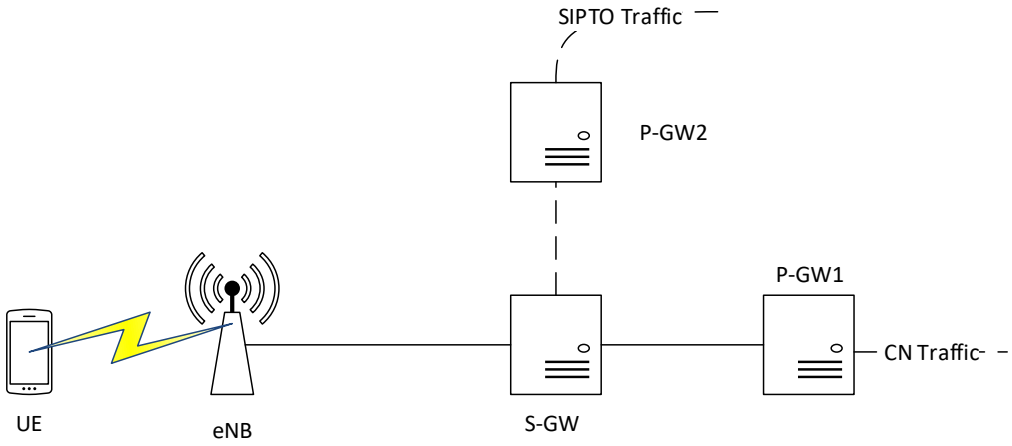


Figure 2.6: Selected IP Traffic Offload

It is important to re-iterate that since traffic offloading procedures implicitly invoke mobility management schemes/policies, it becomes necessary to study them with the perspective of designing mobility management schemes of 5G and beyond networks, which will be done in Chapter 3.

2.2.2 ITU – Vertical multi-homing

The future generation of wireless networks is envisioned to be one that is both dense in users as well as BSs, and heterogeneous. By heterogeneity it is understood that the network will comprise of multiple RATs co-existing in a single domain. This provides an opportunity to the users to utilize multiple RATs in order to improve the total throughput, reduce latency and increase the reliability of their link. And so, the ITU-T through its study on vertical multi-homing [67] provides the requirements, expectations and an architecture to perform the same. Concretely, in vertical multi-homing, each layer has a multi-homing feature and there are many network resources used to establish multiple network connections. In PHY/MAC layer, multiple network access technologies, multiple network interfaces, multiple channels, and multiple radios are network resources. In network layer, multiple IPv6 addresses and multiple prefix information are network resources. In transport layer, multiple transport sessions are network resources. To efficiently establish multiple network connections and manage network resources, it is needed to manage them in an integrated and harmonized fashion.

Moreover, while vertical multi-homing consists of aspects on how to connect with multiple interfaces (and how support across layers for the same is provided), MM in multi-RAT and multi-connectivity scenarios becomes a complex issue and hence, an analysis into the vertical multi-homing concept is well placed in the realm of our current research.

To elaborate, according to [67], vertical multi-homing entails having multi-homing capabilities in each of the layers of the implemented OSI network model, at both the host and the client. The modular approach of the OSI model, although ideal for making modifications without affecting other blocks, is the first challenge that vertical multi-homing encounters. This is so because, inter-layer coordination to optimize resource allocation and utilization is essential for the purpose of vertical multi-homing. As an example, consider the PHY/MAC layer has multiple interfaces active (each interface corresponds to a different technology) and the network layer has multiple active IP prefixes, however if the layers do not interact with each other, then the network layer will not know about the multiple interfaces. Consequently, an optimal association between active IP prefixes and active network interfaces cannot be determined, which will lead to an inefficient utilization of the resources. Hence, it is imperative that the protocol layers interact with each other. Further, as additional requirements, vertical multi-homing necessitates the provision of routing optimization, QoS based connection selection in presence of multiple network connections, bandwidth utilization over multiple network connections (through optimal stream splitting and combining), recovery methods in the event of network interface failure, and network interface selection (in the

event a single or a small subset of available networks can be utilized). Additionally, [67] also presents some methods that can be employed to implement multi-homing. These methods are classified as: 1) Based on correspondence between IPv6 address and interfaces (multiple IP addresses may configure multiple interfaces or a single IP address may be shared amongst multiple interfaces), and 2) Based on the supporting layers for multi-homing.

Through the requirements and methodologies as mentioned above, it is evident that the vertical multi-homing implicitly invokes mobility management schemes/policies. Moreover, as a step to handle vertical multi-homing on the network/host side a vertical multi-homing functions block, represented in Figure 2.7, is defined by ITU.

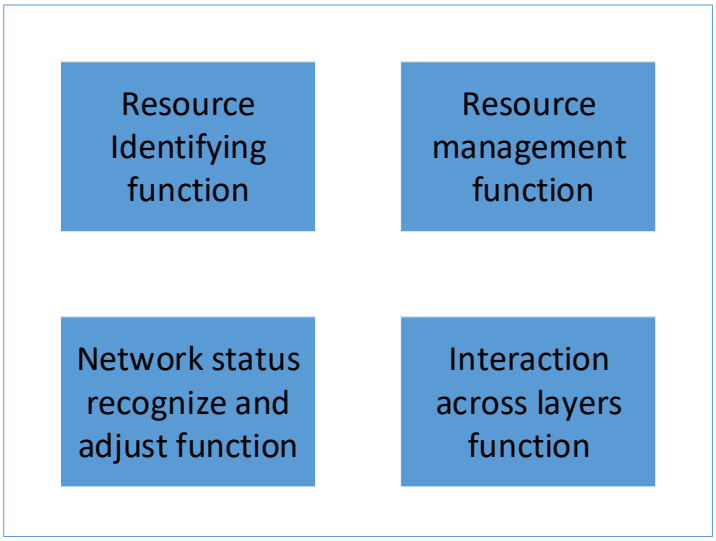


Figure 2.7: ITU-VMH functions block

Concretely, the vertical multi-homing functions such as resource identifying function, resource management function, network status recognize and adjust function, and lastly interaction across layer function provide functionality not only for vertical multi-homing, but they also serve as resources for carrying out mobility management whilst enforcing vertical multi-homing. And hence, the ITU-T, through its study in reference [67], implicitly tackles certain mobility management aspects such as RAT selection, resource identification and management, network status feedback and adaptive management, to mention a few.

2.2.3 CoMP

The Co-ordinated Multipoint (CoMP) strategy involves multiple base stations co-ordinating with each other to serve a given user [68]. Similar to ITU-VMH, CoMP can provision

path redundancy as well as seamless handover capability, owing to its coordinated feature. Further, similar to ITU-VMH, CoMP can configure multi-connectivity alongside per-channel granularity (multiple BSs permit multiple channels for transmission of data and hence, per-channel granularity of service can be provisioned) [68]. However, since CoMP will involve centralized scheduling operations, it will lead to SPoF as well as challenge the scalability of backhaul networks.

2.2.4 IETF based MM techniques

To enable mobility for the networked devices, IETF developed several standards such as the most famous Mobile IPv4 (MIPv4) and the most recent Mobile IPv6 (MIPv6). Since these standards have some drawbacks, there have been many proposed modifications of these standards to overcome the drawbacks and make mobility seamless. Hence, in the ensuing discussions, details with regards to the various mobility management schemes/enablers in the IETF framework are analyzed.

2.2.4.1 MIPv6

The MIPv6 framework is similar to the MIPv4 in that it allows for mobility by defining the MN to have a home address (HA) when attached to its home link and a Care-of-Address (CoA) when attached to a foreign network. The responsibility for allocating HA and CoA lies with the Access Router(AR). And so, according to [69], when a mobile node ventures out of its home area, a CoA is assigned to the MN. Hence, any packet arriving on its HA is automatically re-routed to its CoA wherein the packet then gets delivered to the MN. As an important preliminary step to allow such a routing procedure, a binding between the home address and its CoA is created.

Further, it is essential to note that MIPv6 is different from MIPv4 on several aspects [69]. While MIPv4 provides route optimization as an extension, MIPv6 makes route optimization as a necessary step and also reduces the amount of overhead involved during authentication step for route optimization. Additionally, MIPv6 also utilizes an extended header [69] for forwarding packets to the care-of-address. Hence, it also reduces the amount of time that would have been needed to process an encapsulated packet, as compared to MIPv4.

However, MIPv6 still suffers from link outage when a handover occurs, i.e., there is a service interruption time before it can be restored. Further, route optimization involves setting up new authentication between the network nodes, which will again lead to increased latency during mobility. And so, in the subsequent discussions, methods to tackle these aforementioned issues have been analyzed.

2.2.4.2 FMIPv6

Negotiation of resources, and creating bindings between the HA and CoA is the initial and the most time consuming step during a HO. Further, if the AR is in a separate domain, and if it requires new authentication and authorization steps, then the service interruption time becomes unacceptable. And hence, Fast MIPv6 (FMIPv6) [70] aims at solving the aforementioned issues, although it does not solve the authentication and authorization issue.

FMIPv6 provides a fast handover procedure, and can be essentially termed as a make-before-break approach. When, through layer 2 signaling techniques, a requirement for HO is detected, the current home AR provides the MN with information regarding the neighbouring routers (BS ID, IP prefix, etc.). After this information is made available, the MN configures a tentative CoA and forwards it to the current router in the form of a Fast Binding Update (FBU) message. The current AR then negotiates with the new AR with regards to the validity of the CoA. After the new CoA is considered valid by the new AR, an acknowledgment is sent to the MN. Additionally, during the period when the current AR sends a Fast Binding Acknowledgement (FBAck) and it receives a Fast Neighbour Acknowledgement (FNA) message from the MN (after attaching to the new subnet), the current AR buffers and routes the incoming packets at the new AR. And hence, the latency involved in configuring a new CoA as well as the arrival of packets after performing layer 2 attach at the new subnet is significantly reduced through the above process.

In addition to the aforementioned protocol, [70] also enlists a predictive and a reactive approach, wherein the predictive approach essentially delivers the FBAck message to the MN when it is still attached to its current AR. This demands that a prediction regarding the movement and direction of movement for the MN is made accurately. The other approach, i.e., the reactive approach, is executed when the MN moves to the new subnet before receiving the FBAck message or before it is able to send an FBU message on the current AR's link. In such a scenario, these messages are encapsulated within the FNA message and the protocol operation is carried out. Figures 2.8 and 2.9 depict the predictive and reactive HO signaling diagram, respectively, for reference.

2.2.4.3 HMIPv6

One of the issues encountered during the mobility management process is that the frequent handovers can entail significant delays while updating binding entries as well as performing authentication processes. Hence, in order to reduce the latency due to the aforesaid factors, [71] introduces the Hierarchical MIPv6 (HMIPv6). HMIPv6 introduces an intermediate entity known as the Mobility Anchor Point (MAP). MAP, as shown in Figure 2.10, operates

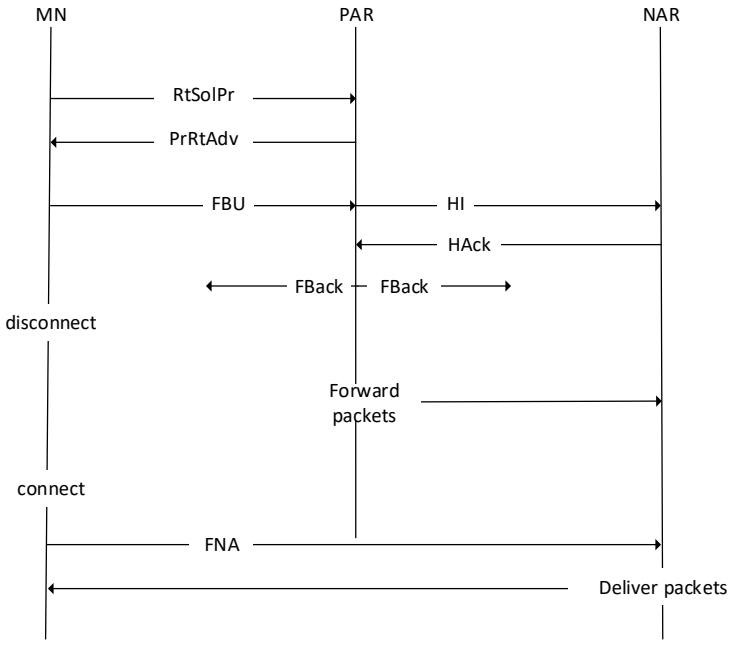


Figure 2.8: FMIPv6 Predictive HO signaling.

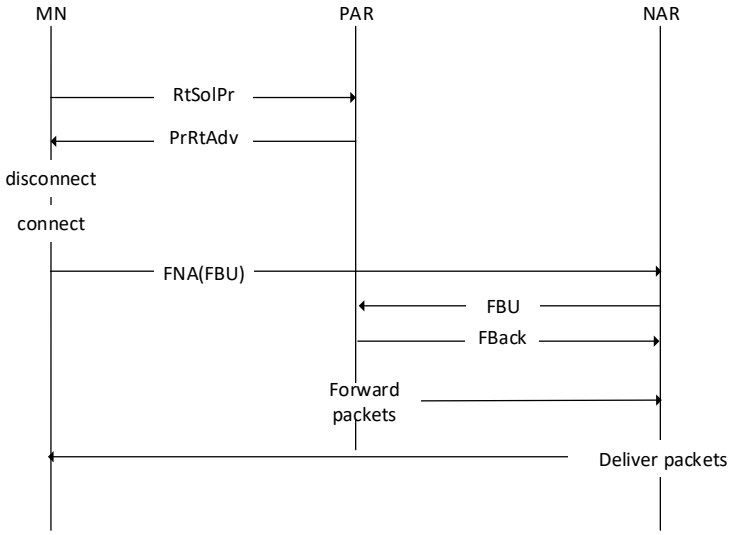


Figure 2.9: FMIPv6 Reactive HO Signaling

between the HA and the MN. The domain of MAP encompasses a collection of wireless BSs, and as can be inferred from the discussion in [71], MAP provides a regional CoA (RCoA) for the MN attached to its domain. Further, within the domain, a Local CoA (LCoA) is configured for the MN, and a binding between the RCoA and the LCoA is created to allow

the packets to be routed to its destination.

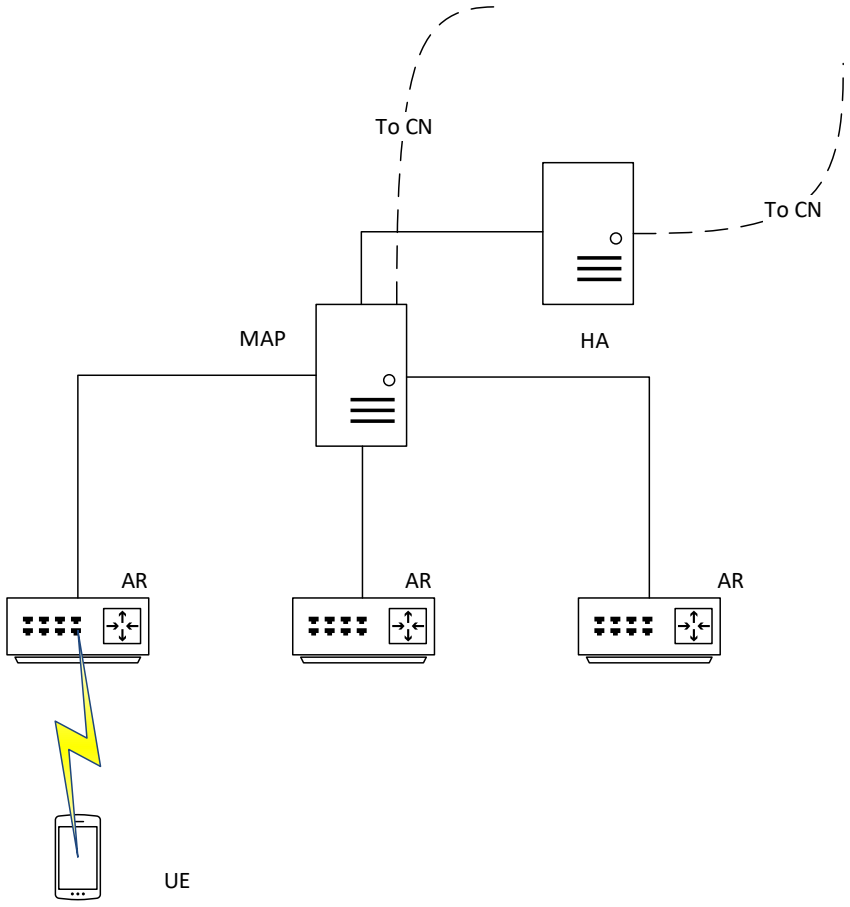


Figure 2.10: HMIPv6 Architecture

Further, when a mobile changes its BS within a MAP domain, it only changes its LCoA, whilst its RCoA is still the same. Hence, the interruption in service is negligible, and the service is agnostic to any movement. Additionally, through route optimization MN can utilize its RCoA to directly forward data to its CrN without having to send it through the home router. And hence, from the above discussion it is clear that HMIPv6 reduces the amount of handover updates, consequently reducing the handover latency.

2.2.4.4 PMIPv6

Almost all mobility management approaches require that the MN is involved in the signaling process to initiate and execute the handover. However, this assumes that the MNs are enabled to work with those mobility management schemes, which essentially requires the

MN to be configured with the necessary software beforehand. And since this might not be true in most cases, Proxy MIPv6 (PMIPv6) was designed in order to provide a network based mobility management scheme. PMIPv6 [72] has been designed such that it does not require the MNs to participate during the handover process, and consequently MNs do not need to be specifically configured to be served by a PMIPv6 enabled domain.

PMIPv6 protocol specifies the provision of two new networking entities, i.e., Local Mobility Anchor (LMA) and the Mobility Access Gateway (MAG) [73, 74]. The LMA is the global mobility anchor for any MN. It is also responsible for issuing IP prefix to the MN in its domain. The MAG on the other hand is responsible for tracking the movement of the MN, and triggering HO as and when it is required [73, 74]. It must be noted that MN is not involved in any HO signaling, and so it is the MAG's responsibility to detect any possibility/requirement of a HO. It is noteworthy that PMIPv6 has also been adopted by 3GPP networks [75], thus reflecting the maturity and reliability of the solution with regards to its utility for future MM solutions.

Figure 2.11 shows an illustration for the PMIPv6 architectural setup. It must be noted that when a node changes BSs, it still is in the MAG domain and hence, no path switch or reconfiguration is required. However, when a MN switches between two MAG domains but within the same LMA domain, then the Binding Cache Entry (BCE) for the MN is updated on the LMA. Further, when the MN attaches to the new MAG it carries the same network prefixes, and hence for the MN and the network, the connection is virtually static. This reduces the prefix configuration latency, and hence reducing the service interruption time.

In addition to the aforementioned features, PMIPv6 also allows for the local traffic, i.e., if the MN and CrN are in the same MAG domain, to be routed through the MAG. This is a form of route optimization as well as traffic offloading, with the benefits being reduced latency and traffic load on the core network. However, being centralized in nature, it can impact the network scalability and reliability in dense and heterogeneous future network environments, as a large volume of the traffic will pass through a single anchor. This can consequently lead to SPoF and congestion [76], thus making it less favorable for 5G and beyond MM mechanisms. And so, certain studies such as [76, 77] have provided discussions on scalable methods for PMIPv6. Specifically, in [77] a PMIPv6 based DMM approach has been proposed. The DMM approach provisions a decentralized method (without any mobility anchors) and helps eliminate the SPoFs. Furthermore, in [76], a cluster based approach was proposed to enhance the scalability of the existing PMIPv6 protocol.

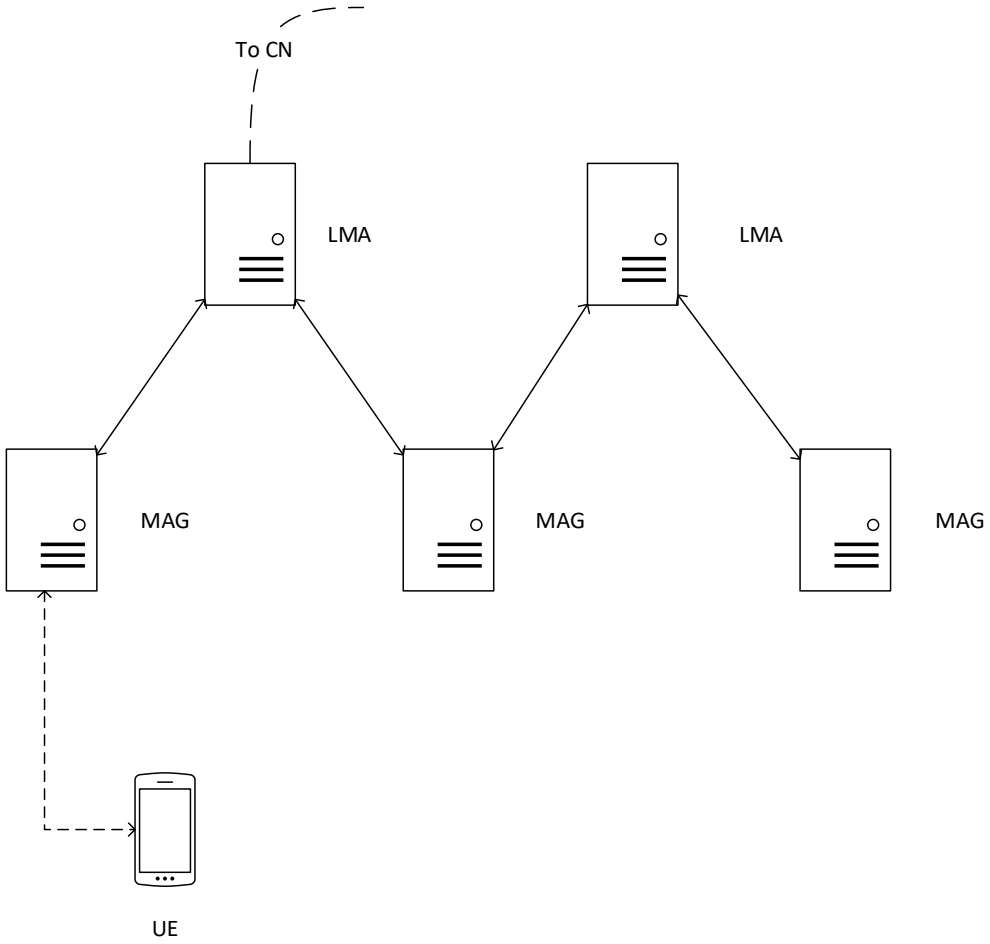


Figure 2.11: PMIPv6 Architecture

2.2.4.5 MPTCP

Whilst mobility management for the most part entails handling the handovers and providing tracking updates in a manner that ensures seamless connectivity, the idea of multiple interfaces and multi-homing should also be considered when designing mobility management schemes. Multipath Transmission Control Protocol (MPTCP) [78–80] allows for multiple TCP paths to operate over single/multiple interfaces as well as multi-homed networks. Generally utilized for increasing data rates [81] and improving the QoS, the provision of multipath redundancy [82–85] and congestion awareness (at the transport layer level) [86–88] will be beneficial for 5G and beyond MM mechanisms. Additionally, MPTCP provisions granularity of service at the flow level, which will be essential for the future MM mechanisms.

However, according to [78, 79, 89], for MPTCP to be implemented without altering the

legacy systems, proxy servers supporting MPTCP will need to be installed in front of the legacy devices, such as the middleboxes installed by service providers. The legacy systems can then communicate with the proxies using the legacy TCP protocol, while the proxies utilize MPTCP for communicating with the destination MPTCP capable device. Such a requirement will potentially impact the scalability of the MPTCP solution for 5G and beyond MM mechanisms.

2.2.4.6 SCTP

Stream Control Transmission Protocol (SCTP) [80], like MPTCP supports multi-homing and allows for separate message streams to be sent simultaneously. However, it differs from MPTCP in the fact that it incorporates qualities of both User Datagram Protocol (UDP) and Transmission Control Protocol (TCP) in how the messages are handled, whereas MPTCP is a direct extension of TCP.

Additionally, and similar to MPTCP, it provisions multipath redundancy and congestion awareness [90]. Moreover, it also facilitates flow level granularity of service, like MPTCP, which will be important for 5G and beyond MM mechanisms. Hence, SCTP through its features is also a potential future MM mechanism enabler.

Note that, for SCTP, both the user and server protocol stacks need to be updated [90]. The aforesaid update will essentially be a software update, wherein the transport layer of the protocol stack is updated. However, given the number of users in future networks, it will pose a scalability challenge for the deployment of SCTP as part of the 5G and beyond MM mechanisms.

2.2.5 IEEE Media Independent Handover 802.21

In order to provide inter-domain mobility, i.e., between various IEEE 802 standards as well as non-IEEE 802 standards such as 3GPP technologies, IEEE standards group came up with the IEEE 802.21 standard. The IEEE 802.21 is a Media Independent Handover (MIH) service, that as the name suggests provides a common intermediate platform and consequently enables the upper layers to interact with the lower layers (layer 2 and below) irrespective of the technology [91].

The MIH service provides the event, command and information services, which form the core of this protocol. This enables the higher layers in the protocol stack to query information that is present on the link and MAC layers for ensuring seamless connectivity in between domains and consequently enhance the user QoE [43, 91–94].

Further, the IEEE 802.21c amendment [93], provides insights as to how UEs with single

radio can perform seamless inter-domain handovers. The suggested approach is a make-before-break approach. This amendment also states that the most time consuming process when involving an inter-domain HO is the authentication and context information exchange. And hence, in order to reduce this latency a proxy connection approach is adopted, depicted in the architecture in Figure 2.12.

In Figure 2.12, the MN, which is undergoing a handover, is initially attached to the source network via a Source Point of Attachment (Source PoA). The MN is consequently being handed over to the Target PoA (TPoA). And so, the MN, instead of traversing the entire network to reach the information server, to retrieve details regarding the candidate target networks and their corresponding handover policies, accesses the same via the Source Point of service (SPoS). Following the handover decision, it is the responsibility of the SPoA to communicate with the TPoA, with regards to resource allocation, and authentication and context information exchange. These processes, according to [93], are performed proactively via the interaction between the SPoS/SPoA and TPoS/Proxy TPoA.

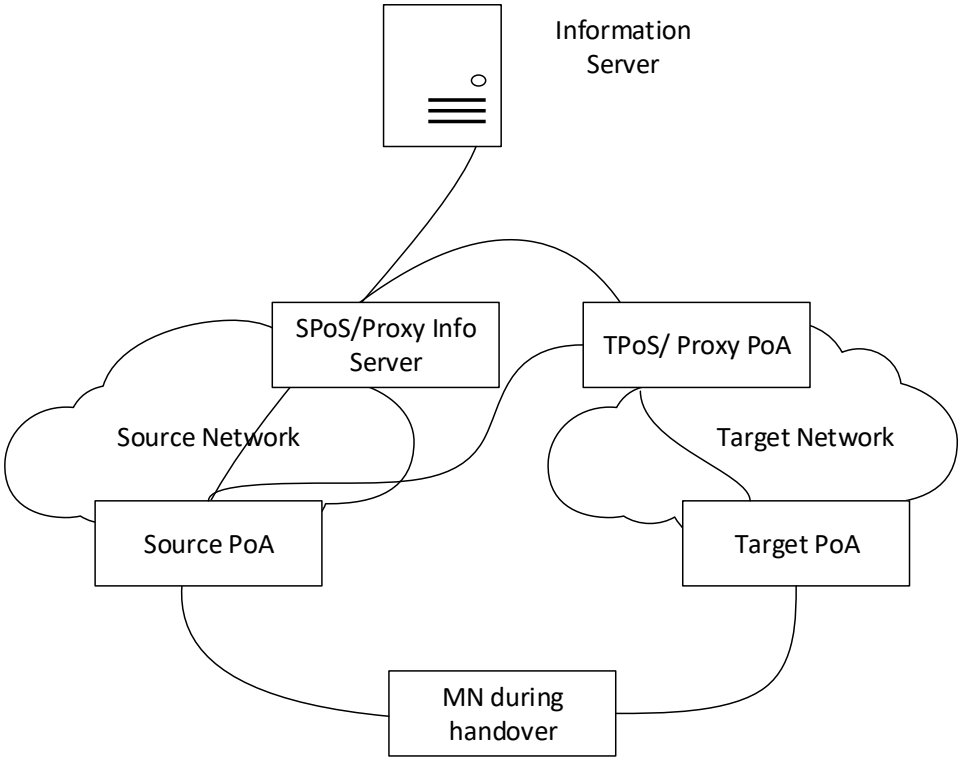


Figure 2.12: IEEE 802.21c – Single Radio handover functional model

Thus, through the provisioned proactive information transfer between the target and source networks, the latency can be significantly reduced. Further, in Figure 2.12, the proxy

connection between target and source networks helps in the secure exchange of context and UE authentication data, while there is no requirement for a layer 2 attach for exchanging such information. This allows for fast and secure handovers and in this way inter-domain handovers for single radio receivers can be made seamless.

It is important to mention that as stated in [95], the initial drafts of IEEE 802.21 did not favour mobile assisted handover mechanisms. However, through the course of its standardization IEEE 802.21 has adopted an approach wherein the decision to perform a handover is taken through collaboration between the network and the UE.

And so, with these provisions, IEEE has facilitated the process of providing a standardized platform for the various heterogeneous technologies to co-operate and allow seamless mobility [43, 91–94]. Moreover, 3GPP technologies can also utilize this information and hence, allow devices to handover from 3GPP to IEEE 802.x RATs and vice versa.

2.2.6 RSS based BS selection methods

The erstwhile Received Signal Strength (RSS) based methods employ a very simplistic approach to BS selection, by comparing the detected BS link quality (RSSI/RSRP/RSRQ) levels [75, 96, 97]. The aforesaid simplistic nature renders them easy to implement, and does not entail a high processing and signaling load either. However, such an approach can be plagued by multiple issues. For example, BSs with a good RSS might be overloaded (as more users will be assigned to them) whilst others maybe under-utilized. Such a scenario also implies that a better RSS does not always guarantee better QoS, since, congestion will lead to degraded QoS. Moreover, in dense scenarios, even with the implementation of a hysteresis, UEs will be subject to FHOs due to the fluctuating RSS and availability of multiple candidate BSs. This exemplifies the unreliable nature of RSS based methods for 5G and beyond MM. Additionally, these methods are one-dimensional, given that they consider only RSS as a decision parameter. The RSS methods also do not provision any granularity of service, context awareness, multiple levels of HO support, etc.

It is imperative to state here that, in Chapter 3, wherein we present the qualitative gap analysis, a subset of the legacy mechanisms discussed in this section have been analyzed. Concretely, we analyze the 3GPP LTE handover mechanisms, 3GPP LTE Traffic offloading, 3GPP Dual Connectivity and LWA, IETF PMIPv6, IETF MPTCP, IETF SCTP, IEEE 802.21 and RSS based BS selection methods. We choose the aforementioned strategies for the qualitative gap analysis due to their wide-ranging acceptance/applicability in wireless networks.

2.3 Mobility Management: Current State of the Art

Global efforts have spinned up consortiums that have provided impetus to the development of 5G, including that of MM strategies. Further, for B5G networks, such as 6G, certain collaborative efforts have already started. References [32,33,35,36,98] highlight the advances that have been made with regards to identifying the enablers and core principles of B5G networks. Hence, in this section we first detail the 5G architecture defined by 3GPP, followed by the discussion on current state of the art in MM mechanisms.

2.3.1 3GPP 5G Architecture Background

We introduce, through Figure 2.13, the 5G architecture standardized by 3GPP [45]. Concurrently, we have also presented the classification of the various mechanisms that we explore in Sections 2.3.2 and 2.3.3 with respect to the 5G architecture in Figure 2.13. This classification is dependent on the portion of the network that is impacted (directly or indirectly) the most by a particular MM scheme. Furthermore, we have illustrated whether the studied mechanisms are either control plane or data plane solutions. Concretely, a CP solution would primarily impact MM via either CP signaling or decisions, while a DP solution would entail provisioning alternate and more efficient data paths. A detailed discussion with regards to these classifications has been provided in Sections 2.3.2 and 2.3.3.

Concretely, the 5G architecture, as shown in Figure 2.13, consists of two main core network functions, i.e., the Session Management Function (SMF) and the Access and Mobility Management Function (AMF). The SMF communicates with the User Plane Function (UPF) over the N4 interface, while the AMF is responsible for communicating with the RAN side over the N2 interface. Furthermore, the AMF and SMF communicate with other network functions, such as the Policy Control Function (PCF), Authentication Server Function (AUSF), etc., to execute their defined functionalities within the ambit of the policies and existing user and network context. For the sake of conciseness, in Figure 2.13 we club all of these functions into a single entity box called *Network Functions*. Moreover, the AMF also has an N26 interface that connects to the EPC to facilitate Inter-RAT mobility, while an N32 interface exists in the event of a change in Public Land Mobile Network (PLMN) with 5G Core (5GC) as the CN for both the visited and home networks. Note that, the interfaces *N2*, *N4*, *N26* and *N32* are all control plane paths, with the AMF, SMF and other network functions forming the control plane entities.

In addition, the AMF in 5G networks is the equivalent of the MME in LTE-4G networks. It focuses on handling mobility at the access network level (such as BS selection, resource allocation, etc.). The SMF on the other hand handles the CN related tasks during mobility

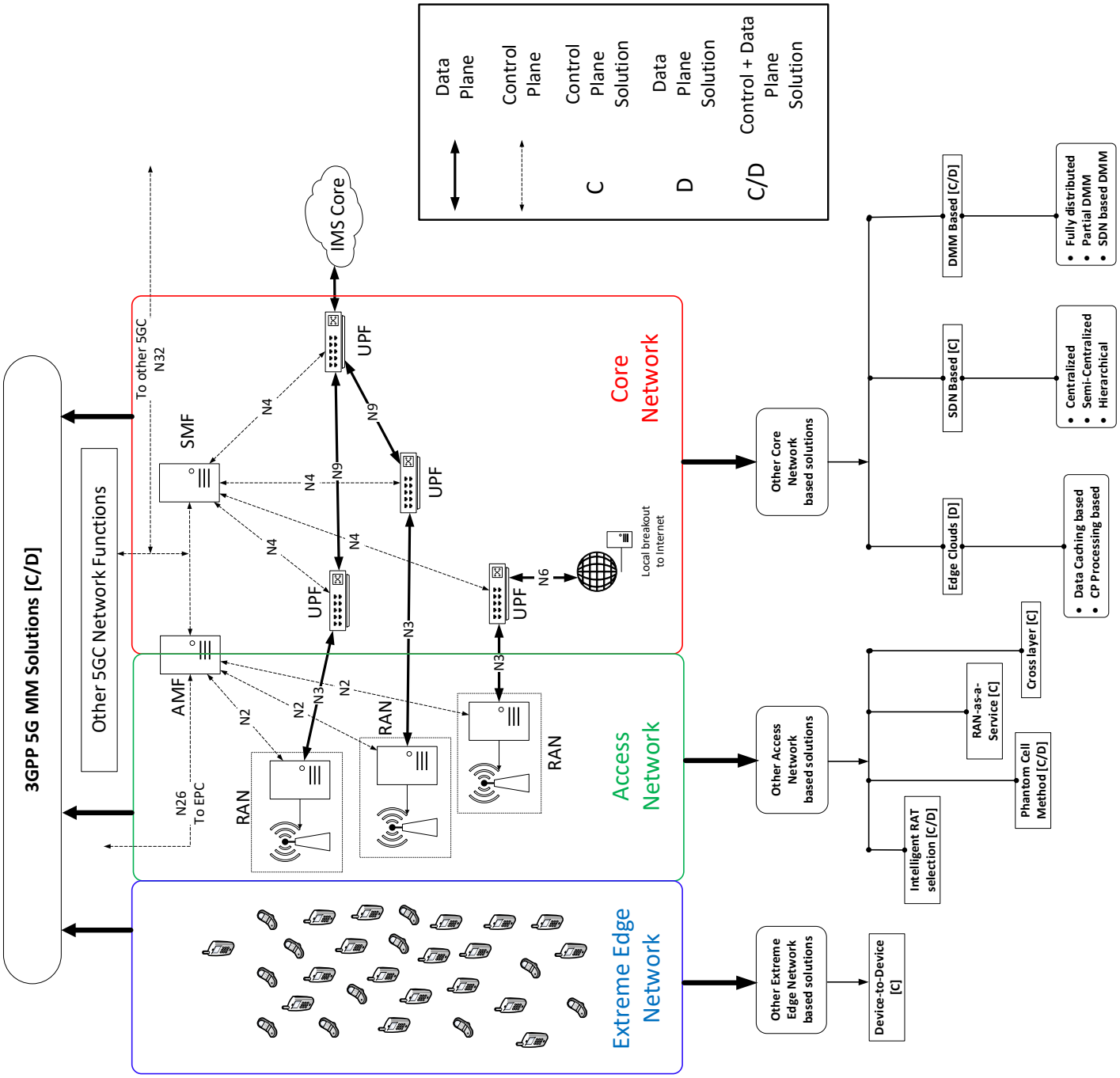


Figure 2.13: Classification of the state of the art in MM strategies on the 5G architecture.

events (such as path re-routing, etc.). Next, in Figure 2.13, it can be seen that the RAN interacts with the UPF through interface N3, and the UPFs use the N9 interface to communicate amongst themselves. Also, the 5G networks provision a local breakout through the N6 interface from an UPF. The interfaces *N3*, *N9* and *N6* constitute the data plane paths, with the RAN and UPF forming the data plane entities. Lastly, the UE, which is also a data plane entity, interacts with the AMF through the N1 interface. However, to maintain clarity, we have omitted the illustration of this interface from Figure 2.13. Thus, with this background, we now explore the current state of the art in MM mechanisms.

2.3.2 3GPP 5G MM Mechanisms

3GPP, through TS 23.501 [45], TS 23.502 [99] and TS 38.300 [100], has provided significant insights into the design and development of 5G MM strategies. New session management methods, service continuity states, UE mobility monitoring, provisioning for multi-homing, load balancing strategies, provision of on-demand MM, resource allocation due to mobility events, the new MM module, i.e., AMF, inter- and intra- next generation core (NGC) handovers, and LTE-EPC 5G-NGC interworking have been introduced in the aforesaid 3GPP specifications. These techniques through the provision of a softwarized solution and a global view of the network scenario alongside user context appear to facilitate the efficient operation of 5G and beyond MM mechanisms. Consequently, in the text that follows, we discuss these newly defined 3GPP MM mechanisms.

A. *UE Mobility monitoring:* In TS 23.501 [45], details with regards to how the UE mobility is monitored and the corresponding actions with regards to resource allocation and context updates have been specified. Concretely, when a UE is mobile, the 5G standards define that the AMF will be responsible for monitoring its movement and hence, its mobility pattern. Furthermore, during a UE mobility event, new resources on the destination BS are managed by the AMF through the RAT and Frequency Selection Parameter (RFSP). Such a process simplifies the identification of the required resources, as well as migration of these resources to the destination network. Moreover, the AMF manages the UE mobility event notification, i.e., it provisions details with regards to the mobility event as well as the areas of interest (Tracking areas, Cells, RANs, etc., to which a UE might migrate to). The other Network Functions (NF), such as the SMF, can subscribe to these notifications so as to employ their decisions and policies.

B. *Session Management:* Through TS 23.501 [45], the various modes that can be utilized

to manage the multiple heterogeneous sessions for a given user has been defined. Notably, if a UE is connected to multiple RATs then, for a given Protocol Data Unit (PDU) session, the UE has the choice to select the access network over which this PDU session will be served. In addition, the UE, in the event of mobility or congestion, can request a PDU session to be transferred from 3GPP to non-3GPP RAT(s). Furthermore, in roaming scenarios, PDU sessions can either avail a local breakout or be routed through the home network. Specifically, each PDU session can be granted, independently, different routing modes. To do so, the SMF in the 5G CN controls and monitors the status of the data paths. Moreover, the SMF also provisions the capability of performing selective traffic routing by the application of Uplink Classifier (UL CL) on certain data plane entities, i.e., UPFs. A UPF essentially performs the function of a router in the 5G network.

C. *IPv6 multihoming:* The new 5G standards, as specified in TS 23.501 [45], have formalized the use of IPv6 multi-homing so as to reap the benefits from the multiple physical channels that will be available for use through multi-connectivity. Specifically, according to TS 23.501, more than one session anchor can be specified for a PDU session. Note that, a PDU session anchor's primary role is to assign the IPv6 prefixes that are used by the UE for a given PDU session to communicate with the public network. However, all these PDU session anchors will have a single UPF as a branching point. Next, during a mobility event, a make-before-break approach for a PDU session is adopted to provision service continuity. It must be stated here that, service continuity is ensured through the Session and Service Continuity (SSC) modes, which we will discuss next.

D. *Session and Service Continuity Modes:* 3GPP, through TS 23.501 defines the SSC modes, which are critical for the network in determining the level of service continuity offered to a PDU session [45]. Concretely, three modes are defined: *SSC mode 1*, *SSC mode 2* and *SSC mode 3*. We briefly describe them as follows:

   – *SSC mode 1:* This mode ensures that the IP address is preserved. Specifically, the PDU session anchor is maintained regardless of the access technology being used by the PDU session after the mobility event. Furthermore, the IP address is maintained throughout the lifetime of the PDU session. Additionally, more PDU session anchors might be allocated for additional IP addresses, however, it is not necessary that they be maintained just like the initial IP address and PDU session anchor.

- *SSC mode 2:* In this mode, if needed, the network can release a PDU session and request the UE to immediately establish a new PDU session with the same network. Moreover, if the UE has multiple PDU session anchors, the additional anchors can be released or allocated (for new IP prefixes/addresses).
- *SSC mode 3:* In this mode, IP address is not preserved. This consequently makes any changes in the user plane visible to the UE. However, to ensure that an acceptable level of QoS, and hence, service continuity is maintained, a *make-before-break* approach is followed. This essentially determines the destination PDU session anchor before relieving the resources the PDU session occupies at its current anchor.

It must be stated here that the SSC mode for a UE is selected by the SMF depending on the UE subscription details as well as the PDU session type.

E. *User Plane aspects:* In 5G networks, UPFs will be utilized to handle the data plane traffic. Concretely, they can be thought of as routers, on whom the routing rules are programmed by the SMF. In TS 23.501 [45], the aforesaid specifics have been defined. However, note that the methodology to establish these paths still involves exchanging Tunnel Endpoint Identifiers (TEIDs) between CN entities. This, as we will state in the analysis, can be a cause of increased network load. Additionally, traffic re-routing, in the event of mobility or load balancing, is handled by the SMF, wherein it sends the necessary information, such as the forwarding target information, to the UPFs. Lastly, in the event of mobility of a UE, packet buffering is also provisioned so as to minimize the loss of packets and hence, QoS.

F. *Dual Connectivity:* Through TS 23.501 [45] and TS 37.340 [101], 3GPP has also concretized and standardized the integration of Multi-RAT Dual Connectivity (MR-DC) into 5G. Concretely, the UEs will now have the capability and possibility to connect to two BSs belonging to the same RAT (LTE-LTE, 5G New Radio (NR) - 5G NR) or to different RAT(s) (LTE - 5G NR). As in LTE-DC, this can be configured to allow fast-switching (fast HO), since control plane is not changed unless the Master Node is changed. Further, it can also be used to increase data rates by using both RATs at the same time. Also, the UP is terminated at Master node, so, no CN signalling is necessary for intra-Master Node HO.

G. *Edge Computing:* TS 23.501 [45] defines the support for edge computing platforms in 5G networks. Concretely, these are utilized in the non-roaming or local breakout

roaming modes. Note that, the 5G CN is responsible for selecting a UPF that is close to the UE and also has access to an edge compute node. Consequently, traffic steering is performed at this UPF towards the edge compute node.

H. *Network Slicing:* The concept of enabling a telecom operator to be able to slice its infrastructure network into logically separated networks and consequently service multiple tenants, e.g., virtual network operators, services (eMBB, URLLC, mMTC), etc., using the same, wherein the logical separation involves dynamic allocation of network resources, is termed as network slicing [57]. 3GPP, in TS 23.501 [45], has discussed network slicing in detail, wherein its support for roaming as well as its involvement in the inter-working process between 5G CN and LTE-EPC have been elaborated. It also defines support for migrating and translating the Single Network Slice Selection Assistance Information (S-NSSAI), which consists of the necessary information with regards to an assigned network slice for a UE, between the Home PLMN (H-PLMN) and the Visited PLMN (V-PLMN) has been detailed. Similarly, for the inter-working process, 3GPP charts out the principles for migration, translation and creation of S-NSSAIs whenever a UE undergoes mobility and changes from a 5G network to an LTE network, and vice versa. Moreover, the support has been defined for scenarios where the N26 interface, which is the standard 5G CN and LTE-EPC inter-working interface, may or may not be present [45].

   On the other hand, and importantly, the concept of network slicing also assists in provisioning tailor-made MM solutions for the tenants that each network slice will cater to. This consequently helps to deploy on-demand MM strategies.

I. *Load Balancing and Congestion Awareness:* In TS 23.501 [45], 3GPP has defined procedures for load balancing at the AMF and SMF, as well as congestion awareness within the core network. Concretely, two specific strategies, i.e., load balancing and load re-balancing, have been provisioned. Within the load balancing paradigm, new users incoming into an AMF region, if necessary, are directed to an appropriate AMF in order to manage the load of the AMFs. To do this, appropriate weights, indicative of the load on each AMF, are assigned and updated at appropriate intervals (typically on a monthly basis). On the other hand, if an AMF becomes overloaded, then load re-balancing is performed. Here, already registered users are migrated to other AMFs that are not overloaded while ensuring minimum service disruption [45]. Note that, the new AMF chosen should belong to the same AMF set. An AMF set is defined as the AMFs which belong to the same PLMN, have the same AMF region ID and the same AMF set ID value [45]. These parameters are pre-configured by the network operator. Lastly,

3GPP also provisions extensive details with regards to handling congestion control for the Non Access Stratum (NAS) messages. This is important from the perspective of MM, as MM messages are carried over NAS to the CN nodes. For further details with regards to the specifics of the congestion control procedures, the reader is referred to TS 23.501 [45].

J. *Cell, Beam and Network Selection:* Through TS 23.501 [45] and in particular through TS 38.300 [100] details with regards to cell, beam and network selection have been specified. For *cell selection* these standards documents, developed by 3GPP, specify support for cell selection procedures given that the UE is in either Radio Resource Control (RRC) idle, or RRC inactive or RRC connected state. Note that, RRC idle state refers to a UE that can listen to paging channels, broadcasts and multicasts, as well as perform cell quality measurements. The RRC inactive state refers to a UE that, in addition to the functionalities specified in the RRC idle state, can roam within the RAN-based notification area (RNA) without informing the NG-RAN. The RRC connected state for a UE implies that it has an active connection and data flow. Most notably, for the RRC connected state, cell mobility and beam mobility have been specified. As the name suggests, a UE can either undergo a cell handover or it can switch between the beams that a given BS uses. To perform this, procedures for beam quality and cell quality measurements have also been defined in [100]. The beam quality measurements are performed in the physical layer for multiple beams being transmitted by a given cell. These measurements are filtered and aggregated at the RRC layer to obtain the cell quality measurements. Note that, these quality measurements are still performed using the RSSI/RSRP/RSRQ/SINR metrics. Furthermore, in [100], procedures for cell selection and handover involving intra- and inter-frequency handover in 5G NR, Inter-RAT handover within 5G CN, Inter-RAT handover from 5GC to EPC and vice versa, have been specified. We refer the reader to TS 38.300 [100] for a more detailed discussion on the same. Moreover for Inter-RAT handovers, procedures for packet buffering and forwarding as well as data path switching, to ensure the requested QoS, have also been defined. Lastly, roaming and access restrictions are also appropriately defined based on the user subscription to both the SMF and AMF. This facilitates the selection of the right BS and PLMN for a given user [45, 100].

K. *Inter-Working, Migration and Handover signaling:* While TS 38.300 [100] specified certain handover procedures for both the CP and DP, a detailed description of the handover signaling, inter-working between 5G CN and EPC, and migration of PDU sessions has been provided in TS 23.502 [99] and TS 23.501 [45]. Concretely, through

[99] the CN signaling process for the various stages in a handover, i.e., handover request, handover preparation, handover complete/cancel/reject, have been presented in detail. These handover signaling procedures have been detailed for Intra-RAT HO (N2 and Xn handovers) as well as for Inter-RAT handovers (involving 5GC and EPC). Moreover, the handover signaling procedures have also been defined for the scenarios wherein the EPC-5GC inter-working interface, i.e., N26, may or may not be present. Also note that, the 5G-N2 handover is similar to the LTE-S1 handover (specified in Section 2.2.1) and the 5G-Xn handover is similar to the LTE-X2 handover (specified in Section 2.2.1). Next, for the 5GC and EPC inter-working, in TS 23.501 [45] the principles for maintaining IP address continuity, i.e., addressing, relaying and routing, in the event of UE mobility from 5GC to EPC or vice versa have been provisioned. However, it is also specified that in the event a UE transitions from 5G to 3G or 2G and vice versa, the IP address continuity might not be maintained. Furthermore, procedures for transferring the PDN/PDU sessions established by a UE over a 4G/5G network, when it transitions to the 5GC/EPC, over the N26 interface have been provisioned in [45]. Also, traffic steering and forwarding procedures have also been elaborated. Lastly, procedures for migrating PDU sessions from non-3GPP access to the 3GPP access, when a UE undergoes a mobility event from 5GC to EPC, is also supported [45].

L. *D2D mobility support:* With the standardization of Proximity Services (ProSe) in 3GPP Release-12 and 13 [15], 5G networks can utilize the capability to orchestrate data forwarding/relaying in both DP and CP. This can consequently enhance the ability of the network to provide a proactive and seamless handover procedure [102].

Given that the 3GPP 5G MM mechanisms provision both CP and DP related strategies as well as the core, access and extreme edge network related mechanisms, in Figure 2.13 they have been classified as illustrated.

2.3.3 Other Research Efforts: Core, Access and Extreme Edge Network Solutions

From the perspective of MM strategies in 5G networks, the main objective of the ongoing academic and industrial research efforts has been to provision mechanisms that cater to the myriad user mobility and application profiles, as well as to ensure context/on-demand based service provision and continuity [103]. For example, in [104], a wide swathe of avenues that exist in the 5G MM design have been explored. It discusses an SDN based framework

that can encompass strategies and techniques which grant certain level of adaptability (feedback based), flexibility (in terms of granularity of service provisions) and reliability (through availability of multiple paths) for 5G MM solutions. Notably, and apart from the aforementioned broad study, specific areas of MM have also been tackled through research efforts such as through our contribution [J1], wherein optimal handover signaling strategies for 4G-5G networks have been proposed.

Hence, given that we will be analyzing a wide range of mechanisms and strategies, we have broadly classified them as being *Core Network*, *Access Network* and *Extreme Edge Network* based solutions, as shown in Figure 2.13. These classifications reflect the regions in the network where the respective mechanisms generate the most impact. Concretely, *Core network* based solutions will invoke solutions that primarily assist in the provision of MM services through the core network. Similarly, the *Access network* and *Extreme Edge network* solutions assist in provision of MM services through the access (RAN side) and extreme edge portion (users/devices side) of the wireless network, respectively.

2.3.3.1 Core Network Solutions

Core network solutions have been categorized further as either being *SDN*, *DMM* or *Edge clouds* based. Solutions that utilize SDN to implement MM can be equipped with a global or locally-global network view. This top-view of the network enables MM solutions to offer a high degree of optimality. However, as a result of the convoluted 5G network scenario, the design of SDN CP also becomes increasingly crucial. Hence, the placement of SDN controller(s) (SDN-C) in the overall network topology is an important factor to consider [105]. Consequently, we present a brief discussion on the SDN based solutions, which might be Centralized, Semi-Centralized or Hierarchical [106–108].

A centralized MM solution will consist of a single global SDN-C which monitors and manages the entire network. With the global view, it enables the formulation of optimal MM solutions. However, the centralized nature might not offer the scalability and reliability (SDN-C can be a SPoF) [106,109] needed by 5G MM solutions. Note that, even though SDN-Cs might appear as SPoFs, corresponding clustering for load sharing and redundancy can help alleviate this issue. Specifically, and similar to the method proposed by 3GPP to pool the Mobility Management entities (MMEs) to avoid SPoF problem and to share the workload between MME instances, SDN-Cs can be clustered together to provision redundancy (and hence no SPoF) and workload sharing. Next, semi-centralized approaches divide the entire geographical region into smaller domains, each managed by a separate SDN-C. This SDN-C, responsible for handling MM in its domain, helps to enhance the network scalability.

However, since each domain still has a single SDN-C managing it, SPoF issue might become a limiting factor. Further, for inter-domain HO, extensive signaling would be required between two SDN-Cs whilst the same would be non-existent in a centralized approach [106]. On account of this trade-off, a semi-centralized approach can be successful if an appropriate number of SDN domains are created, which do not increase the signaling burden while reinforcing the network reliability and scalability characteristics [109]. A combination of the aforementioned approaches, i.e., hierarchical approach, consists of SDN-Cs at multiple levels [106]. Whilst the global SDN-C behaves as a master (tuning HO parameters, manage inter domain HOs, etc.), the SDN-Cs in the lower hierarchical levels manage MM within their domains and function as slaves.

Next, similar to the SDN based solutions, DMM based approaches can contribute significantly to the design and functioning of 5G networks. With the ability to provide a distributed DP in conjunction with a distributed/centralized CP [22–24, 53, 110], DMM can enhance the scalability (by removing anchors prevalent in current MM solutions, i.e., decentralization) and flexibility (by allowing the most optimum access router for each flow independently) of the 5G networks. These approaches can be classified as being fully distributed, partially distributed and SDN based.

The fully distributed approach whilst distributing both DP and CP, will encounter extensive amount of handover signaling between ARs during a mobility event [22, 24]. Note that the DP functionalities and location of ARs are the same as that of the UPFs. However, depending on the type of DMM approach, the CP is fully or partially located on the ARs themselves, instead of being located in a centralized controller. And so, while the fully distributed approach is challenged by the signaling between ARs, the partially distributed (P-DMM) approach centralizes the CP, hence, alleviating this concern [22, 24]. The P-DMM approach also maintains the benefit of avoiding a single mobility anchor. However, an enhancement of this approach is the SDN based approach. Similar to the P-DMM approach, the CP is still at a central controller, i.e., SDN-C, however, the signaling between the controller and the DP devices is far more simplified as compared to the partially distributed approach. The reason being, in an SDN based approach, the ARs are converted to mere forwarding devices and it is the SDN-C that orchestrates the forwarding rules (routing table) on them to realize the data paths for the existing sessions in the network. Concretely, in the SDN based approach the DP devices no longer need to perform a handshake, like in the P-DMM approach, with the central controller to establish a route, instead the routing information is now fed to the DP devices by the SDN controller [22, 110]. These enhancements are further quantified in [22] by the fact that the mean HO latency for SDN based DMM is reduced by 3.94% as compared to P-DMM, while the E2E delay is reduced by 39.55% for

the evaluated scenarios.

Subsequent to these discussions, and given that the current standardization in 5G [45,99] stipulates the functionality for mobility management to be split up between the AMF and the SMF NFs, it is noteworthy that the decoupling of the CP and DP and subsequent utilization of the aforesaid NFs via an SDN-C can provision the capability to implement fast and efficient MM solutions for 5G and beyond networks. However, since CN signaling during mobility events will still be a challenge, given the future network scenario, there remains a possibility for the SDN and DMM based 3GPP 5G MM solutions to be rendered sub-optimal.

Moreover, edge clouds, which essentially refer to data clouds/processing centers close to the RAN within a given network infrastructure, can have a profound impact on the user QoS during mobility scenarios (through fast access to data and compute resources) [111]. Henceforth, several studies such as [25,31,55,112–114] alongside 3GPP and ETSI [115], have studied the fundamental concepts of utilizing the edge clouds for fast data access (via data caching) as well as for processing capabilities (i.e., performing certain MM operations without the messages having to traverse the entire CN). Note that, we classify the edge clouds to be a CN solution, even though we state that they are most likely to be closer to the RAN, because, certain topology designs might entail a hierarchical setup. In this hierarchical setup, there will be some edge clouds that are placed close to the RAN and some of them being placed further away from the RAN, say close to the S-GW and Packet Gateway (P-GW) in an LTE network [111]. Such an approach can help in caching data according to their level of popularity, taking into account CN traffic as well as the latency to retrieve the requested content [111].

Lastly, given the SDN based mechanisms assist in MM through CP procedures, DMM based solutions assist through CP procedures as well as provision alternate and effective DP paths, and Edge clouds provision alternate and effective data paths, they have been classified as being CP, CP/DP and DP procedures, respectively, in Figure 2.13.

2.3.3.2 Access Network Solutions

As part of access network strategies, one of the key approaches that has been proposed, and similar to LTE dual connectivity, is the concept of phantom cell [116]. It allows the UE to camp its CP on a MC, while its DP is being handled at the SCs that lie within the coverage of the earlier mentioned MC. This, in essence offers a low signaling cost regime to perform the intra-MC HOs as the UE does not need to access the CN for radio resource management operations during HO. Concretely, the MC handles the radio resource allocation operations for the phantom cells, and hence, during HOs between the phantom cells the CN signaling

is avoided [62].

Moreover, owing to the softwarization of the complete network, the process of exchanging information between the various OSI layers, i.e., implementation of the cross layer strategy, is eased. This in turn allows the network to formulate solutions that are optimal, taking into cognizance the impact and benefits that the solution will produce at various levels of the network [117–119]. However, to realize cross-layer techniques, significant modifications to the software architecture of the protocol stack will be necessary [117–119]. Another consequence of the softwarization process is the RAN-as-a-Service (RANaaS), also known as Cloud-RAN (C-RAN), which allows on-demand allocation of access network resources (e.g., Baseband unit (BBU) pool, BBU- Remote Radiohead (RRH) functional splitting) depending on the network and user context [120–122]. Additionally, the BBU pool, through close interaction of various RATs at a single location, can orchestrate fast handovers on-demand [123].

However, in order to choose the best BSs to connect to in a multi-RAT scenario, computationally tractable RAT selection mechanisms need to be adopted. The multi-RAT solutions are a broad classification for the myriad RAT selection processes (Optimization based, Fuzzy logic and Genetic Algorithm based, RSSI based, etc. [44, 124–126]) that have been proposed. From our earlier discussions it is evident that RSSI based methods, although simple, do not weigh in other parameters such as network load, backhaul conditions, or user/network policies, for a RAT selection decision. This will most certainly result in sub-optimal solutions. But, optimized mechanisms, that can facilitate closed form solutions and are computationally tractable, will be able to capture more features from the network. Consequently, context aware mechanisms, such as [127] and our contribution [J3], will lead to optimal solutions that can be implemented for real-time scenarios.

It must be stated here that, the aforesaid HO decision may be executed either at the UE (user-centric) [127], at the network, or as a joint effort between the UE and the network (hybrid decision process). Moreover, given the intelligent RAT selection mechanism assists in MM through RAT selection (which is a CP task) and provision of effective and alternate DP paths, the phantom cell method provisions support for MM by handling the CP signaling for SC selection as well as provision alternate and effective DP paths via SCs, and RANaaS and Cross layer strategies assist through efficient resource allocation decisions (which is a CP task), thus they have been classified as being CP/DP, CP/DP, CP and CP procedures, respectively, in Figure 2.13.

2.3.3.3 Extreme Edge Network Solutions

Contrasting to the design and implementation of access and core network based methods, the extreme edge network based solutions consider the potential of utilizing D2D techniques for facilitating seamless HO. Multiple research efforts, such as [128–132], have provisioned methodologies to handle mobility of D2D pairs. Concretely, in [128] two types of handovers for D2D pairs have been provisioned. These are either *D2D aware* and *D2D triggered* handovers. They take into account the fact that the control of the D2D pair can be handed over independently of the actual cellular handover. And so, for the *D2D aware* handover, the D2D pair control (and if possible the cellular control) is handed over from the source eNB to the target eNB only after both the devices in the D2D pair satisfy the conditions to handover to the target eNB. On the other hand, the *D2D triggered* handover mechanism aims at clustering the devices of a D2D group in minimum number of cells. Hence, during mobility events the algorithm tries to determine the cell to which the majority of devices within the D2D group belong too.

Similarly, in [129] two handover management mechanisms have been proposed. While the joint handover strategy aims at migrating both the devices in a D2D pair simultaneously to the target eNB, the half handover stipulates that such a migration can be asynchronous. Furthermore, the D2D handover decision has also been specified in [129]. The Channel Quality Information (CQI) criteria has been utilized for the same. Next, in [130], a Markov chain based model has been proposed for D2D mobility.

Lastly, the work done in [131, 132] develops a model and simulation framework analyzing D2D mobility. Specifically, it considers a D2D pair with one of them being a transmitter (TX) and the other being just a receiver (RX). Thus, a handover procedure is defined for the scenario when the TX moves to the target eNB. In this procedure, the control of the D2D pair is transferred to the target eNB as soon as the TX migrates to it.

And so, given that the other research efforts only define D2D mechanisms that can assist in MM through CP assistance, they have been classified as being CP procedure in Figure 2.13.

To conclude, in Chapter 3 we perform the qualitative analysis for all of the current state of the art methods owing to their recent development in the realm of 5G network solutions.

2.4 Summary

In this chapter, we presented the state of the art mechanisms in mobility management. Specifically, we discussed the functional requirements and design criteria that would be

required with regards to the future MM solutions. We then detailed the various legacy mechanisms, wherein we explored the strategies developed/standardized by the various standardization bodies like 3GPP, IETF and IEEE. We also, studied strategies developed by the academic community for the same. Following this, we then explored the current state of the art in MM strategies, wherein we presented a detailed discussion with regards to the 3GPP 5G MM mechanisms as well as other research efforts in industry and academia. Concurrently, we also presented a novel classification of these studies based on them being either *Core Network based*, *Access Network Based* or *Extreme Edge Network Based*. Notably, our discussions in this chapter have also taken cognizance of the fact that there have already been some meaningful studies towards Beyond 5G networks and their enablers.

Next, we utilize the background developed in this chapter to present a novel Qualitative Gap Analysis for some of the well known/utilized MM strategies in Chapter 3.

Chapter 3

Qualitative Gap Analysis in Mobility Management

Overview

In this chapter, we firstly elaborate upon the three pillars of any future MM strategy, i.e., reliability, flexibility and scalability criteria. We establish a novel relationship between the requirements defined in Chapter 2 and the aforesaid criteria. We then present a novel discussion on the readiness of MM for 5G and B5G networks. We perform this through a novel qualitative gap analysis, wherein we evaluate the pros and cons of the legacy and current MM mechanisms and determine the extent to which they satisfy the aforesaid criteria. Note that, and as we will show in this chapter, a complete agreement towards these three criteria will be equivalent to satisfying the requirements enlisted in Table 2.1. Subsequently, we then determine the persistent challenges that exist towards the development of 5G and beyond MM strategies as well as the potential solutions that will assist in tackling these challenges. We then provision a future framework for the 5G and beyond MM mechanisms. Lastly, we highlight how our contributions, detailed in the Chapters 4-6, aim to realize this framework.

Contributions

[J2] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Are Mobility Management Solutions Ready for 5G and Beyond?", Accepted in Elsevier Computer Communications, pp. 1–36, 2020. (Quartile: Q2; IF: 2.816 (2019))

In Chapter 2, we established a detailed background on mobility management mechanisms and their corresponding utility/impact. Following this, it becomes prudent that we analyze

the gaps that still exist with regards to the MM strategies that will satisfy the 5G and beyond networks' requirements. And so, in this chapter we present a novel qualitative gap analysis, which is also published in part in [J2]. We evaluate certain predominant legacy mechanisms as well as the current state-of-the-art mechanisms on the basis of reliability, flexibility and scalability: the three pillars of any future MM strategy.

3.1 Qualitative Analysis Criteria

As part of this qualitative analysis, we firstly present a detailed description of the three criteria, i.e., reliability, flexibility and scalability, as follows:

- *Reliability* helps to determine whether the MM mechanisms employed will be able to ensure guaranteed and continuous service in any given network topology. Such reliability requirements entail not only continuous connectivity whilst traversing a geographic area, they also include reliability in delivery of packets for critical and delay sensitive services. Further, reliability from a MM mechanism also envelops factors such as tolerance to congestion (through for example, Distributed MM), ensuring faster yet trustworthy re-connection and authentication whilst mobile, ensuring appropriate levels of redundancy in the number of flows, connections, and hosts, and also ensuring appropriate resource allocation for users with myriad mobility and application profiles at the edge, access and core network.

- *Flexibility* as a qualitative analysis metric helps to determine the adaptability that MM mechanisms will provide to the network, which as discussed will be heterogeneous and dense in all perceivable aspects. The flexibility provisioned by MM mechanisms for future networks hence envelops factors such as the ability to formulate and deploy MM policies depending on individual user profiles, flow profiles or based on a slice profile. Further, ensuring the possibility of multi-connectivity through various layers, such as transport layer (SCTP/MPTCP), IP layer (Multi-homing), MAC-PHY layer (Dual Connectivity), will be an important factor for ensuring a flexible MM policy. Additionally, factors such as multi-objective base station selection/user association taking into account factors such as congestion, QoS requirements, backhaul reliability, etc., will be critical to a flexible MM mechanism.

- *Scalability* aspect allows one to determine if the future MM mechanisms can serve the increasing number of user devices with a corresponding increase in requested QoS with heterogeneous mobility profiles. A measure of scalability of MM mechanisms can be

gained by analyzing factors such as number of connections that can be managed given an increasing number of user devices, management of the signaling load generated due to mobility events, management of the increasing load due to processing the many CP messages generated in mobility events, as well as the ability to permit de-centralization (which in essence would ensure scalability) and being easily deployable on a large scale given a new MM mechanism.

We summarize the aforesaid criteria into a list of parameters for each criteria and present them in Table 3.1. Additionally, we also indicate the requirements (from Table 2.1) for whose fulfilment each of these parameters contribute towards. Note that, compliance with each of the stated parameters in Table 3.1 for the reliability, flexibility and scalability criteria will be essential towards ensuring that the MM mechanism under consideration satisfies the requirements defined for the upcoming 5G and beyond networks (Table 2.1). We now elaborate upon the parameter-requirement relationships that have been illustrated in Table 3.1, with the objective of enhancing the comprehensiveness of the evaluation criteria.

3.1.1 Reliability: Parameter to Requirement mapping

The provision of redundancy in the number of flows and connections, i.e., by satisfying parameter *RL1*, can help fulfil requirement *R7* presented in Table 2.1. This is so because, redundancy in connections will help overcome the fragile nature of wireless channels in the frequency bands that constitute VLC and mmWave communications. Next, satisfying the parameter *RL2* will contribute towards fulfilling the requirements *R1*, *R7*, and *R8* (Table 2.1). Here, the ability to provision seamless handover assists in supporting mobility amongst multiple RAT(s) (*R1*), supporting multi-connectivity and thus reliability (*R7*), and utilize enhanced localization capabilities to accomplish the same in dense urban scenarios (*R8*). Additionally, the *RL3* parameter for the reliability criteria, when satisfied, will help to fulfill the *R3* and *R4* requirements (Table 2.1). The reason being, decentralization will allow for efficient handling of the number of devices (*R3*). Moreover, to establish an effective level of decentralization, such as for accessing cached data at the edge and in the IMS core, enablers such as NFV and Mobile Edge Computing (MEC) will be utilized (*R4*). Furthermore, the *RL4* parameter holds significant relevance towards fulfilling the requirements *R5* and *R10* (Table 2.1). Specifically, fast path re-routing in the CN ensures that the increased dynamism, due to the mobility of both the UE and BSs (*R5*), is catered to in the CN. In addition, data path modifications due to service migration and service replications, which do not lead to extensive delays, is also ensured through parameter *RL4*. Lastly, satisfying the *RL5* parameter will help towards fulfilling the *R2* requirement (Table 2.1), since guaranteeing

Table 3.1: Governing Parameters for the Reliability, Scalability and Flexibility of a MM mechanism/standard

#	**Reliability**	**Contr. to Reqs.**	#	**Flexibility**	**Contr. to Reqs.**	#	**Scalability**	**Contr. to Reqs.**
RL1.	Redundancy in the number of flows, connections, etc.	R7	FL1.	Granularity of service. E.g. per flow, per connection, per user, etc.	R9, R11	SL1.	Manageable number of connections with increasing number of users	R3, R9
RL2.	Seamless handover capability[†]	R1, R7, R8	FL2.	Capability to enable connectivity to multiple BSs	R1, R9	SL2.	Manageable signaling load with increasing number of users	R3, R9
RL3.	Decentralization	R3, R4	FL3.	Handover service support at multiple network levels. E.g. Core network, Access network, etc.	R4, R9	SL3.	Manageable processing load with increasing number of users/devices	R3, R9
RL4.	Fast path re-routing at CN	R5, R10	FL4.	Handover decision making utilizing multiple parameters. E.g. network load, requested QoS, etc.	R1, R9	SL4.	Decentralization	R4
RL5.	Congestion aware	R2	FL5.	Context awareness	R2, R9, R10	SL5.	Ease of implementation and integration	R6

[†]Seamless handover capability refers to the ability of a MM mechanism to permit vertical (inter-RAT) as well as horizontal (intra-RAT) handover.

congestion awareness helps service the different QoS requirements of the applications, such as virtual reality and emergency services, with better reliability.

3.1.2 Flexibility: Parameter to Requirement mapping

When a MM mechanism under study satisfies the flexibility parameter *FL1*, it correspondingly helps to fulfil the *R9* and *R11* requirements (Table 2.1). This is so because, *FL1* states that a MM mechanism should support granularity of service. This will correspondingly assist in accommodating the multitude of service requirements independently (*R9*) as well as avoid the *one size fits all* approach (*R11*). Next, *FL2* parameter will help in satisfying the *R1* and *R9* requirements (Table 2.1). Essentially, the capability to be able to connect with multiple BSs will assist in multi-RAT MM (*R1*) as well as in provisioning enhanced agility for MM mechanisms in a dense and heterogeneous network (*R9*). Further, when the *FL3* parameter is satisfied, it helps to fulfil the *R4* and *R9* requirements. The reason being, to enable handover support at multiple levels of the network, usage of SDN, NFV and MEC platform will be necessitated for efficient implementation (*R4*). Moreover, such multi-level handover support will also provision flexibility for the network (*R9*). Additionally, satisfying parameter *FL4* enables the MM mechanism under study to contribute towards satisfying the *R1* and *R9* requirements (Table 2.1). Specifically, having a handover decision mechanism that utilizes multiple parameters aids in handling MM amongst multiple RAT(s) more flexibly and hence, efficiently (*R1*). Also, such strategies will ensure that alongside being flexible, solutions are computationally tractable and energy efficient (*R9*). Finally, parameter *FL5*, when satisfied, will be relevant for the fulfilment of requirements *R2*, *R9* and *R10* (Table 2.1). To elaborate, the context awareness feature of a MM mechanism will assist in provisioning MM support dependent on application, user and network context (*R2*), flexibility to handle the increased heterogeneity in the network (*R9*), and ensure QoS whilst performing complex tasks such as migrating or relocating services based on user mobility events (*R10*) through appropriate path and resource management.

3.1.3 Scalability: Parameter to Requirement mapping

For the scalability criteria, when parameter *SL1*, *SL2* and *SL3* are satisfied by a MM mechanism, they correspondingly also assist in fulfilling the *R3* and *R9* requirements (Table 2.1). Concretely, the ability to be able to manage increasing number of connections, signaling load and processing load with the number of increasing users will correspondingly assist in handling a user density of more than 10^6 devices per km^2 in 5G and beyond networks (*R3*). Also, they will help in ensuring the required scalability to accommodate the increasing

heterogeneity in the network as well as the corresponding tractability of the MM solution (*R9*). Next, when parameter *SL4* for the scalability criterion is met, it helps to fulfil the *R4* requirement (Table 2.1). Specifically, to accomplish decentralization objective the MM mechanism under study will need to utilize enablers such as NFV and MEC. Lastly, satisfying parameter *SL5* will help to meet the requirement *R6* (Table 2.1). The reason being that, ease of implementation usually arises from the fact that a MM mechanism has been used/deployed before, as well as is suitable to accommodate legacy devices whilst catering to a new set of service and devices. Hence, satisfying the *SL5* parameter will assist in ensuring that backwards compatibility requirements (*R6*) are adhered to.

And so, from the aforementioned elaborate understanding of the mapping, it can be deduced that the criteria chosen for our qualitative analysis are comprehensive in nature and approach. Moreover, and considering only the 5G networks since their KPIs have been defined [48], provisioning beyond 99.999% reliability will be ensured through the reliability metric during mobility scenarios. Further, latency less than 5 ms for connected cars and 10 ms for virtual reality and broadband applications, will be guaranteed through the reliability and flexibility metric. Specifically, the reliability metric will help provision congestion awareness, reliable link selection, etc., while flexibility will allow multiple type and number of connections during mobility scenarios. In addition, support for nearly 1 million devices per km^2 with different application and mobility profiles will be ensured through the scalability criterion. Consequently, this further reinforces the comprehensiveness of the criteria chosen for the qualitative analysis that follows.

Before we proceed, we highlight certain specifics with regards to the analysis that follows:

- The goal of the following analysis is not to compare the considered standards and mechanisms against each other but rather to highlight the extent of their suitability for 5G and beyond networks.
- The mechanisms chosen for the qualitative gap analysis are based on their wide-ranging acceptance/applicability in the wireless networks domain.

3.2 Legacy Mechanisms

Utilizing the discussions in Chapter 2 with regards to the legacy mechanisms, i.e., Section 2.2, as well as the MM requirements and evaluation criteria specifics in Section 3.1, we now perform the qualitative analysis for the legacy mechanisms in the text that follows. As part of the analysis, for each of the studied mechanisms we firstly highlight their pros and cons

towards 5G and beyond MM mechanisms. Subsequently, we translate the insights gained from these pros and cons into a summary of parameters satisfied for the reliability, scalability and flexibility criteria.

3.2.1 IETF MPTCP-SCTP

Given our objective of determining the suitability of MPTCP and SCTP for 5G and beyond MM mechanisms, we firstly enlist their *pros* and *cons* as follows:

- MPTCP Pros
  - Allows for multiple data flows at the transport layer level [78, 79, 84], and hence, provisions for resiliency against connection failures, given the multipath feature [82–84]
  - Provisions congestion awareness, with studies such as [86] proposing specific congestion control methods for MPTCP
  - Through its ability to divide a connection into multiple sub-flows, MPTCP provisions the capability to handle each flow independently [84, 88]
- MPTCP Cons
  - The middleboxes installed by service providers are not optimized to support MPTCP [78, 79]
  - MPTCP requires proxies to allow MPTCP enabled devices to take its full benefits [89]
- SCTP Pros
  - Allows for multiple data flows at the transport layer level [85, 90], and hence, provisions for resiliency against connection failures, given the multipath feature
  - Provisions congestion awareness, wherein reference [90] establishes the presence of congestion avoidance methods within the SCTP suite
  - Assists in network level fault tolerance through support for multi-homing [85, 90]
- SCTP Cons
  - Requires both host and destination device protocols stacks to be updated with the SCTP protocol [90]

From the *pros* and *cons* of both MPTCP and SCTP, as listed above, it can be concretely stated that the IETF MPTCP-SCTP methods satisfy parameters *RL1* (allowing for multiple flows over the network for any given user) and *RL5* (provisioning congestion awareness as part of the transport layer characteristic for MM) for the reliability criterion. Further, for flexibility, from our discussion above, it is clear that IETF MPTCP-SCTP only satisfies parameter *FL1* (by allowing for multiple flows, flow level granularity can be induced).

3.2.2 IEEE 802.21

For the purpose of analysis, we list the *pros* and *cons* of the IEEE 802.21 mechanism towards 5G and beyond MM strategies, as follows:

- IEEE 802.21 Pros
  - Provisions seamless handover capability, as it allows users to switch between multiple RATs [43, 91, 94]
  - Provisions the possibility for a UE to connect to multiple BSs [43, 92]
- IEEE 802.21 Cons
  - Requires the protocol stacks of both the host and destination devices to be modified, so as to enable the IEEE 802.21 functionality [91, 93]

And so, given the aforesaid *pros* and *cons* with regards to IEEE 802.21, it can be deduced that it satisfies parameter *RL2* for reliability (allowing for seamless movement between different RATs) and *FL2* for flexibility (allowing for the possibility to connect with multiple RATs) criteria.

3.2.3 IETF PMIPv6

Based on the discussions carried out in Section 2.2.4.4, we now enlist the *pros* and *cons* of the PMIPv6 strategy with regards to its utility for 5G and beyond MM mechanisms, as follows:

- PMIPv6 Pros
  - Given that PMIPv6 is adopted by 3GPP and it forms a relatively agnostic setup for an UE towards its mobility signaling, it can thus provision seamless mobility [73–75]

  - Through the DMM based PMIPv6 approach, decentralization can be introduced [77]. Furthermore, other approaches, such as the clustering based approach in [76], can grant enhanced scalability and reliability to the PMIPv6 approach
  - Given that it has already been adopted by 3GPP for LTE, the available implementational expertise will enhance the ease with which it can be adopted in future networks

- PMIPv6 Cons
  - In its original flavor, PMIPv6 suffers from scalability and reliability issues due to the SPoF formed by the LMA in its architecture [76]
  - An explicit treatment of PMIPv6 with regards to the parameters for flexibility criterion is missing in [73–77]

And so, it can be deduced that the IETF PMIPv6 in its original flavor, given its maturity in development and deployment, satisfies the seamless handover parameter *RL2* in the reliability criteria. Moreover, with enhancements from the use of DMM and cluster based methods, PMIPv6 can be decentralized and scaled thus satisfying parameters *RL3* and *SL4* in reliability and scalability, respectively. Furthermore, since it has already been explored and implemented in the LTE networks, it satisfies parameter *SL5* owing to its relative ease of implementation as against any other new protocol.

3.2.4 3GPP LTE MM Mechanisms

For the 3GPP based MM mechanisms, we firstly highlight the *pros* and *cons* for the handover, traffic offloading and DC and LWA strategies, as follows:

- LTE Handover Pros
  - The LTE-X2 and S1 mechanisms together offer handover support at the access and core network level [133]
  - Through LTE-X2 handover mechanism, CN signaling can be avoided [133]
  - LTE-X2 permits decision making for a handover to be taken at the access network level. Hence, it reduces the processing load on the CN entities as well and also permits fast handover capabilities [133, 134]

- LTE Handover Cons

  - The S1 based handover mechanism involves signaling through the CN, which creates increased load on the CN [134] as well as introduces SPoFs

- LTE Traffic Offloading Pros
  - Provisions a method for managing the traffic load given that the number of users/devices will increase significantly [66]
  - Provisions a method for managing the processing load in the network nodes [66]

- LTE Traffic Offloading Cons
  - LIPA does not support session continuity during mobility events, as well as it requires an additional gateway [66]
  - SIPTO is not helpful in mitigating radio congestion [66]
  - IFOM is significantly harder to implement as it necessitates coordination with the non-3GPP networks [66]

- LTE DC and LWA Pros
  - Provisions the ability to connect to multiple 3GPP as well as Non-3GPP RATs [15, 62, 63, 135]
  - Provisions the capability to have multiple physical paths for data transmission, and thus better fault tolerance [15, 62, 63, 135]

- LTE DC and LWA Cons
  - 3GPP LWA is only applicable for downlink

From the *pros* and *cons* for the LTE MM mechanism, it is clear that they provision redundancy in data paths (through DC and LWA), decentralization (through X2 and traffic offloading) and seamless handover (through X2 and S1 handover), thus satisfying *RL1*, *RL2* and *RL3* parameters for the reliability criterion. Further, for the flexibility criterion, LTE MM mechanisms offer the possibility of a multi-level HO support (through X2 and S1 handover) as well as the ability to connect to multiple BSs/RATs at the same time (through DC and LWA), thus satisfying parameters *FL2* and *FL3* for flexibility. Lastly, LTE MM mechanisms offer enhanced support with regards to the scalability criterion for 5G and beyond MM, as they satisfy parameters *SL2* to *SL5*, given their decentralization, ease of integration, multi-level handover mechanisms (X2 and S1 handover), and traffic offloading characteristics.

3.2.5 Non-3GPP Multi-Connectivity Solutions

For the non-3GPP multi-connectivity approaches we evaluate ITU-VMH and CoMP. We present their *pros* and *cons* as follows:

- ITU-VMH Pros
  - Provisions path redundancy through multi-homing [67]
  - Provisions the capability to connect to multiple RAT(s) at any given time [67]
  - Per-channel granularity of service is possible
- ITU-VMH Cons
  - It will require the transformation of the entire protocol stack [67]
- CoMP Pros
  - Provisions path redundancy through its ability to coordinate data transmission from multiple APs, which may also belong to different RATs [68, 136]
  - Provisions the capability to connect to multiple RAT(s) at any given time [68,136]
  - Through the use of multiple BSs for transmission, per-channel granularity of service is made possible
- CoMP Cons
  - Centralized processing introduces the possibility of SPoF [68, 137]
  - Backhaul networks will need to have extremely high capacity and extremely low latency characteristics, so as to support CoMP whilst maintaining QoS [137]

Concretely, ITU-VMH and CoMP satisfy parameters *RL1* (allowing for the possibility of redundant physical connections) and *RL2* (allowing for seamless mobility) for the reliability criterion, and parameters *FL1* (provisioning the possibility of per-channel granularity for MM) and *FL2* (allowing for the possibility of connecting to multiple RATs/BSs) for the flexibility criterion.

3.2.6 RSS based BS selection methods

Based on the discussions in Section 2.2.6, we present here the *pros* and *cons* of the RSS based BS selection methods as follows:

Table 3.2: Compliance with the Reliability, Scalability and Flexibility criteria for the legacy MM mechanism/standard

		Mechanisms											
		IETF MPTCP-SCTP		**IEEE 802.21**		**IETF PMIPv6**		**LTE MM mechanisms**		**Non-3GPP Multi-connectivity solutions**		**RSS based handover methods**	
		Cnf.†	Refs.δ	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.	Cnf.	Refs.
Reliability	**RL1**	✓	[81–87]	×	[43, 91]	×	[73, 74]	✓	[133] [135] [15, 62, 63]	✓	[67, 68] [136]	×	[75, 96] [97]
	RL2	×		✓		✓		✓		✓		✓	
	RL3	×		×		✓		✓		×		×	
	RL4	×		×	[94]	×	[77]	×		×		×	
	RL5	✓		×		×		×		×		×	
Flexibility	**FL1**	✓	[82, 88] [84, 85]	×	[43, 92] [94]	×	[73–77]	×	[133] [135] [15, 62, 63]	✓	[67, 68] [136]	×	[75, 96] [97, 138]
	FL2	×		✓		×		✓		✓		×	
	FL3	×		×		×		✓		×		×	
	FL4	×		×		×		×		×		×	
	FL5	×		×		×		×		×		×	
Scalability	**SL1**	×	[78, 79] [89]	×	[91, 93]	×	[77] [73, 74]	×	[66] [133]	×	[68] [137]	×	[75, 96] [97, 139]
	SL2	×		×		×		✓		×		✓	
	SL3	×		×		×		✓		×		✓	
	SL4	×		×		✓		✓		×		×	
	SL5	×		×		✓		✓		×		✓	

†The conformance (Cnf.) of a given mechanism for a given criterion.

δThe corroborating references (Refs.), if any, for the specified conformance of a mechanism for a given criterion

- RSS based methods Pros
  - Easy to implement, given that it has already been adopted by 3GPP [75, 96, 139]
  - Relatively low processing and signaling load, owing to its simplicity [139]
- RSS based methods Cons
  - FHOs in ultra dense scenarios is a pertinent issue [138]
  - It is agnostic of other parameters related to the UE and the network, such as the load, UE context, etc., thus making it unreliable and one-dimensional [75, 96, 138]

Given these pros and cons, the erstwhile RSSI based method due to its existence and maturity can ensure mobility between multiple RAT(s), hence, satisfying parameter *RL2* for reliability criteria. Furthermore, owing to the aforementioned simplicity and maturity in development and deployment it also satisfies parameters *SL2*, *SL3* and *SL5* for the scalability criterion.

To summarize, we introduce Table 3.2 wherein we indicate the parameters that each of the explored methods satisfies for the reliability, scalability and flexibility criteria. We also enlist the important references that have lead us to the development of Table 3.2, as presented in this chapter. From the discussions, analysis and Table 3.2, it can be deduced that none of the legacy mechanisms, that have been studied, achieve the requirements as necessitated by 5G and B5G networks. Concretely, none of the studied mechanisms satisfy all the parameters of the criteria utilized for the qualitative analysis. Notably, the 3GPP based LTE MM mechanisms provision the best basis and support for 5G and beyond MM mechanism, given that they collectively satisfy the most parameters amongst other strategies explored.

3.3 Current State-of-the-Art

Based on the discussions in Chapter 2 with regards to the current state-of-the-art mechanisms (Section 2.3), in this section we provision the qualitative analysis for these mechanisms. We develop, firstly, the pros and cons and then subsequently specify the parameters that the mechanism under study satisfies for the reliability, scalability and flexibility criteria, as listed in Table 3.1.

3.3.1 3GPP 5G MM Solutions

Given the extensive overview with regards to the MM solutions that have been provisioned by the 5G standards [45, 99, 100] in Section 2.3.2, we now, as part of our qualitative analysis, present the *pros* and *cons* for the same, as follows:

- 3GPP 5G MM Pros
  - Provisions monitoring of UE mobility, mobility event notifications and resource negotiation mechanisms at destination networks [45]
  - Employs flexible session management strategies, wherein provision of per-PDU session granularity, through path selection, roaming support and traffic steering, has been detailed [45]
  - Support for IPv6 multi-homing [45]
  - Provision for multiple sessions and service continuity modes [45]
  - Support for Multi-RAT DC [45]
  - Support for Edge Computing [45]
  - Network slicing information migration support in the event of inter-/intra- RAT mobility [45]
  - Network slicing support for provisioning on-demand MM
  - Ability to provision context awareness via network slicing
  - Provision for managing core network load by introducing load balancing and re-balancing principles on the AMF [45]
  - Provision of congestion awareness on the CP handling MM messages, i.e., NAS [45]
  - Introduction of beam level MM support [100]
  - Intra-RAT (5GC to 5GC) and Inter-RAT (5GC to EPC and vice versa) HO support [45, 99]
  - Well defined EPC and 5GC inter-working interface, i.e., N26 [45, 99]
  - Mobility support at the D2D level [15, 102]
- 3GPP 5G MM Cons
  - From our contribution [J1], it can be deduced that handover signaling in the CN is extremely sub-optimal

- RAT selection still relies on received signal quality fundamentals only [100]
- A unified framework for cross-layer mechanisms, such as MPTCP-SCTP (transport layer), IPv6 multi-homing (network layer) and MR-DC (Physical and MAC layer) working together, has not been provisioned
- In IPv6 multi-homing, a single point of failure (SPoF) still exists, as the multiple PDU session anchors are still connected to a single UPF from where the paths branch out [45]
- Co-ordination between D2D peers for enacting an efficient MM strategy is not explored explicitly in the standards

From the *pros* and *cons*, it can be deduced that the 3GPP 5G MM mechanisms will be able to support reliability parameters *RL1* (owing to the support for MR-DC and IPv6 multi-homing, and hence, redundancy in the number of connections and flows), *RL2* (owing to the support for MR-DC and handover procedures defined, thus ensuring seamless handover capability), *RL3* (owing to managing mobility at the access, core and extreme edge network levels as well as local breakouts, thus introducing decentralization) and *RL5* (owing to the congestion awareness feature in NAS). Next, for the flexibility parameters, 3GPP 5G MM mechanisms satisfy *FL1* (owing to the granularity of service support per PDU session as well as per mobility level, and the ability to support on-demand MM through network slicing support), *FL2* (owing to the ability to connect to multiple BSs through MR-DC and IPv6 multi-homing support), *FL3* (owing to the handover support at the access, core and extreme edge network levels via the Xn handover, N2 handover and 3GPP ProSe, respectively) and *FL5* (owing to the ability to take into account the context of the tenant via network slicing). Lastly, 3GPP 5G MM mechanisms, for the scalability criterion, satisfy parameters *SL1* (owing to the AMF load balancing strategies, local breakout strategies, multi-level handover support as well as the granularity in service per mobility levels), *SL4* (owing to local breakout and support for edge computing, thus leading to decentralization) and *SL5* (since these are standards, implementation and integration is not a bottleneck).

Note that, scalability parameters *SL2* and *SL3* are not supported owing to the sub-optimality in CN handover signaling as well as the presence of SPoFs, as stated in the *cons* for the 3GPP 5G MM mechanisms.

3.3.2 Other Research Efforts: Core, Access and Extreme Edge Network Solutions

3.3.2.1 Core Network Solutions

For analyzing the core network solutions we utilize the generic classifications, i.e., SDN based, DMM based and Edge Cloud solutions, and firstly list their *pros* and *cons*.

- SDN based mechanism Pros
  - Provisions global view of the network [106, 108]
  - Provisions hierarchical solutions, thus enabling decentralization [106]
  - Provisions the ability to manage CN signaling, and hence, DP paths during mobility events [106–108]
  - Provisions a single point of collection for network statistics thus enabling the design and development of context based MM mechanisms [109]
- SDN based mechanism Cons
  - Extensive CN signaling for managing handovers in a centralized/semi-centralized approach [106]
  - It does not alleviate the issue of mobility anchors which can lead to SPoFs in the DP
- DMM based mechanism Pros
  - Provisions decentralization of the mobility management anchors [22–24, 110]
  - Assists the CN in implementing efficient data paths for UEs undergoing mobility [22, 24, 53]
- DMM based mechanism Cons
  - Fully decentralized solution introduces extensive CN signaling in order to manage the changes in data paths and mobility anchors, and hence, handovers [22]
  - Partially distributed solution, while solving the extensive CN signaling, introduces a central controller, and hence, an SPoF [22]
  - Co-existence and integration with already deployed networks and devices will be a significant challenge [53]

- Edge clouds Pros
  - Ensure data offloading opportunities, and hence, reduction in CN traffic load [25, 115]
  - Facilitate processing of MM related tasks without the messages having to traverse the CN [113]
  - Provisions context awareness, as delay sensitive applications can access edge clouds whilst delay tolerant applications can still access services located in the core network [113, 114]
- Edge clouds Cons
  - Require dedicated infrastructure and appropriate placement [25, 31, 115]
  - Require fast service migration strategies to ensure seamless mobility [55]

From these *pros* and *cons* as well as the preceding discussions, it is evident that the SDN based solutions satisfies parameter *RL2* (allowing for seamless mobility), *RL3* (through the provision of decentralized solutions), *RL4* (through the ability to re-program paths in CN via orchestration of OF rules) and *RL5* (through the ability to utilize network statistics for traffic steering with the CN) for the reliability criterion. For the flexibility criteria, the SDN based mechanisms satisfy the parameters *FL1* (through the capability of orchestrating policies dependent on flow type, slice, etc.), *FL3* (by allowing for CN based MM solutions that will work in synergy with the access network based solutions) and *FL4* (through the global view of the network wherein a variety of parameters such as network load, QoS requirements, etc., are considered). In terms of scalability, SDN based solutions satisfy parameters *SL1* to *SL3* (given the ability to manage and steer traffic flows with the ability of having a distributed, hierarchical or centralized implementation) and *SL4* (due to the possibility of having a decentralized configuration).

The DMM based solutions, however only satisfy parameters *RL2* (allowing for seamless handovers) and *RL3* (due to the decentralized nature) in the reliability criterion. Further, for the flexibility criterion, DMM based solutions only satisfy parameter *FL1*, i.e., they only offer granularity of service by preventing any mobility anchor. It is noteworthy though that, from the scalability aspect DMM based solutions, like SDN based solutions, satisfy parameters *SL1* to *SL4*, and for the same reasons.

Lastly, for the edge cloud based solutions, parameters *RL2* (allowing for seamless mobility through fast access to data/processing capabilities upon migration to the target network) and *RL3* (allowing decentralization of MM based services) are satisfied for the reliability criterion.

For the flexibility criteria, parameters *FL1* (due to the ability to provision services based on mobility and application profiles), *FL3* (by allowing for MM methods at the edge network level in addition to the access and core network based solutions), *FL4* (by provisioning processing capabilities for user association/BS selection services) and *FL5* (by allowing for context awareness in data caching according to user mobility) are satisfied. Additionally, for the scalability criteria, parameters *SL1* to *SL4* are satisfied by the edge cloud solutions. The reason being, they allow for decentralization which can consequently permit better capability to manage connections and control messages due to increasing number of users.

3.3.2.2 Access Network Solutions

As part of the analysis for the access network solutions, we firstly present the pros and cons for each mechanism discussed in Section 2.3.3.2, as follows:

- Phantom Cell method Pros
  - Grants the ability to a UE to connect to multiple BSs simultaneously, thus also granting redundancy in physical layer connections [116]
  - As per reference [116] and our contribution [C1], it provisions the ability to allow per-flow and per-user granularity of service
  - Handover support at access network level [116]
  - Ease of implementation due to existing standards on MR-DC [45, 116]
- Phantom Cell method Cons
  - Handovers between different MC domains will still entail service disruption [45, 116]
  - According to [116] and our contribution [J1], Inter-MC domain handover signaling will still be a significant burden on the CN
- RANaaS Pros
  - Provisions on-demand allocation of network resources at the RAN level [120–122]
  - Provisions the ability to execute on-demand handovers, through close interaction between the various RATs that are integrated at a BBU pool [123]
  - Assists in allowing UEs to camp on more than one BS
  - Introduces support for executing handovers at the access network level [123]

  - Introduces the ability to utilize per-flow/channel granularity of service by being able to manage the physical connections more centrally [120–123]

- RANaaS Cons
  - Requires a complete architectural overhaul at the RAN side of the network [120–122]

- Cross layer Pros
  - Allows for the sharing of network statistics between the various OSI layers [117–119]
  - Allows for interaction between multiple OSI layers, thus facilitating the possibility of efficient utilization of multi-homing [67, 117–119]

- Cross layer Cons
  - Requires significant software modifications to the existing modular nature of the protocol structure [117–119]

- Intelligent RAT selection Pros
  - Optimized RAT selection strategies [44, 124–127]
  - Utilization of multiple parameters, such as BS load, UE context, etc., jointly for RAT selection [44, 124–127]
  - Provisioning the ability to select RATs per-slice/user/flow [127]
  - As per our contribution [J3], it provisions the ability to select multiple BSs (possibly belonging to multiple RATs)

- Intelligent RAT selection Cons
  - Requires rapid collection of network statistics to perform well informed selection
  - Based on our contribution [J3], computational complexity and convergence time of RAT selection algorithms will be critical, given the QoS requirements in 5G

Given the discussions in Section 2.3.3.2 and the *pros* and *cons* listed above, we now determine the parameters, listed in Table 3.1, satisfied by each of the mechanisms explored. Concretely, for the phantom cell method, parameters *RL1* (redundancy in physical layer connections) and *RL2* (seamless mobility) are satisfied for the reliability criterion. For the

flexibility criterion, parameters *FL1* (by permitting the possibility of per-flow and per-user based MM), *FL2* (allowing for connectivity to multiple BSs potentially belonging to different RATs) and *FL3* (provisioning handover support at the access network level that will work in synergy with CN based mechanism) are satisfied. In terms of scalability, the phantom cell method satisfies parameters *SL1* to *SL3* (owing to the handling of handover related computation and decision at the access network) and *SL5* (owing to the existing standards on MR-DC, as discussed in Section 2.3.2).

Next, the RAN-as-a-service concept satisfies parameters *RL2* (allowing for seamless handovers) and *RL5* (the softwarized nature enables dynamic initiation for RAN functionality such as BBU resources, functional splits, etc., depending on the network and user context) for reliability, parameters *FL1* (allowing for per-flow, per-user, per-slice, etc., service granularity through its softwarized nature), *FL2* (allowing the possibility for connecting a user to multiple BSs through its softwarized nature), *FL3* (provisioning handover support at the access network which will work in synergy with the CN and edge network based methods) and *FL4* (enabling the possibility of collection and utilization of RAN based information and generating intelligent BS selection/user association decisions) for flexibility, and parameters *SL1* to *SL3* (by offloading handover decision making and signaling to the access network) for scalability.

On the other hand, cross-layer methods only satisfy parameters *RL2* (allowing for seamless handover) and *RL5* (allows for congestion aware method by sharing statistics about queue lengths, buffer sizes, etc., amongst the various layers) for the reliability criteria. Further, for the flexibility criteria they satisfy only parameters *FL2* (by allowing for the possibility of multi-homing, etc.) and *FL4* (allowing for the possibility of sharing statistics and other information amongst the various OSI layers and enabling joint optimization for BS selection, path re-routing, etc.).

Lastly, for the intelligent RAT selection methods parameter *RL2* (allowing for seamless handover through optimized decisions on RAT selection) is satisfied for the reliability criterion. For the flexibility criterion, parameters *FL1* (allowing for the possibility of flow/user/slice based RAT selection), *FL2* (allowing for the possibility to select multiple RATs for a given user) and *FL5* (via the ability to utilize user and network context for RAT selection) are satisfied, while for scalability only parameter *SL5* (owing to the extensive body of research for optimal RAT selection strategies) is satisfied.

3.3.2.3 Extreme Edge Network Solutions

We firstly present the *pros* and *cons* for the D2D strategies as follows:

- D2D strategy Pros
  - Provisions D2D handover management strategies [128, 129, 132]
  - Provisions MM support at the extreme edge network level [128–132]
  - Provisions the ability to decentralize MM functionality
- D2D Strategy Cons
  - Control signaling overhead will be a challenge [128, 129]
  - The viability with regards to energy efficiency of D2D peers as well as latency incurred in conveying the decisions with regards to MM are un-explored questions

Based on the discussions and the aforesaid *pros* and *cons*, the device-to-device methods satisfy parameter *RL2* (through the provision of various seamless handover management studies) for reliability, parameter *FL3* (provisioning mobility support at the edge network level which will work in synergy with access and core network based methods) for flexibility, and parameter *SL4* (allowing for the decentralization of MM functionality) for scalability.

3.3.3 B5G Networks

In this subsection we present a short study detailing the challenges that current state-of-the-art mechanisms will continue to face for B5G networks. Furthermore, given the special characteristics that B5G networks will pose, as shown in Figure 2.1, we also list potential research areas for MM in B5G networks. Note that, these are then utilized in the subsequent section wherein we define challenges and potential solutions for 5G and beyond MM.

Concretely, while *SDN* and *NFV* will provide the tools for the B5G networks to provision rapid programmability of the meta-surfaces, during mobility scenarios they will be challenged critically. The reason being that, while current networking paradigms permit anywhere between 1 ms–10 ms time interval for performing any programmability task (latency restrictions, as specified in current 5G networks [38], on most services), in B5G networks this will be constrained even further as additional surfaces need to be programmed and orchestrated. Specifically, an increased number of surfaces/network nodes leads to more data to be processed for generating appropriate programmability decisions. These decisions then need to be sent out (orchestrated) to the relatively large number of network nodes (including meta-surfaces), to execute the given task. Hence, this leads to an increased latency constraint on the network programmability aspect. Further, while the meta-surfaces provide a higher degree of freedom to the operator, they need to be programmed, as mentioned above. This

introduces the challenging aspect of managing the SDN domains, NFV orchestration and the related signaling. As a consequence, the compactness as well as the efficiency of the current state-of-the-art SDN and NFV procedures will be challenged.

Next, with techniques such as DC, the challenge will be multi-fold as B5G networks will not just comprise of meta-surfaces, which can also act as a MIMO array, but they will also be equipped with Terahertz and mobile BS based multi-tier networks. And while, DC and multi-RAT procedures, as stated in Sections 3.3.1 and 3.3.2, will aid in ensuring a context-aware network selection procedure, the complexity for the access network techniques will be compounded by the fact that not only will they need to ensure QoS requirements, but they will have to also ensure sufficient available access bandwidth as well as backhaul bandwidth. Note that with the backhaul bandwidth there will be a significant design challenge since VLC technology is capable of carrying data rates of up to 1 Tbps. Current backhaul technologies cannot provision such high bandwidths [52]. Further, it is important to reiterate that the network will be composed of not only 4G-LTE and mmWave BSs, but there will also be VLC and drone based BSs, which essentially are the main reason for the increased complexity as discussed above.

Moreover, for the edge clouds, while they aid in allowing low latency access to cached content as well as the compute resources, the deployment strategies will need to be rethought given the ongoing growth pattern for data usage as well as the number of served devices coupled with more resource hungry services. Certain important recent studies in this direction have been provisioned via references [140, 141].

Given these significant shortcomings in the current state-of-the-art mechanisms towards B5G networks as well as taking into account the seminal works in the area of B5G techniques [32, 33, 35, 36] [39], the potential areas of research in MM for these networks are as follows:

- Characterization of the channel between meta-surface and the users, and meta-surface and the BS, in the event of user/BS being mobile, for the purpose of MM decisions
- Consideration of reliability and coverage of VLC link for MM decisions
- Characterization of the computational complexity for re-calibrating the meta-surfaces alongside the network, during mobility events
- Impact of mobility upon the programmable environment[1] concept, drone based communication and VLC
- Optimal RAT and BS selection with a programmable environment

[1]By environment, we refer to the physical environment that lies between the transmitter and receiver.

- Optimal RAT and BS selection in scenarios where both the UE and BS (drone based) are mobile
- Characterizing the computational complexity of optimization methodologies for user association
- Methods to handle possible increase in handover signaling/messaging during other network processes, such as reprogramming meta-surfaces to serve mobile users
- Formulation of a sound heterogeneous RAT strategy, just like the 4G-5G concept, given mmWave and Terahertz technologies and their associated challenges related to coverage.

Note that, the aforementioned research areas do not form an exhaustive list, but are broadly indicative of what aspects remain to be explored with regards to MM in B5G networks.

To summarize, in this section we firstly introduced the 5G service based architecture and the classification of the various mechanisms that we analyzed, through Figure 2.13. Following this, we qualitatively analyzed the 3GPP 5G MM mechanisms as well as other research efforts with regards to their efficacy towards 5G and beyond MM solutions. Consequently, we introduce Table 3.3 wherein we indicate the parameters that each of the explored methods satisfies for the reliability, scalability and flexibility criteria (Table 3.1). We also enlist the important references that have lead us to the development of Table 3.3, as presented in this chapter. And so, from the capability profiles of each mechanism, as illustrated in Table 3.3, it is evident that even after significant efforts none of them completely meet the specified requirements as expected for the 5G and beyond MM mechanisms. Concretely, neither the 3GPP 5G MM mechanisms nor the other academic and industrial research efforts satisfy all the criteria completely. Subsequently, it is deduced that none of the analyzed mechanisms satisfy the requirements for the future MM mechanisms, as listed in Table 2.1. Hence, through the aforesaid qualitative analysis we have further exposed the gaps in the design and development for 5G and beyond MM mechanisms.

Table 3.3: Compliance with Reliability, Scalability and Flexibility criteria of Current state-of-the-art MM mechanism/standard

		3GPP 5G MM		**SDN based**		**DMM based**		**Edge Clouds**		**Phantom Cell**		**RANaaS**		**Cross layer**		**Intel. RAT sel.**		**D2D**	
		Cf.*	Refs.$^{\delta}$	Cf.	Refs.	Cf.	Refs.	Cf.	Refs.	Cf.	Refs.	Cf.	Refs.	Cf.	Refs.	Cf.	Refs.	Cf.	Refs.
Reliability	**RL1**	✓		×		×		×		✓		×		×		×		×	
	RL2	✓	[45]	✓	[106]	✓	[24]	✓		✓		✓		✓	[67]	✓	[44]	✓	[128]
	RL3	✓	[99]	✓	[107]	✓	[53]	✓	[25]	×	[116]	×	[120]	×	[117]	×	[124]	×	[129]
	RL4	×	[15, 102]	✓	[108]	×	[22, 23, 110]	×	[113–115]	×		×	[121, 122]	×	[118, 119]	×	[125–127]	×	[131, 132]
	RL5	✓		✓	[109]	×		×		×		✓		✓		×		×	
Flexibility	**FL1**	✓		✓		✓		✓		✓		✓		×		✓		×	
	FL2	✓	[45]	×		×		×		✓		✓		✓	[67]	✓		×	[128]
	FL3	✓	[15, 99, 100]	✓	[106]	×	[22]	✓	[25]	✓	[C1]	✓	[120]	×	[117]	×	[124]	✓	[129, 130]
	FL4	×	[102]	✓	[107]	×	[53]	✓	[113–115]	×	[116]	✓	[121–123]	✓	[118, 119]	×	[125–127]	×	[131, 132]
	FL5	✓		×		×		✓		×		×		×		✓		×	
Scalability	**SL1**	✓		✓		✓		✓		✓		✓		×		×		×	
	SL2	×	[45]	✓		✓		✓		✓		✓		×		×	[44]	×	[128]
	SL3	×	[15]	✓	[108]	✓	[53]	✓	[25]	✓	[45]	✓	[120]	×	[117]	×	[124]	×	[129, 130]
	SL4	✓	[100]	✓	[109]	✓	[22, 23, 110]	✓	[113–115]	×	[116]	×	[121–123]	×	[118, 119]	×	[125–127]	✓	[131, 132]
	SL5	✓		×		×		×		✓		×		×		✓		×	

*The conformance (Cf.) of a given mechanism for a given criterion.

$^{\delta}$The corroborating references (Refs.), if any, for the specified conformance of a mechanism for a given criterion

3.4 Mobility Management: Persistent Challenges, Potential Solutions and Next-Generation Framework

From our discussions in Chapters 2 and 3 so far, we have highlighted the requirements from MM mechanisms as well as the criteria that future MM mechanisms should satisfy to meet these requirements in Tables 2.1 and 3.1, respectively. Further, we have analyzed the legacy mechanisms and the current state of the art towards their utility for 5G and B5G networks in Tables 3.2 and 3.3, respectively. However, we have observed that gaps in fulfilling the requirements still persist. Concretely, we have demonstrated that none of the strategies evaluated satisfy the reliability, flexibility and scalability criteria in their entirety. Hence, to be able to design and develop a holistic MM mechanism, it is of substance to our study to understand the challenges/questions that persist. We consolidate, from earlier works in literature and the discussions in Chapter 2 and Sections 3.1-3.3, these key challenges/questions in the text that follows.

3.4.1 Challenges

3.4.1.1 Handover Signaling

Even after the release of 3GPP specifications for 5G [29], HO signaling is still a challenge. Hence, reducing HO signaling to ensure system scalability and reliability will be one of the key challenges. Certain studies such as our contribution [J1], discussed in detail in Chapter 5, have provided methods to help overcome this challenge and thus, can be actively pursued by the research and industrial community.

3.4.1.2 Network Slicing

Network slices have been defined to ensure different service types are served according to their own resource demands. Hence, it will be a key challenge to design MM strategies that either jointly take into account the requirements of multiple network slices or provide individual solutions for each network slice.

3.4.1.3 Integration framework for MM solutions

The state of the art and 3GPP specifications ensure to some extent the provision of flexibility, reliability and scalability for 5G MM solutions, as discussed earlier. However, since these solutions function at different sections of the network (Figure 2.13), the challenge will

be to design them such that collectively they ensure the appropriate levels of flexibility, scalability and reliability in MM mechanisms to cope with the diversity in mobility profiles and applications the devices will access. Also, a part of this challenge will be to ensure that the CAPEX and Operating Expenditure (OPEX), owing to the architectural (software or hardware) transformations stemming from these redesigned MM mechanisms, are manageable.

3.4.1.4 Ensuring Context Awareness

Context based MM solutions accounting for factors such as network load, user preference, network policy, mobility profiles, etc., to ensure best possible provision of requested QoS, will be important. The criticality of this challenge is enhanced by the fact that, low computational complexity whilst executing these solutions will be of the essence to meet the strict latency constraint requirements.

3.4.1.5 Architectural Evolution Costs

SDN and edge cloud capabilities will be important for enhancing the user experience during mobility, as discussed in Section 2.3. However, a key challenge will be to ensure appropriate scalability while maintaining a manageable CAPEX and OPEX.

3.4.1.6 Frequent Handovers

Reducing frequent handovers, ping-pong effects and devising an optimized HO strategy will still be a key challenge, given the dense and heterogeneous future network environment. This is further exacerbated by the fact that current methods, such as IEEE 802.21 and 3GPP specifications, fail to integrate cellular and non-3GPP networks effectively for seamless HO between them. For example, while methods such as LWA have been explored extensively [14, 142], an effective handover methodology between 3GPP and non-3GPP networks still remains elusive.

3.4.1.7 Security

An important challenge for ensuring service continuity and seamless mobility in an extremely dense and heterogeneous network environment, such as 5G and beyond networks, will be to ensure that security related tasks, such as authenticating the user as well as the network, be completed as efficiently as possible. By efficiently here we mean that the authentication should guarantee a required level of security whilst provisioning low computational complexity [143] as well as latency [144]. Again this task will become even more critical in scenarios where mobility occurs between 3GPP to non-3GPP networks.

3.4.1.8 Energy Efficiency

Given that one of the goals of 5G is to ensure enhanced battery lives for the devices, it will be a critical component for 5G MM services to ensure that the mobility of the devices is handled in an energy efficient way [145]. Additionally, 5G MM services will also need to ensure that the energy footprint goal for 5G networks is achieved via techniques such as smart BS selection methodologies [146] and reduced CN signaling [J1, Chapter 5]. By smart BS selection methodologies we refer to being able to not only account for the user energy consumption over the course of its mobility, but also accounting for the energy consumed whilst performing such selections.

3.4.1.9 Meta-surface Reconfiguration for mobility support

For the B5G networks, finding the optimal configuration of meta-surfaces during mobility related scenarios will be challenging. This is because, the physical characteristics of the surfaces will have to be altered rapidly so as to have the signals arriving at the user in a constructive manner.

3.4.1.10 Beyond 5G Network: Handovers

A fundamental question that will be posed in B5G is: how frequently and when will the handovers be needed? The reason this question is a challenge because, up until now the rate of power loss in an urban environment is characterized by a R^4 factor (where R is distance between the transmitter and receiver) given the destructive interference encountered. However, with programmable environments, according to [36], this decay will now be similar to the free space scenario, i.e., R^2, since all signals can be modulated in phase and polarization to interfere at the receiver in a constructive manner. And so, in mobile environments, the power decay will not be significant even at distances further away. Hence, the handover triggering methods and their execution procedure need to be revisited as currently they do not expect such a reliable behavior from the channel.

3.4.1.11 Beyond 5G Network: Protocol stack

A next fundamental question posed in B5G, with reference to meta-surfaces, is: What is the impact on the existing layers? The reason this question is a challenge because, the MAC, Radio Link Control (RLC), PDCP and TCP layers, they all have error control, packet re-ordering, transmission repeat request and other reliability control mechanisms in-built. These were designed keeping in mind that the environment is unreliable and randomly vary-

ing. However, with programmable surfaces the environment will be much more deterministic and reliable. Thus, there arises a case for either eliminating/modifying some of these layers (for example, a lightweight version of TCP may be utilized, as the channel is deterministic and the probability of having lost packets due to error or timeout is significantly lower since the multipaths can be redirected to interfere constructively at the receiver by the meta-surfaces, or the User Datagram Protocol (UDP) can be utilized with much more reliability), which play a critical part in MM procedures, or revisiting their original implementation to adapt to these programmable environments.

3.4.1.12 Dynamic Network Topology

In terms of user association for B5G networks, the challenge will now not be to just choose an BS with the best SINR/RSSI/RSRP/RSRQ, but it will rather be to choose or program an BS/programmable surface configuration/drone, depending on the user mobility, location and coverage from these sources. While it still reduces to the problem presented for 5G networks, the increased dimensionality and heterogeneity of the problem will provide formidable challenges to existing methods.

3.4.1.13 Edge Node configuration in B5G networks

Edge nodes' placement for supporting user mobility will also be challenged. This is so because the possibility of supporting better QoS over longer distances can reduce the requirements for service replication/service migration. This is a consequence of the fact that the handovers would be impacted given the programmability of the environment and the squared decay instead of a fourth power decay in the received signal power.

3.4.1.14 IP address continuity

The vision for near zero latency by 3GPP [147] necessitates that E2E link continuity is ensured given any network and mobility scenario. Hence, maintaining IP address continuity during mobility events will remain a critical challenge as the complexity of the networks increases in 5G and B5G.

The aforementioned key challenges define the technology gap towards fulfilling the MM governing parameters listed in Table 3.1. In the following subsection we list the potential solutions that can fill this technology gap. Note that, we codify these potential solutions as *P1-P7*, which are then utilized in the development of Table 3.4.

3.4.2 Potential Solutions

3.4.2.1 P1: Smart CN signaling

Utilizing the properties of SDN, the signaling performed within the CN for handover and re-routing purposes can be optimized further. This will enable more scalability and better support to users with high mobility. Concretely, techniques such as graph theory, Machine Learning [148] as well as the recently established intelligent Information Elements (IE) mapping methods, as developed in our contribution [J1], etc., can enable faster and efficient CN signaling, as mentioned above. Here by efficiency we imply that the transmission cost, processing cost and other CN signaling related metrics, specified in our contribution [J1] (Chapter 5), are reduced/optimized.

3.4.2.2 P2: On demand MM

Given the functional requirements (Section 2.1), legacy methods (Section 2.2) and the state of the art (Section 2.3), on demand MM strategies (such as in our contribution [C1]) will allow future MM mechanisms to serve users with different mobility profiles, accessing different services and accessing networks with differing loads, more effectively. As an example, slice based MM strategies can enable independent strategies for the various network slices that the 5G networks will serve. This will help cater to the different network slices according to their mobility demands, and avoid the sub-optimal *one size fits all* approach.

3.4.2.3 P3: Deep learning

Learning network parameters such as network load, congestion statistics at access and core network, user mobility trends, etc., enable the network to devise effective and optimal MM strategies for a highly dynamic network environment such as that in 5G and B5G networks. Hence, deep learning methods such as reinforcement learning can assist in such tasks.

3.4.2.4 P4: SDN-NFV integrated DMM

DMM facilitates the distribution of MM functionality throughout the network and avoiding single MM anchors, which consequently assists in alleviating issues such as SPoF and congestion. Note that, SDN and NFV will assist in DMM as network programmability facilitates fast switching while the user/device transits through the network.

3.4.2.5 P5: D2D CP-DP extension

D2D clustering and support for communication with devices in such clusters has been formalized since 3GPP Release-13. Thus, through an extension of CP-DP capabilities of the current D2D framework, i.e., by utilizing the relaying strategies for CP/DP information, handover performance for devices migrating within the network and in such clusters can be enhanced. Further, policy based methods, which take into account the presence of D2D communications between vehicles and other V2X scenarios, will also enable future MM mechanisms to serve the complex scenarios that will prevail in 5G and B5G networks better.

3.4.2.6 P6: Service Continuity through Edge Computing

For serving fast moving users, such as vehicles, and satisfying their latency and bandwidth requirements, edge computing solutions for MM will play a major role in 5G and B5G networks [149]. And while service migration strategies will play a critical role in ensuring seamless connectivity, a fine balance between service replication and service migration will help mitigate the multitude of challenges that arise for such strategies. Further, given that users might crossover to other PLMNs during the duration of mobility [150], which can lead to a change in the edge cloud that serves them, effective service migration strategies will greatly enhance the QoS during mobility.

3.4.2.7 P7: Clean Slate Methods

Current networks rely on resolving the IP addresses of the hosts for the applications requested by the users. However, such a resolution can lead to delays [151]. And so, Information Centric Networking (ICN), and specifically Named Data Networking (NDN) paradigm, avoid this process thus making the network more flexible and faster. Additionally, with the proposition of having in-network caching, ICN and NDN paradigms enable caching capabilities near the users.

Another class of such clean slate methods is MobilityFirst [152]. In MobilityFirst, a new paradigm to networking, like in ICN and NDN, has been proposed. In this paradigm, IP based resolution of nodes has been deprecated, and name based resolution is proposed. Further, concepts similar to ICN and NDN, such as in-network caching etc., have also been proposed. Additionally, and different to the ICN-NDN paradigm, ensuring security in a fully dynamic scenario has been considered as one of the guiding principles of MobilityFirst. Further, MobilityFirst also introduces support for migration of entire networks and not just the end nodes.

Table 3.4: Mapping potential solutions to MM challenges

Challenges	**Reco. Pot. Slns.**	**Comments**	**Param. Satisfied***
Handover Signaling	P1 and P4	Smart CN signaling method, such as in our contribution [J1], will assist in relieving the handover signaling load significantly. DMM strategies will assist in decentralization of MM anchors	RL3, RL5, SL1 – SL4
Network Slicing	P2	On demand strategy will assist the network slices to provision tailor made mobility solutions for the tenants	FL1, FL5
Integration framework for MM solutions	*Design*	This is a design challenge; take into account all the other non-design challenges and other necessary factors (efficacy and delays)	SL5
Ensuring Context Awareness	P2	It will ensure that the user, network and application context is taken into account and appropriate MM solution is provisioned as and when needed	FL5
Architectural Evolution Costs	*Design*	This is a design challenge; take into account all the other non-design challenges and other necessary factors (cost of infrastructure)	SL5
Frequent Handovers	P3	Deep learning can help predict/estimate valuable system parameters, such as SINR, to avoid the frequent handover condition via appropriate BS-user association	RL1, RL2, FL2, FL3, FL4
Security	P1	Effective CN signaling will assist in maintaining/migrating security context when required, thus reducing the latency as well as complexity to ensure the same	RL2, SL3
Energy Efficiency	P1 and P3	Deep learning can provision an optimal MM solution whilst adhering to the energy constraints; Smart CN signaling can enhance energy efficiency during mobility through reduced signaling message exchanges	SL1
Meta-surface Rcfg. for mob. support	P1	Based on the user mobility deep learning algorithms can assist in understanding how the meta-surface configurations have to be adjusted so as to ensure the requested user QoS	RL1,RL2 and FL3
B5G: Handovers	P1, P5 and P6	Edge compute platforms can assist in faster and effective handover decisions. Smart CN signaling can assist in efficient and low latency handover signaling in the CN. D2D networks can assist in smoother handovers	RL2, RL4, FL3, SL1 – SL4
B5G: Protocol Stack	*Design*	This is a design challenge and hence, should collectively take into account all the other non-design challenges as well as other necessary factors, such as efficacy and delays	SL5
Dynamic Network Topology	P3	recognizing complex associations will make deep learning methods essential in determining the optimal user-BS association in dynamic and multi-dimensional networks, such as B5G	RL1, RL2, FL3, FL4
Edge Node Cfg. in B5G Networks	*Design*	This is a design challenge. It should collectively take into account all the other non-design challenges as well as other necessary factors, such as efficacy and infrastructure cost	SL5
IP address continuity	P7	Name based destination resolution allows clean slate methods to assist in maintaining a single IP address throughout with respect to the destination server	RL2, RL4

* Details regarding the parameters and the requirements that they help satisfy are provided in Table 3.1.

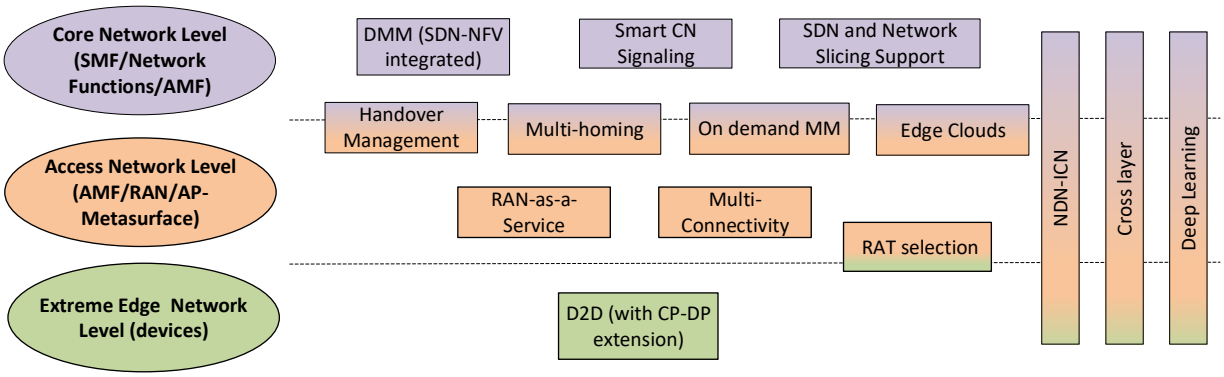


Figure 3.1: Proposed 5G and beyond MM framework.

Consequently, such methods together can provision more scalable, flexible and reliable MM strategies.

And so, up until now in this section, we have highlighted the multiple challenges that the 5G and beyond MM mechanism will face, given our qualitative evaluation for legacy and current state-of-the art methods in Sections 3.1-3.3. We have then provisioned a brief discussion on the potential solutions that can assist in addressing these challenges. We illustrate a novel mapping between these challenges and potential solutions in Table 3.4. Additionally, we have also listed the parameters for the qualitative analysis (and hence the requirements specified in Table 2.1) that they satisfy. This, as a result, reinforces the completeness of our current study. Hence, in the next subsection, utilizing the inferences from Sections 3.1-3.3 and Table 3.4, we propose a framework for 5G and beyond MM.

3.4.3 Proposed 5G and beyond MM framework

We utilize the earlier established classification process for the current state-of-the-art strategies to define our vision for 5G and beyond MM in Figure 3.1. Concretely, we have categorized the MM mechanisms as *Core Network level*, *Access Network level* and *Extreme Edge Network level*, depending on where they will be creating an impact on/from. The specific entities (based on the 5G architecture illustrated in Figure 2.13), to which these aforesaid levels correspond to, have also been mentioned in Figure 3.1.

To elaborate, the core network strategies encompass the DMM, SDN and Network slicing paradigms to provision the necessary reliability, flexibility and scalability from a more global perspective. Additionally, the aforesaid core network strategies need to be well complemented with an efficient CN signaling strategy. Next, handover management, on-demand MM, IPv6 multi-homing and Edge cloud related MM strategies will be enacted not only in the core

network or the access network level, but jointly at both levels thus provisioning the necessary flexibility and reliability. Further, RAN-as-a-Service and Multi-connectivity provisions at the access network level will assist in utilizing the multiple RATs and BSs effectively. Moreover, it is envisioned that the RAT selection process maybe either at the access network or at the device level. The D2D techniques, on the other hand, are expected to provide added assistance for mobility at the device level through DP as well CP functionality.

Complementing these mechanisms, NDN-ICN support will be provisioned at all levels, thus assisting in maintaining IP addresses/prefixes during mobility whilst resolving destinations via names. Note that, traditional IP address/prefix allocation strategies are not intended to be changed. Instead, the NDN-ICN concept provisions an over-the-top assistance. Further, the cross layer strategies, as the name suggests, will spawn across the multiple levels and enact policies, utilizing the available information at each of these levels, which assist in optimal MM related decisions across the network. Lastly, the deep learning strategies will again assist across the multiple levels by learning the complex features about the network context, user mobility and overall QoS requirements, and formulating effective MM related decisions.

Hence, given that we utilize the potential solutions for overcoming the technology gap, specified in Section 3.4.2, alongside certain strategies from the state of the art and legacy MM mechanisms, specified in Sections 3.2 and 3.3, it can be inferred from Tables 3.1-3.4 that our proposed framework will satisfy all the parameters for the reliability, flexibility and scalability criteria. Consequently, it can be stated that the proposed framework in Figure 3.1 will also satisfy all the requirements as defined in Table 2.1, thus provisioning a holistic solution. With this vision, in the following section we summarize the developments of this chapter and establish how the work presented in subsequent chapters aims to realize this framework.

3.5 Summary

In this chapter, we have provided a novel qualitative gap analysis of the legacy and current MM mechanisms with respect to their suitability for 5G and beyond networks. Given the complexity of future network scenarios, i.e., 5G and B5G, a full view of the MM strategies, their capabilities, the persistent challenges and the possible solutions to them, will enable the research community to design better MM strategies.

Concretely, and in continuation with the requirements specified in Table 2.1 in Chapter 2, we firstly defined the three pillars of future MM strategies, i.e., scalability, flexibility and reliability, in-depth in Section 3.1. Further, we also specified the multiple parameters that the

future MM mechanisms will need to satisfy for each of the evaluation criteria, through Table 3.1. Next, from our discussions in Section 3.2 it is clear that the legacy MM solutions fail in provisioning scalability, flexibility and reliability simultaneously. Nevertheless, the current standards and research efforts explored in Section 3.3 are promising as they provide enhanced capabilities towards future MM solutions. We have summarized these conclusions effectively in Tables 3.2 and 3.3. And as a consequence, through this qualitative analysis the various benefits and shortcomings of the legacy and the current state of the art mechanisms, studied in this chapter, can be understood easily by the research community. Subsequently, we established that none of the mechanisms fulfill the complete 5G and beyond MM mechanism requirements.

And so, it is evident that a holistic MM mechanism for 5G and B5G networks remains elusive. Thus, certain challenges that will still persist for the design, development and deployment of future MM mechanisms have been detailed in this chapter in Section 3.4.1. Furthermore, we have provided a concise discussion on the potential MM strategies that the research community can explore so as to solve these persistent challenges and the technological gaps they present, in Section 3.4.2. Following this, we have also provisioned a novel mapping between the potential strategies and the persistent challenges in Table 3.4, thus highlighting the efficacy of our current study. Based on the inferences drawn, we have provisioned a novel framework for the 5G and beyond MM strategies through Section 3.4.3 and Figure 3.1.

Henceforth, in Chapters 4-6, we build upon this future framework and provision an on-demand MM paradigm (Chapter 4); a novel Handover signaling mechanism (Chapter 5); and a novel User Association and Resource Allocation framework (Chapter 6), in order to facilitate *enhanced MM solutions for 5G and beyond networks*. Notably, each of these research works maps to one of the building blocks of the proposed framework presented in Section 3.4.3 and Figure 3.1. Specifically, the on-demand MM paradigm explored in Chapter 4, maps to the *on-demand MM block* in Figure 3.1, the novel handover signaling mechanism explored in Chapter 5, maps to the *Smart CN signaling and Handover management blocks* in Figure 3.1, and the novel User Association and Resource Allocation framework explored in Chapter 6, maps to the *Multi-connectivity, RAT selection and Cross layer blocks* in Figure 3.1.

Chapter 4

Mobility Management as a Service

Overview

Mobility Management (MM) techniques have conventionally been centralized in nature, wherein a single network entity has been responsible for handling the mobility related tasks of the mobile nodes attached to the network. However, an exponential growth in network traffic and the number of users has ushered in the concept of providing on-demand mobility management, i.e., Mobility Management as a Service (MMaaS), to the wireless nodes attached to the 5G networks. Allowing for on-demand mobility management solutions will not only provide the network with the flexibility that it needs to accommodate the many different use cases that are to be served by future networks, but it will also provide the network with the scalability that is needed alongside the flexibility to serve future networks. And hence, in this chapter, a detailed study of MMaaS has been provided, highlighting its benefits and challenges for 5G networks as compared to the 3GPP, IEEE and IETF initiatives. Additionally, the very important property of granularity of service, which is deeply intertwined with the scalability and flexibility requirements of the future wireless networks, and a consequence of MMaaS, has also been discussed in detail.

Contributions

[C1] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Mobility Management as a Service for 5G Networks", IEEE ISWCS 2017, pp. 1–6.

Existing mobility management architectures, such as the one employed by LTE [61], are centralized in nature. To illustrate, the Mobility Management Entity (MME) in the LTE architecture depicted in Figure 2.2 is the central entity which is entrusted with the respon-

sibility of managing mobility of users attached to the network. The aforementioned central architecture suffices current day needs. However, due to an exponential growth in traffic and the number of users, as stated in Chapters 1-3, these architectures will not be viable for the 5G network scenarios. Further, based on the discussions in Chapter 3, it is evident that factors such as lack of scalability and flexibility will render the current strategies insufficient for the scenarios that will prevail in these future networks. It is also important to state here that the scalability and flexibility of MM strategies is very intricately connected to the granularity of service aspect.

In realization of the aforementioned issues, Software Defined Networking (SDN) and Network Function Virtualization (NFV) have been recognized as important enablers for the future networks. Concretely, they enable the implementation of critical network functions, such as mobility management, as applications on top of a central/distributed controller. And, with a global/locally-global perspective of the complete network architecture, the mobility management applications can be employed on an on-demand basis for the user devices. It must be mentioned here that the aforementioned global perspective relates to the scenario when the employed MM application has the complete network view, whilst a locally-global perspective indicates that the employed MM application has a global perspective of only a specific domain, which, for example, may be a geographical area that the SDN-controller, on which it is employed, covers. Further, the granularity of service provided by mobility management applications will also equip the networks with pre-requisites such as flexibility and scalability necessary to meet the demands of the 5G networks. Henceforth, in this chapter we discuss in detail the features of this on-demand mobility management service, better known as MMaaS, as well as the related granularity aspects along with their benefits and challenges for 5G networks.

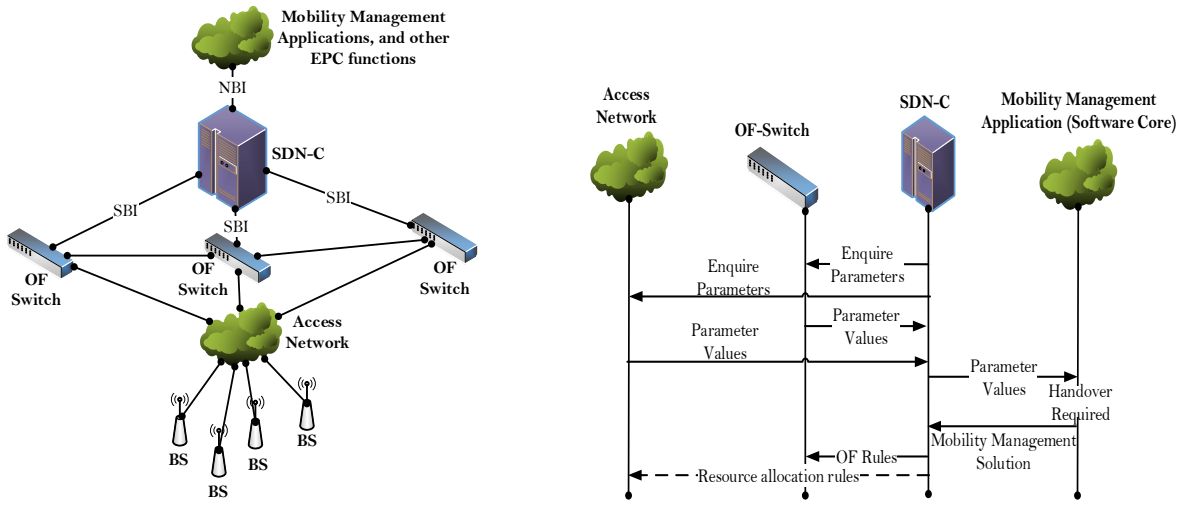
4.1 MMaaS

Existing mobility management strategies have primarily been centralized in nature. But, with the SDN and NFV techniques, mobility management functionalities can now be implemented as an application on top of a controller that provides it with a global view of the domain it serves. This softwarized control over mobility management permits the operators to provide the services on-demand, i.e., MMaaS. It is worth noting that, under the current mobility management strategies, when an MN attaches to the network, a mobility instance is created for it in the MME, and is kept at all times until it de-registers from the network. This leads to the unnecessary utilization of computational resources. However, with MMaaS, mobility management instances can be created on-demand and hence, computational resources

can be allocated likewise. Consequently, MMaaS, through its global view and on-demand computational resource allocation, enables the provision of globally optimized solutions for managing user mobility.

The aforementioned softwarized control allows for the utilization of a versatile set of parameters, which not only provide a globally optimal solution but also permit the self-adjustment of the established mobility management mechanisms. In order to retrieve these parameter values from the network entities, a network controller, i.e., the SDN-controller (SDN-C), has to interact with these entities over the southbound interface (SBI) and then pass on the extracted values to the mobility management application over the northbound interface (NBI) [153]. An illustration of the aforementioned process is provided through Figure 4.1. As can be seen from Figure 4.1(a), the SDN-C is connected to the OpenFlow (OF) switches, which comprise the network data plane. These switches are additionally also connected to the access network, from where values of the parameters such as Signal to Noise Ratio (SNR)/Received Signal Strength Indicator (RSSI) of other and current base stations (BSs) at the MN, types of flows on the MN, MN policies, etc., can be enquired. Further, from the OF switches, information related to the network such as network load, link failure/congestion information, as well as the latency over the links, etc., can be extracted. All of this information is then processed at the SDN-C which then, as is visible in Figure 4.1(a), is sent over the NBI to the mobility management application. These mobility management applications, which may be implemented on a software cloud, after processing this data, provide a solution to the SDN-C, which implements it over the network via the orchestrator through the SBI. Figure 4.1(b) provides a signaling diagram to illustrate the above flow of information to ensure mobility management services to the MNs attached to the network.

It is important to mention here that the message sequence as provided in Figure 4.1(b) might change in practical implementation. However, the overall logical flow, i.e., information enquiry → information reception → information processing → MM rule implementation, is maintained. Further, in Figure 4.1(a), the access network might consist of a Centralized/Cloud RAN (C-RAN) [21], which is primarily composed of a BBU pool and multiple BSs attached to this BBU pool. In the aforementioned scenario, the BBU pool is responsible for handling the access network mobility, i.e., handling the mobility of MNs when they switch BSs within the same BBU domain. Here, a BBU domain specifically refers to a set of BSs controlled by a particular BBU pool. And so, as a consequence, the resource allocation rules message, as shown in Figure 4.1(b), is sent by the SDN-C to the access network only when the scenario demands, for instance: when performing a traffic transfer due to HO. Thus, it can also be inferred that MMaaS is essentially distributed wherein the access network mobility is handled at the BBU pool (or BS in case CRAN is not present), whilst network



(a) (b)

Figure 4.1: (a) Softwarized network control; (b) Signaling diagram for mobility management in SDN based networks.

level mobility (inter-domain mobility) is handled at the SDN-C. Lastly, an important point of consideration with regards to the resource allocation rules procedure is that, in the event legacy RAN deployments exist, the SDN-C, similar to the MME as shown in Figure 2.2, will almost always be in communication with the access network to handle the mobility at the access network level.

Next, in order to further exemplify the advantages that MMaaS provides, a short comparative analysis with respect to the existent/legacy mechanisms has been provided in Table 4.1. The comparative analysis is conducted on the basis of 4 distinct parameters, i.e., *Granularity of service*, *Degree of centralization*, *Network slicing support* and *Self-reorganizing capabilities*. Note that, for this comparative analysis, we have utilized a broader umbrella of definition for the existing mobility strategies by considering them on the basis of the Standards Development Organization (SDO), i.e., IEEE, IETF and 3GPP, that were responsible for their development. Concretely, all the MM strategies discussed in Chapters 2 and 3, can be assimilated into these broad definitions without loss of specificity.

And so, from Table 4.1, it is evident that the existing mobility management mechanisms, designed and developed by IEEE, IETF and 3GPP, do not provide a significant level of service granularity. Whilst these mechanisms provide at best per-MN granularity in service, MMaaS on the other hand has the ability to provide multiple avenues of granularity such as those based on mobility profiles, flows, policies, MN, etc. It is imperative to state here that, in order to provide the level of flexibility and scalability to the future networks, as

would be needed to serve the complex scenarios that will be prevalent, such multi-avenue provision in granularity for mobility management services is an indispensable feature. In addition to the multiple avenues of granularity in service, MMaaS provides a significant advantage over the existing mechanisms by allowing for a de-centralized implementation of mobility management applications. Such flexibility stems from the softwarized control which is a consequence of the SDN and NFV framework. On the other hand, 3GPP due to the centralized MM anchors in the CP, i.e., the MME in 4G and SMF/AMF in 5G, and in DP, i.e., the S-GW in 4G and UPF in 5G, offers a very centralized strategy. However, through the LTE-X2/5G-Xn handover process and the traffic offloading mechanisms, some form of de-centralization can be obtained. IEEE mechanisms, such as 802.21 and 802.11x series, do not provision any specific methods for decentralization. Additionally, while IETF has certain studies on DMM, the overall architecture as defined by the PMIPv6, FMIPv6, etc., is fairly centralized.

MMaaS through its softwarized control and global view will be able to serve multiple network slices whilst existing mechanisms, according to Table 4.1, cannot support network slicing environments. This is so because, existing mechanisms were not designed to logically slice the network infrastructure for the various tenants. By logically slicing the network infrastructure we mean that, resources in the core and access network are reserved independently for each tenant on it. Furthermore, by tenants we refer to services such as voice, broadband and Narrow-Band (NB) IoT, mobile virtual network operators (MVNOs), etc. It must be stated here that, 3GPP based mechanisms, due to the emergence of NB-IoT and already existing QoS request identifier mechanisms, provision certain level of network slicing support.

Lastly, the self-organizing capabilities, i.e., the ability to re-structure the routing rules and the access network resource allocation (if needed) depending on the context of operation, are of prime importance to the future networks as the highly dynamic environment will require the mobility management mechanisms to adapt their solutions according to the scenario without any perceivable latency. MMaaS, owing to its flexibility and granularity characteristics as already mentioned, offers a high degree of self-organizing capabilities. On the contrary, existing solutions as shown in Table 4.1, offer minimal self-organizing features. Concretely, 3GPP, through LTE-X2/5G-Xn, LTE-DC/5G MR-DC and LWA, offers avenues for self-organization in the access network. IEEE, on the other hand, through the 802.21 and 802.11x suite, offers methods for resource allocation and negotiation in heterogeneous RATs. However, these functionalities are minimal as compared to MMaaS. Moreover, IETF, only provisions mechanisms that primarily function on the network layer and above. Hence, the self-organizing capabilities are only limited to provisioning support from the defined

Table 4.1: Comparison between MMaaS and current/legacy architecture

	MMaaS	**3GPP**	**IEEE**	**IETF**
Granularity of service	Multiple avenues [1]	per-MN	per-MN	per-MN [2]
Degree of Centralization	De-centralized	Mostly Centralized [3]	Centralized	Centralized [4]
Network slicing support	Yes	Minimal [5]	No	No
Self-reorganizing capabilities	Very high	Minimal [6]	Minimal	Minimal

mechanisms towards any self-organization rules/policies defined at the access network level.

Thus, this discussion reinforces the belief that existing MM mechanisms are not well-suited to handle the challenging scenarios that future networks will envisage. Furthermore, from the analysis so far, it is evident that granularity of service offers significant benefits to the network, through its provision of scalability and flexibility, as well as to the users, through the provision of optimal mobility management solutions dependent on their context. In addition, there are multiple avenues where granularity in mobility management services can be offered under the MMaaS concept. And so, in the subsequent section, a detailed study on granularity of service and the various avenues, such as mobility profiles, flow types, network load and policies, where the granularity can be offered has been provided.

4.2 Granularity of Service

As stated in the previous section, granularity of service is beneficial to the network through its provision of scalability and flexibility, whilst for the users it formulates optimal mobility management solutions, which in turn benefit them by helping reduce the power consumption as well as improve their perceived Quality of Service (QoS). To better elaborate these afore-

[1]Per-flow, per-mobility profile, policy based, per-MN, etc.

[2]Multi-path TCP (MPTCP) and Stream Control Transmission Protocol (SCTP) allow for multiple paths/flows. However, IETF does not provide per-flow mobility management in these protocols as of yet.

[3]LTE-X2 and 5G-Xn handovers offer some form of de-centralization.

[4]IETF DMM working group presents certain studies on distributed frameworks. However, there are no standards level RFC as of now.

[5]Recently NB-IoT has been standardized, which utilizes LTE bands to serve IoT devices.

[6]LTE-X2/5G-Xn, LTE-DC/5G MR-DC and LWA allow some level of self-organizing capabilities.

mentioned broad benefits, we consider the scenario as specified in Figure 4.2. The scenario specified is a typical mobility scenario wherein an MN migrates from one BS to another, and also switches its access router (AR) in the process. Further, at AR-1 the MN has a certain set of active flows. After moving from AR-1 to AR-2, the MN keeps its current flows active and also initiates other services which consequently result in the creation of new flows. It is imperative to note here that AR-1 and AR-2 are merely data-plane (DP) entities and henceforth, in the architecture mentioned in Figure 4.1(a), they are equivalent to the OF-switches. Next, whilst switching BSs and ARs, the mobility management application analyses parameters and policies which are relevant to both the user and the network. It also checks the context, i.e., the mobility profile, the flow types, etc., and makes a handover/traffic transfer decision.

From the aforementioned decision process, it is clear that the mobility management rules implemented for the MN can provide granularity in terms of mobility profiles, flows, policy, and network load, or a combination of them depending on the context. As an example, in Figure 4.2, the granularity of service from the perspective of flows is illustrated, wherein delay-sensitive and delay tolerant flows are served individually with different MM rules. A more detailed discussion for the same is provided in Section 4.2.2.

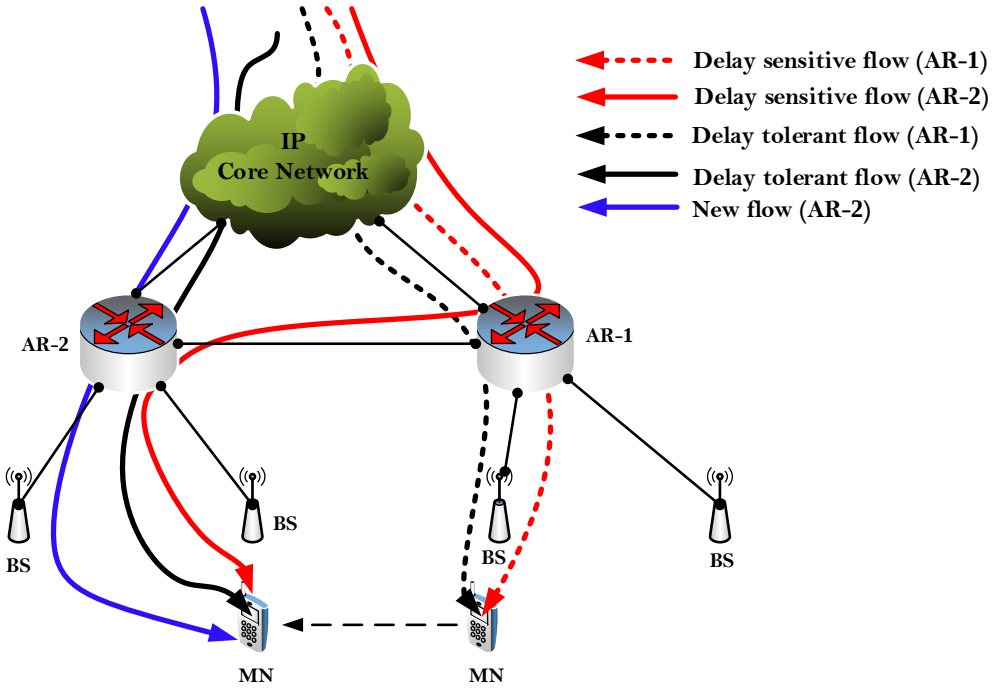


Figure 4.2: MMaaS - Granularity of Service provision example.

Such a holistic and distributed decision process provides the network with the flexibility and also allows it to scale itself as the services offered are on-demand and not centralized to just one entity within the complete network for each user. Further, the decision making process involves analyzing the parameters as well as the context. This enables the network to make an optimized decision on mobility management for each user. From the discussions so far, a deeper inspection into the granularity considerations reveals that there are multiple avenues to provide such discretization in mobility management services. Consequently, in the ensuing discussions granularity of service from the perspective of mobility profiles, flow types, network load and predefined policies (both network and user) has been explored in more detail.

4.2.1 Mobility profile perspective

Users in a network can have varying mobility profiles. While most users will be pedestrians, moving at speeds below 3 km/h, some users such as those in a car or high speed trains might be moving at anywhere between 30-500 km/h. Further, with the future networks slated to support Internet of Things (IoT), support for devices with negligible mobility alongside the aforesaid mobility profiles becomes an important point of consideration for future mobility management solutions. The MMaaS paradigm allows operators to deploy softwarized solutions on top of the core network controller, hence, permitting the network to employ solutions that are tailored for specific mobility profiles. As for example, there may be a scenario where there are several pedestrian users with a few high speed users. These users are then overlaid with a high density of static sensors. In such a scenario, the current networks would assign a mobility profile to every attached device, even if they do not require one (like the static sensors). Further, current networks will also employ the same computational and physical resources for each device attached to the network for the purpose of mobility management. While such a method is simple and easy to implement, it is inefficient as certain device do not require similar levels of mobility management, such as static sensors as already mentioned, compared to the others.

Henceforth, MMaaS provides the opportunity to offer granularity on the basis of mobility profiles, wherein devices based on their mobility profile are allocated appropriate resources within the network. Considering the example as described above, a viable solution based on the MMaaS paradigm would be to assign Macro-cell resources to the high speed users in order to avoid unnecessary handovers. Further, pedestrians can be allocated control-plane (CP) resources at the Macro-cell (to avoid frequent messaging with the core network for resource allocation upon handovers) whilst the data plane can be kept at the SCs. It is similar

to the phantom cell strategy proposed in [116] wherein MNs have CP at the Macro-cell and are subjected to a data shower from the SCs, i.e., DP is at the SCs. Additionally, static sensors, since they are not going to be subjected to any mobility event, do not necessitate a mobility profile. And hence, MMaaS avoids assignment of a mobility profile and subsequently, allocation of network resources to such static users. And so, MMaaS ultimately presents a very resource-efficient and flexible avenue, dependent on the user mobility profile, to employ mobility management services.

4.2.2 Flow perspective

With the smartphone boom, and their myriad capabilities, users have access to a variety of applications ranging from the erstwhile calling and short message services (SMS) to the more recent Voice-over-IP (VoIP) services. The unprecedented growth in Internet traffic in conjunction with the ever increasing diversity of the application types has warranted a rethink on how mobility management mechanisms will be able to deal with such heterogeneity.

From Table 4.1, we know that the existent and legacy techniques do not provide for a per-flow granularity. However, with MMaaS, where the mobility management application has a global view of its domain, MM techniques have the capability to distinguish whether a particular application flow is delay sensitive or delay tolerant. Subsequently, upon the determination of the type of flows associated with a user, mechanisms such as allowing data forwarding and the eventual route optimization process for delay sensitive services; and simple IP switching for delay-tolerant services, can be executed by the mobility management applications. To illustrate the aforementioned capability, consider Figure 4.2 wherein the MN at AR-1 has two flows. A deeper inspection reveals to the network that one of the flows is delay tolerant while the other is delay sensitive. And so upon moving to a new BS under a new router AR-2, the MMaaS paradigm allows the network to provide data forwarding capabilities for the delay sensitive flow whilst simple IP switching is provided for the delay tolerant flow. Additionally, through MMaaS, the new flow originating at AR-2 is provided access to the IP core network through AR-2 and not AR-1, hence, removing any DP anchoring similar to legacy methods.

Further, applications such as enhanced mobile broadband (eMBB), Ultra-reliable low latency communications (URLLC), and Massive machine-type communications (mMTC) [154], can be classified as delay-sensitive and delay-tolerant based on their latency requirements, which also encompasses their critical nature. And so, MMaaS through its ability to process each flow separately, as described above, can serve these aforesaid application types satisfactorily ensuring the required network flexibility and the expected QoS for the user, as defined

under the 5G paradigm.

4.2.3 Network load perspective

Network load perspective in essence involves transfer of traffic which implicitly invokes the mobility management mechanisms. Network intelligence, which is a critical component of future networks, allows the transfer of traffic to some other location thus enabling the network to prevent congestion whilst still ensuring the required QoS to the user. Since, this transfer of traffic involves switching the connectivity of user, through MMaaS, the network provides appropriate mobility management rules dependent on the context. As for example, consider the scenario where users in a particular area are in the coverage of multiple BSs, i.e, the BSs have overlapping coverage areas, and the users can be connected to multiple BSs at any given point in time, i.e., they can experience multi-connectivity. Next, these BSs might belong to the same RAT or to different RATs, thus leading to a heterogeneous and dense environment. In such a scenario, and given that the density of users is exponentially increasing as mentioned before, some BSs or ARs might experience heavy load thus degrading the QoS of the users attached to those network entities.

Henceforth, in order to ease the load on the aforementioned network entities, the network initiates mobility management mechanisms. Subsequently, through the flow and mobility profile based approaches, MMaaS equips the network with the required granularity to transfer certain flows to other points of attachment (in the multi-connectivity scenario) or to completely switch/forward traffic flows depending on their nature. And so, the MMaaS paradigm provides benefits not only to the users, but also extends multiple utilities to the network as discussed above.

4.2.4 Predefined policies perspective

The mobility management mechanisms implement a particular solution by not just analyzing the parameters and context, but they take into consideration the predefined policies of the network and the user as well. The predefined policies may entail features such as network preference, service subscription, roaming policies, etc. Subsequently, depending on the context of the user, MMaaS can decide to give more weight to certain aspects or more formally: *certain components* of the policy vector, as compared to others. This property essentially enables the mobility management mechanisms to provide specific services to individual users depending on their context with respect to the defined policy vectors. In this regard, [103] and [155] propose ideas for mobility management mechanisms based on policy vectors and user context.

To elaborate, [103] proposes an SDN approach on both the multi-mode mobile terminal (MMT) as well as in the core network. The MMT with assistance from the network gathers information such as available BSs and their APIDs, RSSI, network load, etc. This enables the MMT to compare these registered parameter values against its pre-defined policy vectors. Subsequently, it enables the MMT to perform network selection, which is then communicated to the core network. The core network then through its SDN-C, and in co-ordination with the C-RAN, orchestrates the required operations in order to provide resources to the MMT over its selected set of BSs as well as within the core network. On the other hand, [155] firstly allows the MN to implement its policies when determining the BSs it can attach to. After informing the core network about its choice of BSs, the network implements its policies and then prunes the list of BSs further. The core network subsequently informs the MN about the BS it should attach to. And hence, through the MMaaS paradigm, policy based granularity of service can also be extended towards the users thus emphasizing the utility of MMaaS.

4.3 Related Work

Surveys such as [43] and [44] provide an important basis for understanding the challenges and opportunities that will exist when implementing current/legacy mobility management solutions in a dense and heterogeneous network environment, such as the 5G wireless networks. While they provide detailed analysis on the many efforts that have been made to support seamless mobility, this thesis extends the aforementioned studies by providing significant insights into the new paradigm of MMaaS. Further, in [103] the SDN and NFV techniques have been implemented not just at the network side, but also at the MN and the CN. A policy based mobility management technique on a per-flow basis has been proposed. While, through the proposed technique, a flexible and granular mobility management strategy (on the basis of flows) has been proposed, our work studies multiple other avenues (such as mobility profiles, flow types, network load and predefined policies) where granularity of service can be offered under the MMaaS paradigm. Additionally, certain projects such as 5G-NORMA [104] provide for an SDN and NFV enabled self adjusting mobility framework. Further, [104] also proposes granularity of service based on the required cell size (which indirectly connects to the mobility of the user and the required quality of service), as well as discusses the on-demand mobility management for different slices and sessions. In contrast, through this chapter, in this thesis not only have we built upon the aforementioned approaches to mobility management, but we have also provided for a detailed description on the various advantages, challenges and the distinct avenues for a flexible mobility management strategy for 5G networks under the MMaaS paradigm.

4.4 Summary

MMaaS through its software control, global view, and on-demand service will be an important enabler for the future wireless networks, i.e., the 5G networks. Through its granularity of service provisions, as studied in detail in this chapter, it provides the networks with the flexibility and scalability features so as to cater the highly dense, heterogeneous and dynamic environments that the 5G networks will encounter. Additionally, in this chapter, with the help of certain scenarios, the multiple avenues where granularity of services can be provided have been explored in detail, and subsequently, their advantages have been presented as well. To reinforce the capabilities of the MMaaS paradigm, we qualitatively evaluated it against the 3GPP, IEEE and IETF mechanisms on the basis of granularity of service, de-centralization capabilities, network slicing support and self re-organization capabilities. From this analysis it was determined that the MMaaS paradigm is far more suited to serve the future wireless networks as compared to the 3GPP, IEEE and IETF counterparts.

And so, to conclude, MMaaS, although faced by multiple challenges, will become an important pillar for the future wireless networks, thus enabling them to provide features such as low latency, high data rates, multi-slicing, etc., which are importantly also a part of the broader 5G objectives. Furthermore, the MMaaS paradigm will help to fulfill the on-demand MM block component for the MM framework specified in Figure 3.1.

Chapter 5

Enhanced Handover Signaling Method and System

Overview

As we know, future wireless networks are expected to be ultra-dense and heterogeneous not just in terms of the number and type of base stations, but also in terms of the number of users and application types they access. Such a network architecture will require MM mechanisms that adapt rapidly to its highly dynamic characteristics. In particular, the optimality of the handover signaling within these future network architectures will be extremely critical given their density and heterogeneity. By handover signaling we refer to the exchange of messages that occurs between the various network entities to successfully execute any phase of a handover process. Further, here the optimality is relevant for both the total amount of signaling created and the total delay per handover process. Thus, in this chapter we firstly present a novel and optimized message mapping and signaling mechanism for the handover preparation and failure phases. We also develop a novel handover failure aware preparation signaling methodology, which accounts for the possibility of a handover failure and grants additional enhancements to the handover preparation and failure signaling phases. Through the analytical framework provided in this chapter, we conduct studies to quantify the performance gains promised by the proposed mechanisms. These studies cover myriad handover scenarios as identified by 3GPP, and use the statistics from cellular network operators and vendors. We then develop the idea and analytical framework for network wide analysis, wherein the network wide processing cost and network occupation time for various handover failure rates are computed. Lastly, we propose an evolutionary network architecture that facilitates the proposed signaling mechanism as well as assists operators in maintaining a manageable CAPEX. It combines the current day and 3GPP proposed 5G network architec-

ture with the Software Defined Networking (SDN) approach. As a result, we argue that the proposed mechanisms are viable and outperform the legacy handover signaling mechanisms in terms of latency incurred, total network occupation time, number of messages generated and total bytes transferred.

Contributions

[PT1] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Handover Method and System for 5G Networks", WO 2019/229219 A1 (WIPO PCT), pp. 1-98, 2019. (Positive International Search Report)

[J1] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Evolutionary 4G/5G Network Architecture Assisted Efficient Handover Signaling", IEEE Access, vol. 7, pp. 256–283, Dec. 2018. (Quartile: Q1; IF: 4.098 (2018))

[C4] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Improved Handover Signaling for 5G Networks", IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2018, pp. 164–170.

[C3] **A. Jain**, E. Lopez-Aguilera, and I. Demirkol, "Enhanced Handover Signaling through Integrated MME-SDN Controller Solution", IEEE 87th Vehicular Technology Conference VTC Spring 2018, pp. 1–7.

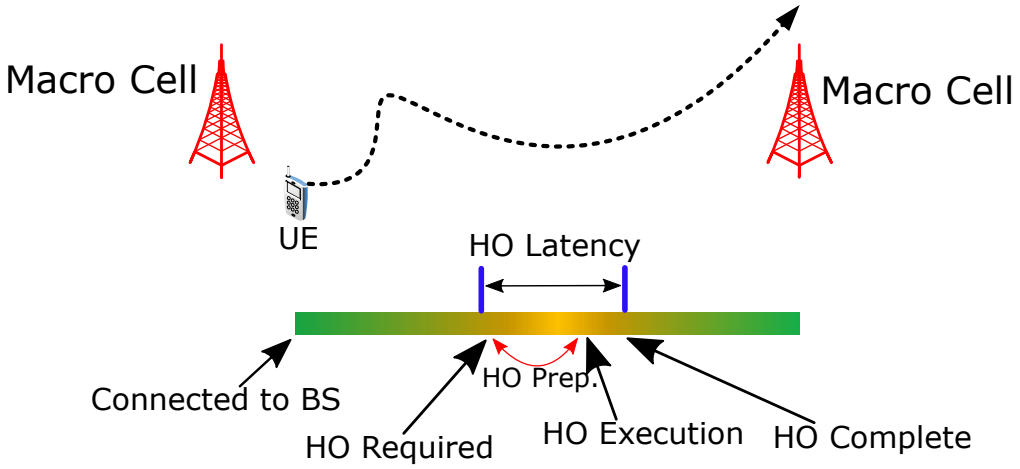
Central to the solutions that offer seamless mobility support in wireless networks are the handover mechanisms, which allow a user to change its physical point of attachment within the network when it is subject to a mobility event and certain pre-programmed conditions are satisfied [133, 156, 157]. For example, if the RSSI or the received signal power from the current serving base station goes below a particular threshold and, simultaneously if the same parameter for another base station in the vicinity goes above a certain threshold, then a decision to change the point of attachment, i.e., the base station, can be taken by the network or the user.

Further, given the highly heterogeneous scenario that will be prevalent in 5G networks, in this chapter we revisit the legacy handover mechanisms which form a critical part of MM. These legacy handover mechanisms are composed of four phases, i.e., handover decision (parameter values, such as RSSI, etc., based decision for BS selection), handover preparation (resource negotiation and allocation involving source and target networks), handover execution/rejection/cancel (path re-routing with the user transitioning from source to target network, or issuance of a cancel/reject indication due resource allocation failure) and

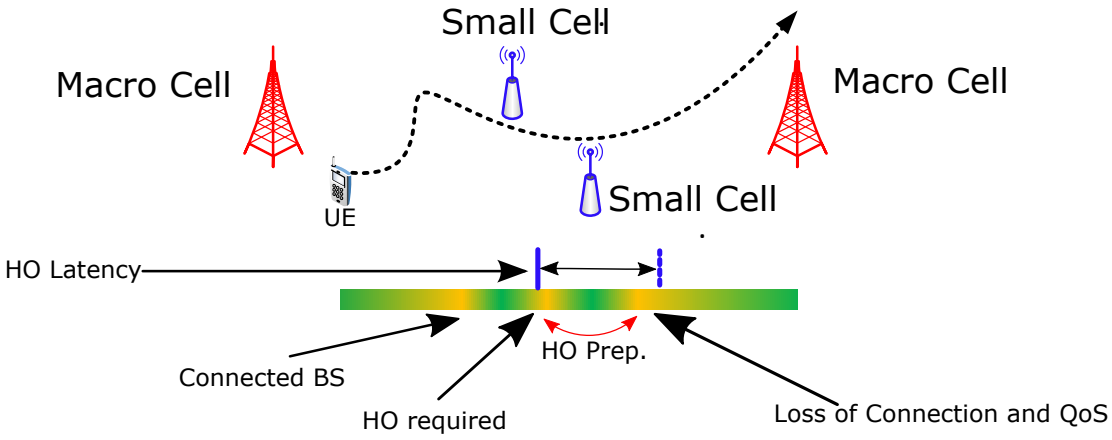
handover complete (release of source network resources upon successful handover to target network). Each of these stages contribute towards the overall latency and signaling cost to execute the handover. Hence, optimizing/enhancing them will facilitate in improving the QoE and QoS to the device/user.

Consequently, many current research efforts, such as [44, 103, 146, 158–163], have provided studies and methods that will facilitate the enhancement/optimization of the aforesaid handover phases. However, the handover preparation and failure phases remain relatively unexplored in the studies referenced above and similar to them. It is during the handover preparation phase that the negotiation and allocation of resources for an impending HO is carried out. Further, during the handover failure phase, signaling that involves sending an indication to the source network and the user undergoing HO with regards to the failed HO attempt is performed. Thus, fast execution of the aforesaid signaling, in a markedly more complex network environment, will be a vital requirement for an efficient next-generation HO management framework. This requirement is further elaborated via the current and future network scenarios, and their corresponding HO phases, illustrated in Figure 5.1.

In the current network scenarios, depicted in Figure 5.1(a), a user has significant time to trigger, prepare, execute and complete a handover. However, the same is not true for the future scenarios, as illustrated in Figure 5.1(b). In the future network scenarios, the density of base stations will be high, i.e., base stations with smaller coverage areas (Small-cells) and higher bandwidths will be packed more closely in a given area. In addition, Macro-cells with significantly large coverage areas will be existent as well, to assist the Small-cells. Hence, if current handover management strategies are utilized, the time taken to complete the handover will be much greater than the dwell time of the user (with its mobility profile) at the desired base station whilst the conditions are still favorable to establish a link. Specifically, the time available to perform the resource allocation and negotiation process will be shorter. Such a scenario would thus lead to loss of connectivity and hence, a poor network performance. Moreover, the HO signaling overhead will also be of critical importance for the network performance because of the FHOs caused by cell densification, and the diverse RATs used resulting in many inter-RAT HOs. Hence, an optimized handover process, where the HO latency and signaling overhead are reduced, will be an extremely vital component of future handover management strategies. Concretely, and from the discussion above, a fast and efficient handover preparation phase signaling will be important for an optimized handover process. Further, in the event that the HO has failed, the CN has to ensure a fast release of the allocated resources so that they can be reused, as well as the user under consideration is free to choose another base station. Thus the HO failure signaling process should also be optimized.



(a)



(b)

Figure 5.1: a) Handover scenario in current wireless networks; b) Handover scenario in future wireless networks.

Hence, in this chapter we have introduced a novel message mapping and parallelized control signaling methodology for the preparation and failure phase scenarios studied by the 3GPP [29, 133]. We follow this approach for both the legacy as well as the 5G networks. This approach subsequently results in a reduction of the overall transmission cost, processing cost, latency as well as the number of bytes transferred during a handover event. Further, the proposed approach has been designed not only for the Intra-RAT HO scenarios, but

also for the scenarios involving inter-RAT HOs including 5G HO scenarios discussed by 3GPP. In addition, in this chapter we have introduced a novel HO failure aware preparation phase signaling mechanism. The proposed mechanism accounts for the possibility of a HO failure during the design and execution of the HO preparation phase. And as will be seen in Section 5.3.3, the HO failure aware mechanism not only enhances the HO failure step but it also presents additional enhancement for the HO preparation signaling step. Further, in this chapter, we have provided a simple yet rigorous analytical methodology to validate the proposed mechanism. The aforementioned methodology enables the reader to compare the performance of the proposed mechanism with the legacy mechanisms, i.e., 3GPP standards, on the basis of latency incurred, transmission cost (i.e., total network occupation time), processing cost (number of messages generated) and the overall amount of bytes transferred. Note that, through the message size analysis we are able to determine the reduction in the overall amount of bytes transferred through the network during a given HO preparation or failure signaling sequence. It is important to state here that the packet or system level simulations would not be able to derive realistic network parameter values, since the network topology, the transport technology used, queueing at the network elements, etc. is dependent on the specific operator scenario and cannot be modeled accurately. Hence, we have utilized real data from network operators and vendors and have attempted to provide a simplistic, realistic, and yet holistic analysis. We have also introduced a novel network wide analysis. Through this we establish the fact that, given any distribution over the number of HOs for the studied HO types and for any HO failure rate, the proposed methodology greatly improves the system performance in terms of overall processing cost and total network occupation time, as compared to the legacy mechanisms.

Next, the aforesaid message mapping and parallelized control information transfer is facilitated via an evolutionary network architecture. This network architecture establishes evolved CN entities wherein they are integrated with an SDN agent. Moreover, the MME in the 4G network and the SMF in the 5G network are evolved to SDN enabled CN entities. We refer to them as SeMMu. The reason for such an integration being that the MME and SMF are responsible for the CN signaling during a mobility event in their respective networks. Hence, this allows the SeMMu to facilitate the proposed HO signaling mechanism. Further, in this work, instead of the the AMF (which 3GPP defines as the mobility management unit in the 5G NGC), we exploit the idea of SDN-enabled SMF because it is the SMF which is involved in HO-related CN signaling, whereas the AMF is connected to the access network only. Thus, the HO-related CN signaling is not influenced by the AMF. Accordingly, in this chapter, we propose such evolutionary network architecture, by also describing the implementation aspects of such SeMMu entities.

Through this chapter, we propose an evolutionary architecture, considering the co-existence of the legacy networks (4G, 3G, 2G) and the newly proposed 5G networks by 3GPP, that enables a manageable CAPEX for the operators whilst also enhancing the handover preparation and failure phase performance. To summarize, in the current work we advance the state of the art by:

- Introducing enhanced HO preparation and failure signaling phases for the myriad 5G and legacy networks inter-RAT HO scenarios.
- Introducing a novel HO failure aware preparation phase signaling sequence. This approach will provision additional optimization to the legacy HO failure signaling step as well as the HO preparation phase.
- Presenting performance analysis, based on latency, transmission cost and processing cost, of the proposed and legacy signaling mechanisms for the myriad 5G and legacy network HO scenarios specified by 3GPP.
- Introducing a novel message size analysis for the proposed as well as legacy HO scenarios.
- Introducing a novel network wide analysis in terms of the number of messages processed as well as the total network occupation time.
- Presenting a novel 5G NGC and legacy inter-working architecture with and without the capabilities of an N26 interface. Note that, the N26 interface, defined by 3GPP, allows for the inter-working between the 5G and legacy networks. It allows for reduced signaling to prepare or reject a HO between 5G and LTE-EPC networks.
- Presenting a novel interfacing mechanism between the MME/SMF and the SDN agent.

5.1 Legacy Handover Preparation and Failure Signaling

Erstwhile standardization efforts by 3GPP [28, 29, 133, 164–172] have led to the formulation of the handover signaling mechanisms currently being utilized in cellular networks [133] and to be used in future wireless networks [29]. Specifically, handover preparation and failure signaling phases will be critical to the overall system performance during mobility events. Figures 5.2, 5.3 and 5.4 illustrate the corresponding legacy handover preparation and failure signaling phases.

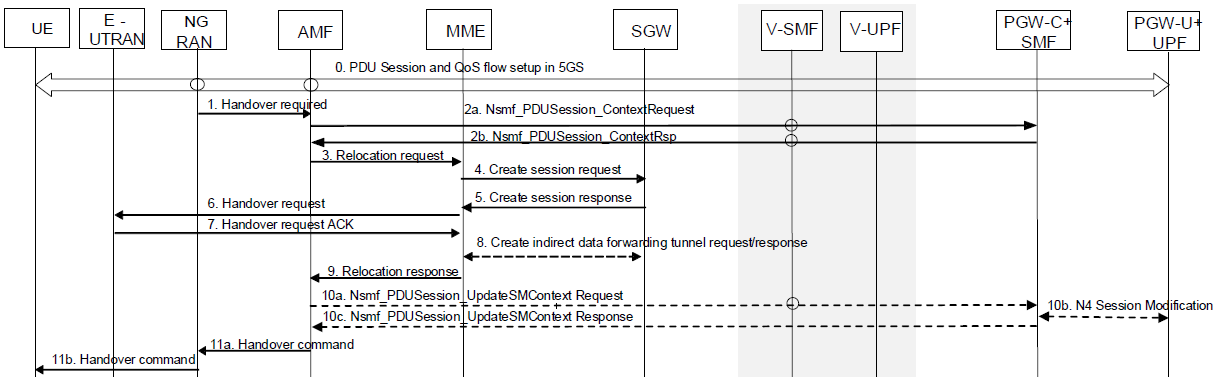


Figure 5.2: Legacy handover preparation signaling for Inter-RAT HO (5G-NGC to EPS) [29].

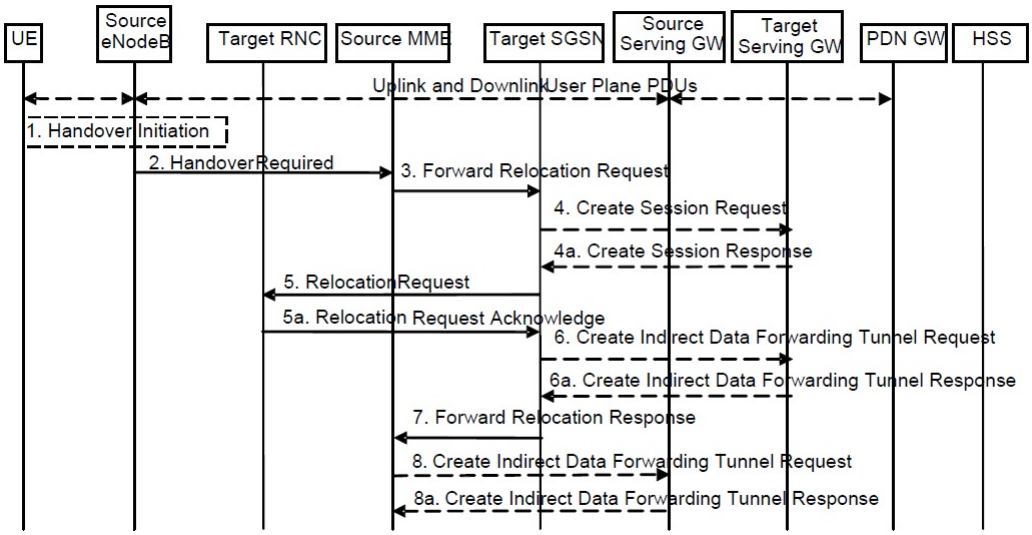


Figure 5.3: Legacy handover preparation signaling for Inter-RAT HO (LTE to 3G/2G) [133].

The legacy handover preparation signaling, exemplified in Figure 5.2 and 5.3, is initiated by a *handover decision* made by the source network. This is followed by a *handover required* message (#1 in Figure 5.2, and #2 in Figure 5.3). Following these initial stages, the handover preparation phase is comprised of resource negotiation and allocation through the RRM operations (messages 6 and 7 in Figure 5.2; messages 5 and 5a in Figure 5.3), as well as CP signaling to establish GTP tunnels. These GTP tunnels require the entities at either end of the tunnel to have the TEIDs and transport layer addresses of each other. Hence, the preparation step also encompasses the creation and exchange of TEIDs and transport layer addresses between the core network entities. In order to realize a successful handover preparation, handshakes between the core network entities, i.e., messages 4, 5, 8 and 10b in Figure 5.2; messages 4, 4a, 6, 6a, 8 and 8a in Figure 5.3, are required. Next, the legacy

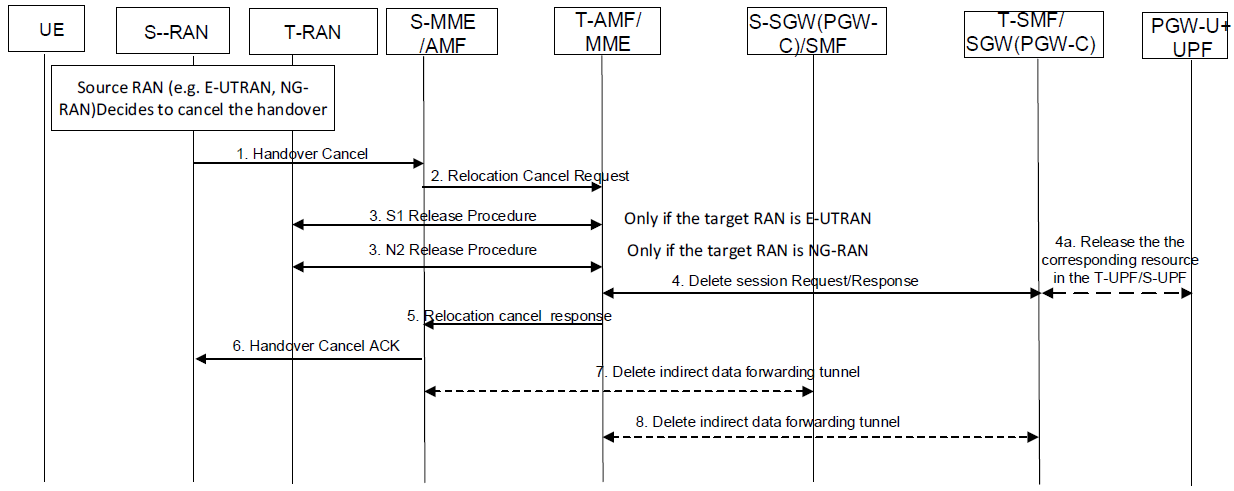


Figure 5.4: Legacy handover failure signaling for Inter-RAT HO (5G NGC and EPS) [29].

handover failure phase signaling for inter-RAT HO (5G to EPS[1]) has been illustrated in Figure 5.4. For the 5G NGC, the HO failure phase signaling currently only includes the HO cancel mechanism. In 3GPP specifications, the handover failure phase signaling encompasses two different types of signaling methods, i.e., Handover Cancel and Handover Rejection. We define them as follows:

- *Handover Cancel:* A handover cancel mechanism has been defined both for the 5G NGC as well as for the legacy networks, i.e., 4G, 3G and 2G. The cancel method is event based, i.e., it is initiated by a trigger event such as expiration of a timer, etc. It may be invoked only by the source network and at any point before the command to handover from the source network to the target network is sent from the MME/AMF to the source eNB/next generation NodeB (gNB).

- *Handover Reject:* A handover reject mechanism is currently defined only for the legacy networks, i.e., 4G, 3G and 2G networks. Similar to the handover cancel phase, the handover rejection method is event based, i.e., it is triggered by an event such as failure to allocate sufficient resources at the target access network. Hence, upon reception of a rejected request to reserve resources, the target MME/SGSN informs the source eNB/RNC about the rejected requested and hence, a handover reject.

And so based on the above definitions, due to certain network conditions, such as the expiration of a timer, etc., the source network may decide to cancel the HO (Figure 5.4).

[1]The EPS consists of EPC and E-UTRAN. Note that, the standard documents by 3GPP utilize EPS and EPC interchangeably while defining HO scenarios. Hence, in this chapter we utilize the same principle.

Consequently, the source MME/AMF informs the source base station with regards to the canceled handover (message 1 in Figure 5.4). Further, the source and target MME/SMF delete the sessions that had already been created with the target and source S-GWs/UPFs (messages 4, 4a, 7 and 8 in Figure 5.4), respectively. The creation and deletion of these sessions with other core network entities involves handshakes, which, as we will discuss in the following subsection (5.1.1), are a significant source of inefficiency in CN handover signaling.

Note that the signaling schemes illustrated through Figures 5.2, 5.3 and 5.4 are representative and other HO preparation and failure phase scenarios explored by 3GPP in [29, 133] are also of the same nature, wherein handshakes are utilized to accomplish the signaling procedures.

5.1.1 Signaling Inefficiency

From our discussions and Figures 5.2, 5.3 and 5.4, it can be deduced that during the legacy handover preparation and failure phases, handshakes will be required to exchange the required CP information between the core network entities. For instance, in the handover preparation phase in Figure 5.2, to establish a session between the Target MME and the Target S-GW a handshake, i.e., messages 4 and 5, is required between these respective entities. Such handshakes, whilst being a reliable methodology, will occupy the network for a long period as opposed to a mechanism that does not involve any handshakes. Further, it will also lead to higher latency, signaling cost, processing cost and total bytes of data transferred. And given the future network scenario depicted in Figur 5.1(b), wherein the network will be dense and heterogeneous, the legacy mechanisms will be rendered inefficient. Thus, we define a new principle that is utilized to create a compressed message ensemble and an enhanced signaling method, showing increased performance in terms of latency, signaling cost, processing cost and total amount of bytes transferred. The principle is as follows:

"Identify the sequence of messages, such as the handshakes, where the performance of the 3GPP defined methods can be improved/enhanced. Then re-shuffle the information elements (IEs), if possible, to form a compressed message ensemble such that the sequence of messages under scrutiny are executed efficiently, if possible in parallel, but with the desired functionality."

And so, we next discuss the novel message mapping and signaling strategies that alleviate the deficiencies mentioned above.

Table 5.1: Different handover scenarios analyzed

		Target Network		
		5G NGC†	**EPC**	**3G/2G**
Source Network	**5G NGC†**	N2 based HO: UE migrates from one NG-RAN to another	Inter-RAT HO involving an N26 interface	
			Inter-RAT HO without an N26 interface	
	EPC	Inter-RAT HO involving an N26 interface	Intra-RAT HO involving MME relocation but no S-GW relocation	Inter-RAT HO involving S-GW relocation and an Indirect tunnel
				Inter-RAT HO involving S-GW relocation and a Direct tunnel
		Inter-RAT HO without an N26 interface	Intra-RAT HO involving MME and S-GW relocation	Inter-RAT HO without any S-GW relocation but with an Indirect tunnel
				Inter-RAT HO without any S-GW relocation but with a Direct tunnel
	3G/2G		Inter-RAT HO involving an S-GW relocation and Indirect tunneling	
			Inter-RAT HO without any S-GW relocation but with Indirect tunneling	

†3GPP standards document only discuss 5G NGC to EPS HO and vice versa.

5.2 Proposed Handover Preparation and Failure Signaling

The proposed handover preparation and failure phases consist of a novel message mapping and signaling mechanism, wherein a compact and intelligent mapping of IEs from the legacy to the proposed signaling messages has been provided. Additionally, the proposed mechanism

also involves parallel transfer of CP information. To facilitate these capabilities, we utilize the SDN enabled CN entities including the SeMMu, defined earlier. Specifically, the SeMMu through its SDN capabilities and centralized location facilitates:

- Parallel transfer of CP information to other CN entities.
- Allocation of TEIDs and transport layer addresses for the other CN entities at the SeMMu itself.

Thus, through the use of compact message ensemble and parallelization of information transfer, the handover preparation and rejection signaling for the various 3GPP HO types are, as we will discuss in this section, optimized. The HO scenarios that have been analyzed in this work are presented in Table 5.1. Concretely, the various networks, i.e., 2G, 3G, 4G-LTE and 5G, have been considered in this table and all the possible HO scenarios among them have been enlisted.

To evince the optimization achieved for the aforementioned 3GPP HO scenarios, we consider a representative HO scenario, i.e., Inter-RAT HO from 5G NGC to EPS network wherein the serving gateway is relocated. The optimized/enhanced message maps and signaling sequence for other scenarios have been illustrated in the *Appendix A*.

And so, for the representative scenario considered, a user undergoes an Inter-RAT handover with the source system being 5G and the target system being an EPS network. Further, serving gateway relocation defines that during the handover process the gateway that is serving the user is changed. In this particular scenario, the gateway in the source network is a UPF which upon handover to the EPS is switched to a target S-GW. Also, note that the considered scenario consists of an N26 interface which facilitates the inter-working between the NGC and EPS. However, in the analysis, we have presented results for the scenarios wherein the N26 interface does not exist. Such a scenario will be prevalent in the initial 5G standalone (SA) deployments, i.e., an interface between 5G NGC and EPS will be missing in some of the initial deployments, due to cost/compatibility reasons. Subsequently, we also provision the proposed HO signaling diagrams for this scenario in Appendix A.

Note that, in addition to the proposed signaling method, in this section we also present a novel handover failure aware handover preparation method. This novel approach enhances not only the handover failure method but also improves the handover preparation method further. We defer the detailed discussion on this method to Section 5.2.3.

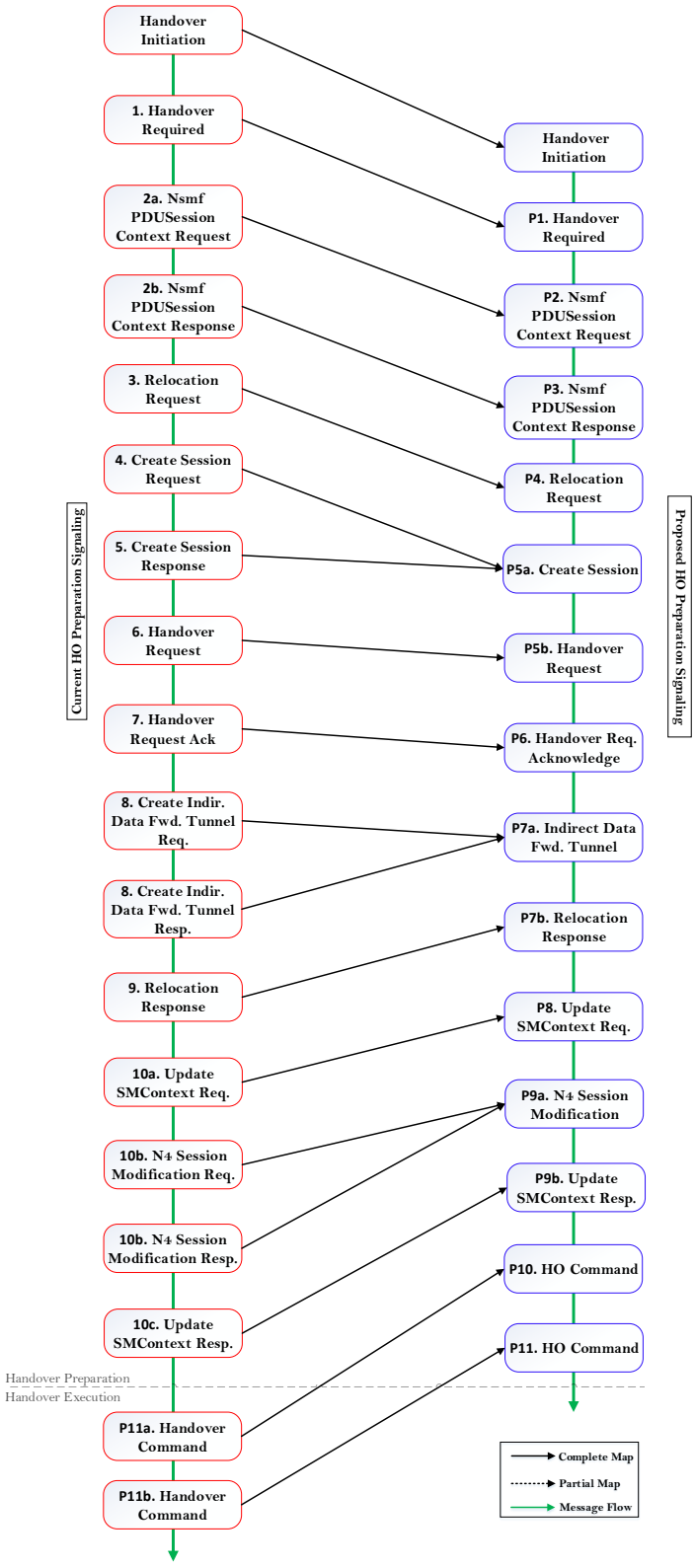


Figure 5.5: Proposed Handover Signal mapping for Inter-RAT HO from 5G NGC to EPS.

5.2.1 HO Preparation: Optimal Message mapping and Signaling

An illustration of the proposed message mapping and signaling sequence for the handover preparation phase of the representative HO scenario has been presented in Figures 5.5 and 5.6, respectively. By message mapping (Figure 5.5), here we refer to a graphical representation of how messages from the legacy message ensemble are mapped to the proposed message ensemble. Specifically, during the mapping process, IEs, which are the building blocks of these messages, are re-organized in a way that helps to reduce the message ensemble. This consequently also aids in an improved performance in terms of latency incurred, transmission cost, processing cost and the overall number of bytes transferred, as will be seen in Section 5.3. It must be stated here that, the mapping process is performed without transforming the format and contents of the IEs, as it ensures an evolutionary approach that is easy and fast to adapt for the operators and vendors alike. Next, the signaling sequence presented in Figure 5.6, is an illustration of the sequence in which the messages from the proposed ensemble are executed.

Additionally, and before delving deeper into the discussion with regards to the mapping and signaling process, it must be noted that throughout the text Legacy messages are assigned only numbers, while the proposed mechanism messages are assigned numbers beginning with letter "P". e.g. a legacy message would be numbered as *7*, while a proposed mechanism message would be numbered as *P6*.

And so, from Figure 5.5, it can be observed that the proposed message mapping reduces the size of the message ensemble to 15 as compared to the 18 required during the legacy handover preparation signaling. In the legacy and proposed message ensembles, the *Handover Command* messages (P11a-11b and P10-11 respectively) have also been included instead of considering them in the handover execution phase. This is so because, unless the RRM information from the target network is delivered by the source network to the user, from the user's perspective the network is still in a handover preparatory phase. Next, in the message mapping presented in Figure 5.5, the *Handover Initiation*, *Handover Required* and *HO Command* messages are left unchanged from the legacy message ensemble. Concretely, the IEs in the aforementioned messages are left unchanged from the legacy mechanism. Further, and in accordance with the discussion in Section 5.1, the messages involving RRM operations are left unaltered as the primary aim of the proposed strategy is to reduce the CN signaling during the handover process. Hence, *Handover Request* and *Handover Request Acknowledge* messages (6 and 7) are unaltered. The messages that are modified (enhanced) have been explored below:

- *Create Session (P5a):* This message is composed of the IEs from messages 4 and 5

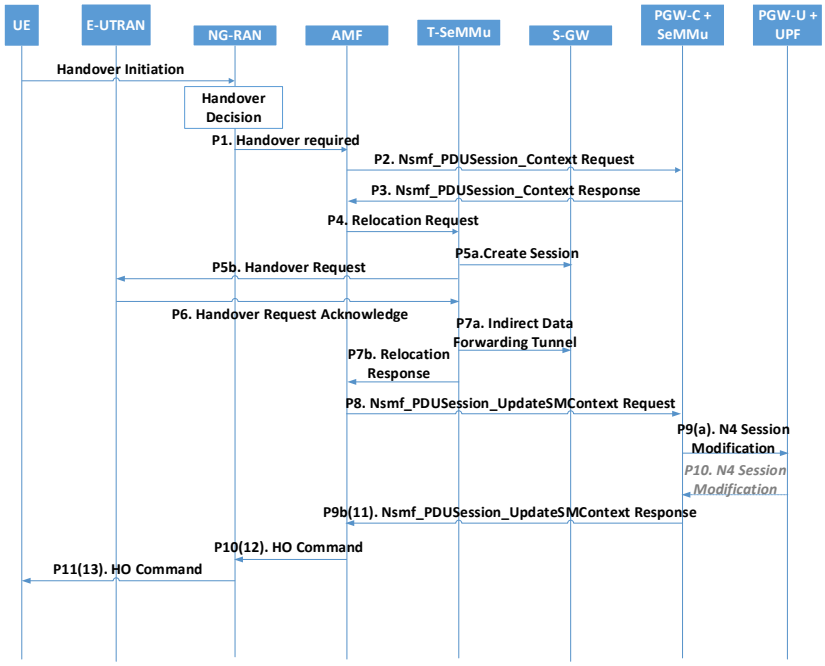


Figure 5.6: Proposed Handover signaling sequence for Inter-RAT HO from 5G NGC to EPS.

both. Whilst through message 4 the SeMMu provides the S-GW with the necessary information about the PDN connections that are going to be handed over, message 5 allows P5a to allocate the S-GW its own resources such as TEID and transport layer addresses. This re-arrangement of the IEs helps to eliminate the requirement of a handshake, which consequently enhances the handover preparation phase signaling.

- *Indirect Data Forwarding Tunnel (P7a):* The given message is composed of the two sub–messages that are contained within the handshake labeled message 8 in the signaling defined by 3GPP [29]. The forward message of the aforesaid handshake enables the MME to specify the TEID(s) and address(es) for the indirect data forwarding tunnel to the S-GW. Next, the S-GW specifies its own TEID(s) and address(es) for the indirect data forwarding tunnel in the response part of the specified handshake (message 8) to the MME. However, the proposed message mapping enables the SeMMu to allocate the required TEIDs and transport layer addresses, including that of the S-GW itself, for the indirect forwarding tunnel without the requirement of a handshake. Concretely, the IEs from messages involved in handshake 8 are re-organized such that the TEID(s) and address(es) of the other CN entities, as well as of the S-GW itself, for the indirect data forwarding tunnel are specified to the S-GW in a single step, i.e., through message P7a.

- *N4 Session modification (P9a):* In the signaling specified by 3GPP, the N4 modification request message (#10b in Figure 5.2) permits the SMF to apprise the UPF about the TEID(s) and address(es) of the S-GW for setting up a data forwarding tunnel. Further, the UPF responds to this message with an N4 modification response message (#10b in Figure 5.2) with its own CN tunnel info consisting of TEID(s) and address(es). However, in our proposed approach, the IEs of the aforesaid modification request/response messages are mapped to message P9a, such that the TEID(s) and address(es) of the S-GW and UPF for the data forwarding tunnel are specified to the UPF. This eliminates the handshake and hence, enhances the handover signaling process.

Next, in the proposed handover preparation signaling, presented in Figure 5.6, the sequence and operation of *Handover Initiation, Handover Required* and *Handover Command* messages remains unaltered from the legacy signaling [29]. Further, the messages that are associated with the RRM operations, i.e., *Handover Request* and *Handover Request Acknowledge*, also remain unaltered in their operation. However, these messages (6 and 7 in the legacy signaling, i.e., Figure 5.2) have been re-assigned as messages P5b and P6 in the proposed signaling approach. Additionally, utilizing the already stated capability of parallel information transfer through the SDN agent on the SeMMu, messages P5a-5b, P7a-7b and P9a-9b, are executed simultaneously pairwise. Lastly, the HO command message is forwarded by the AMF to the source NG-RAN (S-NG-RAN). The S-NG-RAN then forwards it to the UE, marking the end of the HO preparation phase.

5.2.2 HO Failure: Enhanced process

Recall from our discussion in Section 5.1 that, the handover failure phase signaling consists of two methods, i.e., Handover Cancel and Handover Rejection. Considering the representative HO scenario, i.e., 5G NGC Inter-RAT HO scenario, the enhanced handover cancel phase signaling for the same has been illustrated in Figure 5.7. Concretely, we utilize the principle used in Section 5.2.1 to compress the message ensemble and enhance the signaling process for the HO cancel phase as well.

In the proposed signaling presented in Figure 5.7, the source RAN firstly decides to cancel the HO. Thus, following this decision, a *HO cancel* message (P1) is sent to the AMF, which then issues a *Relocation Cancel request* message to the T-SeMMu. The T-SeMMu then utilizes its SDN capabilities to simultaneously:

- Delete the sessions it created during the HO preparation phase via the *Delete Existing Session* message (P4a).

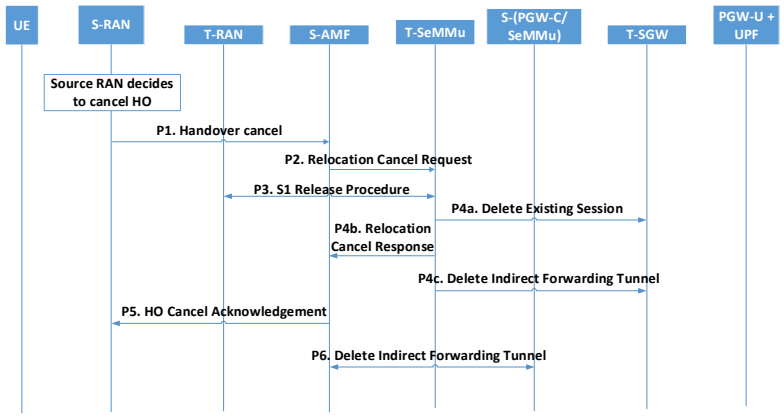


Figure 5.7: Proposed Handover cancel phase signaling for Inter-RAT HO from 5G NGC to EPS.

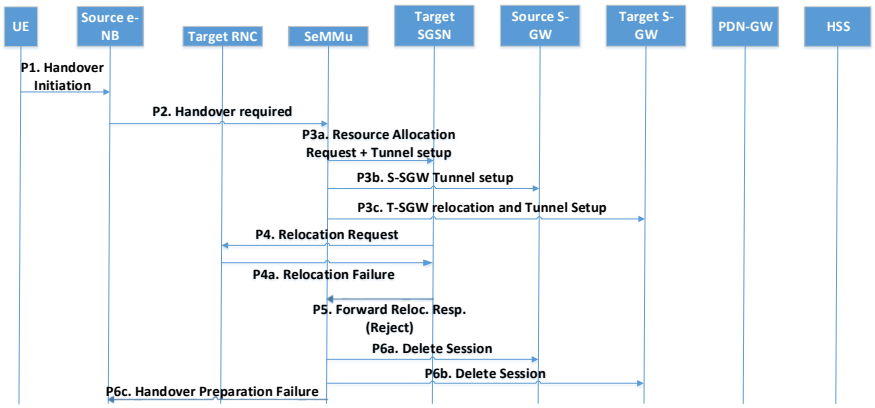


Figure 5.8: Proposed Handover rejection phase signaling for Inter-RAT HO from LTE to 3G/2G network when there is a S-GW relocation and indirect tunneling exists.

- Issue a *Relocation Cancel Response* message (P4b) to the source S-AMF (Source AMF).
- Delete the indirect forwarding tunnels that it created during the HO preparation phase via the *Delete Indirect Forwarding Tunnel* message (P4c).

Concretely, the aforementioned parallelization of messages provides the claimed optimization in the HO cancel signaling phase. Subsequently, the S-AMF sends a HO cancel acknowledgement message (P6) to the source RAN. Lastly, the S-AMF also performs a handshake (P7) with the source SeMMu for the deletion of any indirect tunnels that were setup during the HO preparation phase in the 5G NGC. Since, the AMF is not equipped with the capabilities of allocating TEID(s) and address(es) like the SeMMu, it has to perform the

handshake P7, instead of transferring just a single message with all the CP information to the source SeMMu.

Recall from our discussion in Section 5.1 that for the LTE-EPC to 3G/2G and vice versa handover scenarios, both the handover cancel and handover rejection methods are defined [133]. However for analytical reasons, elaborated in Section 5.3, in this work we only consider the HO rejection phase signaling for this case scenario. Further, we utilize the same principle as the HO cancel phase and HO preparation phase (Section 5.2.1), for the HO rejection phase signaling enhancement.

As a representative example, through Figure 5.8, we present the enhanced HO rejection signaling for LTE-EPC to 3G/2G HO scenario when there is an indirect tunnel and a S-GW relocation exists. Briefly, after the initiation of the Handover required message (P2), the SeMMu communicates with the Target SGSN for allocating resources as well as setting up the tunnel (message P3a). Simultaneously, the SeMMu also sets up tunnels with the Source and Target Serving GWs through messages P3b and P3c, respectively. Next, the Target SGSN requests the Target RNC to setup access network resources for the impending handover via the Relocation Request message (P4).

However, message P4a indicates to the Target SGSN that a HO is being rejected (possibly due to lack of physical layer resources) and hence, forwards the same indication to the SeMMu in the *Forward relocation response* message. Following this message, the SeMMu deletes the sessions it had created during messages P3b and P3c with the Source and Target S-GWs respectively. The SeMMu performs this operation via messages P6a and P6b. To conclude the HO rejection phase, the SeMMu parallely also issues a *Handover Preparation Failure* to the source eNB.

And so, based on these proposed signaling improvements, in Section 5.2.3 we present the novel *Handover Failure aware Preparation signaling* process.

5.2.3 Handover Failure aware preparation signaling

For the HO cancel phase presented in Figure 5.7, it can be observed that the *Delete Session* and *Delete Indirect Tunneling* messages are sent to the target S-GW and source UPF by the target and source SeMMus respectively, as a consequence of the sessions that were established in messages P5a, P7a and P9a (Figure 5.6). These messages release the CN resources that are reserved by the SeMMu during the HO preparation phase. Further, from [29] it is understood that the HO cancel phase can be executed at any point within the HO preparation phase before the *HO Command message*, or even after the HO preparation phase if the UE fails to attach or register to the target network.

With this background, and through Figures 5.6, 5.7 and 5.10, it must be noted that if the HO cancel phase is executed before message P5a (Figure 5.6) is sent, then it will be fully optimal, i.e., there will be no requirement for handshakes to delete the tunnels. This is so because, there will be no CN resources that would have been allocated during the HO preparation steps at that point, and hence, there will be no requirement of messages P4a, P4c and P7 (Figure 5.7) to release those resources. However, if the HO cancel phase is executed at any time instance after the execution of message P5a (Figure 5.6) within the HO preparation phase, then the cancel phase signaling will require messages that will help release the allocated CN resources. Thus, the proposed HO cancel phase signaling presented in Figure 5.7 is capable of adapting itself depending on when it is initiated, and as a consequence we consider it to be near-optimal.

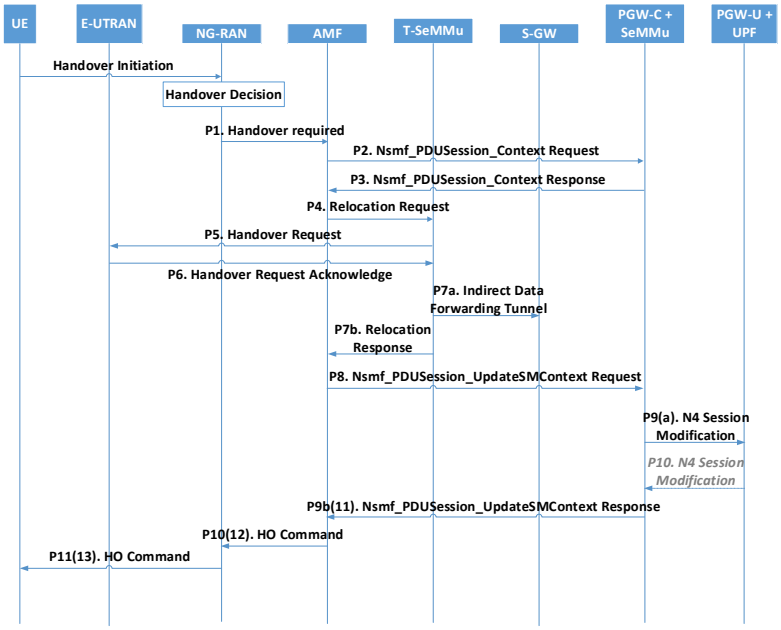


Figure 5.9: Handover failure aware Handover preparation Signaling for Inter-RAT HO from 5G NGC to EPS.

Upon performing a deeper analysis into the IEs that constitute messages P5a, P7a and P9a (Figure 5.6), we observe that the IEs of messages P5a and P7a can be re-shuffled and mapped to create a single message P7a, i.e., *Indirect Data forwarding tunnel*, as shown in the Handover failure aware preparation phase signaling (Figure 5.9). Further, the new message P7a is executed after messages P5 and P6 (Figure 5.9), which facilitates the HO cancel phase with a greater chance of achieving its optimal state presented in Figure 5.10. The reason being, the later the tunnels are setup, the higher is the chance of a HO cancel event not

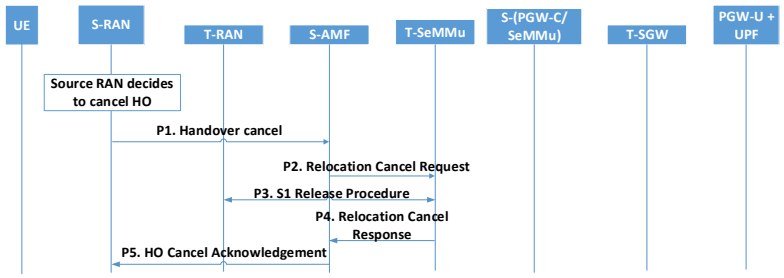


Figure 5.10: Optimal proposed Handover rejection phase signaling sequence for Inter-RAT HO from 5G NGC to EPS.

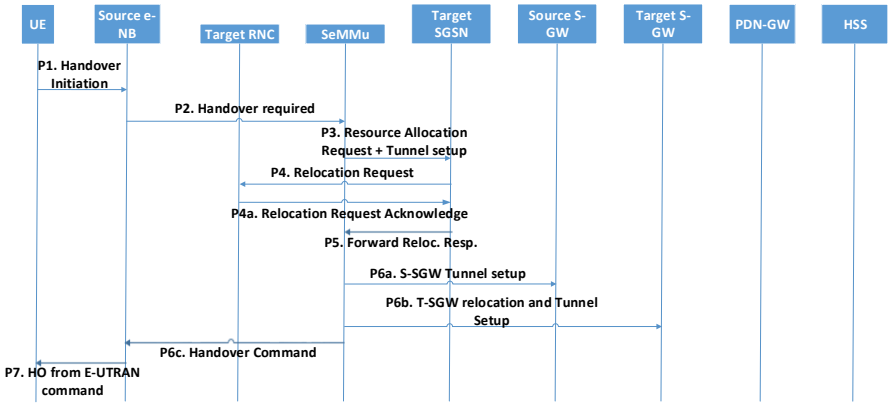


Figure 5.11: Handover failure aware Handover preparation Signaling for Inter-RAT HO from LTE-EPC to 3G/2G when there is indirect tunneling and S-GW relocation occurs.

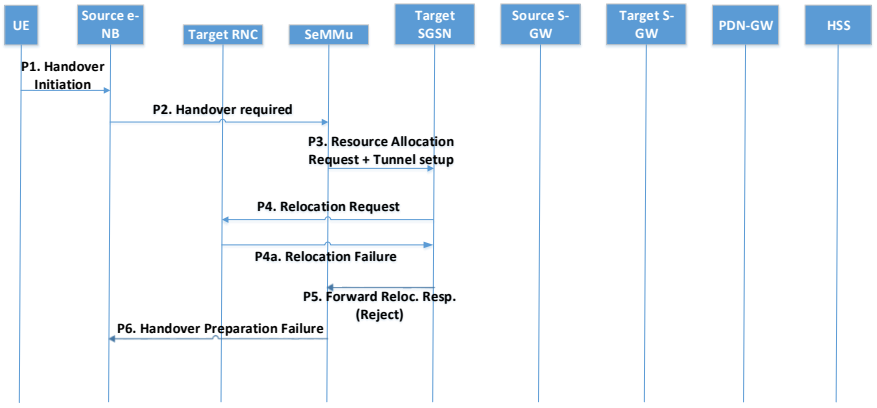


Figure 5.12: Optimal proposed Handover rejection phase signaling sequence for Inter-RAT HO from LTE-EPC to 3G/2G when there is indirect tunneling and S-GW relocation occurs.

requiring to delete these tunnels as it may be initiated before they are setup. Additionally,

and in the event, the HO cancel phase is executed after message P9a (Figures 5.6 and 5.9), the aforesaid enhancement would require the HO cancel signaling to delete 2 tunnels (created through messages P7a and P9a in Figure 5.9) instead of the 3 (created through messages P5a, P7a and P9a in Figure 5.6). However, given the dynamic nature of HO cancel phase, in the analysis we only consider the enhanced signaling specified in Figure 5.7. This is also the worst case HO cancel phase scenario as it is executed after all the tunnels have been setup.

Additionally, utilizing this novel signaling approach, the handover rejection phase for the LTE to 3G handover scenario, illustrated in Figure 5.8, has been further enhanced (Figure 5.12). To achieve this enhancement, the HO preparation phase for the given scenario is first modified (Figure 5.11) such that all the tunnel and session creation messages are executed after *Relocation Request Acknowledge* message (5a in Figure 5.3). The reason being, in the event there is a HO rejection, the *Relocation Request Acknowledge* message issues an indication of the rejection to the SGSN. The SGSN then passes this indication to the SeMMu, which instantly passes the reject indication to the source eNB, without the requirement of any session and tunnel deletion messages. Hence, the HO rejection phase signaling is further enhanced as compared to signaling proposed in Figure 5.8. Further, given that the resource allocation process is successful and a positive indication is received from the *Relocation Request Acknowledge* message, the tunnel and session creation messages are executed simultaneously with the *HO command* message. This parallel execution with the *HO command* grants additional enhancement to the HO preparation phase as it helps to reduce the latency further.

5.2.4 Xn, X2 and S1 Interface based Handover Signaling

As an evolution from the EPC architecture, the 5G NGC specifies an Xn interface between two gNBs. In scenarios, wherein the two gNBs involved in the HO process are connected via an Xn interface, the given interface facilitates a faster handover process. Subsequently, upon deeper exploration of the handover preparation signaling mechanism for an Xn based HO from [173], it is evident that the existing mechanism is optimal. Concretely, since the signaling does not involve any handshakes and any significant interaction with the CN entities, the proposed handover preparation signaling process will neither provide any gains nor will it lead to any regressive effects on the performance of the Xn based HO mechanism.

Further, the LTE-EPC defines two specific interfaces through which the Intra-RAT handovers can be executed, i.e., the X2 and the S1 interface [133]. Whilst, the X2 interface is defined between two eNBs (like the Xn between two gNBs in 5G NGC), the S1 interface involves the CN. The legacy X2 and S1 handover preparation signaling mechanisms have

been presented in [174]. Analyzing the signaling mechanisms presented in [174], it can be concluded that the existing X2 and S1 handover mechanisms, similar to the Xn handover mechanism for 5G NGC, are optimal. Concretely, while the X2 HO is similar to the Xn HO in 5G NGC (wherein the HO signaling is only at the access network), for the S1 HO the only CN signaling is that between the MME (SeMMu) and the source and target eNBs (wherein RRM operations take place). Hence, both X2 and S1 HO signaling scenarios are considered to be optimal, as stated above.

Note that, here for the S1 handover we consider the scenario wherein the user does not switch its MME (SeMMu) and S-GW. In the event either is changed, the proposed handover mechanism leads to immediate gains, which have been presented via the analysis in Section 5.3. Additionally, it is important to state that the Xn, X2 and S1 interfaces are an integral part of the evolutionary network architecture proposed in this work, and are agnostic to the implementation of the architecture and signaling methodology.

Thus, with the aforesaid principles, processes and methodologies, in the following section we present a detailed quantitative performance improvement analysis for the myriad scenarios that have been listed in Table 5.1.

5.3 Performance Analysis

To analyze the proposed handover preparation and failure signaling phases, we use latency, processing cost, transmission cost and amount of bytes transferred within CN as evaluation metrics. Note that, as in [175], utilizing these metrics for evaluating new handover strategies is standard practice.

5.3.1 Analytical Formulation

For the analysis we first define a set $S = \{s_1, ..., s_N\}$ corresponding to all the link delays encountered within a given signaling sequence. Then the set $J = \{j_1, ..., j_K\}$ is the set of all the parallel link delays, where $K \leq N$. By *Parallel link delay* we mean that, if x messages are to be executed simultaneously then the overall delay incurred will be the maximum of the delays experienced by the messages under observation. It is then computed as

$$\begin{aligned} \text{Parallel Link Delay} \;=\; & \max(\textit{Link delay msg 1}, \ldots, \\ & \textit{Link delay msg } x), \end{aligned} \tag{5.1}$$

Additionally, we consider only a single processing delay for the group of messages that are being executed in parallel. Hence, the set of processing delays can be defined as $D = \{d_1, ..., d_K\}$. It is imperative to state here that the assumption for the aforesaid processing delays, mentioned in Section 5.3.2, is conservative in nature. Hence, any SDN agent processing delay is also included within the utilized assumptions. Also, in HO procedures no routing table updates would be necessary, as they are already configured during the network setup phase, hence no signaling overhead is created by SDN agents for HOs.

And so, for the computation of latency incurred during the handover preparation and handover failure signaling phases, we consider the contributions from parallel link delays and processing delays as:

$$\text{Latency} = \sum_{i=1}^{K} \{j_i + d_i\} \tag{5.2}$$

The Transmission Cost computation, on the other hand, requires that each link delay be considered for the evaluation, and is computed as

$$\text{Transmission Cost} = \frac{\sum_{l=1}^{N} s_l}{1\text{ms}}. \tag{5.3}$$

Concretely, the Transmission Cost analysis represents the amount of time the CN links are occupied during the complete HO signaling process. Next, for the processing cost analysis, we utilize the analytical methodology in [134] and define it as the number of messages generated during the HO preparation/failure signaling phase. We then compute the percentage processing cost saving as

$$\text{Proc. Cost Saving} = \frac{MSG_{Legacy} - MSG_{Proposed}}{MSG_{Legacy}} * 100\%, \tag{5.4}$$

wherein, MSG_{Legacy} is the number of messages in the legacy approach for HO preparation/failure and $MSG_{Proposed}$ is the number of messages in the proposed approach for HO preparation/failure.

In addition, in this chapter we also present an analysis for the network wide processing cost and occupation time. Whilst the network wide processing cost will reflect the network wide reduction in the processing cost through the proposed method, the network occupation time will be reflective of the reduction in the amount of time the CN links are occupied due to the handover signaling across the network.

To conduct the aforesaid analysis we introduce the formulations in equations (5.5) and (5.6). $\text{NPC}_{DHfS_{1p}}$ and $\text{NOcT}_{DHfS_{1p}}$ are the Network wide processing cost and Network occu-

pation time, respectively, given a HO distribution (distribution of percentage of total users undergoing a particular handover), HO failure (rejection/cancellation) rate and percentage of users undergoing S1 HO, respectively. Note that, we consider only the Intra-MME/S-GW scenario for S1 HOs as it is not impacted by the implementation of the proposed mechanism but it still involves CN signaling. Moreover, we do not consider the X2 and Xn handovers for the analysis, as they do not involve any significant CN signaling. The rest of the notations in (5.5) and (5.6) are as follows: $\mathbf{H_{pscost}}$ is Handover preparation processing cost vector; $\mathbf{Dist_{HO}}$ is the handover distribution vector; HO_{sperc} is the handover success percentage; $\mathbf{H_{fcost}}$ is the processing cost vector during Handover failure; HO_{fperc} is the HO failure percentage; N_{HO} is the number of users undergoing handover in the network; S_{1p} is the percentage of S1 handovers (Intra-MME/S-GW); S_{1scost} is the processing cost for a successful S1 HO preparation; S_{1fcost} is the processing cost for a failed S1 HO; $\mathbf{H_{tscost}}$ is the transmission cost vector for a successful HO preparation; $\mathbf{H_{tsfcost}}$ is the transmission cost vector during a HO failure scenario; $S_{1tscost}$ is the transmission cost for a successful S1 HO preparation; and $S_{1tsfcost}$ is the transmission cost incurred when a S1 HO fails.

5.3.2 Parameter Specification and Assumptions

As part of the analytical framework, the parameter values that will be utilized to conduct the analysis are provided in this subsection. Firstly, the one-way delays for each CN link, necessary for the latency and transmission cost analysis, have been defined in Tables 5.2 and 5.3 by utilizing the data from a Japanese cellular operator [176], Cisco [177] and a Greek cellular operator. Further, the delays presented for each link are considered to be symmetric, i.e., if a delay of 1ms is incurred for the link from AMF to SeMMu, then the same link delay is assumed from SeMMu to AMF.

In Table 5.2 the link delays presented are derived from a Japanese operator deployment data [176] and CISCO data [177]. In addition, the link delays are computed considering

$$\begin{aligned} \mathrm{NPC}_{DH_fS_{1p}} &= \left\{(\mathbf{H_{pscost}} * \mathbf{Dist_{HO}}^{\mathbf{T}}) * HO_{sperc} + (\mathbf{H_{fcost}} * \mathbf{Dist_{HO}}^{\mathbf{T}}) * HO_{fperc}\right\} * (1 - S_{1p}) * N_{HO} +, \\ & S_{1p} * N_{HO} * \left\{S_{1scost} * HO_{sperc} + S_{1fcost} * HO_{fperc}\right\} \end{aligned} \tag{5.5}$$

$$\begin{aligned} \mathrm{NOcT}_{DH_fS_{1p}} &= \left\{(\mathbf{H_{tscost}} * \mathbf{Dist_{HO}}^{\mathbf{T}}) * HO_{sperc} + (\mathbf{H_{tsfcost}} * \mathbf{Dist_{HO}}^{\mathbf{T}}) * HO_{fperc}\right\} * (1 - S_{1p}) * N_{HO} +, \\ & S_{1p} * N_{HO} * \left\{S_{1tscost} * HO_{sperc} + S_{1tsfcost} * HO_{fperc}\right\} \end{aligned} \tag{5.6}$$

Table 5.2: Link Type and Corresponding Delays in Proposed Architecture (Derived from a Japanese Operator [176] and Cisco data [177]

	Link Type	**Link Delay**
1.	UE to NG-RAN	1ms
2.	NG-RAN to AMF	7.5ms
3.	AMF to SeMMu (PGW-C + SMF)	1ms
4.	AMF to SeMMu	1ms
5.	SeMMu to S-GW	7.5ms
6.	SeMMu (PGW-C + SMF) to PGW-U + UPF	7.5ms
7.	SeMMu (PGW-C + SMF) to PCRF+PCF	7.5ms
8.	AMF to AMF	15ms
9.	SeMMu to PGW	7.5ms
10.	SeMMu to E-UTRAN	7.5ms
11.	E-UTRAN to UE	1ms
12.	PGW to PCRF	7.5ms
13.	S-GW to PGW	7.5ms
14.	SeMMu to SGSN	1ms
15.	SGSN to RNC	6ms
16.	SGSN to S-GW	7.5ms
17.	SeMMu to SeMMu	15ms

that the MME (SeMMu in this study) and the SGSN are co-located, as specified in [178]. Utilizing this co-location principle, we also establish the link latency between AMF and SeMMu. Further, the 15 ms SeMMu-SeMMu and AMF-AMF delay is based on the premise that the delay between the SeMMus/AMFs will be greater than the largest CN delay within a SeMMu/AMF domain. Hence, for the purpose of analysis in this chapter and for the data provided from the Japanese operator and Cisco, an assumption of two times the greatest link delay within a SeMMU/an AMF domain has been considered.

On the other hand, the values of delays obtained from the Greek operator (Table 5.3) correspond to eNBs from two different networks and CN elements from 3 different MME domains. Consequently, for the chosen network and its MME domain, the link delays are computed as the average of all the delay values provided by the network operator for that specific link. Further, the UE-eNB and the eNB-SeMMu delay for both data sets is derived from the Cisco framework in [177]. Additionally, for the latency analysis, we consider the processing delay to be 4 ms in all CN entities, as in [177].

For the network wide analysis, we consider that the number of users undergoing handover

Table 5.3: Link Type and Corresponding Delays in Proposed Architecture (Derived from a Greek Operator and Cisco data [177])

	Link Type	**Link Delay**
1.	UE to NG-RAN	1ms
2.	NG-RAN to AMF	19ms
3.	AMF to SeMMu (PGW-C + SMF)	0.5ms
4.	AMF to SeMMu	0.5ms
5.	SeMMu to S-GW	1ms
6.	SeMMu (PGW-C + SMF) to PGW-U + UPF	1ms
7.	SeMMu (PGW-C + SMF) to PCRF+PCF	1ms
8.	AMF to AMF	2ms
9.	SeMMu to PGW	1ms
10.	SeMMu to E-UTRAN	19ms
11.	E-UTRAN to UE	1ms
12.	PGW to PCRF	1ms
13.	S-GW to PGW	1ms
14.	SeMMu to SGSN	0.5ms
15.	SGSN to RNC	2ms
16.	SGSN to S-GW	1ms
17.	SeMMu to SeMMu	2ms

at any given time in the considered network, i.e., the parameter N_{HO} in (5.5) and (5.6), is 3 million. The analysis does not take into consideration the users that undergo an X2 or Xn based handover, i.e., they are not included amongst the 3 million users that we include in our analysis, as they do not involve any HO-related CN signaling. In addition, and based on discussions in Sections 5.2.2 and 5.2.3, the HO cancel phase is considered only for the 5G networks, while for the legacy networks (4G/3G/2G) we only consider the HO rejection phase signaling. Recall that, for the 5G networks the rejection phase signaling does not exist. Further, the considered HO cancel phase for the 5G NGC is as shown in Figure 5.7, which is also the worst case enhanced signaling for the same. However, for the legacy networks (4G/3G/2G) we do not consider the HO cancel phase since:

- The HO cancel signaling process for the legacy networks is fundamentally the same as that in the 5G NGC. Hence, considering the HO rejection signaling phase for the legacy networks aids in the completeness of analysis and study.
- Given the dynamic nature of HO cancel phase (Section 5.2), considering the HO rejection phase signaling also facilitates the ease of analysis.

We then develop five randomly distributed settings over the HO types (Table 5.1) for the computation of network wide processing cost and network occupation time. Concretely, we define the HO distributions that will be utilized for the analysis through equations (5.5) and (5.6), i.e., the parameter $\mathbf{Dist_{HO}}^{\mathbf{T}}$. The distributions are generated using Algorithm 1, wherein one of the distributions is predefined to be uniform across the HO types. Through uniform we mean that the percentage of users experiencing a particular handover scenario is the same for all HO types. It is imperative to state here that, the premise behind considering random distributions over the HO types is the lack of availability of real data from network operators.

Algorithm 1 Distribution Generation

1: **procedure** DISTRIBUTIONGENERATOR
2: $iter \leftarrow 5$
3: $i \leftarrow 1$
4: $mprct \leftarrow 0.2$
5: $NoH \leftarrow$ Number of Handover Types
6: **for** $i < iter$ **do**
7: $maxper \leftarrow mprct$
8: $minper \leftarrow 10^{-4}$
9: $j \leftarrow 1$
10: **for** $j <=$ NoH **do**
11: $Distper(i,j) \leftarrow U[minper, maxper]$
12: $maxper \leftarrow min(1 - sum(Distper(i,:)), mprct)$
13: $j \leftarrow j + 1$
14: $Distper(5,:) \leftarrow ones(1, NoH)/NoH$

And so, in Algorithm 1 we first define the maximum percentage of users ($maxper$) that undergo a particular HO type to be 20%, whereas the minimum percentage ($minper$) of users that undergo a particular HO type is 0.01%. Next, to generate the random distribution, we utilize the uniform probability distribution (U), with its upper and lower bounds being specified by the maximum and minimum percentage, respectively. We continually update the maximum percentage so as to prevent any skewness in the nature of distribution. The update rule is defined as the minimum value amongst 20% (initial maximum percentage value) and the percentage of users that remain to be associated to a particular HO type. We then define the last distribution as being uniform across all the HO types (Algorithm 1: Line 14).

5.3.3 Performance Analysis

In this section, utilizing the formulation presented in Section 5.3.1, we present and discuss the analytical results for the latency, processing cost and transmission cost of the new signaling framework for the handover preparation and failure phases presented in Section 5.2. For the analysis, we utilize the link latency data shown in Tables 5.2 and 5.3. The analytical methodology undertaken here is used to compare the performances of the proposed approach and the current 3GPP defined approach.

5.3.3.1 Latency analysis

Utilizing equation (5.2), as well as the cellular operator data from Section 5.3.2, we present the analytical results for the latency improvement for the handover preparation phase in Tables 5.4 and 5.5.

Table 5.4: Preparation Phase: Handover Latency Improvement Analysis (Cisco and Cellular Operator-Japan)

Handover Type	**Legacy Mechanism**	**Proposed Mechanism**	**Percentage Latency Reduction**
1.U$^{\rho}$	155 ms	95 ms	38.71%
1.U$^{\triangle}$	138.5 ms		31.41%
1.V$^{\rho}$	181 ms	89 ms	50.82%
1.V$^{\triangle}$	171.5 ms	138.5 ms	19.24%
1.W	179 ms	123 ms	31.28%
1.X.a†	128 ms	65.5 ms	48.83%
1.X.b†			
1.Y.a†	82 ms	65.5 ms	20.12%
1.Y.b†		58 ms	29.27%
1.X.a*	129.5 ms	65.5 ms	49.42%
1.Y.a*	82 ms	65.5 ms	20.12%
2.y	113 ms	90 ms	20.35%
2.x	159 ms	90 ms	43.40%

1: Inter-RAT HO; **2:** Intra-RAT (LTE) HO; **a:** Indirect Tunnel; **b:** Direct Tunnel; **U:** with N26 interface
V: without N26 interface; **X:** with T-SGW; **Y:** without T-SGW; $^{\rho}$5GS to EPS; $^{\triangle}$EPS to 5GS
y: inter-MME and intra-SGW; **x:** inter-MME and S-GW; *3G/2G to LTE; †LTE to 3G/2G
W: Intra-NG-RAN N2 based HO in 5G NGC

We show through this analysis that the proposed mechanism reduces the latency as compared to the legacy mechanism for both sets of operator data and all HO types considered. Note that, while the proposed mechanism helps reduce the latency by more than 19% for all HO types over the Japanese operator data (Table 5.4), the latency reduction over the Greek operator data (Table 5.5) ranges from 8.3% to 35.80%. Such differential behavior is

Table 5.5: Preparation Phase: Handover Latency Improvement Analysis (Cisco and Cellular Operator-Greece)

Handover Type	**Legacy Mechanism**	**Proposed Mechanism**	**Percentage Latency Reduction**
1.U$^{\rho}$	155 ms	126 ms	18.71%
1.U$^{\triangle}$	155.5 ms		18.97%
1.V$^{\rho}$	162 ms	104 ms	35.80%
1.V$^{\triangle}$	151 ms	132 ms	12.58%
1.W	157 ms	129 ms	17.83%
1.X.a†	103 ms	73.5 ms	28.64%
1.X.b†			
1.Y.a†	83 ms	73.5 ms	11.45%
1.Y.b†		73 ms	12.05%
1.X.a*	103 ms	73.5 ms	28.64%
1.Y.a*	83 ms	73.5 ms	11.45%
2.y	120 ms	110 ms	8.33%
2.x	140 ms	110 ms	21.43%

1: Inter-RAT HO; **2:** Intra-RAT (LTE) HO; **a:** Indirect Tunnel; **b:** Direct Tunnel; **U:** with N26 interface
V: without N26 interface; **X:** with T-SGW; **Y:** without T-SGW; $^{\rho}$5GS to EPS; $^{\triangle}$EPS to 5GS
y: inter-MME and intra-SGW; **x:** inter-MME and S-GW; *3G/2G to LTE; †LTE to 3G/2G
W: Intra-NG-RAN N2 based HO in 5G NGC

a consequence of the varied deployment scenarios for different operators dependent on their requirements.

From the analytical results it is evident that the gains obtained for the 5G HO scenarios (1.U$^{\rho}$, 1.U$^{\triangle}$, 1.V$^{\rho}$, 1.V$^{\triangle}$, 1.W) is significant. The reason being, the prevalence of handshakes that involve the exchange of tunnel setup information and their consequent optimization by the proposed mechanism. Specifically, the gains obtained for the scenarios that do not involve the N26 interface (scenarios 1.V$^{\rho}$, 1.V$^{\triangle}$) are significant not only due to quantitative reasons, but also because scenarios without an N26 interface will be prevalent during initial deployment scenarios. Further, we also show that the LTE-3G/2G HO scenarios, wherein a S-GW is being relocated (scenarios 1.X.a†, 1.X.b†), the percentage reduction in latency via the proposed mechanism is the highest amongst any other LTE-3G/2G HO scenarios. The aforesaid characteristic is observed because the number of messages that can be optimized through parallel message transfer and intelligent IE mapping is higher as compared to other scenarios. Concretely, during S-GW relocation process more handshakes are performed as compared to the scenario where there is no relocation, which consequently results in more avenues for optimization of the signaling messages.

Next, the handover latency improvement analysis for the handover failure phase, corresponding to the data from both operators, has been presented in Tables 5.6 and 5.7. Whilst

Table 5.6: Failure Phase: Handover Latency Improvement Analysis (Cisco and Cellular Operator-Japan)

Handover Type	**Legacy Mechanism**	**Proposed Mechanism**	**Percentage Latency Reduction**
1.Z$^{\rho}$	112 ms	72.5 ms	35.27%
1.Z$^{\triangle}$	122 ms		40.57%
1.X.a†	104 ms	64.5 ms	37.98%
1.X.b†			
1.Y.a†	58 ms	64.5 ms	-11.20%
1.Y.b†		58 ms	0.00%
1.X.a*	104 ms	58 ms	44.23%
1.Y.a*	58 ms	58 ms	0.00%
2.y	89 ms	89 ms	0.00%
2.x	135 ms	89 ms	34.07%

1: Inter-RAT HO; **2:** Intra-RAT (LTE) HO; **a:** Indirect Tunnel; **b:** Direct Tunnel; $^{\rho}$5GS to EPS
$^{\triangle}$EPS to 5GS; †LTE to 3G/2G; *3G/2G to LTE; **X:** with T-SGW; **Y:** without T-SGW
x: inter-MME and S-GW; **y:** inter-MME and intra-SGW; **Z:** 5G HO Cancel

Table 5.7: Failure Phase: Handover Latency Improvement Analysis (Cisco and Cellular Operator-Greece)

Handover Type	**Legacy Mechanism**	**Proposed Mechanism**	**Percentage Latency Reduction**
1.Z$^{\rho}$	130 ms	110.5 ms	15.00%
1.Z$^{\triangle}$	139 ms		28.50%
1.X.a†	92 ms	72.5 ms	21.20%
1.X.b†			
1.Y.a†	72 ms	72.5 ms	-0.69%
1.Y.b†		72 ms	0.00%
1.X.a*	92 ms	72 ms	21.73%
1.Y.a*	72 ms	72 ms	0.00%
2.y	109 ms	109 ms	0.00%
2.x	129 ms	109 ms	15.50%

1: Inter-RAT HO; **2:** Intra-RAT (LTE) HO; **a:** Indirect Tunnel; **b:** Direct Tunnel; $^{\rho}$5GS to EPS
$^{\triangle}$EPS to 5GS; †LTE to 3G/2G; *3G/2G to LTE; **X:** with T-SGW; **Y:** without T-SGW
x: inter-MME and S-GW; **y:** inter-MME and intra-SGW; **Z:** 5G HO Cancel

for the Japanese operator data, the latency improvement ranges from 34.07% to 44.23%, for the Greek operator data the latency reduction is between 15.50% to 28.50%. The differen-

tial performance behavior, as mentioned earlier, is representative of the variable deployment scenarios dependent on operator requirements. Specifically, for the 5G NGC network, the HO cancellation phase signaling between NGC and EPS (scenarios $1.Z^{\rho}$, $1.Z^{\triangle}$) observes a latency reduction by upto 40.57%. The aforesaid improvement is as a consequence of the presence of handshakes, whose composition and execution have been enhanced by the proposed mechanism. Recall that, as per discussions in Sections 5.2.2, 5.2.3 and 5.3.2, the worst case scenario for the HO cancel phase signaling has been considered, i.e., the HO cancel phase (Figure 5.7) is executed after all the resources have been setup in the HO preparation phase.

Further, the rejection phase signaling in LTE and 3G/2G networks for scenarios that do not involve S-GW relocation (scenarios $1.Y.a^{\dagger}$, $1.Y.b^{\dagger}$) is already optimal and does not incur any improvement or degradation through the implementation of the proposed mechanism. However, for the scenario where there is no S-GW relocation but an indirect tunnel is utilized during the LTE to 3G/2G handover ($1.Y.a^{\dagger}$), the proposed mechanism results in a degraded performance for the handover rejection phase as compared to the legacy approach. A deeper analysis (through Figure 5.3 but without the presence of T-SGW) reveals that while in the legacy mechanism the source S-GW tunnel (message 8 and 8a in Figure 5.3) is not setup until the resource negotiation phase (message 5 and 5a in Figure 5.3) is accomplished, the proposed mechanism, in order to obtain the advantages of parallelization, performs the source S-GW tunnel setup (message P3b in the signaling scenario specified by Figure 5.8 without target S-GW) before resource negotiation. Consequently, when the RRM operation results in a handover failure, the extra source S-GW tunnel setup message in the proposed setup leads to the aforementioned performance degradation. It is important to state here that, for the corresponding scenario the handover preparation phase signaling incurs an improvement of 20.12% over the legacy approach.

Further, to de-register the resources setup by the source S-GW setup message in the proposed mechanism, a delete session request message (P6a in the signaling scenario specified by Figure 5.8, but without target S-GW) from the SeMMu will also be required. And as will be seen in the next subsection, this will lead to a degraded performance in terms of incurred transmission cost as compared to the legacy mechanism. However, through the *Handover Failure aware Preparation Signaling* analysis, wherein the novel handover failure aware method is used, we will observe that this performance degradation is alleviated whilst also benefiting the handover preparation phase further.

5.3.3.2 Transmission Cost Analysis

Utilizing equation (5.3), we present the transmission cost analysis for the handover preparation and failure signaling phases for the different HO types specified by 3GPP (Table 5.1) and the different deployment scenarios presented by the Japanese and Greek telecom operators (Tables 5.2 and 5.3). Concretely, a comparative performance analysis between the legacy and proposed mechanism for the handover preparation phase signaling has been presented in Figures 5.13 and 5.14. Further, a similar comparative analysis for the handover failure phase signaling has also been provided through Figures 5.15 and 5.16.

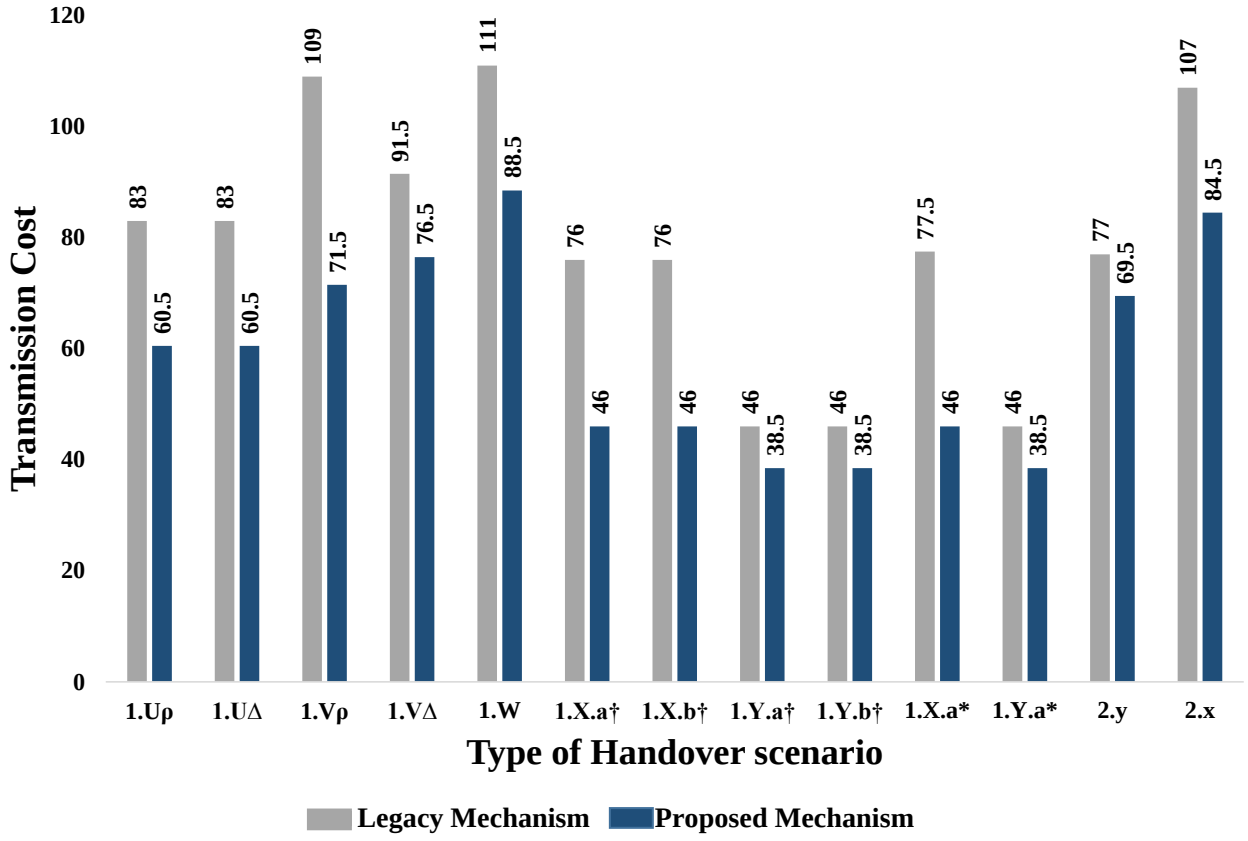


Figure 5.13: Handover preparation scenario: Transmission cost analysis for the Japanese operator deployment (X-axis notations have been re-utilized from Tables 5.4-5.7).

From the analytical results presented in Figures 5.13 and 5.14, it is established that the proposed mechanism enhances the handover preparation phase signaling compared to the legacy mechanism, by reducing up to 40.67% in transmission cost incurred to complete the signaling process for all the considered HO scenarios. For the scenarios involving 5G NGC, the gain characteristic, i.e., the trend in performance gains, is similar to that observed for the latency improvement analysis presented earlier. Further, in scenarios involving 4G/3G/2G

networks, where S-GW relocation occurs (scenarios 1.X.a†, 1.X.b†, 1.X.a*), the transmission cost reduction obtained is higher than that obtained in the other legacy HO scenarios. However, as per our discussions in latency improvement analysis subsection, the gains obtained for the Greek and Japanese operator deployments are different due to the difference in the resources and requirements presented by the operators.

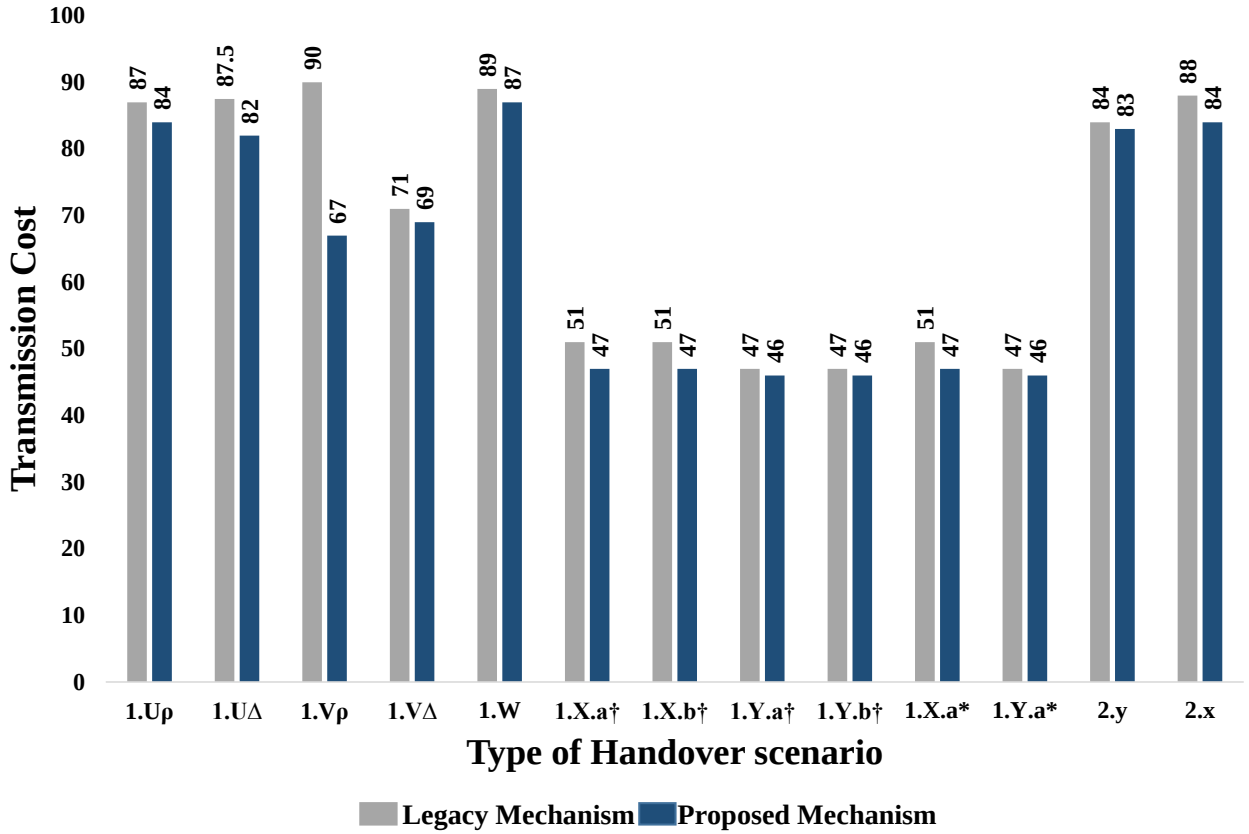


Figure 5.14: Handover preparation scenario: Transmission cost analysis for the Greek operator deployment (X-axis notations have been re-utilized from Tables 5.4-5.7).

Next, through Figures 5.15 and 5.16, it can be observed that the proposed mechanism either improves or does not degrade the performance of the HO scenarios considered for the HO failure phase signaling, except when there is no S-GW relocation with an indirect tunnel for a LTE to 3G/2G Inter-RAT HO (scenario 1.Y.a†). The reason being, for the purpose of parallelization of CP information transfer from the SeMMu, the Source S-GW tunnel setup message is executed before the RRM procedure (message 3b in Figure 5.8 but without the T-SGW). And, since the RRM procedure at the target network results in a handover failure, an extra message to de-register the allocated resources is required (Message 6a in Figure 5.8 but without the T-SGW). Hence, these extra messages contribute towards the aforesaid

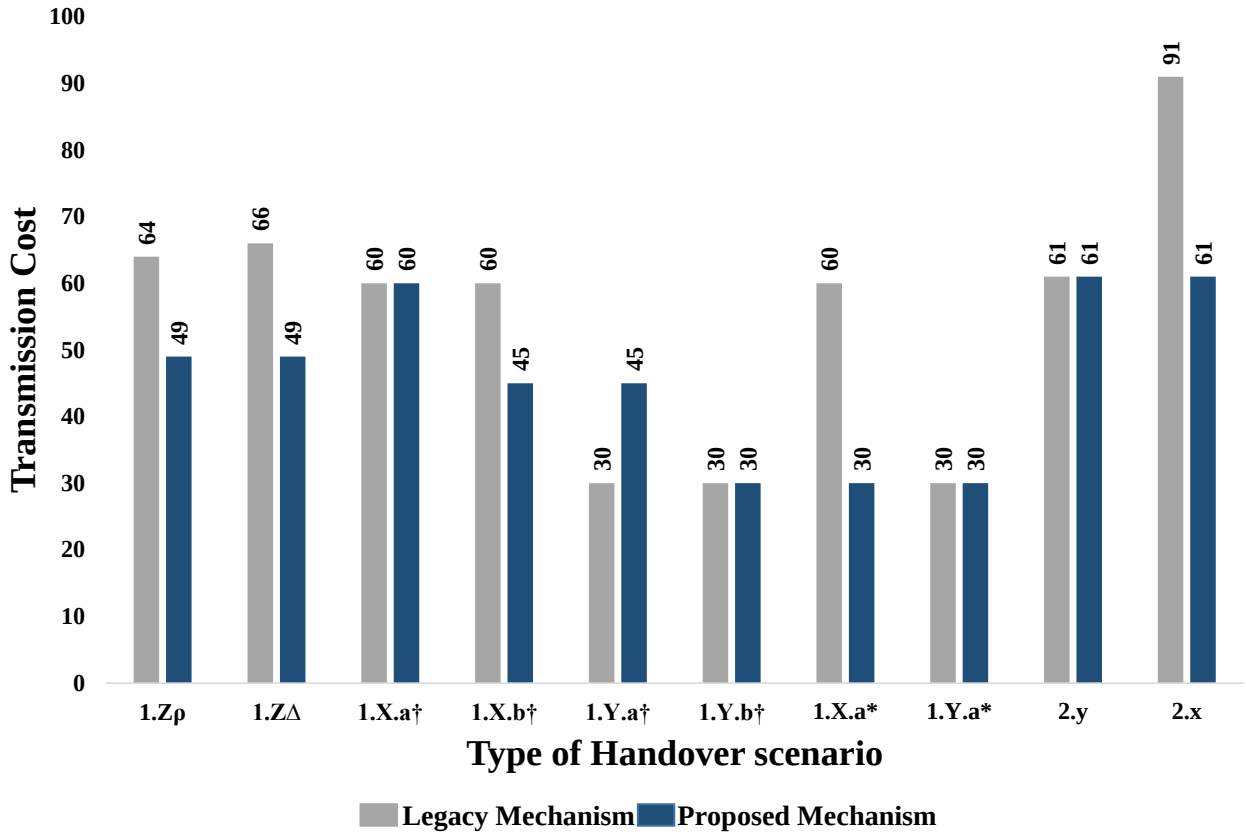


Figure 5.15: Handover failure scenario: Transmission cost analysis for the Japanese operator deployment (X-axis notations have been re-utilized from Tables 5.4-5.7).

degradation in performance. However recall that in the later subsections we will discuss the analysis where the novel HO failure aware preparation signaling has been utilized. Through the analysis we establish that the concerns of degraded performance are alleviated by this novel strategy.

5.3.3.3 Processing Cost Analysis

Unlike the transmission cost and latency, the processing cost is unaffected by the change in operator deployment scenarios as it solely depends on the number of messages that will be processed within the CN. Hence, in this subsection, utilizing the formulation in Section 5.3.1 as well as the proposed and legacy signaling sequences, a comparative analysis with regards to the processing cost savings offered by the proposed and legacy mechanisms has been presented via Tables 5.8 and 5.9, respectively.

Through the analytical results in Table 5.8, it can be observed that the proposed algorithm reduces the processing cost for the handover preparation phase signaling for all

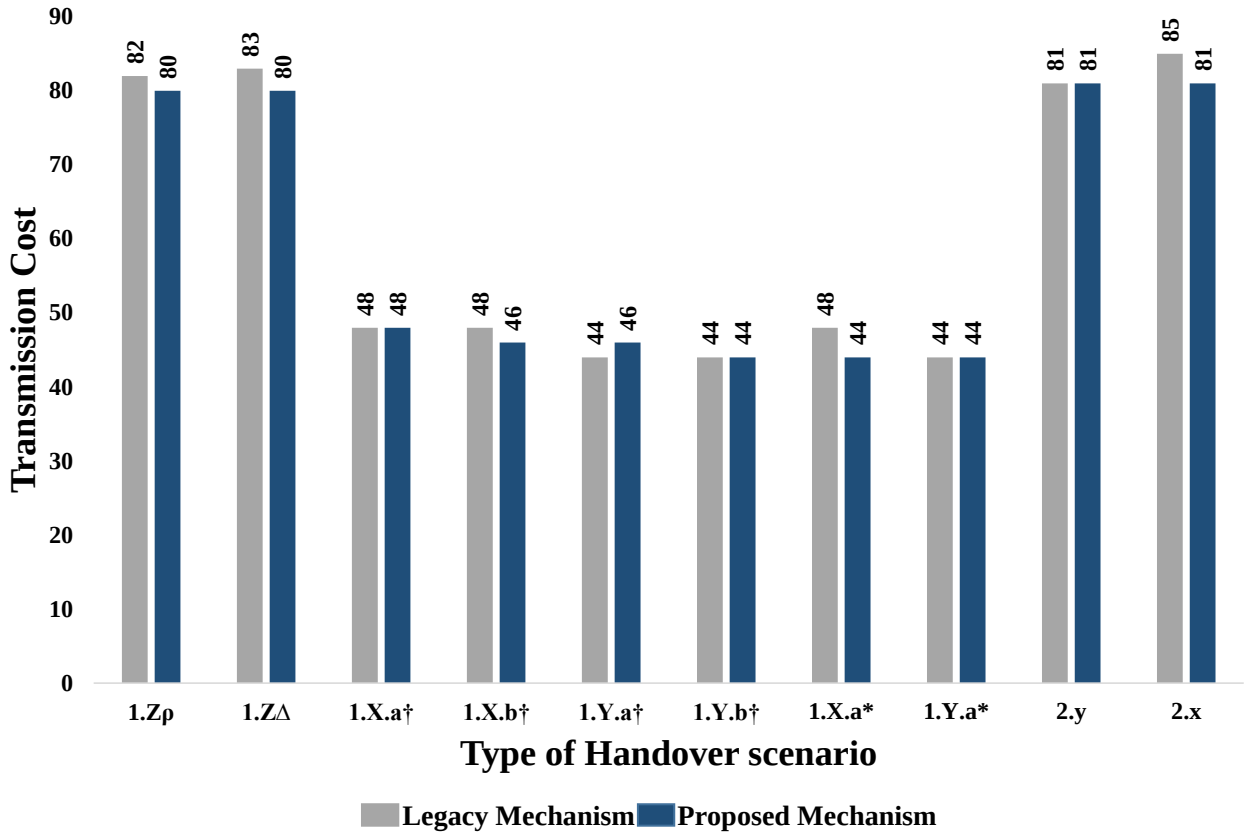


Figure 5.16: Handover failure scenario: Transmission cost analysis for the Greek operator deployment (X-axis notations have been re-utilized from Tables 5.4-5.7).

the considered handover scenarios. Quantitatively, the performance enhancement for the 5G NGC HO scenarios (1.U$^{\rho}$, 1.U$^{\Delta}$, 1.V$^{\rho}$, 1.V$^{\Delta}$) ranges from 10.00% to 27.77%, with the scenarios without an N26 interface (1.V$^{\rho}$, 1.V$^{\Delta}$) also showing improvement. In addition, the savings offered over legacy scenarios where a S-GW relocation occurs (1.X.a†, 1.X.b†, 1.X.a*) is above 20.00%, while that offered in other legacy scenarios is 10%. Next, for the handover failure phase signaling (Table 5.9), the proposed mechanism enhances the signaling for both the HO cancel phases in 5G NGC (1.Z$^{\rho}$: 16.67% and, 1.Z$^{\Delta}$: 28.57%) as well as in two other specific scenarios, i.e., HO rejection in LTE to 3G/2G Inter-RAT HO with Target S-GW and direct tunnel (1.X.b†: 18.18%), and HO rejection in 3G/2G to LTE Inter-RAT HO with a Target S-GW (1.X.a*: 36.36%). The proposed mechanism neither enhances nor degrades the performances of the failure phase signaling for other HO scenarios, except when there is an Inter-RAT HO from LTE to 3G/2G involving an indirect tunnel and without S-GW relocation (scenario 1.Y.a†). The degradation in performance for the aforesaid scenario stems from the reasons discussed in latency and transmission cost analysis subsections, i.e.,

Table 5.8: Processing Cost Analysis for Handover Preparation phase

$^{\alpha}$**Handover Type**	**Processing Cost**		
	Legacy Mechanism	**Proposed Mechanism**	**% Saving**
1.U$^{\rho}$	18 messages	15 messages	16.67%
1.U$^{\triangle}$	20 messages		25.00%
1.V$^{\rho}$	18 messages	13 messages	27.77%
1.V$^{\triangle}$	20 messages	18 messages	10.00%
1.W	18 messages	15 messages	16.67%
1.X.a†	14 messages	10 messages	28.57 %
1.X.b†			
1.Y.a†	10 messages	9 messages	10.00 %
1.Y.b†			
1.X.a*	14 messages	10 messages	28.57 %
1.Y.a*	10 messages	9 messages	10.00 %
2.y	10 messages	9 messages	10.00 %
2.x	14 messages	11 messages	21.43 %

$^{\alpha}$The notations have been re-utilized from Tables 5.4-5.7

Table 5.9: Processing Cost Analysis for Handover Failure phase

$^{\gamma}$**Handover Type**	**Processing Cost**		
	Legacy Mechanism	**Proposed Mechanism**	**% Saving**
1.Z$^{\rho}$	12 messages	10 messages	16.67%
1.Z$^{\triangle}$	14 messages		28.57%
1.X.a†	11 messages	11 messages	No Change
1.X.b†		9 messages	18.18%
1.Y.a†	7 messages	9 messages	-22.22%
1.Y.b†		7 messages	No Change
1.X.a*	11 messages	7 messages	36.36 %
1.Y.a*	7 messages	7 messages	No Change
2.y	7 messages	7 messages	No Change
2.x	7 messages	7 messages	No Change

$^{\gamma}$The notations have been re-utilized from Tables 5.4-5.7

the execution of source S-GW tunnel setup message before the RRM process requires an extra delete session message from the SeMMu towards the S-GW to de-register the allocated resources. Hence, this increases the number of messages to be processed by the CN, which consequently leads to the degraded performance.

5.3.3.4 Handover Failure aware Preparation Signaling

The analytical evaluation presented in the previous subsections reveals that the proposed mechanism, while enhancing the handover preparation phase, can slightly under-perform for certain handover scenarios during the handover failure phase. And so, utilizing the discussions as well as the novel handover failure aware signaling from Section 5.2.3, we show that at least eight HO scenarios can be optimized further in terms of latency, transmission cost and processing cost. The rest of the scenarios are already optimal and hence, are not impacted by the proposed enhancement. The analytical results, presented in Table 5.10, show that the HO failure aware approach not only alleviates the performance degradation issue in the handover failure phase, but it also enhances the handover performance phase signaling.

Note that, the analytical results presented in Table 5.10 utilize the delay values from the Japanese operator deployment. Moreover, it is clear from our analysis so far that, the trend for the gains obtained by utilizing our methodology is the same irrespective of the operator. And so, quantitatively, for the HO preparation scenarios in 5G NGC, a reduction of up to 15.38% in the processing cost over the values in Table 5.8 is observed. However, for the 4G/3G/2G (legacy) HO scenarios, the number of messages required to complete the entire signaling process does not change, and thus, the processing cost remains unchanged. Further, for scenarios 1.X.a†, 1.X.b†, 1.Y.a†, the number of messages required to execute the handover failure phase, i.e., HO rejection phase, is reduced significantly as compared to that specified in Table 5.9. In addition, for the 5G NGC scenarios (1.U$^{\rho}$, 1.U$^{\Delta}$, 1.V$^{\rho}$, 1.V$^{\Delta}$, 1.W), we consider the handover failure phase, i.e., HO cancel phase, to be near optimal owing to its sensitivity to the time at which it is initiated during an ongoing HO preparation phase, as discussed in Section 5.3.2.

Next, the added enhancement over the proposed mechanism improves the processing cost saving for HO scenarios 1.X.a† and 1.X.b† during a handover failure phase by 36.36%, as compared to the values in Table 5.9. Further, for the HO scenario 1.Y.a† in Table 5.9, the processing cost saving performance is no longer degraded. Instead, the novel handover failure aware method reduces the number of messages required from 9 to 7, i.e., by 22.22%, for the proposed mechanism.

Further, the latency analysis presented in Tables 5.4, 5.6 and 5.10 establishes that the HO preparation and failure phases can be enhanced further with the novel handover failure aware method proposed here. Quantitatively, the HO preparation phase corresponding to the first eight HO scenarios are further enhanced by up to 9.92% with the maximum gains being obtained for the 4G/3G/2G HO scenarios. The HO failure phase signaling for the

Table 5.10: Handover failure aware signaling design analysis

$^{\beta}$**Type of Handover**	**HO preparation**						**HO Failure**					
	Proc. Cost$^{\eta}$	% Saving$^{\gamma}$	Latency	% Saving	Trans. Cost	% Saving$^{\gamma}$	Proc. Cost$^{\eta}$	% Saving$^{\gamma}$	Latency	% Saving$^{\gamma}$	Trans. Cost	% Saving$^{\gamma}$
1.U$^{\rho}$	14	6.67%	95 ms	0.00%	53	12.39%	The HO cancel phase, as discussed in section 5.2.3, is near optimal owing to its adaptability depending on when it is invoked.					
1.U$^{\triangle}$			88.5 ms	7.34%								
1.V$^{\rho}$	11	15.38%	89 ms	0.00%	56.5	20.97%						
1.V$^{\triangle}$	17	5.56%	132 ms	4.69%	69	9.80%						
1.W	14	6.67%	116.5 ms	5.28%	81	8.47%						
1.X.a†	10	0%	59ms	9.92%	46	0.00%	7	36.36%	58ms	10.07%	30	50.00%
1.X.b†								22.22%				33.33%
1.Y.a†	9	0%	59ms	9.92%	38.5	0.00%	7	22.22%	58ms	10.07%	30	33.33%
1.Y.b†	Optimal											
1.X.a*	Optimal											
1.Y.a*	Optimal											
2.y	Optimal											
2.x	Optimal											

$^{\beta}$The notations have been re-utilized from Tables 5.4-5.7; $^{\eta}$The processing cost, defined in Section 5.3, is the number of CN messages

$^{\gamma}$ The percentage saving over the values obtained with the proposed mechanism in Tables 5.5, 5.7, 5.9, 5.10 and Figures 5.15 and 5.16.

corresponding 4G/3G/2G HO scenarios are also enhanced further by 10.07%. Consequently, the improvement in the handover failure phase signaling also alleviates the drawback of degraded performance.

Lastly, for the transmission cost, the analytical evaluation reinforces the trend of added enhancement to the performance of the handover failure phase. The transmission cost for the HO failure phase signaling in scenario 1.X.a† of Table 5.10 is halved compared to the cost presented in Figure 5.15. Further, the scenarios 1.X.b† and 1.Y.a† also experience an improvement of 33.33% in the transmission cost for their corresponding HO failure phase signaling. As a consequence, the drawback of performance degradation is also mitigated. Additionally, for the handover preparation signaling, the added enhancements facilitates an improvement ranging from 8.47% to 20.97% for scenarios 1.U$^{\rho}$, 1.U$^{\Delta}$, 1.V$^{\rho}$, 1.V$^{\Delta}$ and 1.W in Table 5.10, whilst the transmission cost performance of the remaining scenarios remains unaffected as compared to that presented in Figure 5.15.

It is important to state here that, in Table 5.10, scenarios 1.Y.b†, 1.X.a*, 1.Y.a*, 2.x and 2.y are not impacted by the Handover failure aware method and, hence, are referenced as *Optimal*. Concretely, the proposed mechanism without the handover failure aware methodology is already optimal for the aforementioned scenarios.

5.3.4 Message Size Analysis

The mechanism that has been proposed in this work utilizes the fact that the IEs can be intelligently re-packaged to create lesser number of messages and hence, enhance the signaling that is performed at the CN. This restructuring of the messages will also alter the size of the messages, i.e., the number of bytes carried per message, as well as the overall bytes transferred per signaling sequence. Thus, through Tables 5.11 and 5.12 we provide a comparative analysis between the legacy and proposed mechanisms for the message sizes and the overall bytes transferred in a single sequence of handover signaling.

Note that for the analysis, we do not consider the 5G HO scenarios as the message sizes for 5G HO signaling are still not completely defined. Hence, we consider four representative scenarios from the LTE-EPC and 3G/2G HO signaling, shown in Table 5.12. The chosen scenarios encompass Inter- and Intra-RAT HO, relocation/no relocation of S-GW/S-GW and SeMMu, and indirect and direct tunneling, thus ensuring completeness to the analysis. Further, it must be stated that the current analysis is independent of the operator deployment scenario, and hence, it is valid for both the operator deployments considered in this work.

Table 5.11 presents a detailed breakdown of the messages and their sizes in bytes for the HO preparation signaling corresponding to the scenario when there is an Inter-RAT

Table 5.11: Message size Computation: Inter-RAT HO from LTE to 3G/2G when S-GW is relocated and indirect tunneling exists

Msg. num.	**Legacy messages**	**Size (Bytes)**	**Msg. num.**	**Proposed messages**	**Size (Bytes)**
1	Handover Initiation	62	P1	Handover Initiation	62
2	Handover Required	302	P2	Handover Required	302
3	Forward Relocation Request	762	P3a	Resource Allocation Request + Tunnel Setup	838
4	Create Session Request	288	P4	Relocation Request	335
4a	Create Session Response	117	P4a	Relocation Request Acknowledge	130
5	Relocation Request	335	P5	Forward Relocation Response	128
5a	Relocation Request Acknowledge	130	P6a	S-SGW Tunnel Setup	105
6	Indirect Data Forwarding Tunnel Req. T-SGW	86	P6b	T-SGW Tunnel Setup	345
6a	Indirect Data Forwarding Tunnel Resp. T-SGW	111	P6c	Handover Command	166
7	Forward Relocation Response	147	P7	HO from E-UTRAN Command	63
8	Indirect Data Forwarding Tunnel Req. S-SGW	86			
8a	Indirect Data Forwarding Tunnel Resp. S-SGW	111			
9	Handover Command	166			
10	HO from E-UTRAN Command	63			
Total bytes		2766	Total bytes		2474

HO from LTE to 3G/2G and S-GW relocation along side indirect tunneling occurs. For the analysis, the message sizes corresponding to the legacy and proposed mechanism were constructed utilizing the data provided in 3GPP specifications [133, 164–166, 168, 169, 171, 179, 180], wireshark traces [181] and ITU ASN.1 specifications [182] (for the data types and sizes of the IEs). Through the analysis, it was deduced that since the number of messages in the proposed mechanism is reduced as compared to the legacy mechanism, the number of bytes for message headers is also reduced. Concretely, since each message that is passed through the network consists of a message header, specifying source and destination addresses/identifiers, etc., a reduction in the number of messages will also mean that there is a corresponding decrease in the amount of header that traverses through the network.

Quantitatively, the largest message in the proposed mechanism, i.e., message P3a, is 838 bytes long, while the largest message in the legacy mechanism (message 3) is 762 bytes long. Thus, the proposed restructuring process maintains the message sizes near the range of message sizes in the legacy mechanism. Consequently, it can be said that the proposed mechanism does not present any significant challenge for the reliable transmission and processing of CP messages within the network. In addition, the total amount of bytes transferred within the proposed mechanism will be 2474 as compared to 2766 bytes in the legacy mechanism, to complete the HO signaling. And hence, through the non-repetitive and intelligent repackaging of IEs into the proposed messages, the number of bytes that have to be transported across the CN for the HO scenario under observation is reduced by 10.56%. Next, we present the analysis for the total message bytes transferred during the legacy and proposed mechanism for the scenarios under observation (Table 5.12).

Table 5.12: Message size analysis

$^{\zeta}$**Type of Handover**	**Total bytes for Legacy Messaging**	**Total bytes for Proposed Messaging**	**Percentage Reduction**
1.X.a†	2766	2474	10.56%
1.Y.b†	2164	2072	4.25 %
1.X.a*	2766	2493	9.87%
2.x	2817	2544	9.69%

$^{\zeta}$The notations have been re-utilized from Table 5.4-5.7

The analytical results in Table 5.12 reinforce the fact that HO scenarios in which S-GW or S-GW and SeMMU relocation occurs are optimized more than the other scenarios. While scenario 1.Y.b† in Table 5.12 has a 4.25% reduction in the total bytes that would be transferred over the CN to complete the signaling sequence, scenarios 1.X.a†, 1.X.a* and 2.x register a reduction of 10.56%, 9.87% and 9.69%, respectively. The aforementioned

results also illustrate that the proposed mechanism, irrespective of the HO scenario, reduces the number of bytes that would be transferred over the CN, thus enhancing the network performance as it will have lesser bytes to transfer across the network as well as to process.

5.3.5 Network Wide Analysis

In this subsection, we present an analysis for the network occupation time and network wide processing cost savings by utilizing equations (5.5) and (5.6) from Section 5.3.1, and the parameter framework presented in Section 5.3.2. Figures 5.17 and 5.18 illustrate the network wide occupation time and processing cost performance, respectively, for the legacy and proposed mechanisms.

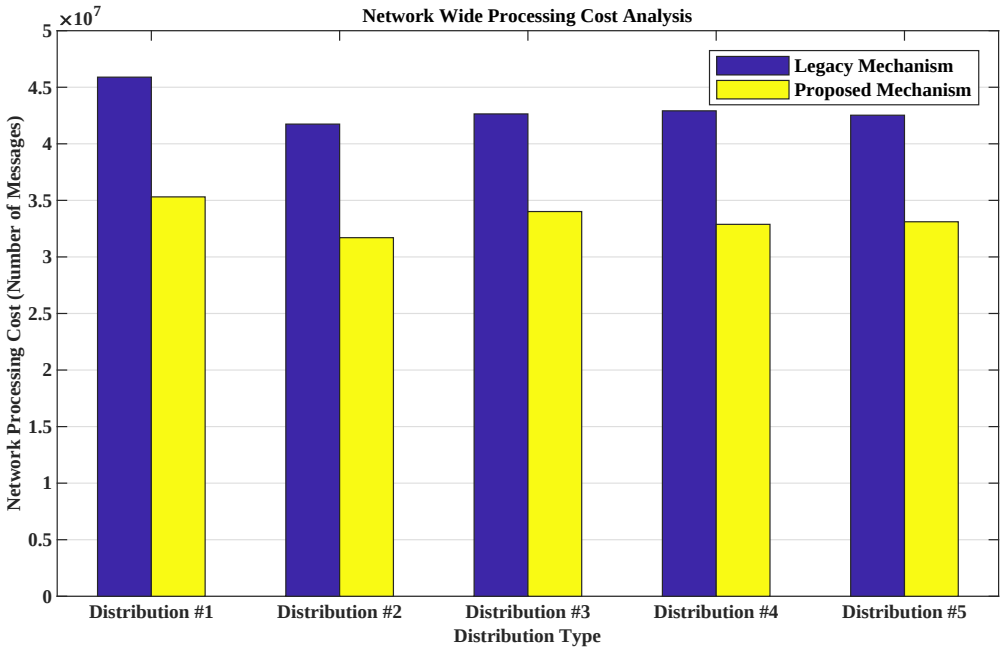


Figure 5.17: Network wide processing cost analysis.

Note that, the processing cost analysis is independent of the operator deployment scenario. Also, it is important to reiterate that, the trend for the gains established throughout our analysis is independent of the chosen operator. Hence, for the network wide occupation time analysis, we only consider the delay values as obtained from the Japanese cellular operator (Table 5.2). We consider HO failure rates from 0.1%-0.5% and also vary the percentage of S1 HO (intra-MME/S-GW) from 10%-50%. Given the lack of availability of real data from the telecom operators, we randomly select a particular HO failure rate and S1 HO percentage alongside a distribution, and then compute the two metrics utilizing equations

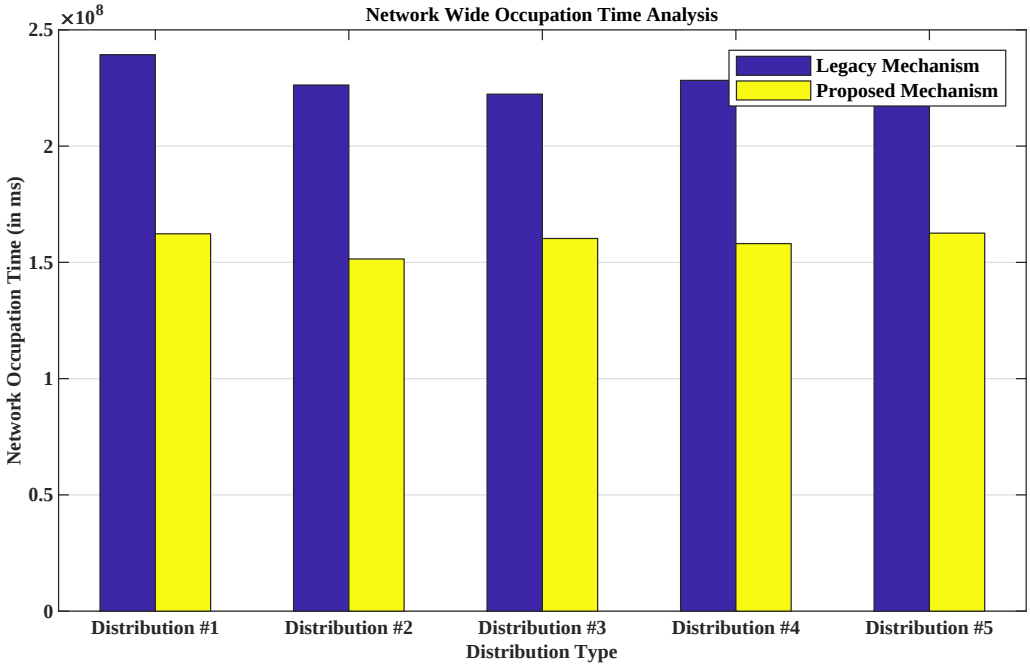


Figure 5.18: Network wide occupation time analysis.

(5.5) and (5.6). Such an evaluation process helps to eliminate any possible bias in specifying the prevalent handover scenario, and thus aids in the completeness of the analysis.

Figures 5.17 and 5.18 show that given any prevalent HO scenario and distribution of HO types, the proposed mechanism outperforms the legacy mechanism. Concretely, the proposed mechanism provisions a saving of 27.90%-33.06% over the legacy mechanism for the network occupation time, while for the network wide processing cost, the proposed mechanism provides a saving of 20.24%-24.05% over the legacy mechanism. And, given these significant savings in the processing cost and link occupation time, it will help the future networks, such as 5G, to be more time and resource efficient. By resource efficient here we mean that, the network will be more scalable in terms of computational and physical resources.

5.4 Evolutionary 4G/5G Network Architecture

Given the performance analysis results, and specifically the network wide analysis, the benefits offered by the proposed HO mechanisms, utilizing the SeMMu, are compelling. Henceforth, in this section we present an exemplary evolutionary network architecture that not only facilitates the execution of the proposed mechanism, but also provides the operator

with an avenue to have a manageable CAPEX towards evolving their networks to being fully softwarized. Thus, through Figure 5.19 we illustrate the proposed evolutionary core network architecture.

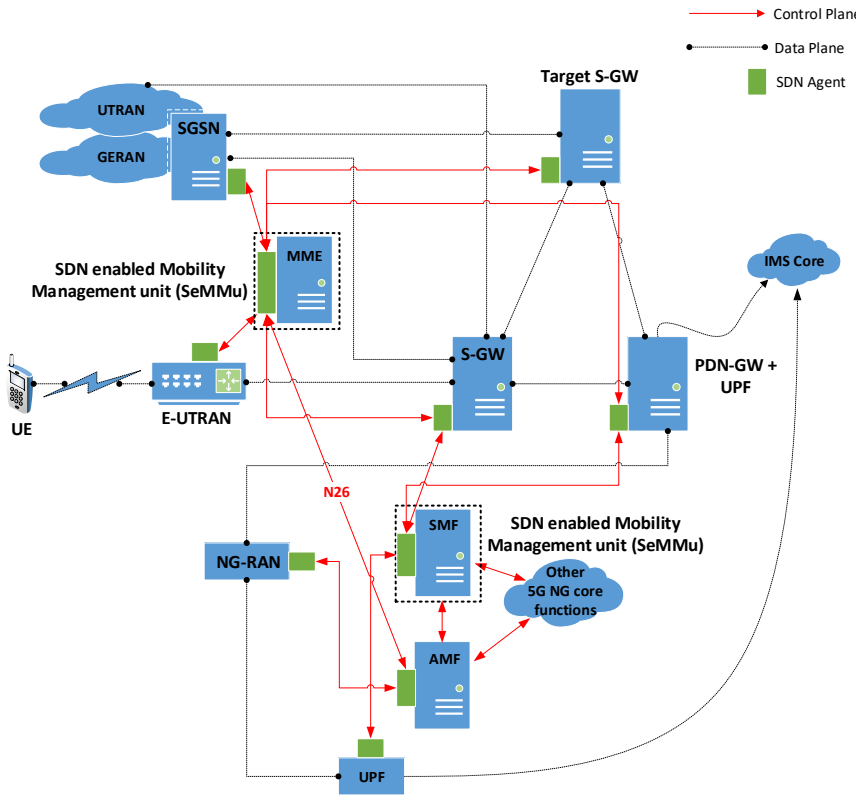


Figure 5.19: Proposed evolutionary network architecture.

The proposed network architecture is evolutionary with respect to the fact that, it firstly introduces an evolved core network entity, namely, the SeMMu. The SeMMu combines the functionalities of the MME/SMF and the SDN-C. Recall that, in the proposed architecture we consider the SMF as the main 5G mobility management unit instead of the AMF, as done by 3GPP. The reason being, the SMF is involved in the CN signaling during a HO whilst the AMF is only limited to the access network resource management. Moreover, the functional integration is carried out such that the SeMMu only modifies the CP between the network entities and itself, while avoiding any impact on other core network operations (such as the DP). Additionally, while in the 3GPP defined network architecture specific interfaces are utilized to connect the network entities, in the proposed network architecture, for the SeMMu to communicate with the other core network entities, an SDN agent needs to be integrated with these other entities (such as SGSN, S-GW, UPF, etc.). Such an integration, whilst maintaining the smooth inter-working between 5G and legacy networks, also enables

the operators to evolve their legacy networks towards a completely softwarized architecture. As a consequence of this evolutionary framework, the proposed architecture will help to facilitate a reduction in the CAPEX for the operators, which is a major 5G objective. The SDN capabilities also enable the proposed architecture to execute the optimized handover signaling, discussed in Sections 5.2 and 5.3, as it allows the SeMMu to push the required CP information to other CN entities. It is imperative to state here that, although we introduce an SDN agent overlay, we only transform the 3GPP defined functionalities of the MME/SMF whilst preserving the functionalities of all the other CN entities. Further, given any of the proposed signaling sequences and mapping, the network architecture remains the same. Concretely, the proposed evolutionary network architecture is consistent for any HO scenario. And, given the results in Section 5.3 as well as the fact that distinct flow rules per user do not require separate SDN agent threads, the network will be scalable.

Moreover, the proposed architecture is designed such that the RRM interactions are left unaltered. The reason being, if the RRM procedures are handled at the SeMMu, then while it would enable enhanced decision making given the global view the SeMMu has, it will introduce additional delays, and hence, increased latency for the handover process. Subsequently, the SeMMu is neither connected directly to the Radio Network Controller (RNC) nor to the NG-RAN. Instead, the SeMMu communicates with the SGSN/AMF, which is responsible for managing the session as well as the CP signaling with the RNC/NG-RAN. Further, within the EPC, the SeMMu allows the eNB to perform the RRM operations, even though it is directly connected to it. Lastly, the interworking framework, presented in Figure 5.19, is facilitated by the presence of an N26 interface between the AMF in the NGC and the SeMMu in the EPC. Note that the interworking between 5G NGC and EPC can be established even in the absence of the N26 interface, as discussed in our contribution [C4] and reference [29].

5.4.1 Benefits and Challenges

The aforesaid integration has multiple benefits as well as certain design and implementation challenges. The benefits of the SeMMu based network architecture include:

- The ability to access system parameters, which will allow the SeMMu to establish optimized MM solutions through the virtualized functions in a fully SDN architecture, via a global or locally global view of the network domain. Here domain refers to the geographical area of the network that is administered/controlled by the SeMMu.
- Introduction of SDN agents to the CN entity is a first step towards the fully SDN

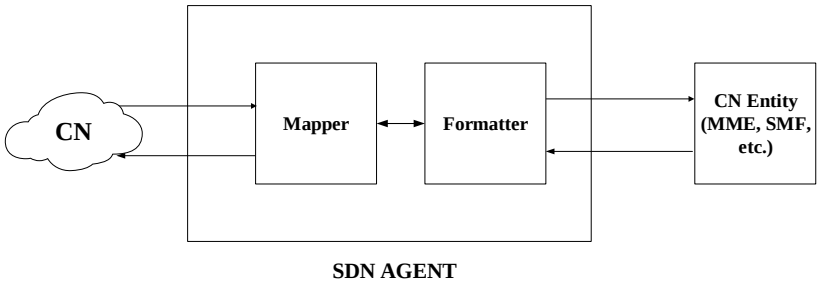
architecture that is envisioned for the future networks. Given the ability of an SDN controller based entity to decouple the CP from DP (and implement the rules on the DP entities), the SDN agents are utilized to push the CP information necessary for the handover related signaling to the CN entities.

- The given framework establishes an evolutionary path towards a fully softwarized network architecture. Thus, the given framework assists in reducing the CAPEX for the operators as it helps them evolve their current architecture towards a fully softwarized architecture.
- With the SDN based architecture presented in this section, the handover preparation and failure signaling phases can be optimized (Section 5.2) as compared to the legacy mechanisms, i.e., 3GPP standards. The optimizations obtained via signaling re-sequencing and message mapping have been elaborately discussed in Section 5.3.

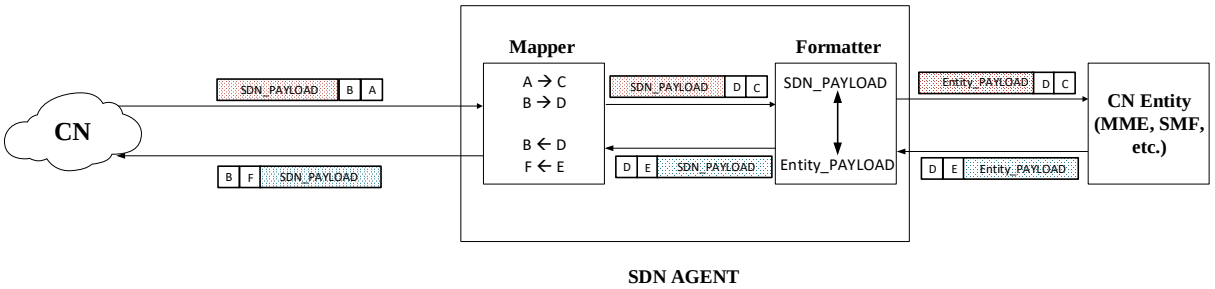
Next, the main implementation challenge arising as a consequence of the SeMMu based evolutionary network architecture, is the integration of the SDN agent to the CN entities. On one hand it will include an initial CAPEX to integrate the SDN agents and, on the other hand, new interfaces need to be defined so as to allow the MME/SMF and the SDN agent to communicate with each other. Additionally, advanced software mechanisms to identify and pack the IEs into the proposed message ensemble, discussed in Section 5.2, will be required. Whilst the CAPEX incurred will not be significant given the benefits offered by the SeMMu solution, in the subsequent discussion we provide a brief insight into the approaches that can be utilized to overcome this implementation challenge.

5.4.2 SDN agent integration

The integration process of the SDN agent should not disrupt the overall network functioning, design and architecture. Further, the DP operations should be agnostic to the proposed integration process. In order to realize this seamless integration, we introduce a novel setup wherein the MME/SMF entail a software modification and the SDN agent is composed of two components (illustrated in Figure 5.20(a)). Concretely, the two components that constitute the SDN agent are the *Mapper* and the *Formatter*. Given that, there is an SDN agent overlay on top of the 3GPP defined network architecture, the SDN agents will view the messages and destination address in a different format as compared to the CN entities defined by 3GPP. Note that, the mapper is connected to the external network through a communication interface through which the SDN agent transmits/receives the data. In addition, the formatter is connected to the CN entity through a bidirectional communication



(a) SDN agent architectural framework.



(b) SDN agent operational framework.

Figure 5.20: SDN agent for the evolutionary network architecture.

interface for exchanging the CP information messages. The detailed functioning of both these components is provided as follows:

- *Mapper*: The mapper essentially performs a mapping and de-mapping of the address that the CN entities would observe without the SDN agent overlay to the addresses as observed by the SDN agents on the CN, and vice versa. Thus, when the mapper receives a message frame from the CN, it first removes the frame header. During this process, it identifies the message source and destination, i.e., SDN-enabled CN entity addresses, and then maps these addresses to the address of the source and destination as would be seen by the CN entities, if there were no SDN agents integrated with them. Next, it transfers the message payload along with the source and destination address to the formatter. On the other hand, when the mapper receives the messages from the formatter, it identifies the type of message, the source address and its destination address. It then maps these addresses, i.e., address that would be observed by the CN entities in the absence of the SDN agent overlay, to the address in the external CN (observed by SDN agents) and transmits it to the intended CN entity. Lastly, the application level scheduling of the messages to be sent to other CN entities is done by the scheduler present in the transformed MME/SMF CN entity, discussed later in this

section.

- *Formatter*: The main task of the formatter is to transform the format of the incoming and outgoing messages according to the format expected by the SDN agent and the CN entity, respectively. For a message coming from a CN entity, the formatter changes the formatting applied by the CN entity to the one understood by the SDN agent and then passes it to the mapper. Conversely, when a message arrives at the formatter from the mapper, it formats the payload along with the source and destination address into a format that can be deciphered by the CN entity.

Next, a graphical illustration of the entire message processing chain within the SDN agent has been presented in Figure 5.20(b). Upon the reception of a message from another CN entity, it is passed onto the mapper. Here, the mapper firstly resolves the source and destination SDN-enabled CN entity address A and B, respectively. Concretely, the source address A is mapped to the actual source CN entity address C and, similarly, the destination address B is mapped to the actual destination CN entity address D. Upon performing this mapping, the message is then sent to the formatter. The formatter converts the message payload alongside the source and destination addresses to a format that is understood by the MME/SMF. This is then passed to the modified MME/SMF modules. On the other hand, for an outgoing message, the formatter is the first entity of the SDN agent to process it. The aforesaid processing involves transforming the outgoing message to a format that is understood by the SDN agent. It is then passed onto the mapper wherein the actual source and destination CN entity address, i.e., D and E, is mapped and replaced by its SDN-enabled CN entity address, i.e., B and F, respectively.

This discussed SDN agent architecture can be implemented as a software within the existing CN entities (in which case the mapper in the SDN agent would not be required as the address of both the SDN agent and the CN entity will be the same) or on a generic hardware platform which is interfaced with the existing CN entity hardware. While the former process can be accomplished as a software upgrade at the CN entities, the latter will require additional hardware interfacing and CAPEX for installation. Next, the MME/SMF will entail an additional software upgrade irrespective of the type of SDN agent integration. Note that, we only introduce a software upgrade on the MME/SMF since, it is one of the components of the SeMMu and hence, it will be required to execute the proposed signaling mechanism that involves transformed and compressed (in terms of number of messages) message ensemble, as discussed in Section 5.2. Thus, a message analyzer-generator and a scheduler component have been introduced within the MME/SMF. Figure 5.21 illustrates a block diagram of the transformed MME/SMF. The message analyzer-generator component performs the

function of analyzing the type of message received as well as its IEs, and then generates the appropriate response to the received information in the form of messages from the new message ensemble (Section 5.2). It also generates metadata that informs the scheduler about the possibility of parallelization with a given set of outgoing messages. Subsequently, the scheduler at the MME/SMF determines whether a certain set of messages have to be parallelized or not, depending on the metadata received from the message analyzer-generator block. Here parallelization refers to the fact that messages to multiple CN entities can be executed simultaneously. Hence, the scheduler in the MME/SMF determines the possibility of parallelization, and accordingly passes the set of messages to the formatter entity of the SDN agent. Given the aforesaid functionality, architecturally we define the scheduler in an MME/SMF to perform a bi-directional exchange of information with the message analyzer-generator within that MME/SMF as well as the formatter of the SDN agent integrated with its MME/SMF.

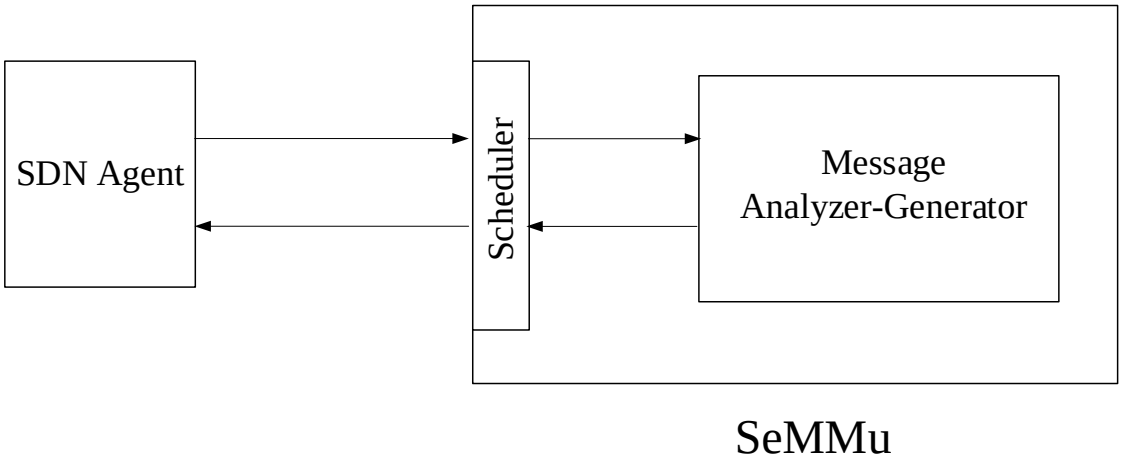


Figure 5.21: SDN-enabled Mobility Management unit (SeMMu) architectural framework.

Note that, we have not provided a graphical illustration of the message processing chain, similar to the SDN agent, for the SeMMu. The reason being, the discussions in Section 5.2 with regards to the compressed message ensemble creation and parallel transfer of HO-related CP messages, essentially presents the main functionalities of the message analyzer-generator and scheduler components, respectively, of the modified MME/SMF in the SeMMu. And so, when a message is received from the SDN agent at the MME/SMF, it is first processed by the scheduler. For the incoming message, the scheduler simply removes the headers within the received frame and passes its payload to the message analyzer-generator module. The message analyzer-generator module then:

- Analyzes the IEs of the received message.
- Determines the response message(s) and generates the required IEs.
- Generates the metadata to be forwarded to the scheduler indicating whether the outgoing message(s) can be parallelized or not.
- Formats the IEs into a message payload.
- Passes the payload along with the destination address to the scheduler.
- Passes the metadata to the scheduler.

The scheduler then forwards the messages accordingly to the SDN agent, where they are further processed according to the process illustrated in Figure 5.20(b). Thus, as a consequence of this integration process, the proposed enhanced HO signaling approach can be executed, while the DP remains agnostic to these transformations.

To conclude this section we note that the SDN capabilities, provisioned to the CN entities for enhancing CP signaling, can be extended to DP functionalities such as data forwarding, path switching, etc. The provision of such an extension enables the proposed architecture to be evolutionary in nature, acting as a bridge between current and envisioned future networks.

5.5 Related Work

Myriad current and past research efforts that provision a comprehensive study into the main stages of handover management, i.e., *handover decision*, *handover preparation*, *handover execution/failure* and *handover completion*. Notably, [44, 103, 161–163] provide sufficient background and analysis into these different stages. In [44] a detailed survey on the various aspects of handover management such as execution phase, decision phase and system information collection has been provided. Concretely, for the network discovery phase, which is the same as acquiring measurement reports from users for handover decision phase, various parameters such as network congestion, channel conditions, etc., have been discussed. Next, for the handover decision making phase, techniques involving multi-attribute decision making, user-centric decision making, etc., have been explored in detail. Following this, for the handover execution phase, methods such as mobile-assisted, network-assisted, etc., have been considered in [44].

Further, in [161] an analysis of the interruption time during the handover phase in an LTE-Advanced network has been performed. Note that, the specific stage of handover phase

that has been enhanced in [161] is the handover execution stage. Further, an analysis with TDD and FDD modes has been considered for the same.

Next, in [103,162,163], SDN based approaches have been considered for mobility management. Specifically, in [162], the SDN controller along side a double V-LAN tagging approach is utilized to minimize the path switching operation which would reduce the latency and core network signaling. In [103] a policy and per-flow based mobility management approach has been presented, wherein the flow level granularity of service provision along side policies specified by network, user, applications, etc., are considered for executing the mobility management task. By policies, here we mean a collection of network, user and application parameters that are utilized in generating a handover decision. Additionally, in [163], an approach towards the seamless mobility management between LTE and WLAN networks has been provided. This approach involves splitting the CP and DP, via SDN based approach, and migrating the CP to the cloud based infrastructure. Hence, with the help of the global view of the network, the controller can facilitate seamless mobility for the user between heterogeneous networks, i.e., LTE and WLAN, through the specified lightweight route reconfiguration procedures.

However, most research efforts similar to [44, 103, 161–163] do not emphasize on the criticality of handover preparation and failure phase. Additionally, they do not explore their latency, transmission cost and processing cost contribution to the overall handover management operations. And hence, this reinforces the novelty of the aspects related to handover management that we explore as well as that of the enhanced signaling strategies for handover preparation and failure signaling phases that we propose. Further, while [183] proposes an integration of the MME with the SDN-C and its utility for handover management using IEEE 802.1ad CN signaling, it does not present an evolutionary mechanism such that the operator CAPEX can be manageable. In addition, the system design focuses on the 3GPP-LTE architecture and does not consider the currently proposed 5G network architecture. Further, research efforts such as [184, 185], present an SDN and NFV based evolutionary network framework for the LTE-EPC. However, like [183], they do not encompass the inter-working architecture with other 3GPP defined technologies such as 3G and the newly defined 5G architecture as well.

5.6 Summary

In this chapter, we have firstly proposed the enhanced messaging mechanism, wherein we transform the critical HO preparation and failure signaling phases for the various 5G NGC and LTE-EPC Inter- and Intra-RAT HO (involving 5G, 4G, 3G and 2G networks) scenarios.

We establish a set of principles that allows us to restructure the messages corresponding to the aforesaid signaling phases. This restructuring helps in compressing the message ensemble and in enabling parallel execution of the messages. Further, a latency, transmission cost, processing cost and message size analysis is conducted, which concludes that the proposed mechanism enhances the legacy handover signaling significantly. We also provision a novel HO failure aware signaling methodology, which accounts for the possibility of a HO failure in the design of the HO preparation signaling. The aforesaid novel strategy is proven to enhance both the preparation as well as the failure phase signaling. Further, and as a means to exemplify the superiority of the proposed mechanism, we present a network wide analysis. Through this analysis we have demonstrated that, for large number of users, the proposed mechanism outperforms the legacy mechanism both in terms of the total processing cost as well as the amount of time the network is occupied to transfer the HO preparation/failure messages.

Lastly, we have proposed an exemplary novel evolutionary architecture that consists of an evolved CN entity, namely, the SeMMu. The evolutionary characteristic of the proposed mechanism helps to maintain a manageable CAPEX. It also facilitates the execution of the aforementioned enhanced HO signaling.

Thus, to conclude, in this chapter we have advanced the work in the area of handover signaling by accomplishing, and verifying analytically, strategies that enhance the process of handover management in terms of latency, processing and transmission overhead. Given the fact that handover management is a critical component of mobility management, the work done in this chapter provisions enhanced mobility management mechanisms, that can cater to future network requirements, for the operators and vendors. Moreover, as part of our MM framework, illustrated in Figure 3.1, the proposed handover signaling methodology will cater to the *Smart CN signaling* and *Handover Management* solution components.

Chapter 6

User Association and Resource Allocation Framework (AURA-5G)

Overview

5G wireless networks, being dense and heterogeneous, will need efficient MM, and specifically user association, strategies to provision the QoS of diverse applications and hence, users that it will serve. Whilst determining the most suitable BS for the users, multiple constraints such as available backhaul capacity, link latency, etc., will need to be accommodated for. Hence, to provide an optimal user association solution, in this chapter we present a joint optimization framework, namely AURA-5G. Under this framework we formulate our user association strategy as a Mixed Integer Linear Program (MILP) that aims to maximize the total sum rate of the network whilst optimizing the bandwidth assignment and base station selection. We analyze multiple active application profiles simultaneously, i.e., enhanced Mobile Broadband (eMBB) and massive Machine Type Communication (mMTC), in the network, and study the performance of AURA-5G. Additionally, we provision a novel study on the multiple dual connectivity modes, wherein the user can be connected to either one Macro-cell and a possible Small-cell, or with any two favorable candidate base stations. Utilizing the AURA-5G framework, we perform a novel comparative study of all the considered scenarios on the basis of total network throughput, performance against baseline scenario and system fairness. We show that the AURA-5G optimal solutions improve the different network scenarios in terms of total network throughput as compared to the baseline scenario, which is a conventional user association solution. Further, we also present a fidelity analysis of the AURA-5G framework based on the user throughput distribution, backhaul utilization, latency compliance, convergence time distribution and solvability. And since, a given network cannot always guarantee to satisfy the future network loads and application constraints, we show that AURA-5G can be

utilized by the operators/vendors to evaluate the myriad network re-dimensioning approaches for attaining a feasible and optimal solution. Henceforth, we then explore the possibility of network re-dimensioning and study its impact on system performance for scenarios where the performance of AURA-5G is severely impacted due to the extremely strict nature of the constraints imposed in the MILP.

Contributions

[J3] **A. Jain**, E. Lopez-aguilera, and I. Demirkol, "User Association and Resource Allocation in 5G (AURA-5G): A Joint Optimization Framework", Submitted to Elsevier COMNET, pp. 1–35, 2020. (Quartile: Q1; IF: 3.03 (2018))

The upcoming 5G networks will be characterized by an extremely dense and heterogeneous amalgamation of BSs with different Radio Access Technologies (RATs), users as well as application types [7]. Such a network environment will present significant challenges to the complex task of mobility management (MM) [C1]. As part of the MM objective for 5G networks, seamless mobility with extremely low latency will need to be provisioned. However, and according to our contribution [J2], to guarantee such latency and reliability characteristics, 5G MM mechanisms will be broadly required to ensure fast handover methods, efficient signaling during mobility events, optimal and fast user to BS associations, reliable and fast path re-configuration as well as service migration.

Specifically, to ensure that the QoS requested by an application on a given user is met, a user does not experience FHO as well as the network capabilities and capacities are respected, efficient user association techniques will be critical. And given the exponentially increasing number of users that 5G networks will cater to [6], finding the optimal user equipment (UE)-BS association will present a significant challenge for these user association techniques. This will be further exacerbated by the fact that 5G networks will be constrained by a multitude of requirements imposed for ensuring application QoS as well as limitations with regards to technological capabilities.

To elaborate, in [186], 3GPP established that to be able to provision services such as VR/AuR among others, a minimum rate requirement of 100 Mbps would be required for enhanced Mobile Broadband (eMBB) services. Further, it was also determined that such services would necessitate anywhere between 5-10 ms latency (or round trip delay) [186]. Alongside these requirements, massive machine type communication (mMTC) services would need to be serviced anytime and anywhere, even though they do not communicate as regularly as the eMBB services. Further, the network would have to accommodate very high density of

mMTC devices that will be prevalent in 5G networks, e.g., 24000 users per km^2 according to [187]. Moreover, the ultra-reliable low latency communication (URLLC) services will require latency within the range of 1-3 ms as well as extreme network reliability [186]. Coupled with these aforesaid requirements, 5G networks will also be challenged by the amount of available resources. Concretely, while mmWave SCs will help resolve the lack of resources in current sub-6 GHz access network, the corresponding backhaul links will become increasingly strained. Further, the availability of BSs with links that satisfy the latency requirements will be critical.

Henceforth, these aforesaid requirements and technological challenges will make the problem of user association alongside resource allocation in 5G heterogeneous networks (HetNets) important to explore and address. Consequently, in this work we aim to do the same. Notably, a broad spectrum of strategies/methods to accomplish the task of user association in 5G HetNets have been discussed in the literature, which we have also taken cognizance of in Section 6.6. However, certain gaps still exist with regards to the aforesaid problem. And so, we elaborate them as follows:

1. Most of the research works discussed in the literature present a user association method that allows the user to connect to only one BS at most [188–191]. Certain works such as [192, 193], etc., discuss the problem of user association in a Dual Connectivity (DC) scenario. However, the analysis is limited to scenarios where SCs are tightly coupled to Macro-cells (MCs). Concretely, this means that the choice of an SC is governed by the choice of the MC. While this is inline with the current 3GPP DC standards in Release-15 [101], it is in general a very restrictive choice. Henceforth, we state that a gap exists here where none of the works in the state of the art, to the best of our knowledge, consider that an independent choice of MC and SC, or even two SCs or MCs, to serve a user can be made. Note that a relatively tangential work to ours in [194] discusses such a possibility for slice level mobility in 5G networks. However, they do not present any concrete methodology or analysis for the purpose of user association.

2. None of the works present in the literature provide an application aware strategy [188–191, 195–204]. Concretely, they do not consider or analyze the impact of the prevalence of eMBB and mMTC services together. This is critical for user association in 5G HetNets, since the presence of different services will lead to different bottlenecks within the same system as shown in this study, thus making the process of finding an optimal association even more complicated.

3. Delving deeper into the analysis presented in the literature, it is evident that for the computation of the signal quality in terms of Signal to Interference and Noise ratio

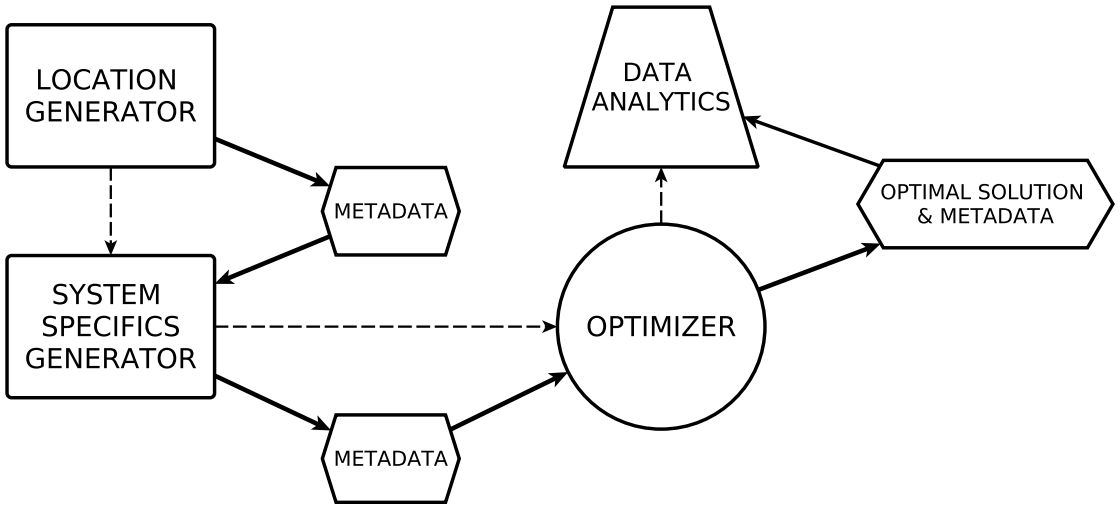


Figure 6.1: AURA-5G Framework. The logical flow, i.e. flow of control, within the developed tool is depicted using dashed arrows, whilst solid arrows indicate the data flow in the program.

($SINR$) an isotropic transmission and reception model is assumed. However, none of the user association works explore the impact of transmit and receive beamforming on the overall system performance, as well as for the complexity of computation of the $SINR$.

4. While certain research efforts discuss the computational complexity of the non-linear optimization framework for the user association problem, e.g. [205–208], none of these works provision a detailed analysis with regards to the computation time, solvability, as well as other network parameters such as achieved latency and backhaul utilization.

Given the aforesaid deficiencies, to the best of our knowledge, we present the very first study in literature with regards to application aware user association in 5G HetNets in this chapter. Concretely, we have explored the prevalence of multiple services and their impacts on the user association problem. Henceforth, for our study we have considered the scenarios where there are only eMBB services, and where both eMBB and mMTC services co-exist. We characterize the performance of our Joint Optimization framework, i.e., AURA-5G, in both these setups and present insights, which currently are not provided by any other research effort. Note that, we leave the study involving URLLC services as part of the future work to this thesis. Additionally, for our evaluation process we utilize realistic network scenarios and parameters. This consequently, helps establish the efficacy of the AURA-5G framework. However, a detailed discussion with regards to the scenarios and parameters is deferred until Sections 6.2 and 6.3.

Furthermore, we also consider the DC scenarios wherein we explore the futuristic trends of having *independent choices of MC and SC, i.e., the choice of SC is not geo-restricted to the coverage of the chosen MC* and *the possibility of selecting either two SCs or two MCs.* In addition to the DC scenarios, we also study the single association strategies (which most research works in the state of the art consider) and a baseline strategy (discussed in detail in Section 6.2). These latter scenarios provide us a basis for comparison for the *AURA-5G* framework, that we have developed.

Moreover, in this chapter we have also presented a detailed study into the performance of such joint optimization strategies when the environment is interference limited due to omnidirectional antennas or when beamforming is utilized. The motivation behind exploring the aforesaid scenarios, is the fact that while most of the current day radio antennas do not utilize beamforming, networks such as 5G and beyond 5G will utilize massive MIMO setups that will support beamforming. Hence, it becomes imperative to study both these scenarios, by virtue of the algorithm being deploy-able irrespective of the infrastructural setup.

Next, as part of the contributions of this chapter, we also emphasize on the AURA-5G software framework which we have developed[1] to obtain the aforementioned deep insights into the user association problem. As can be seen from the framework diagram in Figure 6.1, AURA-5G is basically composed of four building blocks. The very first block, i.e. *location generator* block, takes care of the generation of location specific information for the users and BSs to be utilized during the analysis. Concretely, it generates the location coordinates for the users and BSs within the topology. Next, the *system specifics generator* block creates the backhaul link based details for the system, computes the SINR matrix for the system (according to whether we are in an interference limited scenario analysis or a beamforming based scenario analysis) and saves the necessary metadata that would be required by the subsequent *optimizer* block. For the *optimizer* block we utilize the **Gurobi** toolbox [209] and solve the Mixed Integer Linear Programming formulation (MILP), discussed in Section 6.1, to compute the optimal solution. Note that, in the *optimizer* block we also specify the particular scenario (described in Section 6.2) that has to be evaluated on the system specifics generated by the preceding framework boxes. The *optimizer* block then saves the optimal solution and supporting metadata, which are consequently utilized by the *data representation/analytics* box for gaining insights into the obtained solution for the given scenario.

[1]The complete framework has been developed using Python. It can be found at: *the github repository link will be provided after the review process.*

6.1 The Optimization Framework: Mathematical Formulation and Solver Implementation

In this section, based on the gaps in user association and resource allocation strategies elaborated earlier, we present the mathematical formulation of our joint optimization strategy. Subsequently, we also discuss the challenges that need to be addressed for an efficient implementation of the AURA-5G framework.

And so, we consider a wireless heterogeneous network scenario, wherein 4G-LTE eNBs provide the role of MCs and the 5G gNBs function as the SCs. We denote the set of MCs as $\mathcal{M}$ and the set of SCs as $\mathcal{N}$. We consider that the SCs are connected to the MCs via a backhaul link. This backhaul link can be either wired (fiber based) or wireless (mmWave based). Further, the MCs have backhaul links to the core network. However, these backhaul links are only wired (fiber based) in nature. We denote these aforementioned backhaul links as, B_m and B_n (where $m \in \mathcal{M}$ and $n \in \mathcal{N}$) for the m^{th} MC and n^{th} SC, respectively. Further, the capacity of each of the backhaul links in the considered scenario is denoted as C_m and C_n for the m^{th} MC and n^{th} SC, respectively. Given, the heterogeneous characteristic of the backhaul technologies, their corresponding capacities will also be different. We elaborate more on this in Section 6.3, wherein we describe the system model in detail.

Next, we specify the delay imposed by the backhaul links as D_t, where $t = 1 \dots d$. Here, d is the number of links in the considered scenario. Further, $d > (|\mathcal{M}| + |\mathcal{N}|)$, where $|\cdot|$ denotes the cardinality of the set, because each MC is defined with a backhaul network that has one or more hops to the core network. However, for the purpose of backhaul utilization analysis, the multiple hops from any given MC can be considered together as a single link. This is so because, all the wired hops from MCs are defined to have the same capacity. In addition, for the SCs there is an additional hop (link), i.e., the connecting link to the MC, which may be wired or wireless depending on the operator deployment strategy. Similar to the MC backhaul hops, the wired SC to MC links also have the same capacity but less than that of the MC to CN links. We provide numerical details with regards to this in our system model description in Section 6.3.

The users within the HetNet are deployed using a homogeneous poisson point process (HPP), and are denoted as U_f, where $f = 1 \dots u$, and u being the total number of users within the scenario. For the ease of understanding, we introduce Table 6.1, which contains a list of all the variables, constants and notations that have been utilized for this work. Given these preliminaries, we state the objective of our user association strategy. Concretely, our objective is to maximize the overall system throughput in the downlink, whilst adhering to the various constraints that the 5G HetNets will impose. It is imperative to state here

that, the *AURA-5G* framework is also applicable to the uplink. Specifically, in this work we consider the backhaul capacity, minimum required rate and path latency (one-way downlink delay) as the constraints for our joint optimization problem. Thus, we frame our user association strategy as a MILP problem as follows:

$$\max_{x_{ij},g_{ijk}} \sum_i \sum_j \sum_k x_{ij} g_{ijk} w_k log_2(1+\Psi_{ij}) \tag{6.1}$$

$$\text{s.t.} \sum_i \sum_k x_{ij} g_{ijk} w_k \leq W_j \qquad \forall j \tag{6.2}$$

$$\sum_j x_{ij} \leq 2 \qquad \forall i \tag{6.3}$$

$$\sum_k g_{ijk} \leq 1 \qquad \forall i,j \tag{6.4}$$

$$\sum_j \sum_k x_{ij} r_{ijk} \geq R_i \qquad \forall i \tag{6.5}$$

$$\sum_i \sum_k x_{ij} r_{ijk} \leq C_j \qquad \forall j \in \mathcal{N} \tag{6.6}$$

$$\sum_i \sum_k x_{ij} r_{ijk} + \sum_t \xi_{tj} \leq C_j \qquad \forall j \in \mathcal{M} \tag{6.7}$$

$$p_j x_{ij} \leq l_i \qquad \forall i \tag{6.8}$$

$$x_{ij}, g_{ijk} \in \{0,1\} \qquad \forall i,j,k \tag{6.9}$$

where, x_{ij} indicates the association of user i to BS j. A value of 1 signifies an active association and 0 defines that there is no association. Further, g_{ijk} defines the bandwidth assignment to a user i at BS j, which has k different available bandwidth options. A value of 1 for any given i,j and k combination defines the fact that the bandwidth option k at BS j for user i has been selected, while a value of 0 for the same defines vice versa. In equation (6.1), which defines our total sum rate maximization objective, w_k is a constant value that indicates the actual bandwidth resource in *MHz* for the option k. Next, the constraint defined in equation (6.2) specifies that the total bandwidth resources allocated to all the users associated with BS j cannot exceed the total available bandwidth W_j at BS j. In equation (6.3), we define the dual connectivity constraint wherein a user can select a maximum of two BSs. As we will see later in this section, we modify this constraint to study both the single and dual connectivity scenarios. Subsequently, the constraint in equation (6.4) guarantees that no more than one bandwidth option can be chosen by a user i at an BS j.

Next in equation (6.5) we specify the minimum rate constraint wherein, the sum rate

Table 6.1: Definitions list for Notations, Variables and Constants

i, j, k, f, t	Index variables
$\mathcal{M}$	Set of Macro-cells
$\|\mathcal{M}\|$	Total Macro-cells in the system
$\mathcal{N}$	Set of Small-cells
$\|\mathcal{N}\|$	Total Small-cells in the system
B_m	Backhaul link for Macro-cell m
B_n	Backhaul link for Small-cell n
C_m	Capacity of Backhaul link for Macro-cell m
C_n	Capacity of Backhaul link for Small-cell n
D_t	Delay imposed by link t where $t \in (1, \ldots, d)$
d	Total number of links in the scenario
U_f	User f where $f \in (1, \ldots, u)$
u	Total number of users in the scenario
x_{ij}	Binary variable indicating association of user i with BS j
g_{ijk}	Binary variable indicating selection of bandwidth option k at BS j for user i
w_k	Bandwidth option k at an BS
Ψ_{ij}	SINR registered by user i for BS j
W_j	Total available bandwidth at BS j
r_{ijk}	Composite variable defining the rate offered by BS j to user i with bandwidth option k
ξ_{tj}	Backhaul resource consumption by Small-cell t associated to Macro-cell j
p_j	collection of links defining a path to the core network from BS j
R_i	Minimum Rate constraint for user j
C_j	Collection of backhaul capacities for the Macro and Small-cells
l_i	Maximum bearable downlink latency [delay] for a user i
Γ_{ijk}	Binary variable, introduced for linearization (Section 6.1.1), indicating association of user i with BS j where bandwidth option k has been assigned.
V_{ijk}	A constant, representing $w_k log_2(1 + \Psi_{ij})$

for a user i from the BSs and the corresponding bandwidth options it selects at those BSs has to be greater than minimum rate requirement R_i, for every user in the scenario. We also define a composite variable r_{ijk}, computed as $g_{ijk} w_k log_2(1 + \Psi_{ij})$, which corresponds to the rate BS j offers to user i at bandwidth option k in case there is an active association between them, i.e. $x_{ij} = 1$. Note that, R_i will depend on the type of application a user

accesses, i.e., eMBB, mMTC and URLLC applications will have different minimum rate requirements. In this work, we only consider the eMBB and mMTC applications, and utilize the 3GPP 5G specifications [9] and other literature works such as [210] for their minimum rate requirements.

In equations (6.6) and (6.7), we introduce the backhaul capacity constraint, wherein the total allocated link rate to all the users associated to a given BS j cannot exceed the available link bandwidth C_j. It is important to state here that, the backhaul capacity constraints for the SCs (equation (6.6)) and MCs (equation (6.7)) are characteristically different. This is so because, in the considered scenario, an SC always has a backhaul link, either wired or wireless, to an MC. Henceforth for the MC, it is mandatory that we consider the contribution of the SCs as well in order to ensure that the backhaul capacity constraint is not violated. Consequently in our MILP formulation, in equation (6.7) we introduce the term ξ_{tj}, which specifies the rate consumption by SC t at MC j. It is expressed by the left hand side of equation (6.6), and is equivalent to the capacity of the backhaul link utilized by all the users associated with SC t.

Next, in equation (6.8) we introduce the path latency constraint for each user i. We define l_i as the downlink latency (delay) that an application on user i can permit, based on its QoS requirements. We also introduce an additional system variable, p_j, which specifies the cumulative latency offered by the links that connect BS j to the core network. Here, by a link we specifically mean a wired/wireless hop towards the core network from the BS j. Hence, and as we will observe in further detail in Section 6.3, different BSs will offer different latency (delay) as is the case in real networks. Consequently, the constraint in equation (6.8) will assist the algorithm in selecting an association for all applications in the system that assures that their latency requirements are satisfied.

Lastly, in equation (6.9), we state that x_{ij} and g_{ijk} are both binary variables. Henceforth, this preceding discussion concretizes our joint optimization objective, wherein we not only aim to find the right user-BS association, i.e. x_{ij}, but also the possible bandwidth allocation through g_{ijk}. However, it is important to note here that the multiplication of x_{ij} and g_{ijk} in our objective function, i.e., equation (6.1) and subsequently in the constraints in equations (6.2), (6.5), (6.6) and (6.7), introduces a non-linearity. To resolve this we perform a linearization operation, which is detailed in the following text.

6.1.1 Linearization

To avoid the non-linearity introduced by the multiplicative term involving two binary decision variables, i.e. x_{ij} and g_{ijk}, in our optimization problem formulated in equations (6.1)-(6.9),

we perform a simplistic linearization operation that will enable us to apply our proposed user association strategy as a MILP.

Firstly, we introduce the linearization term in equation (6.10) wherein we replace the multiplicative quantity by a single binary variable.

$$\Gamma_{ijk} = x_{ij} g_{ijk} \forall i, j, k \tag{6.10}$$

where $\Gamma_{ijk} \in \{0,1\}$. A value of 1 denotes the active association of a user i with BS j with bandwidth option k allocated at this BS, while 0 indicates vice versa. Subsequently, we replace $x_{ij} g_{ijk}$ in equations (6.1)-(6.9) with Γ_{ijk}. In order to make this linearization functional, we will also need additional constraints that establish a relationship between Γ_{ijk}, x_{ij} and g_{ijk}. These additional constraints are as follows:

$$\Gamma_{ijk} \leq g_{ijk} \qquad \forall i, j, k \tag{6.11}$$

$$\Gamma_{ijk} \leq x_{ij} \qquad \forall i, j, k \tag{6.12}$$

$$\Gamma_{ijk} \geq g_{ijk} + x_{ij} - 1 \qquad \forall i, j, k \tag{6.13}$$

The aforesaid equations establish the necessary relationship required between the linearizing variable, and the variables comprising the term that is being linearized. Henceforth, we now present our modified MILP formulation, as a result of the aforesaid linearization, in equations (6.14)-(6.25) as follows:

$$\max_{\Gamma_{ijk}} \sum_i \sum_j \sum_k \Gamma_{ijk} w_k log_2(1 + \Psi_{ij}) \tag{6.14}$$

$$\text{s.t.} \sum_i \sum_k \Gamma_{ijk} w_k \leq W_j \qquad \forall j \tag{6.15}$$

$$\sum_j x_{ij} \leq 2 \qquad \forall i \tag{6.16}$$

$$\sum_k g_{ijk} \leq 1 \qquad \forall i, j \tag{6.17}$$

$$\sum_j \sum_k \Gamma_{ijk} V_{ijk} \geq R_i \qquad \forall i \tag{6.18}$$

$$\sum_i \sum_k \Gamma_{ijk} V_{ijk} \leq C_j \qquad \forall j \in N \tag{6.19}$$

$$\sum_i \sum_k \Gamma_{ijk} V_{ijk} + \sum_t \xi'_{tj} \leq C_j \qquad \forall j \in M \tag{6.20}$$

$$p_j x_{ij} \leq l_i \qquad \forall i \qquad (6.21)$$

$$\Gamma_{ijk} \leq g_{ijk} \qquad \forall i,j,k \qquad (6.22)$$

$$\Gamma_{ijk} \leq x_{ij} \qquad \forall i,j,k \qquad (6.23)$$

$$\Gamma_{ijk} \geq g_{ijk} + x_{ij} - 1 \qquad \forall i,j,k \qquad (6.24)$$

$$x_{ij}, g_{ijk}, \Gamma_{ijk} \in \{0,1\} \qquad \forall i,j,k \qquad (6.25)$$

where $V_{ijk} = w_k log_2(1+\Psi_{ij})$, and is a constant since the values for both w_k and $log_2(1+\Psi_{ij})$ are defined/computed beforehand. Further, ξ'_{tj} represents the modified variable for expressing the contribution of the SCs towards the backhaul utilization to an MC. We introduce this modified variable to account for the linearization operation, since the computation of ξ_{tj} in equation (6.7) involves the multiplicative term $x_{ij}g_{ijk}$, as observed from our discussions regarding equation (6.6) and equation (6.7).

6.1.2 Solver Implementation Challenges

While we may have linearized the system of equations for our optimization framework in Section 6.1.1, unfortunately non-linearity still exists given that the variables x_{ij}, g_{ijk}, and consequently Γ_{ijk} are binary in nature. However, we establish that a simplistic approach, wherein we – a) relax the binary nature of the aforesaid variables to bounded constraints, and b) threshold the solution values of these integral variables; can help us avoid such non-linearities. Moreover, solvers such as Gurobi allow the users to program optimization problems, such as ours, and solve them using LP relaxation, branch-and-bound and other advanced mixed integer programming techniques [209]. And so, we utilize this powerful characteristic of Gurobi to solve our optimization framework, and consequently, determine the optimal user association strategy.

In addition, we have developed an implementation framework named *AURA-5G* that also undertakes the tedious task of computing the link SINR matrix. The complexity of this process is highlighted by the fact that in scenarios where there is transmit and receive beamforming, the computation of the link SINR matrix will require the system to know beforehand the beam directions of all the BSs. Concretely, for a UE of interest, all the other UE-BS associations must be known so as to be able to compute the interference from the BSs other than the BS of interest. Note that an BS will only create interference at the UE when the transmit beam of the BS and receive beam of the UE are aligned with each other, whole or in part.

And so, in the following text, through a hypothetical scenario, we show the complexity of computing the aforementioned link SINR matrix. Let us consider the scenario where there

is a UE and Z possible BSs to which this UE can attach to in any receive beam direction. Let us also define a binary variable $\delta \in \{0,1\}$ which indicates whether an BS, through its transmit beam, creates interference for the BS of interest at the UE under observation, i.e., $\delta = 1$, or it does not, i.e., $\delta = 0$. Thus, the total number of combinations of interfering beams (BSs) that needs to be explored to determine the value of SINR, for an BS of interest, is given as $2^{(Z-1)}$.

Next, and for the sake of simplicity, we quantize the number of possible receiver beam directions as Φ. As a consequence, number of computations required to determine the vector of SINR values for a given UE can be expressed as:

$$[(2^{Z-1} + 2)Z]\Phi \tag{6.26}$$

where the additive term of 2 indicates an addition operation for computing *interference plus noise* term and a division operation for ultimately computing the *SINR*. Thus from equation (6.26), it can be seen that the number of computations, and hence the number of combinations that need to be explored, grows exponentially with the number of candidate BSs Z, in scenarios where there is receive and transmit beamforming. This validates our earlier claim regarding the tediousness of computing the SINR matrix.

Certain works in literature, such as [211–213], provide insights as to how a statistical estimate for the SINR can be obtained in a beamforming scenario given a user and multiple candidate BSs. However, these works do not account for the possibility of multiple users in the vicinity of the user of interest. Thus, we utilize a simplified process to determine the SINR at any given UE in a beamformed regime, wherein we only consider the receiver beamforming at the UE and allow the BSs to transmit in an omnidirectional manner. This reduces the number of computations significantly, because now the remaining $Z-1$ BSs will create an interference for the BS of interest. Hence, for Φ quantized receive beam directions, the number of operations required are:

$$[(Z + 1)Z]\Phi \tag{6.27}$$

Comparing equations (6.26) and (6.27), we establish that for a given UE our method utilizes significantly less number of operations to compute the SINR, and hence overcome the earlier said challenge of computing the SINR matrix in a beamformed environment. However, it must be stated that the computed SINR estimate will be a lower bound on the actual SINR value. This is so because, we do not consider the transmit beamforming on the BSs. Consequently, we increase the number of interferers in our computation compared to those

where both transmit and receive beamforming is utilized. Notably though, the efficacy of our analysis for the optimization framework is further enhanced as it utilizes the lower bound, i.e. worst case scenario, for the SINR according to the preceding discussions.

Following this optimization framework, in the next section we introduce the various scenarios that have been explored in this work. We also introduce the necessary modifications to the constraints in the MILP framework to study the corresponding scenarios.

6.2 Scenarios Evaluated

The optimization framework developed in Section 6.1 presented the objective and the multiple real-network constraints that will be utilized when deciding the most optimal BS selection for a given set of users and their corresponding locations. Based on this framework, in this section we introduce the myriad scenarios that have been explored in this work. We also present the necessary modifications, if required, in our optimization framework to study the corresponding scenarios. Table 6.2 illustrates all the scenarios that have been discussed.

6.2.1 Deployment Strategies

For the analyzed scenarios, we generate a set of topologies by deploying the MCs, SCs and users based on the parameters defined in Table 6.4. Of specific interest amongst these is the deployment of SCs within the scenario map. While MCs are at fixed locations, governed by the scenario map size and the MC inter-site distance, the SCs are distributed based on an HPP around each MC. The density is defined in [187] based on the Metis-II project guidelines.

Given these deployment characteristics, we undertake a study on scenarios where these SCs are deployed in a circle of radius $0.5 \times \boldsymbol{ISDMC}$ (see Table 6.4), termed as *Circular Deployment* from here on, and scenarios where they are deployed in a *Square Deployment*. In the latter scenario, the SCs are deployed in a square whose center is at the MC location and the length of each edge is equal to the MC inter-site distance. Note that, while actual deployments will vary depending on operator requirements, *Circular Deployment* provides a realistic and simple deployment strategy for the SCs. Further, and as we will see in Section 6.4 (Figure 6.5), a *Circular Deployment* strategy will lead to areas around MC edges where there will be no coverage via SCs. Hence, to circumvent this issue, we also explore the *Square Deployment* scenario.

The goal of including these deployment strategies into our study is to give the operators

Table 6.2: Analyzed Scenarios

Composite Scenario Name	Circular Deployment	Square Deployment	AnyDC	MCSC	Interference Limited	Beamformed	eMBB	eMBB + mMTC
CABE	✓	–	✓	–	–	✓	✓	–
CMBE	✓	–	–	✓	–	✓	✓	–
CAIE	✓	–	✓	–	✓	–	✓	–
CMIE	✓	–	–	✓	✓	–	✓	–
SABE	–	✓	✓	–	–	✓	✓	–
SMBE	–	✓	–	✓	–	✓	✓	–
SAIE	–	✓	✓	–	✓	–	✓	–
SMIE	–	✓	–	✓	✓	–	✓	–
CABEm	✓	–	✓	–	–	✓	–	✓
CMBEm	✓	–	–	✓	–	✓	–	✓
CAIEm	✓	–	✓	–	✓	–	–	✓
CMIEm	✓	–	–	✓	✓	–	–	✓
SABEm	–	✓	✓	–	–	✓	–	✓
SMBEm	–	✓	–	✓	–	✓	–	✓
SAIEm	–	✓	✓	–	✓	–	–	✓
SMIEm	–	✓	–	✓	✓	–	–	✓

an insight as to how different deployment characteristics can impact the system performance whilst defining a UE to BS association map. This will allow them to understand the benefits and drawbacks of each of these deployment strategies with regards to the joint user association and resource allocation problem. Moreover, it also provides a framework for the operators to introduce their own custom deployments, and analyze the behavior of the user association strategy.

6.2.2 Service Classes

5G, as has been discussed throughout this thesis, will cater to the multiple service classes, i.e., eMBB, mMTC and URLLC [9]. As a consequence, in this work, we study the performance of *AURA-5G* in the presence of only *eMBB service* requests, as well as for the case where *eMBB and mMTC service requests* are generated simultaneously within the topology under study, to show the impact of provisioning of diverse services. Note that, while the eMBB services will request significantly higher throughput (we study the impact of the minimum rate requirements of eMBB services in 5G, as detailed later), mMTC services due to their

relatively higher density but low individual data rates will create bottlenecks in the access and the backhaul networks for the eMBB service requests. However, in this work, for mMTC devices we consider the guard band mode of operation. Hence, the mMTC devices do not consume resources in the access network and just contribute towards the consumption of BH resources. In addition to these services, URLLC services present the unique challenge of ensuring not only low latency but also significantly high levels of reliability from the network. Nevertheless, we postpone our discussion with regards to URLLC services to our future works.

6.2.3 Directivity Regimes

In Section 6.1.2, we briefly commented upon the fact that in this chapter two scenarios, depending on whether beamforming is used, are explored. Concretely, we consider one scenario wherein all transmit and receive antennas have a 360° transmit and receive pattern, respectively. As we will detail next, this will be an *Interference Limited* regime. And so, we also study the behavior of AURA-5G in *beamformed regimes*, wherein, and for the sake of simplicity (See Section 6.1.2 for details), we only consider receiver beamforming at the UE. We elaborate further on the aforesaid *directivity regimes* as follows:

- **Beamformed Regime:** In the scenarios where there is beamforming we consider only receive beamforming at the UE, and utilize it for the purpose of calculating the values of the SINR, i.e. Ψ_{ij}, for all user i and Small-cell BS j pairs[2]. For the Macro-cells we employ sectorization, details of which are specified in our system model in Section 6.3, and allow the UEs to have isotropic reception for the frequencies at which the MCs operate.

- **Interference Limited Regime:** Without beamforming, the environment is flooded with multiple interfering signals, which can deeply degrade the performance of the system. Concretely, for this scenario neither the SC and MC BSs employ any sort of beamforming/sectorization nor do the UEs employ any receiver beamforming. Thus, we also evaluate our MILP framework for user association in such a challenging network environment.

[2]Note that, transmit beamforming at the Small-cell BSs can also be considered here. However, this complicates the computation of the SINR as discussed in Section 6.1.2.

6.2.4 Dual Connectivity Modes

With the current 5G standardization, i.e., Release-15, EN-DC and MR-DC have been formalized as one of the critical features for provisioning higher data rates for users. However, current standards constrain the choice of Small-cells severely, by limiting them to the secondary cell group (SCG) specified by the MC [115]. Hence, in this chapter we outline two DC strategies that build on the current standards and explore the possibility of either – a) MCSC: having an MC and a possible SC (not geo-restricted by the choice of MC), or b) AnyDC: choosing any two possible BSs. We elaborate further on the aforementioned DC connectivity modes, as follows:

- **MCSC:** In this mode of DC, a UE is required to attach to an MC and select at most one SC. The choice of MC does not geo-restrict the choice of SC, as we attempt to go beyond the existing standards on SCG. Additionally, it must be reiterated here that, a connection does not guarantee bandwidth allocation as it is subject to the available physical resources at that moment. Such a scenario can be seen as equivalent to the one where a UE is in a *RRC connected inactive state* at that BS [214].

  And so for this DC mode, the dual connectivity constraint [equation (6.16)] in our optimization framework in Section 6.1 is modified to:

$$\sum_{j \in \mathcal{M}} x_{ij} == 1 \qquad \forall i \tag{6.28}$$

$$\sum_{j \in \mathcal{N}} x_{ij} \leq 1 \qquad \forall i \tag{6.29}$$

  recall that $\mathcal{M}$ and $\mathcal{N}$, as shown in Table 6.1, represent the set of MCs and SCs, respectively. Concretely, equation (6.28) ensures that each UE selects an MC, and equation (6.29) enables them to select at most one SC.

- **AnyDC:** This mode for DC will permit the users to select any two BSs irrespective of the fact that whether they are an MC or an SC. Consequently, we also incorporate this scenario in our study, which, as we will see in Section 6.4, rightly points towards the potential for improved performance but at a higher computational cost as compared to *MCSC* scenarios. By higher computational costs here we refer to the convergence time to the optimal solution, which we study later in Section 6.4.6.

  Hence, to study the *AnyDC* scenario, the dual connectivity constraint in our optimization framework in Section 6.1 is modified as follows:

$$\sum_{j \in \mathcal{M} \cup \mathcal{N}} x_{ij} == 2 \qquad \forall i \tag{6.30}$$

6.2.5 Baseline and Single Association

In our study, we also analyze the *single association* and *baseline association* scenarios as benchmark solutions. Consequently, we conduct a performance comparison of our user association and resource allocation strategy, i.e., *AURA-5G*, with these scenarios based on the obtained performance metrics from the DC modes discussed in Section 6.2.4.

We further elaborate on these two association strategies in the text that follows.

- **Single Association:** As the name suggests, in this scenario we enable the UE to connect to at most one BS. This is in essence what current day wireless networks offer. And so, we modify the dual connectivity constraint in equation (6.16) to:

$$\sum_{j \in \mathcal{M} \cup \mathcal{N}} x_{ij} \leq 1 \qquad \forall i \tag{6.31}$$

  Note that, the Single Association (SA) scenarios along side the DC mode scenarios are of significant interest since they will be prevalent in situations where it is not possible for the network to allocate resources on two BSs for a given UE. Henceforth, we observe and analyze their performance along side the DC mode scenarios and compare it with the baseline scenario, which we elaborate upon next.

- **Baseline Association:** For the baseline scenario, we adopt the user association strategy that is being used by current day mobile networks, Wi-Fi, etc. Concretely, we utilize Algorithm 1, wherein we first compute the SNR that the users would observe from each BS. Based on this observed SNR, we associate the users to the BS with the best SNR. Moreover, to compute the achieveable data rate we utilize the SINR ($\Psi(i,j)$) at the UE for the chosen BS. Given these UE-BS pairs, the bandwidth (B) at any given BS is then divided equally amongst all the UEs associated to it.

6.2.6 Constraint Based Scenarios

In addition to different combinations of topology, DC mode, Directivity and Service Class based scenarios, in our study we introduce an amalgamation of different network constraints, listed in Table 6.3, as well. These myriad combination of constraints are then combined with

Algorithm 2 Baseline Scenario Generation

procedure BASELINEGENERATOR
  $N\_User \leftarrow$ Number of Users; $N\_BSs \leftarrow$ Number of Base Stations
  $R \leftarrow$ Vector of data rates for all users; $N_{BS} \leftarrow$ Vector for number of users per BS
  $m\_id \leftarrow$ Index of the BS with the highest SNR for user i
  $N_{BS} \leftarrow zeros(N\_BSs)$; $R \leftarrow zeros(N\_User)$
  $i, j \leftarrow 1$; $iter\_user \leftarrow N\_User$; $iter\_BS \leftarrow N\_BSs$
  for $i < iter\_user$ **do**
    for $j < iter\_BS$ **do**
      $SNR(i, j) \leftarrow$ SNR of user i from BS j
      $\Psi(i, j) \leftarrow$ SINR of user i from BS j
    $m\_id \leftarrow find(max(SNR(i, :)))$
    $N_{BS}(m\_id) \leftarrow N_{BS}(m\_id) + 1$
  for $i < iter\_user$ **do**
    $idx \leftarrow find(max(SNR(i, :)))$
    $R(i) \leftarrow (\frac{B}{N_{BS}(idx)}) log_2(1 + \Psi(i, idx))$

Table 6.3: Constraint Combinations for Scenarios

Constraint Combination	**Description**
MRT	Minimum Rate Constraint for eMBB services.
CB	The wired backhaul link capacity is capped. For SCs, we cap the capacity of the backhaul to the MC at 1 Gbps, while for the MC to the CN it is 10 Gbps [52].
CPL	eMBB applications will also have latency constraints, although not as strict as the URLLC applications. However, taking the requirements into account, we also explore the impact of constrained path latency.
MRT + CB	Minimum Rate Requirements and Constrained Backhaul together.
MRT + CPL	Minimum Rate and Constrained Path Latency constraints together.
MRT + CPL + CB	Minimum Rate Constraint, Constrained Backhaul and Path Latency constraint, all need to be satisfied simultaneously.
CB + CPL	Backhaul and Path Latency constraints are employed together.

the scenarios in Table 6.2, following which they are optimized and analyzed by the *AURA-5G* framework for user association and resource allocation.

As we will see from our observations in Section 6.4, different combinations of these constraints on the scenarios analyzed have significant impact on the performance metrics. Further, interesting insights that can be utilized by the operator to enhance the performance for the purpose of UE-BS association as well as resource allocation, have been outlined.

6.3 Evaluation Framework

In this section we establish the evaluation framework that we have considered for analyzing the multiple scenarios, described in Section 6.2, by utilizing the optimization framework developed in Section 6.1. Concretely, we consider the topology, as shown in Figure 6.2, with a geographical area of 600 m $\times$ 600 m. The scenario under investigation consists of a heterogeneous multi-layered radio access deployment. For this, we consider the 4G-LTE eNodeB as the MCs operating at the sub-6GHz frequency range, specifically at 3.55 GHz with an inter-site distance of 200 m [187]. Further, we deploy SCs utilizing a homogeneous poisson point process (HPP) within the vicinity of each MC. The number of SCs per MC is chosen from a uniform distribution between 3 to 10. Note that, we repeatedly generate the location coordinates for the SCs, using the aforementioned HPP, until they have a minimum of 20 m inter-site distance [187]. In addition, they operate on the mmWave frequency range of 27 GHz for the access network, i.e., from SC to user, and at 73 GHz for the possible wireless backhaul to the MC in accordance with [215, 216].

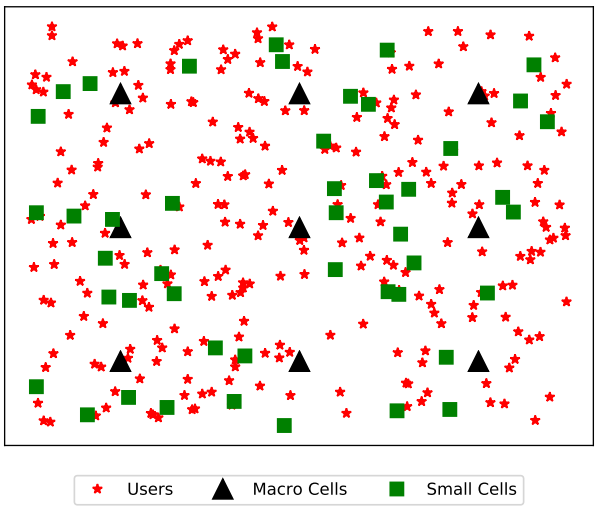


Figure 6.2: Illustrative example of the network topology under study

These SCs are then connected to an MC either via a wireless or a wired backhaul link. We utilize the fact that SCs operating in the mmWave frequency range, due to the operational

Table 6.4: Evaluation Parameters

Parameter Description	**Value**	**Parameter Description**	**Value**
Speed of Light (c)	3×10^8 m/s	LTE eNB (Macro-cell) operating frequency	3.55 GHz
Small-cell access network operating frequency	27 GHz	Small-cell backhaul operating frequency	73 GHz
Height of user	1.5 m	Height of Small-cell	10 m
Height of Macro-cell	25 m	Simulation Area	$0.36 \times 10^6 \text{m}^2$
Number of eMBB users	[150, 175, 200, 225, 250, 275]	Density of mMTC users	24000 per MC
Transmit Antenna Gain for MCBS	17 dBi	Transmit Antenna Gain for SCBS	30 dBi
Macro-cell Transmit Power	49 dBm	Small-cell Transmit Power	23 dBm
UE Receive Gain for Small-cell	14 dBi	UE Receive Gain for Macro-cell	0 dBi
Small-Cell Bandwidth for access network	1 GHz	Macro-cell Bandwidth for access network	80 MHz
Number of hops from MC to core network	$U \in$ [1,4]	White Noise Power	-174 dBm/Hz
Break point distance for wireless backhaul between SC and MC	25 m	Macro-cell Intersite distance (ISDMC)	200 m
Small-cell Intersite distance	20 m	Minimum Rate for eMBB users	100 Mbps
Wired Backhaul Capacity from SC to MC	1 Gbps	Wired Backhaul Capacity from MC to core network	10 Gbps
Wireless link delay	1 ms	Wired link delay	1 ms
Maximum Permissible latency for eMBB services	3 ms	Number of Iterations for evaluation	100
Small-Cell bandwidth options for users (BWSC)	[50 MHz, 100 MHz, 200 MHz]	Macro-cell bandwidth options for users (BWMC)	[1.5 MHz, 3 MHz, 5 MHz, 10 MHz, 20 MHz]
Data rate range for mMTC services	$U \in$ [1,1000] Kbps	UE receive beam Half Power Beamwidth (HPBW)	45°
Pathloss Exponent (Small-cell and LOS condition)	2.1	SF Std. deviation (Small-cell and LOS condition)	4.4
Pathloss Exponent (Macro-cell and LOS condition)	2.0	SF Std. deviation (Macro-cell and LOS condition)	2.4
Pathloss Exponent (Small-cell and NLOS condition)	3.2	SF Std. deviation (Small-cell and NLOS condition)	8.0
Pathloss Exponent (Macro-cell and NLOS condition)	2.9	SF Std. deviation (Macro-cell and NLOS condition)	5.7
Minimum Rate Requirement (eMBB services)	100 Mbps	Latency Requirement (eMBB services)	3 ms
Optimizer Cutoff Time	600 seconds	Number of SCs per MC	$U \in$ [3,10]

characteristics of mmWave, i.e., high atmospheric absorption and severe blockage from the various objects in the transmission (TX) path, will have a significantly reduced TX range as compared to the sub-6 GHz band [217]. Henceforth, we specify a breakpoint distance of 25 m from the MC, beyond which SCs are linked to the MC using a wired backhaul link. This in turn implies that the SCs within the aforementioned distance of the MC connect to it via a wireless link. Further, we specify an out of band operation regime for these wireless backhaul links, wherein the total available bandwidth is divided equally amongst the SCs attached to a given MC. Thus, to compute the available capacity on this link, we utilize the *Shannon-Hartley* theorem specified in equation (6.32).

$$C = W \times \log_2(1 + \Psi') \tag{6.32}$$

where C is the channel capacity, W is the transmission bandwith and Ψ' is the calculated signal to noise ratio between the SC and MC. Additionally, for the backhaul network we consider relevant wired technologies as specified in [52] and deploy a 10 Gbps capacity fiber link from the MC to the core network. The wired backhaul link between SC and MC has a capacity of 1 Gbps. Note that, we dimension the backhaul link capacities such that an MC is able to serve all the SCs connected to it.

Next, we specify that each MC is connected to the core network via wired links. The number of hops for a given MC to the core network is chosen from a discrete uniform distribution over 1 to 4 hops. Further, each of the wired links within our defined topology imposes a delay of 1ms [52]. Additionally, since the SCs can have a wireless backhaul link to the MC, we define that a wireless link also imposes a 1 ms delay [177].

We then deploy the users in the scenario area by utilizing a HPP. As specified in our discussions in Section 6.2, we consider scenarios where either only eMBB devices or both eMBB and mMTC devices exist. Hence, for the purpose of analysis we consider the various user densities for both eMBB and mMTC devices, as listed in Table 6.4. We simplify our evaluation framework by utilizing the fact that mMTC devices operate in the guard band mode. Hence, they consume only backhaul resources in our evaluation framework. Further, in [210], it is stated that mMTC devices generate traffic between 1 Kbps and 1 Mbps mostly. Consequently, we consider a uniform distribution between 1-1000 Kbps and utilize it to compute the backhaul network resources consumed by the mMTC devices.

We also specify the bandwidth options that a UE has from a given SC as well as an MC [218]. In addition, for the channel model we adopt the NYU CI model [215, 219], which is expressed as follows:

$$PL = FSPL + 10n \log_{10}(d/d_0) + X_\sigma \tag{6.33}$$

where PL defines the pathloss in dB, $FSPL$ [in dB] is computed as $20log_{10}\left(\frac{4\pi f d_0 \times 10^9}{c}\right)$, d_0 is 1m, and X_σ denotes the shadow fading component with a standard deviation of σ. Based on the experiments carried out in [215] we adopt the pathloss coefficient, i.e., the value of n, and the standard deviation σ for shadowing. These values have been specified in Table 6.4. Note that, we consider both the Urban Micro (U-Mi Street Canyon) and the Urban Macro (U-Ma) scenarios in [215, 219] as they are reflective of the scenarios that will prevail for SC and MC, respectively, in a dense urban environment. Moreover, we also take into account the possibility of encountering obstacles in such dense urban environments by simulating the LOS-NLOS probability models for U-Mi Street Canyon and U-Ma scenarios as specified by 3GPP in [220]. For U-Mi Street Canyon the LOS probability model is expressed as shown in equation (6.34), where d_{2D} is the two dimensional distance between transmitter and receiver. Further, the U-Ma LOS probability model has been expressed, in a manner similar to the U-Mi model, in equations (6.35) and (6.36), where h_{UT} represents the height of the user terminal, i.e. UE. Note that, we compute the NLOS probability by simply utilizing the fact that $P_{NLOS}^{MC/SC} = 1 - P_{LOS}^{MC/SC}$.

Lastly, we also provision parameters such as the MC height, SC height, UE height, transmit and receive gains, intersite distances as well as QoS requirements for the services discussed in this chapter (minimum rate and latency). These parameters have been derived utilizing 5GPPP project proposals [187], 3GPP specifications [221] and other relevant research efforts [216].

Given the setup detailed so far, we perform 100 Monte Carlo runs for each scenario and constraint combination. These Monte Carlo trials help us attain a certain measure of confidence over our observations. Additionally, we also define a cutoff period of 600 seconds for our optimizer to determine a solution for the user association problem. The reason for

$$P_{LOS}^{SC} = \begin{cases} 1 & ,d_{2D} \leq 18m \\ \frac{18}{d_{2D}} + exp^{(-\frac{d_{2D}}{36})(1-\frac{18}{d_{2D}})} & ,18m < d_{2D} \end{cases} \tag{6.34}$$

$$P_{LOS}^{MC} = \begin{cases} 1 & ,d_{2D} \leq 18m \\ \left[\frac{18}{d_{2D}} + exp^{(-\frac{d_{2D}}{63})(1-\frac{18}{d_{2D}})}\right]\left[1 + C'(h_{UT})\frac{5}{4}\left(\frac{d_{2D}}{100}\right)^2 exp^{(-\frac{d_{2D}}{150})}\right] & ,18m < d_{2D} \end{cases} \tag{6.35}$$

$$C'(h_{UT}) = \begin{cases} 0 & ,h_{UT} \leq 13m \\ \left(\frac{h_{UT}-13}{10}\right)^{1.5} & ,13m < h_{UT} \leq 23m \end{cases} \tag{6.36}$$

such a cutoff timer being, in any dynamic network environment such a time period would be more than sufficient to determine an optimal association. And so, we now list all the other simulation parameters, along side the parameters discussed thus far in this section, and their corresponding description and values in Table 6.4.

With this background, in the next section we evaluate the performance of the AURA-5G framework based on *Total network throughput*, *System fairness*, *User Throughput distribution*, *Backhaul utilization*, *Latency compliance*, *Convergence time* and *Solvability*. It is important to state here that, the scenarios, detailed in Section 6.2, along side the parameters, elaborated upon earlier in this section, provision a very realistic scenario. As a consequence, this accentuates the efficacy of our framework in provisioning a realistic and implementable framework for industry and academia.

6.4 Results and Discussions

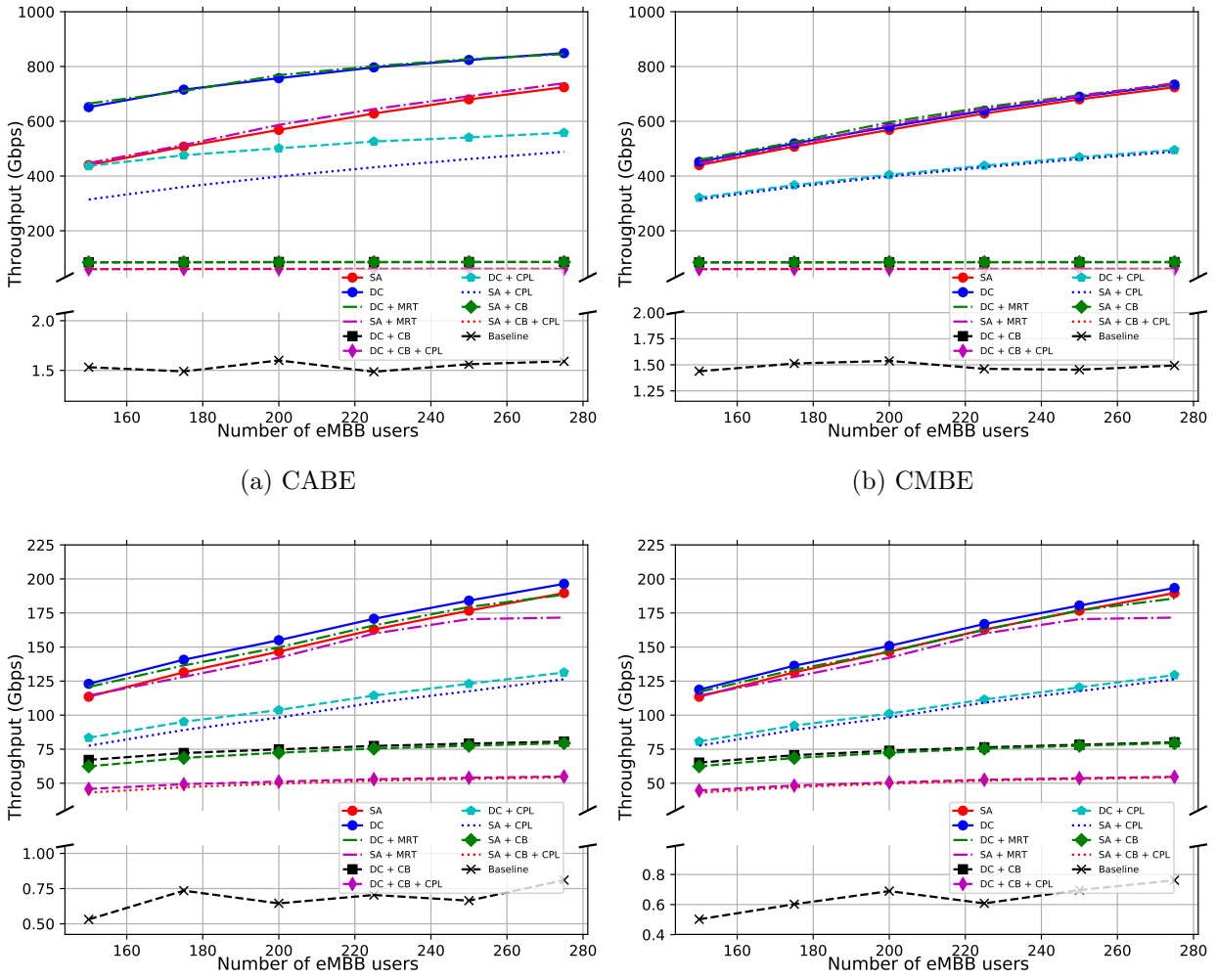
Based on the evaluations performed by utilizing the AURA-5G framework, in this section and Section 6.5 we consolidate and discuss our findings in detail. We structure our discussion into two phases wherein at first (in Section 6.4) we consider the setup that utilizes the evaluation parameters in Table 6.4 as is. The observations for this setup considers scenarios with only eMBB users and, eMBB and mMTC users together. Secondly, based on crucial observations from the first phase, in Section 6.5 we perform network re-dimensioning and then present details on how AURA-5G can assist the operator in gaining insights that will ultimately lead to improved system performance through re-dimensioning.

We now proceed towards our discussion, and reiterate that we utilize the notations presented in Tables 6.2 and 6.3 for the myriad scenarios explored in this chapter.

6.4.1 Total Network Throughput

6.4.1.1 eMBB services based scenarios

For the scenarios where users with only eMBB services are considered, we present our observations with regards to the total network throughput in Figures 6.3 and 6.4. For the purpose of comparison, we also include the observations from the baseline scenario in Figures 6.3 and 6.4. Note that, the observations presented have been obtained after averaging the total network throughput over 100 Monte Carlo trials. In addition, the Minimum Rate (MRT) constraint (see Table 6.3 for description), due to its strict nature, can lead to circumstances where either an optimal solution does not exist or it takes too long (greater than 600 seconds) for the optimizer to search for one. In either case, we consider these simulation trials to be



(a) CABE

(b) CMBE

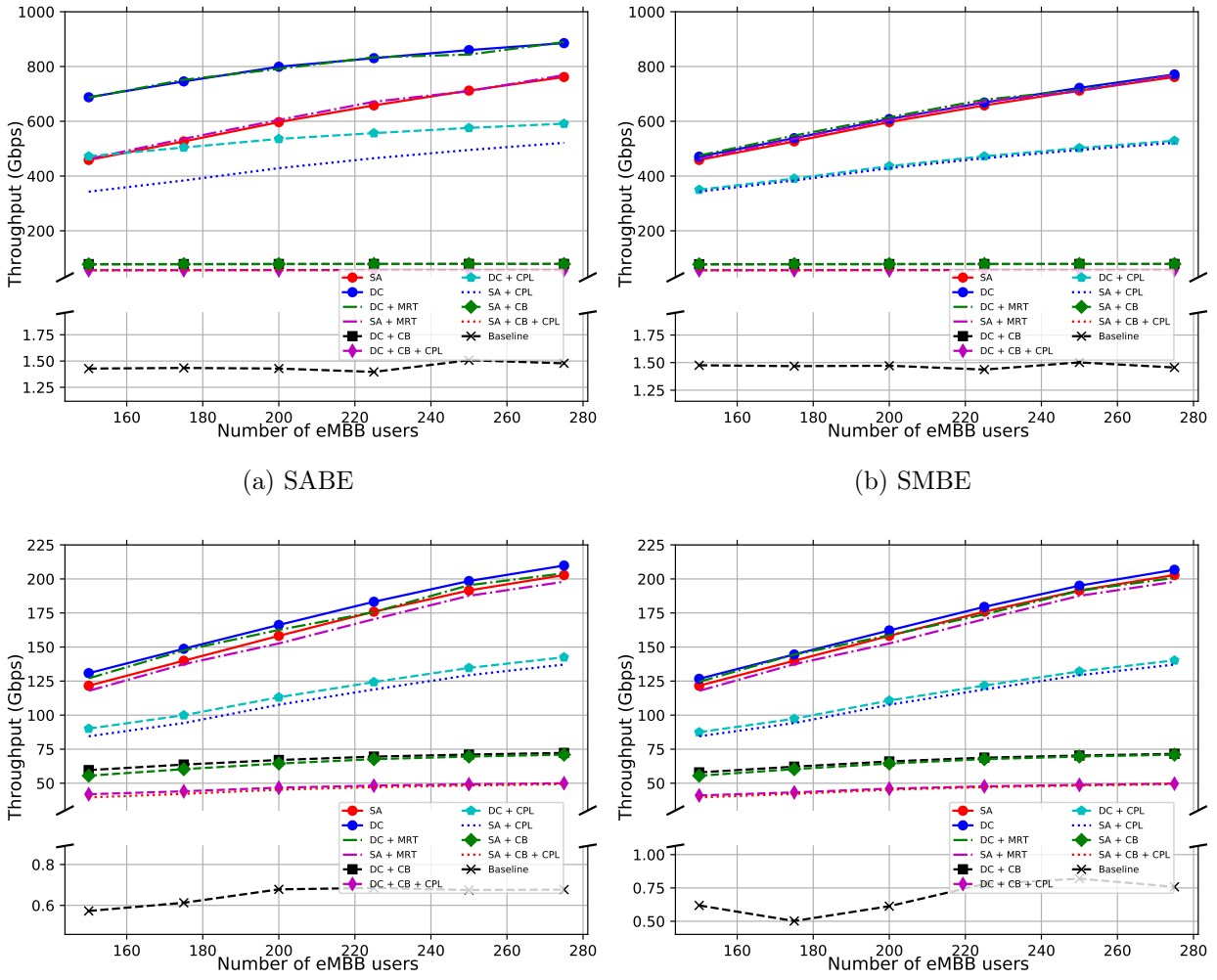
(c) CAIE

(d) CMIE

Figure 6.3: Total Network Throughput for multiple combination of constraints being employed on (a) CABE, (b) CMBE, (c) CAIE and(d) CMIE scenarios.

unsuccessful with regards to finding an optimal solution, and hence, exclude them from the evaluation of the AURA-5G framework for the *total network throughput*, *system fairness*, *backhaul utilization* and *latency compliance* metrics. We refer the reader to Section 6.4.7 and 6.5, wherein a more detailed discussion with regards to the issue of *solvability* and how it is addressed has been provided. Henceforth, for the *total network throughput* analysis, we now evaluate the *CABE*, *CMBE*, *CAIE* and *CMIE* scenarios in Figures 6.3(a)-(d), where multiple combination of constraints (specified in Table 6.3), beamformed and interference limited regime, and circular deployment have been considered.

From Figures 6.3(a)-(d), firstly we deduce that the AURA-5G framework outperforms the



(a) SABE

(b) SMBE

(c) SAIE

(d) SMIE

Figure 6.4: Total Network Throughput for multiple combination of constraints being employed on (a) SABE, (b) SMBE, (c) SAIE and (d) SMIE scenarios.

baseline scenario for all set of constraints and scenarios. Next, the DC scenarios outperform the corresponding SA scenarios when *AnyDC* is employed (Figure 6.3(a)). However, when *MCSC* is employed the gains are not as significant (Figure 6.3(b)), because with DC in *MCSC* the UEs are connected to one MC and can additionally connect to at most one SC. Further, in SA, due to the nature of our optimization methodology being to maximize the total sum rate, the UEs are associated mainly to the SCs. Hence, the gains for DC scenarios in the *MCSC* setup are not as significant as those in *AnyDC*. Moreover, from Figures 6.3(a) and (b), it can be observed that constrained path latency (CPL) and backhaul (CB) severely impact the overall network throughput. The reason being that the BSs with the best available SINR,

and hence capacity, might not be able to satisfy these latency and backhaul constraints.

We then consider the interference limited regime based scenarios, i.e. *CAIE* and *CMIE*, in Figures 6.3(c) and (d), respectively. Immediately we can observe a significant reduction in the total network throughput as compared to that in the beamformed regime (Figures 6.3(a) and (b)). This is inline with our expectations, since the SINR in an interference limited scenario will be significantly degraded as compared to that observed in the beamformed regime. Further, due to this interference limited nature of the scenarios, in Figures 6.3(c) and (d), we do not observe a significant gain in performance from *AnyDC* over *MCSC*.

Next, we analyze the square deployment based scenarios, i.e. *SABE*, *SMBE*, *SAIE* and *SMIE*, in Figures 6.4(a)-(d). To reiterate, in square deployment based scenarios, the SCs are distributed in a square geometry around each MC to which they have a backhaul link. Given these scenarios, from Figures 6.4(a)-(d), the generic trend of observations does not change compared to the circular deployment. Concretely, we observe that the AURA-5G framework, given any set of constraint combinations and scenarios always outperforms the baseline scenario. Further, beamformed regime scenarios perform better than their interference limited regime counterparts (Figures 6.4(a)-(b) and 6.4(c)-(d)), *AnyDC* based DC scenarios have a significant performance gain over SA scenarios, which is not the case with *MCSC* based DC scenarios (Figures 6.4(a)-(b) and 6.4(c)-(d)) and, latency and backhaul constraints significantly reduce the total network throughput (Figures 6.4(a)-(d)).

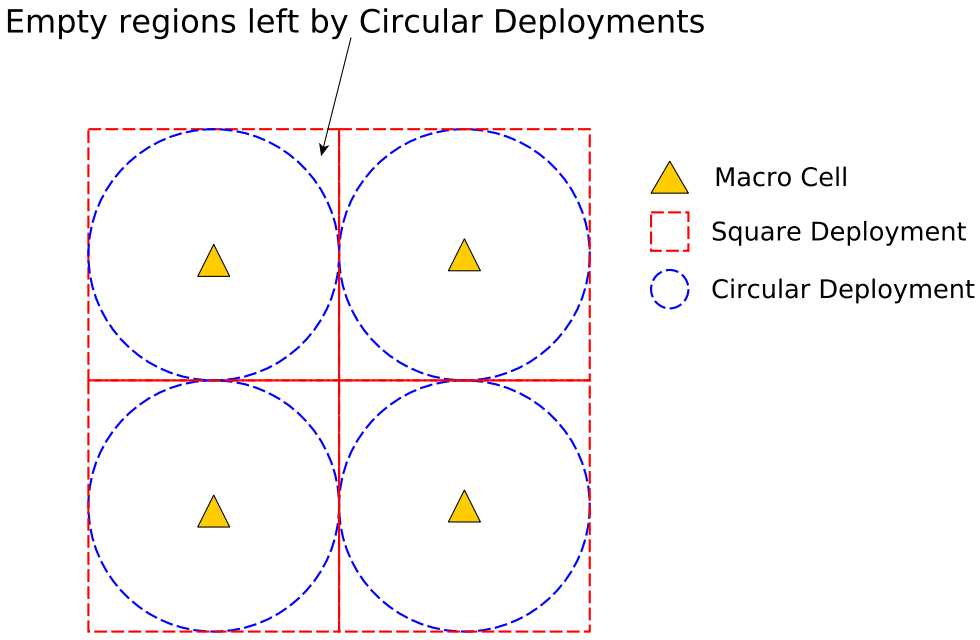
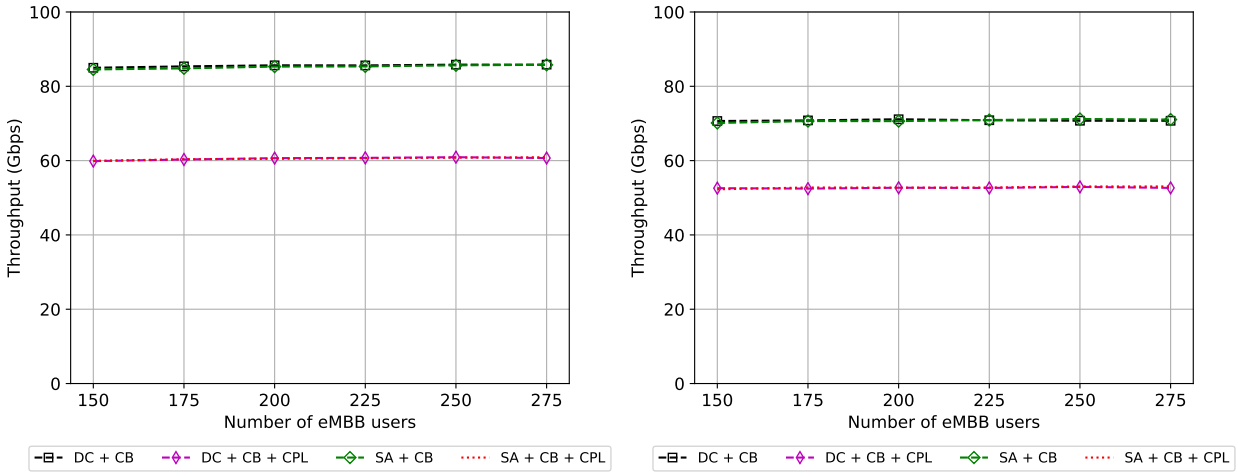


Figure 6.5: Circular and Square deployment characteristics for SCs around MCs.

However, in addition to these aforesaid observations, the square deployment based scenar-



(a) CABE

(b) CABEm

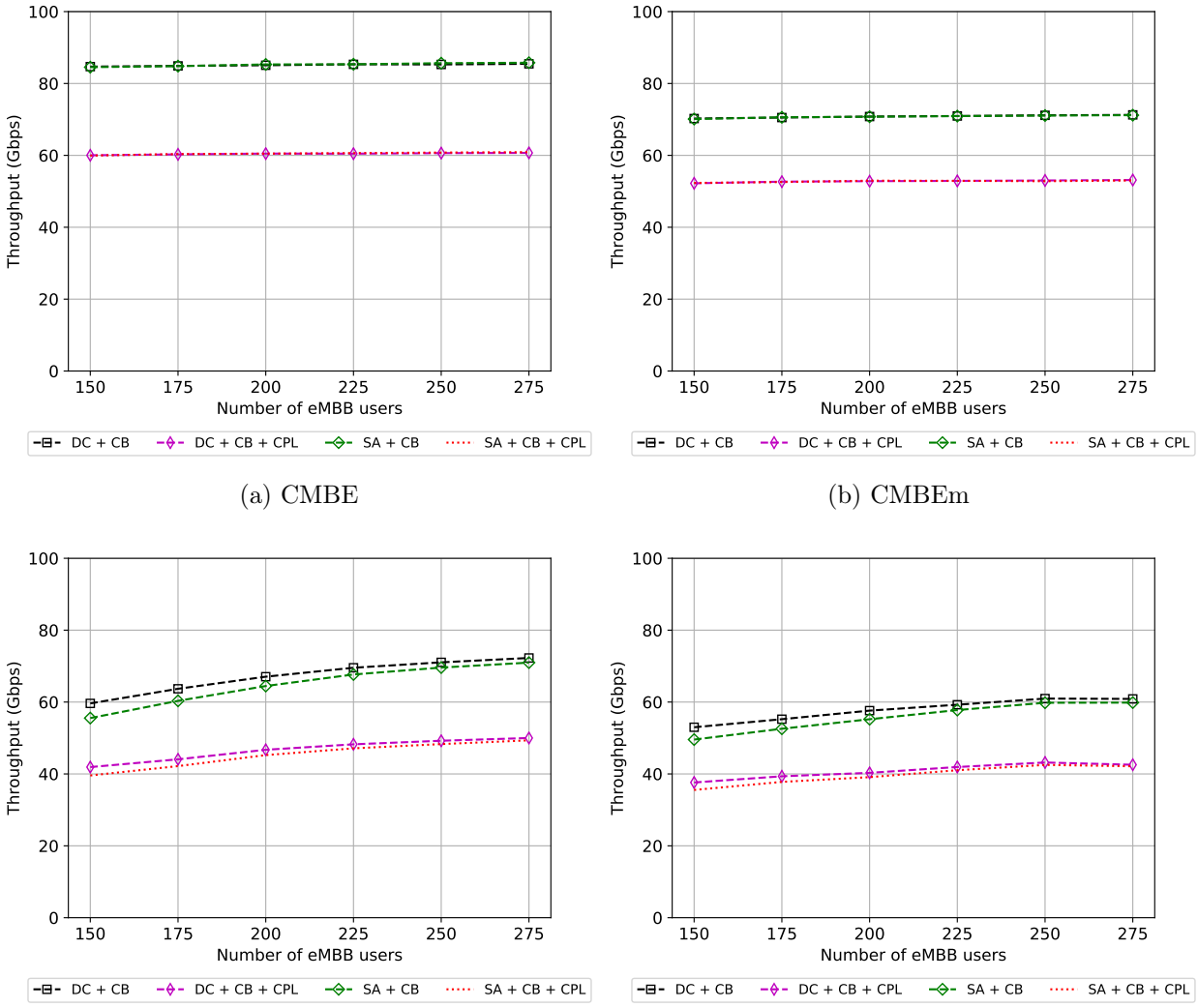
(c) CAIE

(d) CAIEm

Figure 6.6: Total Network Throughput for multiple combination of constraints being employed on (a) CABE, (b) CABEm, (c) CAIE and (d) CAIEm scenarios.

ios provision approximately 6% increase in total network throughput across all the constraint combinations explored, except when backhaul capacity constraints are applied. The reason being:

- In a circular deployment based scenario, the SCs are deployed around the MCs such that there will be blind spots, i.e., there will be areas where there is weak or no SC coverage at all, since circular geometries leave empty spaces where their edges meet. However, with a square deployment scenario the probability that there are such blind spots is less, as square geometries do not leave any empty spaces like their circular



(a) CMBE

(b) CMBEm

(c) SAIE

(d) SAIEm

Figure 6.7: Total Network Throughput for multiple combination of constraints being employed on (a) CMBE, (b) CMBEm, (c) SAIE and (d) SAIEm scenarios.

counterparts. This consequently gives users in these geographically transitional areas a better opportunity to connect with an SC. An illustration to highlight the aforesaid characteristic of circular and square deployments has been shown in Figure 6.5.

- A square deployment configuration however means that the probability that SCs are further away from the MC is higher. This consequently increases the probability of employing a wired backhaul from SC to MC which, as can be seen from our evaluation framework in Section 6.3, might have lower capacity than the mmWave based wireless backhaul. Hence, this reduces the overall data carrying capacity of the network which

corresponds to the stated reduced total network throughput as compared to the circular deployment when backhaul capacity constraints are employed.

6.4.1.2 mMTC with eMBB services based scenarios

For the scenarios where both mMTC and eMBB services co-exist, we firstly re-iterate the fact that the mMTC devices only consume backhaul resources (see Section 6.3). Hence, the main total network throughput characteristics stay similar to those observed for the scenarios where only eMBB services exist. However, certain scenarios where we take into account the backhaul capacity constraints show interesting observations. Consequently, in Figures 6.6(a)-(d), for the *CABE*, *CABEm*, *CAIE* and *CAIEm* scenarios we illustrate the total network throughput only when CB and CPL (see Table 6.3 for descriptions) constraints have been imposed. Note that, the above mentioned scenarios are all circular deployment based.

Concretely from Figures 6.6(a)-(d), we observe that the presence of mMTC devices leads to a reduction in the overall network throughput experienced by the eMBB services. This is along expected lines, given that the mMTC devices, as stated before, consume a portion of the available backhaul capacity. In addition, it can be seen that the total network throughput for the beamformed regime (Figure 6.6 (a) and (b)) is much higher than that observed in an interference limited regime (Figure 6.6(c) and (d)).

Next we consider, through Figures 6.7(a) and (b), scenarios *CMBEm* and *CMBE*, where *MCSC* configuration for DC modes is utilized. We observe that, for the scenarios where mMTC services also exist along side the eMBB services, the total network throughput achieved by the eMBB services is lower. The reasoning, as has already been stated before in this section, is that the mMTC devices consume a portion of the available backhaul capacity thus reducing the overall achievable throughput for the remaining services in the system. We further present the square deployment scenarios in an interference limited regime in Figures 6.7(c) and (d), and compare them with their circular deployment counterparts presented in Figures 6.6(c) and (d). We deduce that, as compared to the circular deployment scenarios in Figures 6.6(c) and (d), the total network throughput observed for the square deployment scenarios is lower when CB and CPL constraints are considered. The reason being, and we re-iterate, that a square deployment configuration leads to a higher probability of the SCs being further away from the MC. As a consequence, this increases the probability of employing a wired backhaul from SC to MC which might have lower capacity than the mmWave based wireless backhaul. Hence, this reduces the overall data carrying capacity of the network which corresponds to the stated reduced total network throughput as compared

to the circular deployment.

6.4.2 System Fairness

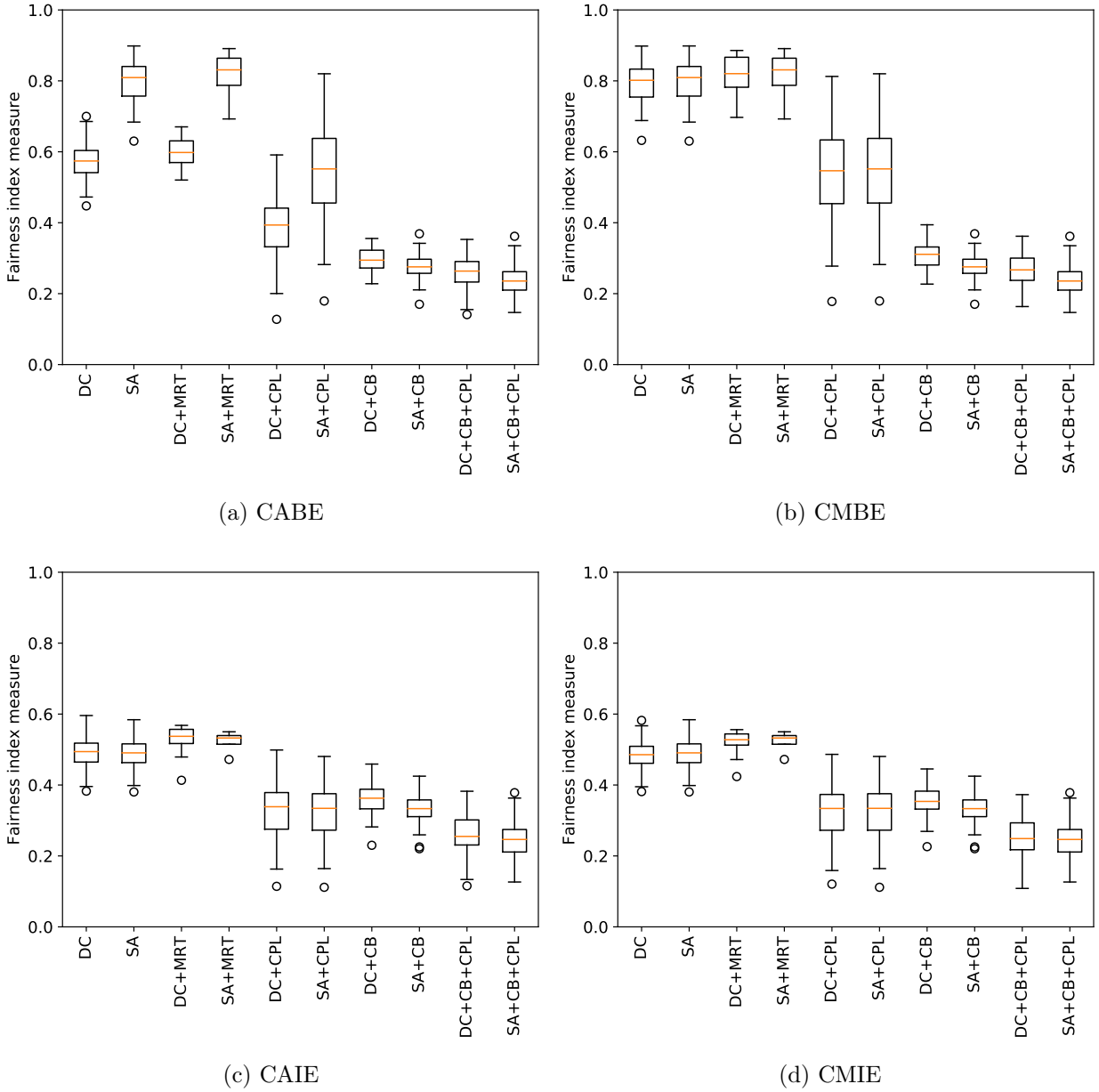
We analyze the fairness of our optimization based framework through the Jain's Fairness Index [222] for the scenarios explored in this work. The fairness index is computed for the individual user throughputs, covering the constraint combinations and scenarios that have been discussed in Section 6.4.1, and then a detailed discussion has been provided. It is important to state here that the objective of evaluating the AURA-5G framework for fairness measure is, *to be able to study the impact of various constraints and scenarios, prevalent in the 5G networks, on the fairness offered by the system, given the objective function of total sum rate maximization* (Section 6.1).

6.4.2.1 eMBB service based scenarios

In Figures 6.8(a)-(d), the Jain's fairness index for the scenarios *CABE*, *CMBE*, *CAIE* and *CMIE* with the different constraint combinations have been illustrated. It is important to state that, the boxplot based representations of the system fairness index, in Figures 6.8-6.10, have to be interpreted as follows:

- The white box represents the user throughput values encompassing the first to the beginning of the last quartile, i.e., from 25% to 75% of user throughput values.
- The red line represents the median value of the user throughput values for a given scenario.
- The whisker extend to the top and bottom by 1.5× the value of the quartile range in point# 1, for any given scenario, respectively.
- The remaining values are plotted as outliers, and represented by circles.

Specifically, from Figure 6.8(a) we observe that the AURA-5G framework in the Single Association (SA) setup provisions a higher fairness as compared to the Dual Connectivity setup, except when backhaul capacity constraints are considered. The reason being, since SA allows the UEs to connect to at most one BS, network resources are more evenly distributed and hence more users are able to connect and reserve resources in the network. However, since in DC the UEs have the possibility of connecting to two BSs, the amount of resources available per user in the network is significantly less. Hence, the disparity in the amount of resources reserved by the users in DC modes is much higher. This, as a consequence, results



(a) CABE

(b) CMBE

(c) CAIE

(d) CMIE

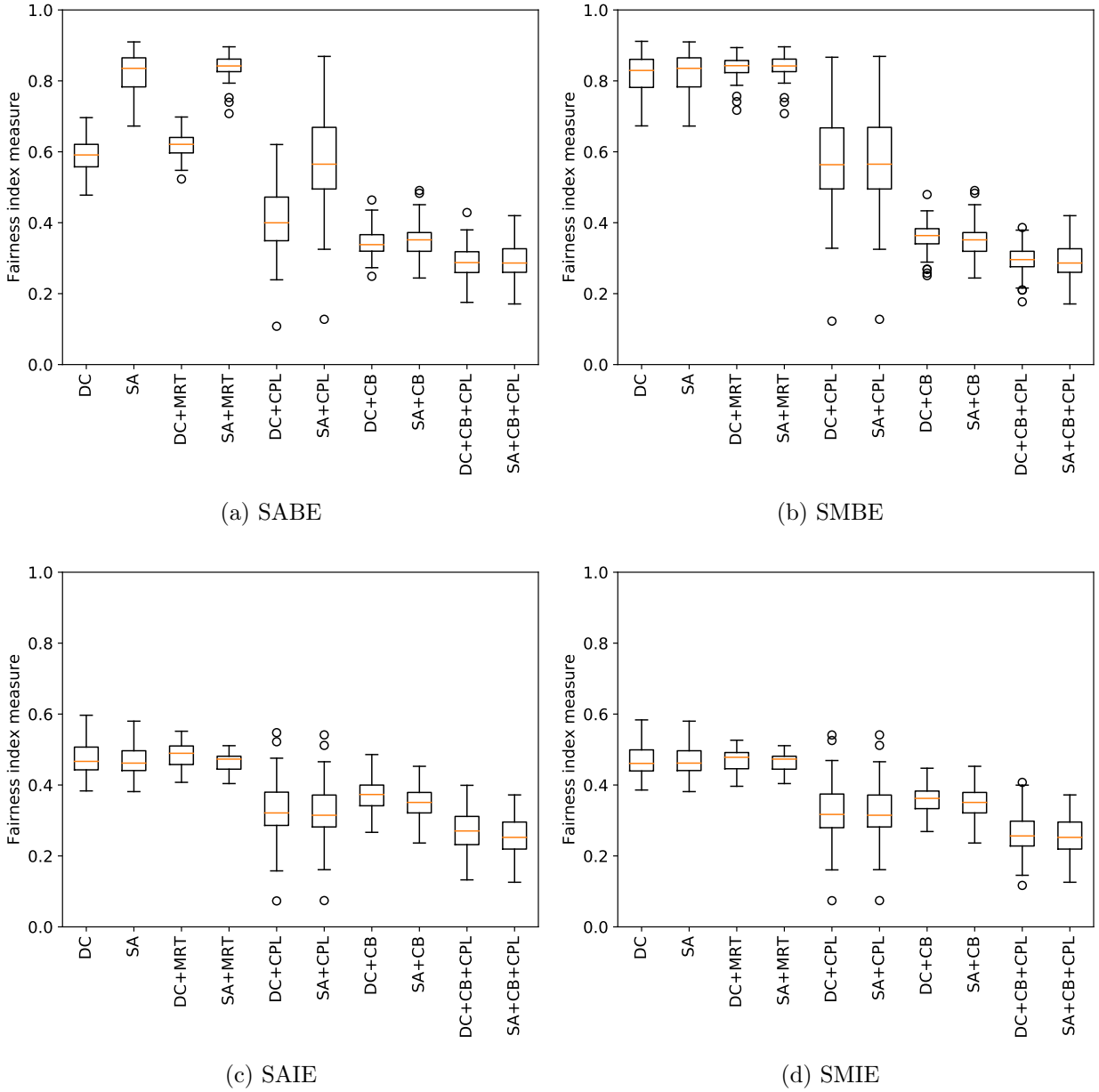
Figure 6.8: Jain's Fairness index deviation measure for user throughputs over multiple combination of constraints being employed on (a) CABE, (b) CMBE, (c) CAIE and (d) CMIE scenarios.

in the aforementioned fairness characteristic. However, as can be seen in Figure 6.8(a), when we consider the MRT requirement constraint there is a slight improvement in the fairness, because the algorithm tries to find a solution wherein each user is allocated at least 100 Mbps. This forces the system to allocate resources more fairly so as to satisfy the minimum rate constraint and thus results in the marginal increase in fairness, as observed.

Further, in Figure 6.8(a), we observe that the imposition of CPL and CB constraints results in significant lowering of the overall system fairness. This is as a consequence of only a small subset of BSs and backhaul paths being able to satisfy the set constraints. Hence, in order to satisfy these requirements the UEs share the limited available resources on these candidate BSs and backhaul links. Moreover, given that the algorithm aims to maximize the total sum rate of the system, UEs with better SINR are allocated better bandwidth resources from the limited available system resources. Hence, this creates significant disparities between the throughput of different users, thus leading to a reduction in system fairness. Lastly, DC scenarios provision better fairness than the SA scenarios when CB constraint is applied. This is so because, the users have the opportunity to compensate for the lack of backhaul capacity resources in one link by acquiring bandwidth resources in the other link connected to the second BS selected. However, in SA, the lack of backhaul capacity resources combined with the nature of the optimization algorithm to maximize the total sum rate leads to a significant disparity in how the system resources are allocated to the users.

Next, in Figure 6.8(b), for the *CMBE* scenario wherein *MCSC* setup is utilized, an overall improvement in the system fairness in the DC modes is observed. This is as a result of the fact that the users are now forced to select one MC amongst the two BSs they choose. Hence, this relieves resources from the SCs which are the main drivers for maximizing the total sum rate of the system. However, this was not the case in the *AnyDC* scenario, wherein users could select even two SCs. As a consequence, *MCSC* provides more users with the opportunity to select an SC and maximize their possible data rate, which leads to the improvement in the system fairness, as stated before. Moreover, and as expected, for the interference limited regime scenarios shown in Figures 6.8(c) and (d), the fairness measures are significantly lower as compared to the beamformed based scenarios (Figures 6.8(a) and (b)). Given the severe interference and the objective of maximizing sum rate, only a few users will have a good SINR, which as a consequence will receive the maximum share of the network resources. Hence, this leads to the severe disparity in the achievable rate per user, which subsequently explains the drop in the system fairness. Note that, rest of the trends in system fairness measures for the interference limited regime scenarios follow those already observed for the beamformed regime scenarios.

Following the discussion for circular deployment based scenarios, we next consider the square deployment based scenarios, i.e., *SABE*, *SMBE*, *SAIE* and *SMIE*, in Figures 6.9(a)-(d). From these figures, we observe that the generic trend for the fairness measure is similar to those observed for the circular deployment scenarios (discussed above). However, the square deployment for certain constraint combinations and scenarios enhances the overall system fairness. An example being the SABE scenario, wherein for all constraint combinations we



(a) SABE

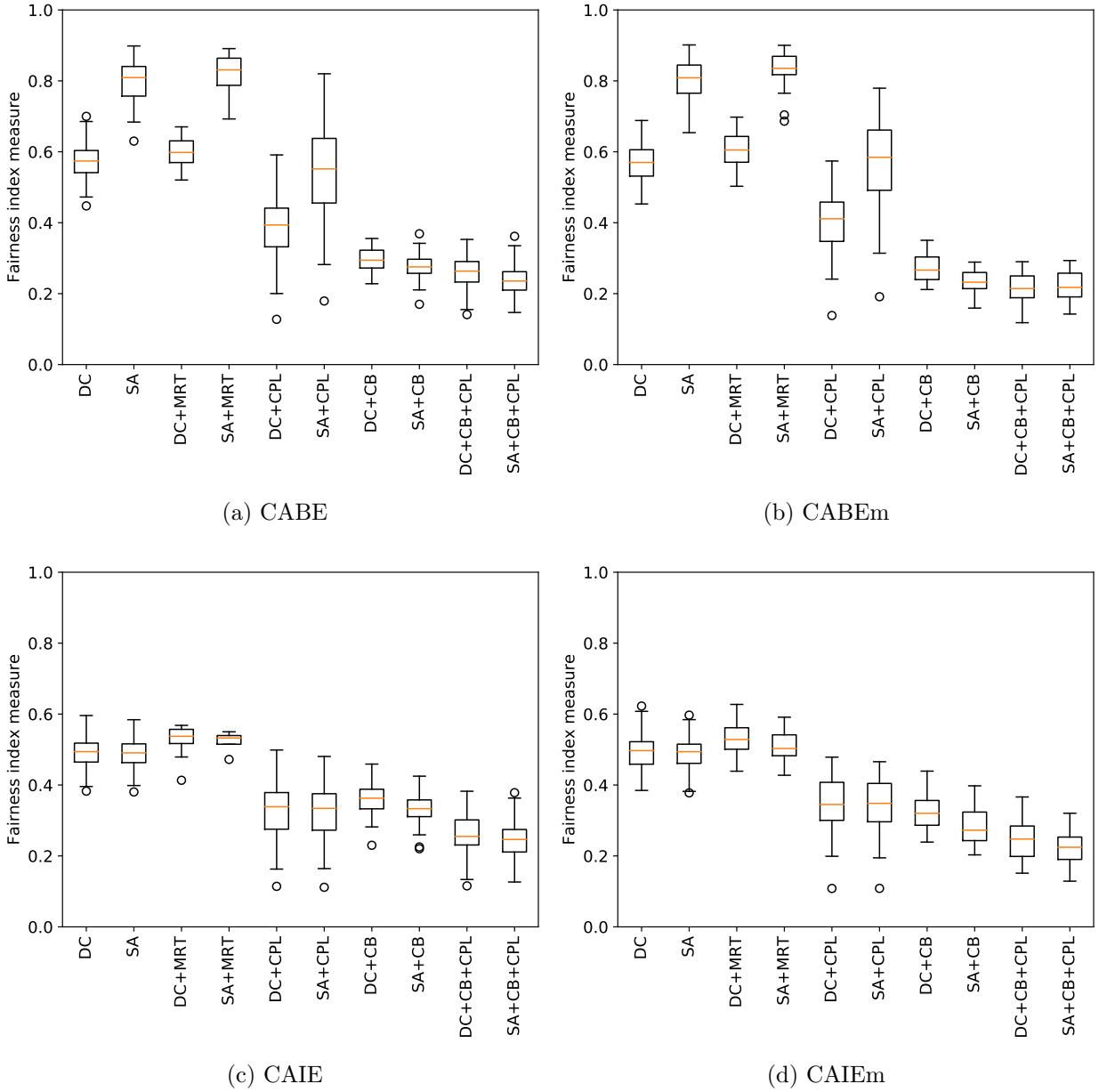
(b) SMBE

(c) SAIE

(d) SMIE

Figure 6.9: Jain's Fairness index deviation measure for multiple combination of constraints being employed on (a) SABE, (b) SMBE, (c) SAIE and (d) SMIE scenarios.

observe between 5-6% improvement in system fairness. This is because of the reasons we have already elaborated in Section 6.4.1.2, i.e., square deployments result in less blind spots within the deployment, hence resulting in a fairer allocation of resources to the users as compared to the circular deployment.



(a) CABE

(b) CABEm

(c) CAIE

(d) CAIEm

Figure 6.10: Jain's Fairness measure for multiple combination of constraints being employed on (a) CABE, (b) CABEm, (c) CAIE and (d) CAIEm scenarios.

6.4.2.2 mMTC with eMBB based scenarios

For the scenarios where mMTC and eMBB services are considered together, we present our observations through Figure 6.10. It can be seen from Figures 6.10(a)-(d) that the fairness index does not change significantly as compared to the fairness measure observed in eMBB only scenarios, even though we consider the mMTC devices within our framework. The

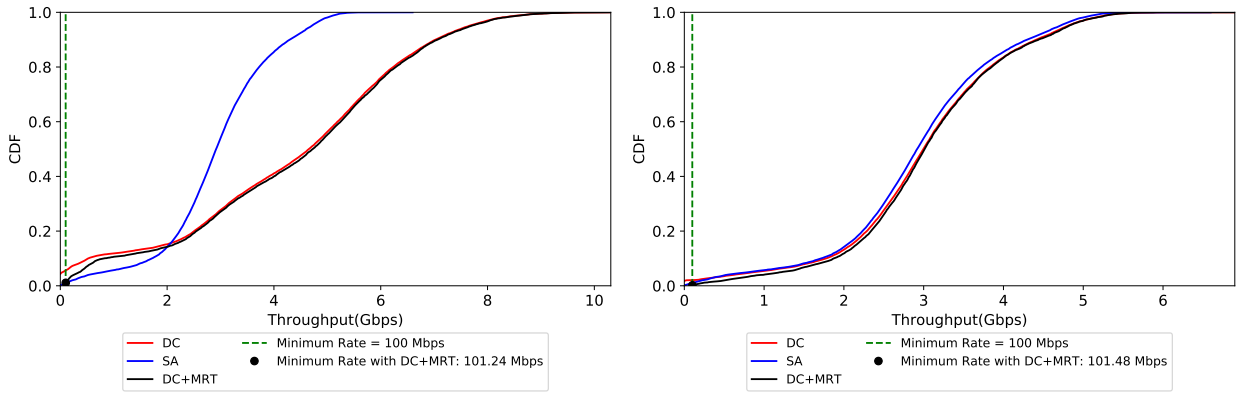
reason for such a behavior is two-folds. Firstly, the fairness is computed by utilizing the throughput experienced by each individual eMBB user in the system, which is a function of the access network resources. And secondly, the mMTC users operate in the guard band, thus not consuming any access resources from eMBB users. Henceforth, there are very slight variations in the fairness index measure as mMTC users only impact the system performance through backhaul resource consumption.

Further, we also considered scenarios with square deployment and circular deployment along side the *MCSC* setup. And similar to the aforesaid deductions, we observed negligible change in the fairness index when mMTC and eMBB services are considered together as compared to when only eMBB devices are considered. Note that, given the complex and diverse nature of the scenarios that we have explored in Figure 6.10, we assert that they are sufficient to establish and understand the performance trends in general.

6.4.3 User Throughput Distribution

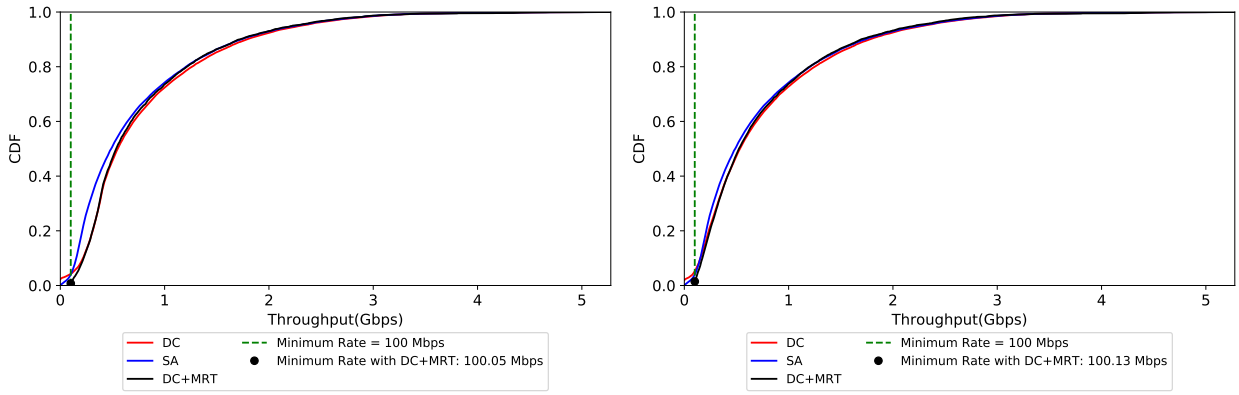
While we observe that the AURA-5G framework outperforms the baseline scenario in terms of the total network throughput, it becomes important to understand the fidelity of the proposed framework. Consequently, interesting insights can be obtained when we observe the user throughput distribution. For this we consider the throughput observed for all the users across all the Monte Carlo trials and represent them in the form of an empirical distribution, represented in Figures 6.11 and 6.12. Note that these distributions are for the cases where DC mode with and without the Minimum Rate constraint, as well as the SA mode without any constraints over the *CEBAS*, *CEBMS*, *CEIAS*, *CEIMS*, *SEBAS*, *SEBMS*, *SEIAS* and *SEIMS* scenarios have been considered. Further, the choice of the aforesaid representative scenarios is two folds – a) we do not consider any backhaul capacity constraints in this analysis, thus eliminating any requirement to analyze the scenarios with mMTC devices, and b) the chosen scenarios encompass all the possibilities with regards to eMBB devices only setup (Table 6.2).

And so, firstly from Figures 6.11(a)-(d) we observe that the minimum rate requirement is satisfied by the AURA-5G framework for all the scenarios under consideration. Next, we observe that for the *AnyDC* case (Figure 6.11(a)) the throughput per user is concentrated at data rates much higher than those observed for the *MCSC* case (Figure 6.11(b)). For example, if we consider the 95^{th} percentile of the users, then for the *AnyDC* case it lies closer to 8 Gbps whilst that for the *MCSC* case is closer to 5 Gbps. However, the users for the *AnyDC* setup in Figure 6.11(a) are distributed much more as compared to the *MCSC* setup in Figure 6.11(b), where the users seem to be more closely packed in terms of the rates they



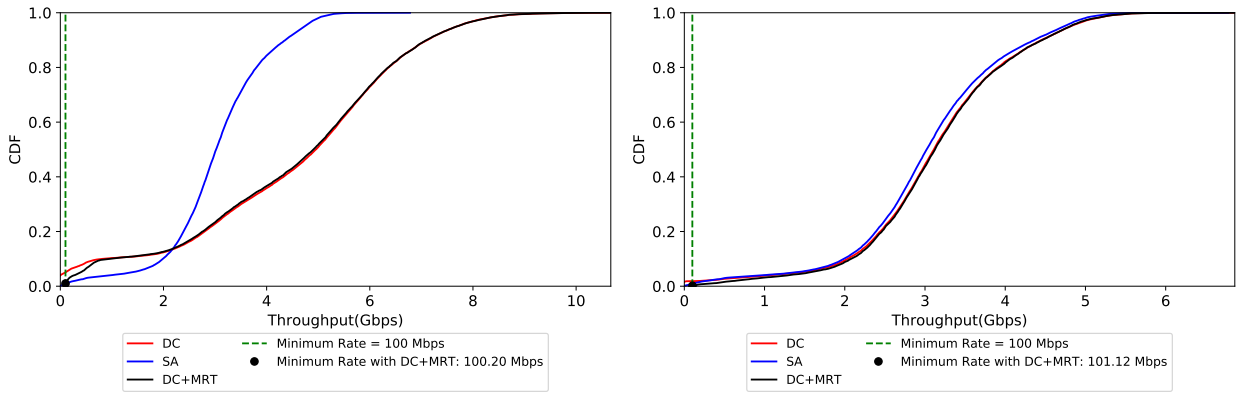
(a) CEBAS

(b) CEBMS



(c) CEIAS

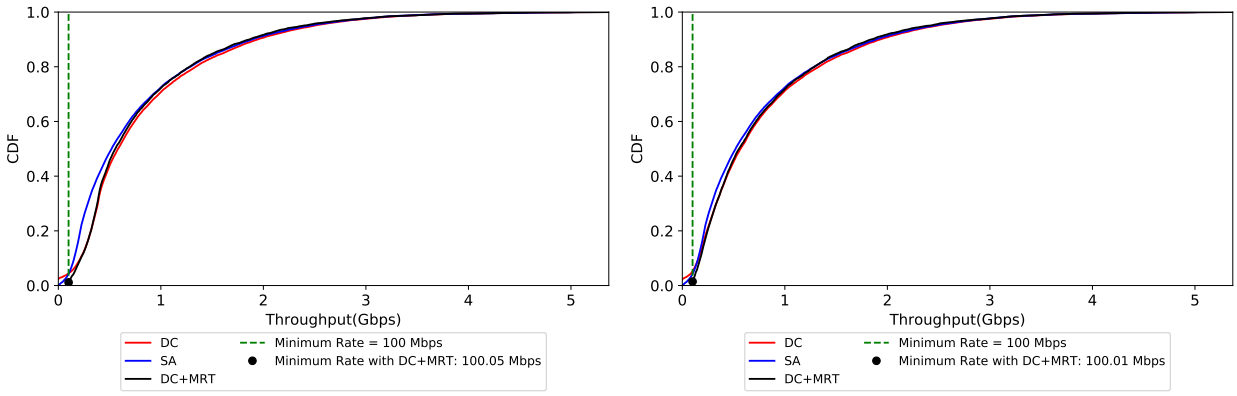
(d) CEIMS



(e) SEBAS

(f) SEBMS

Figure 6.11: User Throughput Distribution for Dual Connectivity (DC) with Minimum Rate (MRT) constraints in (a) CEBAS, (b) CEBMS, (c) CEIAS, (d) CEIMS, (e) SEBAS and (f) SEBMS scenarios.



(a) SEIAS (b) SEIMS

Figure 6.12: User Throughput Distribution for Dual Connectivity (DC) with Minimum Rate (MRT) constraints in (a) SEIAS and (b) SEIMS scenarios.

obtain. This is inline with our discussions in Sections 6.4.1 and 6.4.2, wherein the *AnyDC* setup outperforms the *MCSC* setup for DC modes in terms of overall network throughput but with a lower system fairness.

Next we consider the interference limited scenarios (Figures 6.11(c) and (d)) and immediately we can observe that the per user throughput values are concentrated more towards the 1-2 Gbps range. This is validated by the fact that in interference limited regimes the SINR experienced by users is significantly less as compared to that in the beamformed regime. As a consequence, the throughput experienced per user is also affected, which leads to the aforesaid observations. In addition to the circular deployment scenarios, we also explored the user throughput distribution for the square deployment scenarios through Figures 6.11(e)-(f). We see that the user distribution characteristics follow a similar trend as to those observed for the circular deployment scenarios. Concretely, and as discussed earlier, with the beamformed regime the throughput performance of *AnyDC* setup for the 95^{th} percentile users is far greater than that of the *MCSC* setup. However, this comes at a cost of reduced system fairness.

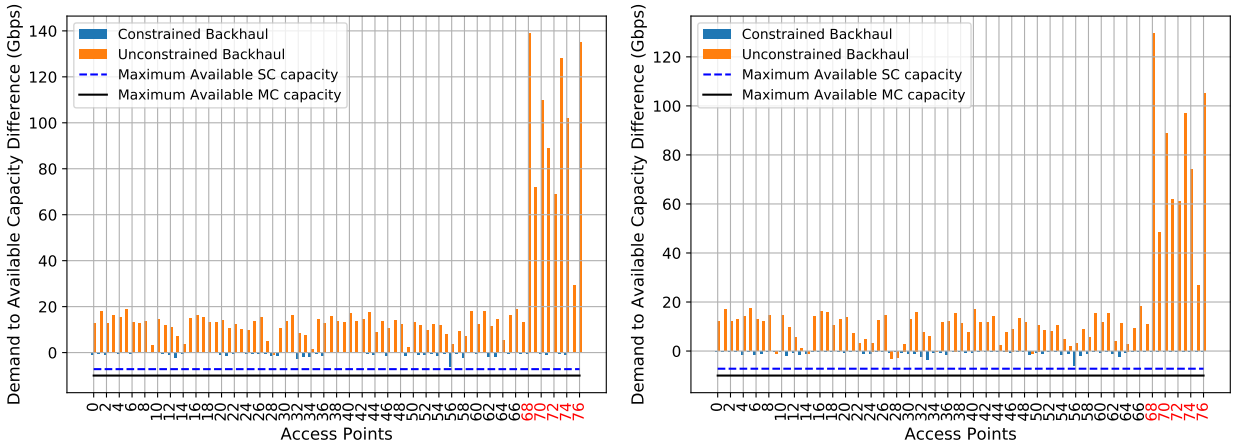
Lastly, the user throughput distribution curves as seen from Figures 6.12(a) and (b) show that, for the DC modes with Minimum Rate requirement constraints and with mMTC and eMBB services co-existing together, they are invariant when compared with the user distribution curves obtained for scenarios with only eMBB services. This is so because, based on our evaluation framework defined in Section 6.3, the mMTC services only impact the available backhaul capacity. And so, in the absence of backhaul capacity constraints, the mMTC services will not impact the user throughput distribution. Further, given the

extremely challenging nature of the scenarios where Minimum Rate and Backhaul Capacity constraints are imposed in conjunction (see Section 6.4.7), we do not illustrate the results for the same.

6.4.4 Backhaul Utilization

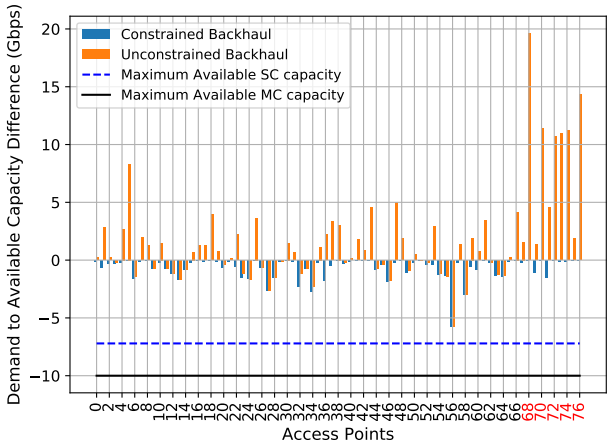
The primary goal for analyzing the backhaul utilization after the AURA-5G framework has been implemented on certain scenarios, is to determine if the existing backhaul setup, wherein we consider a combination of wired and wireless backhaul links with the wired links having capacities of 1Gbps and 10Gbps (Section 6.3), is a bottleneck. Further, we also utilize this analysis to understand the compliance of the AURA-5G framework with the backhaul capacity constraints. For this analysis we select a subset of representative scenarios, which we believe will help concretize the understanding with regards to the performance trends. Hence, through Figures 6.13 and 14, we depict the backhaul utilization as observed for CABE, CMBE, CAIE, CMIE, SABE, SAIE, CABEm and CAIEm scenarios. The choice of the aforesaid scenarios stems from the fact that these selected scenarios include the *MCSC* and *AnyDC* setup, beamformed and interference limited regime, square deployment setups, as well as the mMTC and eMBB services together in the beamformed and interference limited scenarios alongside *AnyDC* setup. This set of scenarios owing to their diversity and challenging nature give the necessary and sufficient idea with regards to the backhaul utilization characteristics.

From Figures 6.13(a)-(f) and 6.14(a)-(b), we firstly observe that the AURA-5G framework is successful in satisfying the backhaul capacity constraints as and when they have been imposed. Here by backhaul capacity constraints we mean that the wired backhaul links are capped by their designated capacities as stated in Table 6.3. Further, the wireless backhaul links are constrained by the capacity computed utilizing the *Shannon-Hartley* formula based on their corresponding SINR value. It is also important to state here that on the vertical axis in all the subplots presented in Figures 6.13 and 6.14, we represent the difference between the demand and the available capacity. Hence, a negative value on the vertical axis indicates that the backhaul resources on the corresponding BS have not been fully utilized, whilst a positive value indicates over-utilization by the corresponding amount. And so we can see that for the unconstrained scenarios the backhaul resources are always over-utilized. However, for the backhaul capacity constrained scenarios, we observe that our framework succeeds in finding an optimal solution without over-utilizing the total available backhaul resources. This significant difference in backhaul utilization also reflects the greedy nature of the optimization framework, whose objective is to maximize the total network throughput. Note that we have

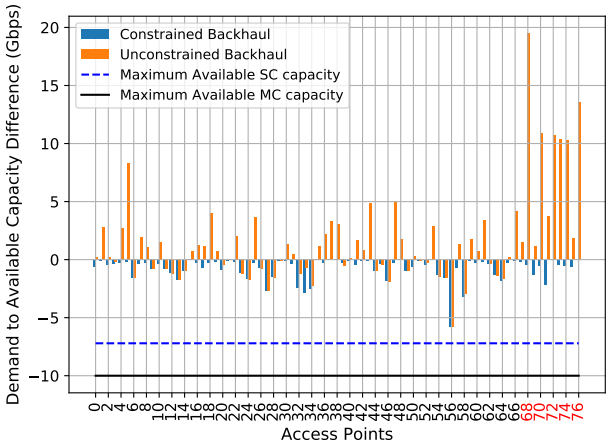


(a) CABE

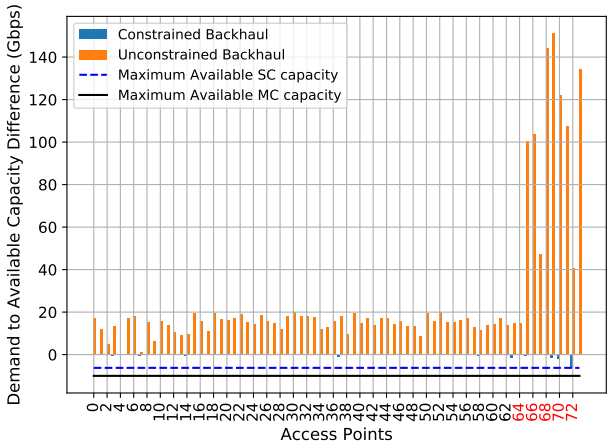
(b) CMBE



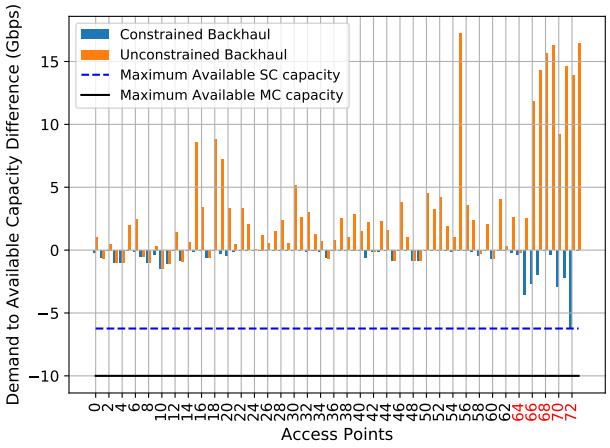
(c) CAIE



(d) CMIE

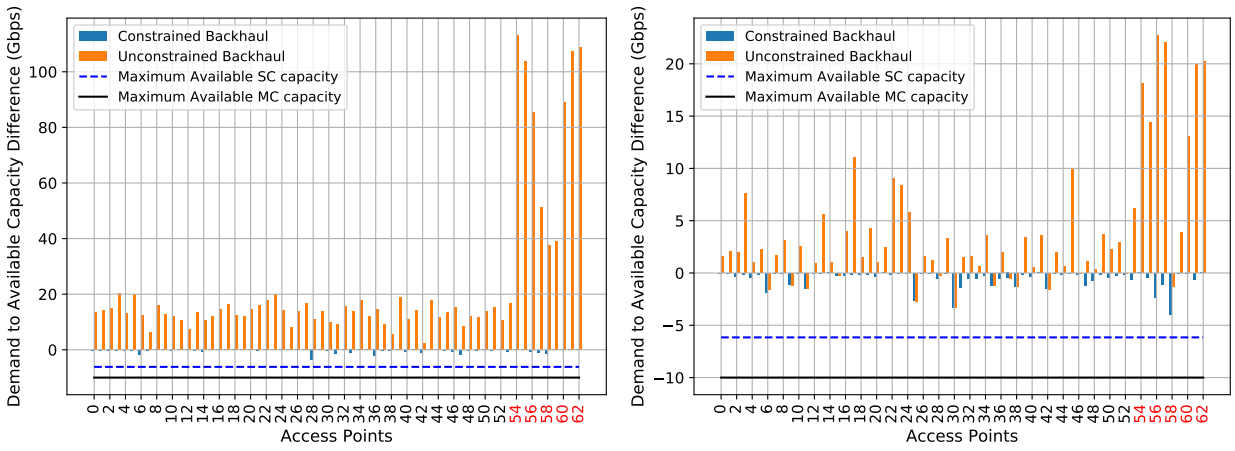


(e) SABE



(f) SAIE

Figure 6.13: Backhaul Utilization for Dual Connectivity (DC) and DC with Backhaul Capacity constraints in (a) CABE, (b) CMBE, (c) CAIE, (d) CMIE, (e) SABE and (f) SAIE scenarios. Red colored BS indices are for MCs and the rest for SCs.



(a) CABEm (b) CAIEm

Figure 6.14: Backhaul Utilization for Dual Connectivity (DC) and DC with Backhaul Capacity constraints in (a) CABEm and (b) CAIEm scenarios. Red colored BS indices are for MCs and the rest for SCs.

also indicated the maximum available backhaul capacity for the SCs (broken blue line) and MCs (black line). This assists the readers in understanding the maximum data carrying capacity that the set of SCs and MCs in the network have, as well as in exemplifying the fidelity of the AURA-5G framework. Concretely, the maximum capacity on an SC should not exceed that on an MC. This is so because, all SCs, in our framework, route their traffic through the MC. Hence, a higher maximum bandwidth availability on the SC as compared to an MC would be equivalent to trying to balance a big box on a thin needle. Additionally, and for the reasons as stated above, from the backhaul utilization values for the unconstrained setup we observe that the backhaul through the MCs (the BS indices marked in red in Figures 6.13 and 6.14, i.e., the last 9 BS indices on the x-axis) is significantly over-utilized as compared to the SCs.

Next, we observe that scenarios wherein beamforming has been applied, i.e., Figures 6.13(a), (b), (e) and 6.14(a), are severely limited for backhaul resources. The reason being that, the blue bars (the darker bars, if viewed in black and white), which indicate the available demand to backhaul capacity difference in the constrained setup, are extremely small. This indicates that nearly all of the available capacity has been utilized, even without the Minimum Rate constraints being applied. The reason being that, beamforming results in an improved SINR measure within the system. This consequently enables the users to achieve a better throughput, and hence, the aforementioned backhaul utilization characteristic. Thus, an important insight for network operators, suggesting a requirement for

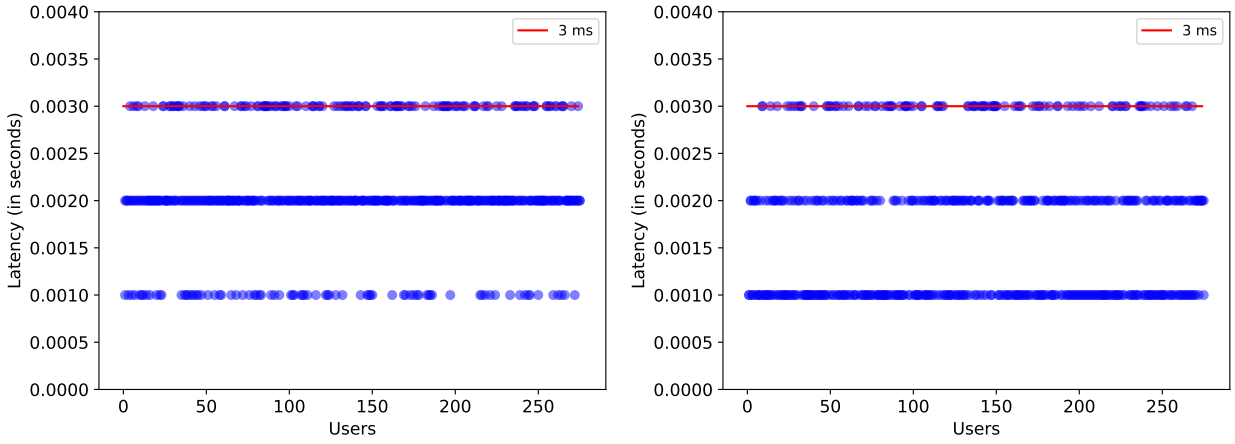
network re-dimensioning (Section 6.5), can also be drawn from these observations. Further, in Figures 6.13(c), (d), (f) and 6.14(b), wherein the interference limited regime has been adopted, the overall backhaul utilization in the unconstrained setup is much lower than that observed for the scenarios involving beamformed regime. This is as a result of the severe interference causing a significant loss in SINR, and hence, per user throughput. This claim is also corroborated by the reduction in network throughput observed in Section 6.4.1 for interference limited scenarios.

Lastly, through Figures 6.14(a) and (b), wherein the mMTC services have been considered alongside the eMBB services, it can be observed that the AURA-5G framework is able to provision optimal user-BS associations whilst adhering to the backhaul capacity constraints. Furthermore, as compared to the corresponding scenarios where only eMBB services are present, i.e., CABE and CAIE, the backhaul utilization for the constrained backhaul case in CABEm (Figure 6.14(a)) and CAIEm (Figure 6.14(b)) scenarios is slightly higher. This is so because, in addition to the eMBB services, the mMTC services also consume a portion of the backhaul resources. Hence, the overall increase in backhaul utilization.

6.4.5 Latency Requirement Compliance

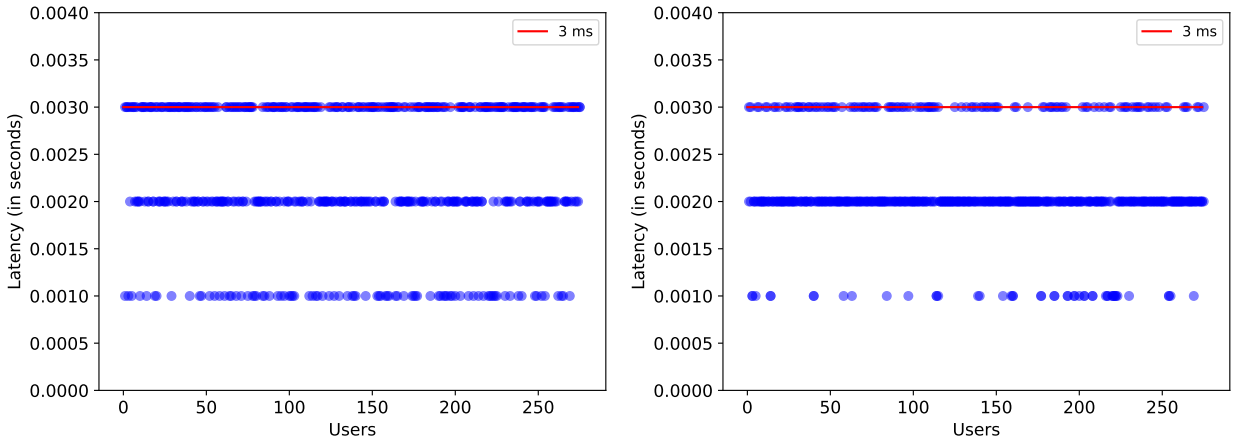
As part of our fidelity analysis for the AURA-5G framework, we delve into how it satisfies the specified service latency requirements through Figures 6.15(a)-(f). It is important to state here that, the latency (or the downlink delay which we define as latency in our work) is governed by the number of hops the data has to traverse to reach the user from the core network. Moreover, as defined in Table 6.4, the imposed latency constraint upon the eMBB users is 3 ms. Hence, we consider certain representative scenarios such as *CABE*, *CMBE*, *SABE*, *SMBE*, *CMIEm* and *CAIEm* for our analysis. These scenarios encompass the *AnyDC* and *MCSC* setup, the beamformed and interference limited regimes, as well as the eMBB only and eMBB with mMTC services based setups. Note that, the downlink delay for the eMBB services is considered to be 3 ms, based on the scenario parameters defined in Table 6.4. The mMTC services are considered to be delay tolerant in our framework. Further, we do not include the last wireless hop, i.e., MC or SC to the UE, in our optimization framework as it does not induce any variability within the scenario given that it will be omnipresent whatever the given association be. Hence, we focus on the number of hops that the data has to traverse from the CN to the BS.

From Figures 6.15(a), (c) and (e), wherein the *AnyDC* setup is employed, we observe that the density of users with 3 ms latency is higher as compared to the ones where *MCSC* setup is employed, i.e., in Figures 6.15(b), (d) and (f). This is so because, in *AnyDC* to maximize



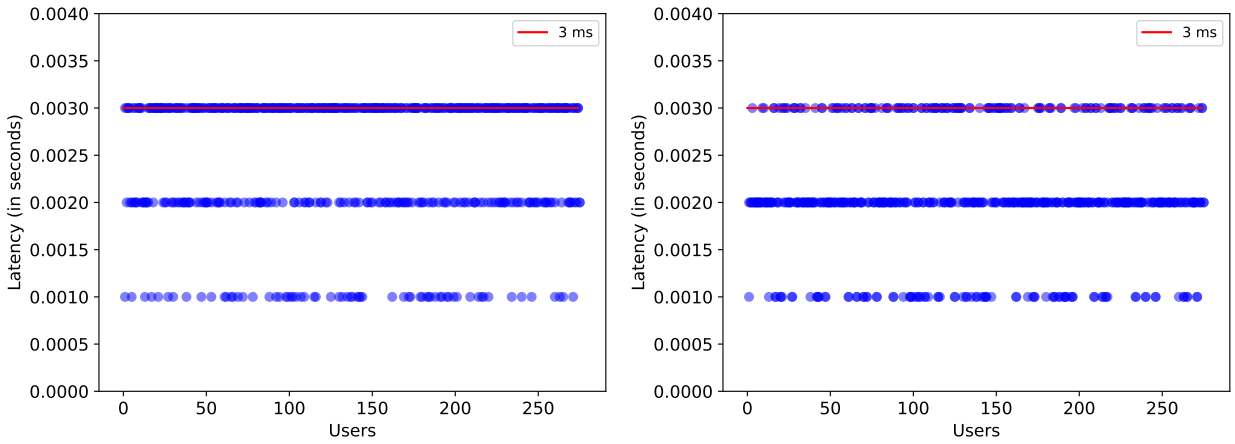
(a) CABE

(b) CMBE



(c) SABE

(d) SMBE



(e) CAIEm

(f) CMIEm

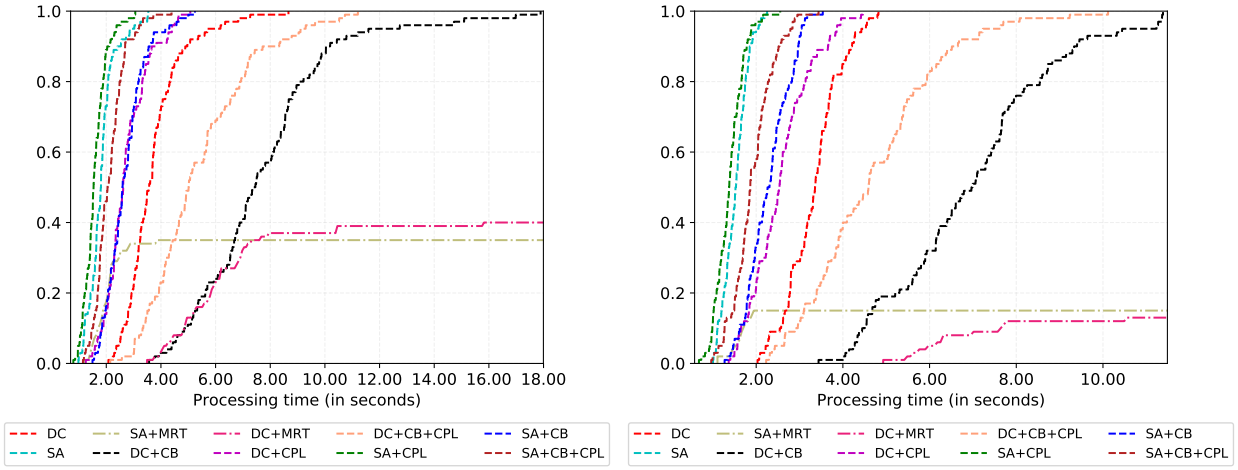
Figure 6.15: Observed Latency for (a) CABE, (b) CMBE, (c) SABE, (d) SMBE, (e) CAIEm and (f) CMIEm scenarios.

the total sum rate our algorithm tries to find SCs first for all the users. However, in the *MCSC* setup, we force our algorithm to find atleast one MC for each user. Hence, given the fact that an SC is connected to the CN through a corresponding MC, the latency incurred by the users in the *MCSC* scenarios is comparatively less as compared to the *AnyDC* scenario. The operators can utilize this insight to design base station selection schemes wherein for services that can tolerate higher delays the *AnyDC* setup maybe employed, whereas for services with extreme latency constraints, an *MCSC* setup could be employed. Further, for the square deployment scenarios (Figures 6.15(c) and (d)) and mMTC based scenarios (Figures 6.15(e) and (f)) the trend for latency compliance follows that of the *CABE* (Figure 6.15(a)) and *CMBE* (Figure 6.15(b)) scenarios, as discussed above. Hence, through this analysis we reinforce fidelity of the AURA-5G framework towards the joint optimization problem that we explore in this chapter.

6.4.6 Convergence Time Distribution

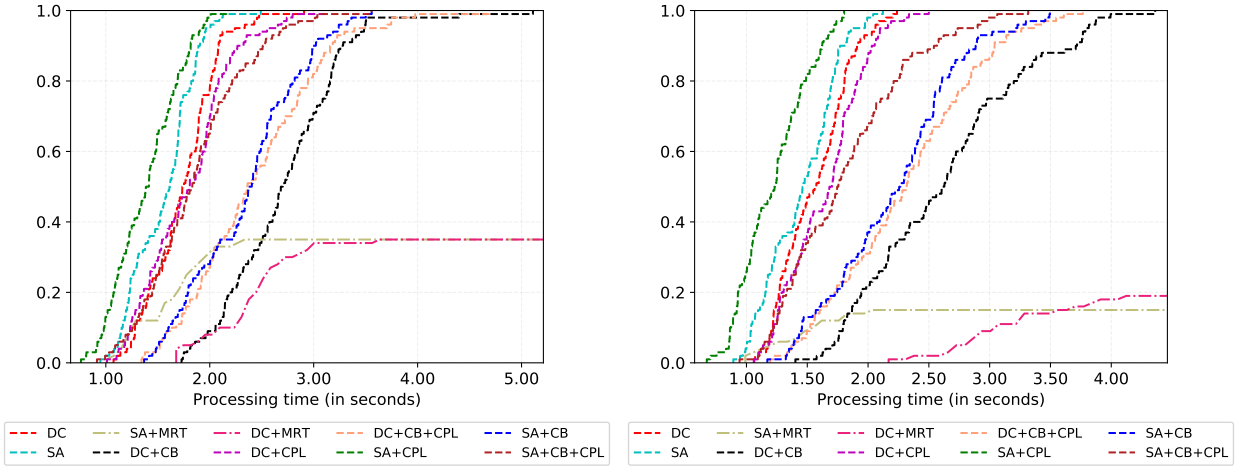
Next, we study the convergence time to the optimal solution for the AURA-5G framework. This will be critical for real time implementation, and hence is of immense value to not only the academic but also to the industrial community. The reason being, network scenarios with the combination of constraints discussed in this work, will be prevalent in 5G networks. And given that there will be a central controller within a local area of these networks [C2, J1, J2], the newly designed mobility management algorithms, such as the AURA-5G framework, will be placed on these controllers to enhance the QoE for the users. Consequently, through Figures 6.16 and 17, we evaluate the convergence time parameter for the various constraint combinations imposed on the myriad scenarios explored in this chapter. From the CDFs presented in Figures 6.16 and 6.17, a probability measure of 1 indicates that all the 100 iterations (Monte Carlo trials) for the specific constraint combination over the scenario under study have converged. On the other hand, a probability measure of 0 indicates that none of the iterations converge. Note that the simulations were performed on a commodity server with 20 cores (with each being an i9-7900x at 3.3GHz core), Ubuntu 16.04 LTS OS, and 64GB of RAM.

From Figures 6.16(a)-(d) and 6.17(a)-(b) we observe that for all scenarios and most constraint combinations, the AURA-5G framework is able to determine an optimal solution. It is worth mentioning that, the AURA-5G framework is able to provision an optimal solution in an acceptable time frame, given the density and heterogeneity of 5G networks. This is of extreme importance for real-time implementation because of the elevated level of dynamics in 5G networks as compared to its predecessors.



(a) CABE

(b) CAIE

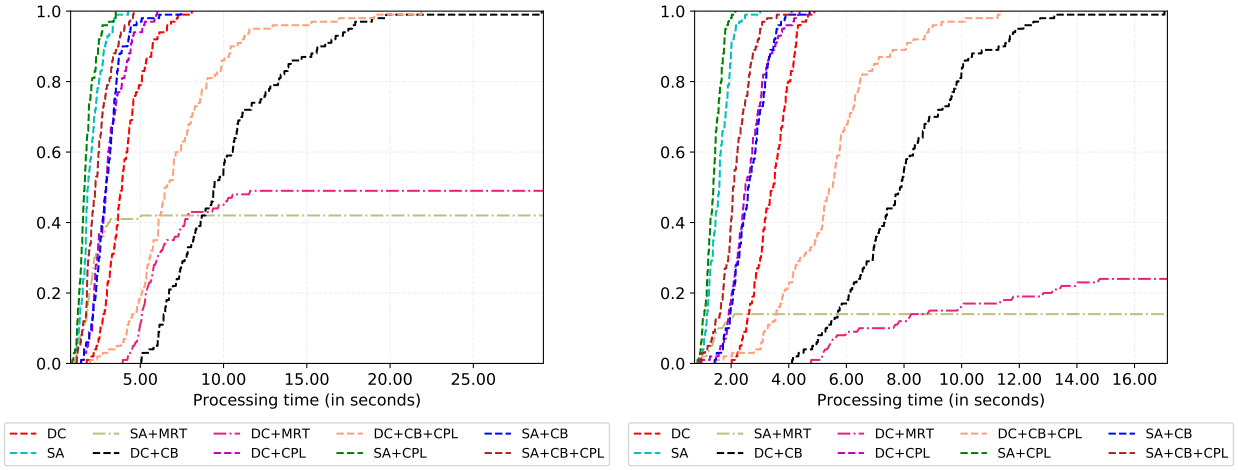


(c) CMBE

(d) CMIE

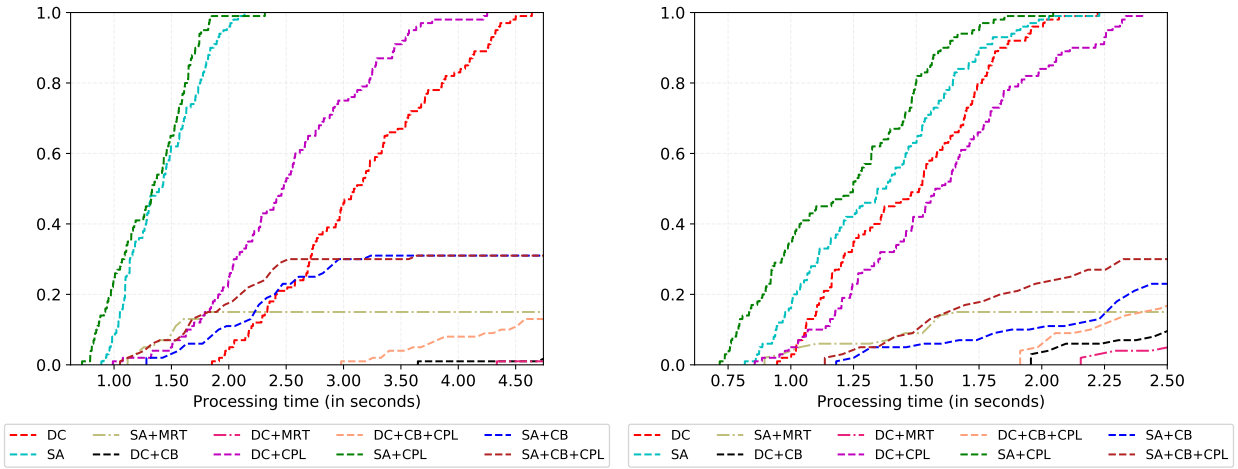
Figure 6.16: Convergence time CDF (Empirical) for (a) CABE, (b) CAIE, (c) CMBE and (d) CMIE scenarios.

Next, we observe that for the Single Association (SA) scenarios the time required for obtaining an optimal solution is significantly less as compared to the Dual Connectivity (DC) mode scenarios. This is so because, the solution space for an SA scenario will be much smaller than that for a DC scenario, hence the time required to search for the optimal solution is correspondingly also reduced. We further observe that as constraints are imposed, the amount of time required to search for the optimal solution increases. This is inline with our intuition, since addition of constraints adds extra dimensions to the search space. Most notably scenarios with the Minimum Rate (MRT) constraints for both the SA and DC modes do not converge to an optimal solution in the given timeframe (we set a 600 seconds cutoff



(a) SABE

(b) SAIE



(c) CAIEm

(d) CMIEm

Figure 6.17: Convergence time CDF (Empirical) for (a) SABE, (b) SAIE, (c) CAIEm and (d) CMIEm scenarios.

time) for all the Monte Carlo trials carried out. This reflects the complexity introduced by the MRT constraint and a possible requirement to re-dimension the network so as to be able to accomodate the rate requirements for any given topology. We refer the reader to Section 6.5 for further details on the network re-dimensioning aspects.

Further, in Figures 6.16(a)-(d) and Figures 6.17(a)-(b), we also highlight an exception to the generic trend stated above. The Path latency (CPL) constraint when imposed on SA and DC leads to a faster search time as compared to their respective SA and DC counterparts in most scenarios. This is due to the fact that while most constraint combinations in our work lead to an increasingly complex search space, and hence an increased convergence time as

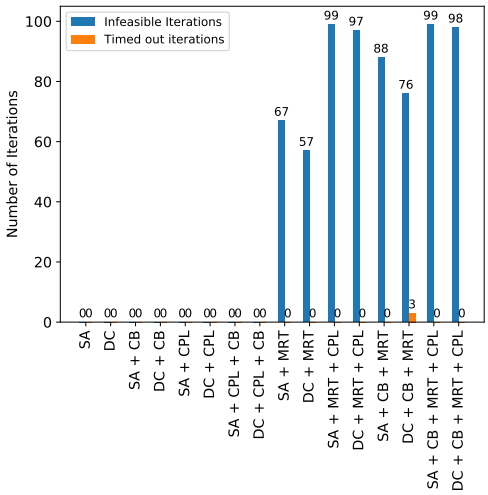
corroborated by our results in Figures 6.16 and 6.17, the addition of path latency constraint creates a cut in the solution hyperspace that reduces the overall complexity of the search space and consequently the convergence time. This is also indicative of the fact that very few BSs in the topology are actually able to satisfy the path latency constraint when imposed in combination with other network constraints. Thus, this gives an insight into the contrasting behavior of different constraints, and their overall impact on the system performance.

Lastly from Figures 6.17(c) and (d) wherein the mMTC services are considered as well, it can be observed that most of the iterations for the scenarios, in which the backhaul capacity is constrained, do not converge to an optimal solution in the stipulated time. This is so because the mMTC services place an additional burden on the backhaul by consuming a portion of their available capacity. This, as a result, leads to a more challenging scenario for the AURA-5G framework to determine an optimal solution as the eMBB services have less amount of available backhaul capacity. Consequently, we observe the non-convergent behavior of the scenarios with the backhaul capacity constraint.

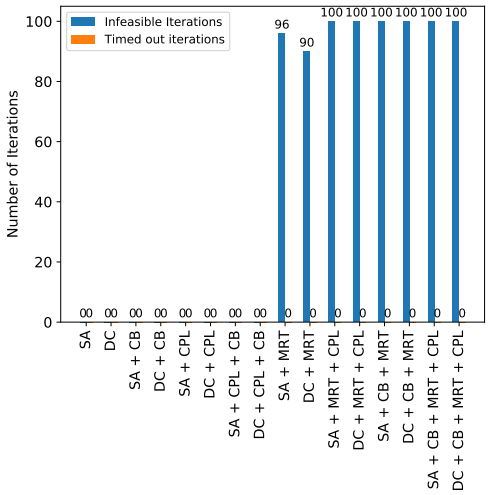
6.4.7 Solvability Analysis

In Section 6.4.6 we observed that certain scenarios with backhaul capacity and minimum rate constraints do not converge to an optimal solution in the cutoff time period of 600 seconds, as defined in our evaluation framework. However, it might very well be possible that either a solution does not exist or the optimizer got timed out, i.e., it might or might not have a feasible solution, but, it was not able to determine the same up until the 600 seconds timeframe. Hence, with this background, in this section we undertake a solvability analysis with the specified time limit parameters and aim to understand the bottleneck constraints for the AURA-5G framework, given the various scenarios we have studied. Moreover, for the analysis we have only considered scenarios wherein there are 275 eMBB users, which is the maximum number of users from the range of values we evaluate upon, i.e., 150-275 eMBB users, as mentioned in Table 6.4.

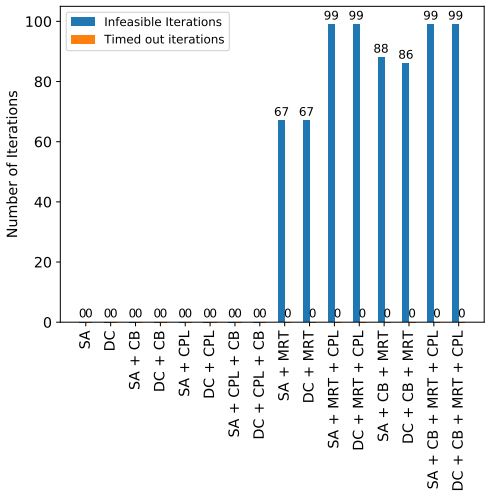
For the solvability analysis we introduce Figures 6.18(a)-(f), wherein we have also provisioned an analysis for the most complex combination of constraints, i.e., CB + MRT, CPL + MRT and CB + MRT + CPL (see Table 6.3 for description). From Figures 6.18(a)-(f) we observe that for all the scenarios explored, the Minimum Rate (MRT) constraint behaves as a bottleneck constraint for the optimizer in the AURA-5G framework. This is also reflected from the time convergence plots in Figures 6.16 and 6.17. The reason being that there is limited access bandwidth available. In addition, given the nature of the scenario, i.e., if its beamformed or interference limited, the SINR characteristics transform and subsequently



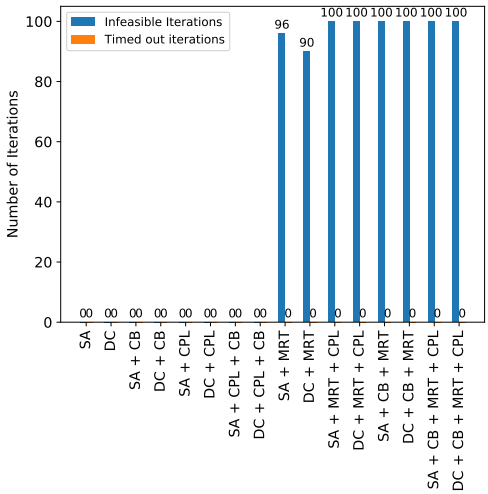
(a) CABE



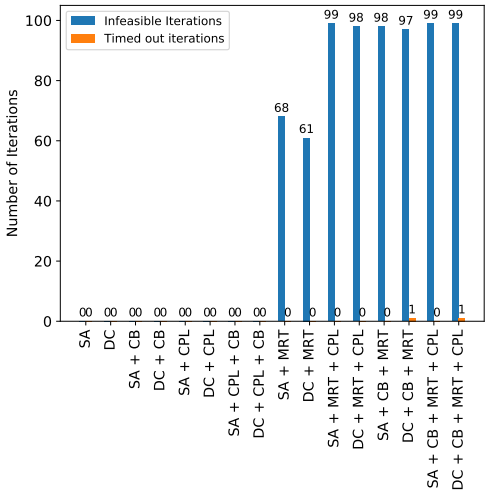
(b) CAIE



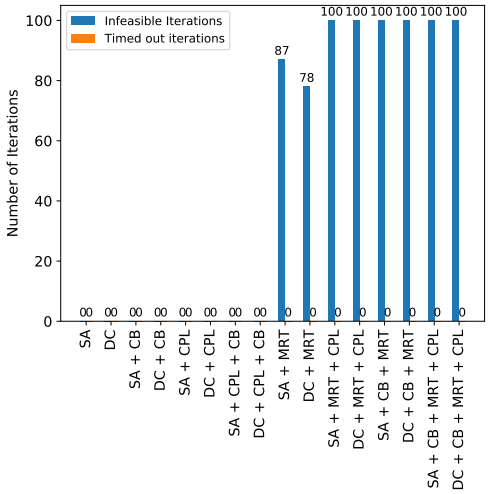
(c) CMBE



(d) CMIE

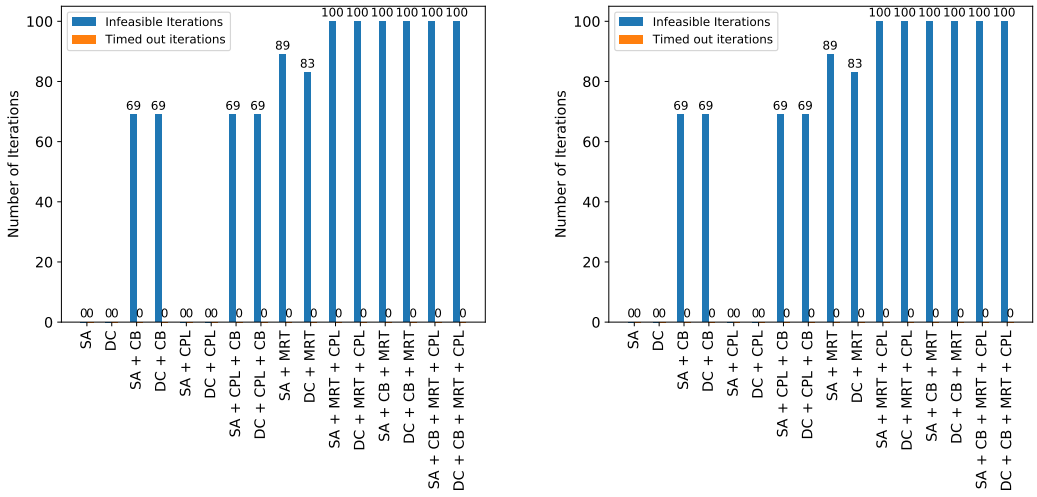


(e) SABE



(f) SAIE

Figure 6.18: Optimizer Status for (a) CABE, (b) CAIE, (c) CMBE, (d) CMIE, (e) SABE and (f) SAIE scenarios with 275 eMBB users.



(a) CAIEm (b) CMIEm

Figure 6.19: Optimizer Status for (a) CAIEm and (b) CMIEm scenarios with 275 eMBB users.

impact the decision of the optimization framework. Such system based variability in SINR also limits the achievable per user rate, hence, rendering the MRT constraint as a bottleneck.

Further, from Figures 6.18(a)-(f) we see that in the interference limited regime the optimizer performance is much more severely affected as compared to the beamformed scenario, which is in line with the rest of our analysis so far. Moreover, for the square deployment scenario (Figures 6.18(e) and (f)) the backhaul constraints are even more restrictive given the fact that the probability of the SC BSs being more distant from the MC is higher. Hence, the probability of having a wired backhaul, which has a 1 Gbps capacity and is in most cases much lower than what is offered by the mmWave wireless backhaul, is also subsequently higher. As a consequence, from Figures 6.18(a), (c) and (e) it can be deduced that for the square deployment scenario with CB and MRT constraint, the system performance is impacted severely with at least 10 more iterations without a solution compared to those in the circular deployment scenario.

Next, we observe the performance for the scenarios wherein both the mMTC and eMBB services have been considered (Figures 6.19(a) and (b)). Recall that the analysis provisioned here is for the case when we consider 275 eMBB users in the system, i.e., the maximum number of eMBB users evaluated within our evaluation framework (Table 6.4). From Figures 6.19(a) and (b), it can be seen that the backhaul capacity constraint also emerges as a bottleneck. This is corroborated from the time convergence curves in Figures 6.17(c) and (d), where the scenarios with the CB constraints do not illustrate convergence for all the

iterations. This is because of the fact that the mMTC devices consume a portion of the available backhaul capacity. Consequently, due to the reduced backhaul capacity, the optimizer finds itself in a situation wherein these very backhaul resources transform into a bottleneck.

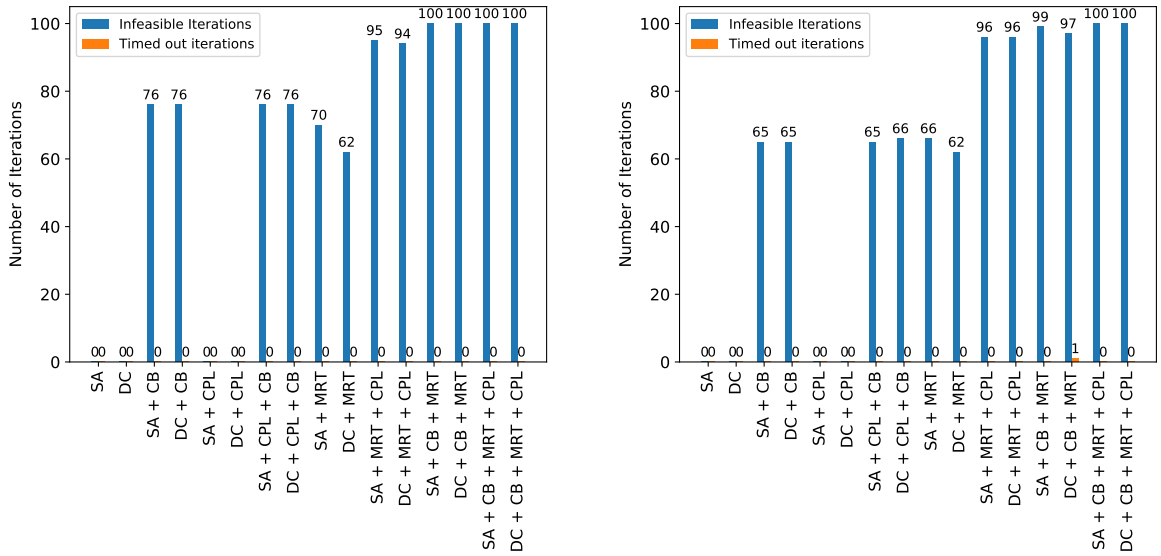
Lastly, from Figures 6.18(a)-(f) and 6.19(a)-(b) we deduce that as the complexity of the constraint combinations increases, the AURA-5G framework finds it increasingly challenging to determine the optimal solution. In particular, the MRT and CPL constraints appear to be fairly challenging for our user association methodology. Further, as has already been stated above, for the scenarios with mMTC services included, constraint combinations with CB also transform into being extremely challenging ones to satisfy. Consequently, in the next section we explore certain network re-dimensioning methodologies that assist the optimizer to determine an optimal association.

6.5 Network Re-dimensioning

From the analysis presented in Sections 6.4.6 and 6.4.7, we have observed that certain constraint combinations for the scenarios analyzed prove to be significantly difficult for the MILP framework to satisfy. These insights can be a useful network designing tool for the operators, and subsequently they can re-dimension/upgrade their networks to meet the demands. Hence, in this section through Figures 6.20-6.31, we discuss certain network re-dimensioning options and their corresponding results. We present the fact that re-evaluating and re-defining appropriate network parameters results in an improved performance by the AURA-5G framework. However, for our analysis, we consider only SABEm and CABEm scenarios in this section as they encompass all the complexities of the scenarios that we have studied in this chapter.

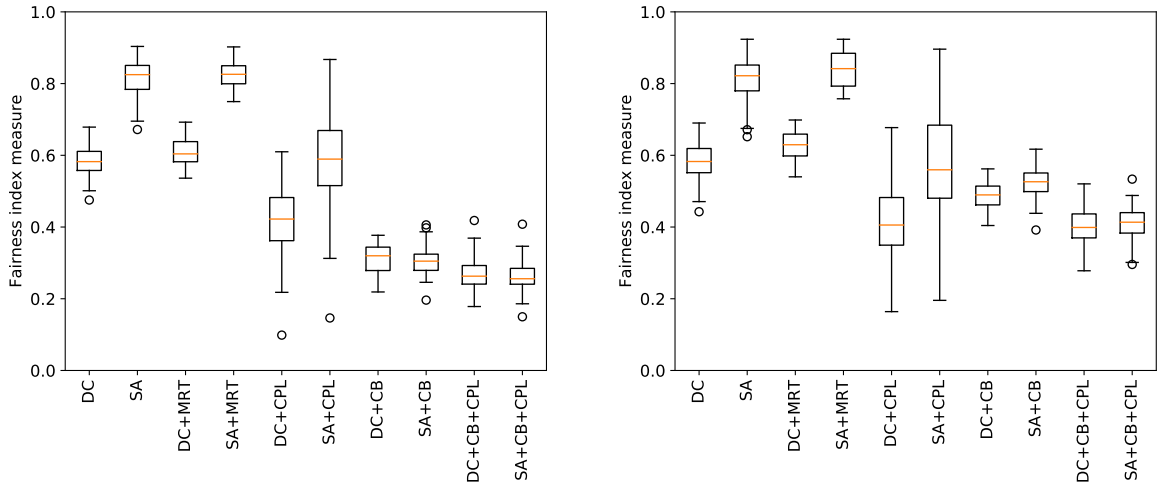
And so, the analysis presented thus far has led to the conclusion that one of the constraints that has proven to be extremely difficult to satisfy, especially when mMTC and eMBB services are considered together in the system, is the backhaul capacity constraint. Moreover, scenarios wherein *beamforming* and *AnyDC* modes have been utilized will prove to be particularly challenging, given the lack of backhaul resources and the throughput maximization nature of the optimizer. In addition, due to the lack of access resources as well as the prevailing SINR characteristics, the MRT constraint also imposes a severe challenge for the AURA-5G framework. Hence, through Figures 6.20-6.25, we analyzed scenarios with a re-dimensioned backhaul and access network wherein both mMTC and eMBB users are considered alongside the circular and square deployment, and the beamformed and AnyDC regime.

For the re-dimensioning we firstly calculated the average amount of backhaul utilized



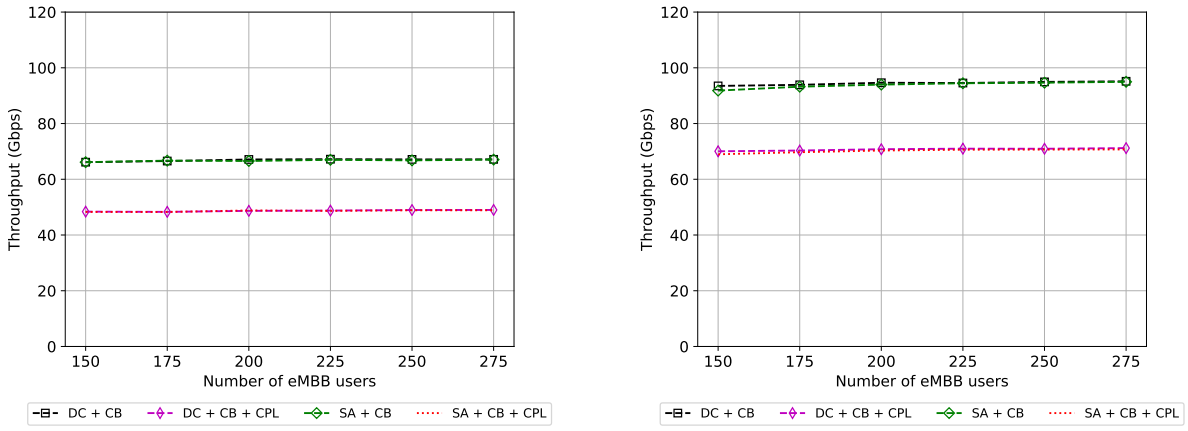
(a) (b)

Figure 6.20: Optimizer Status for (a) SABEm without Relaxed Backhaul, and (b) SABEm with Relaxed Backhaul scenarios with 275 eMBB users.



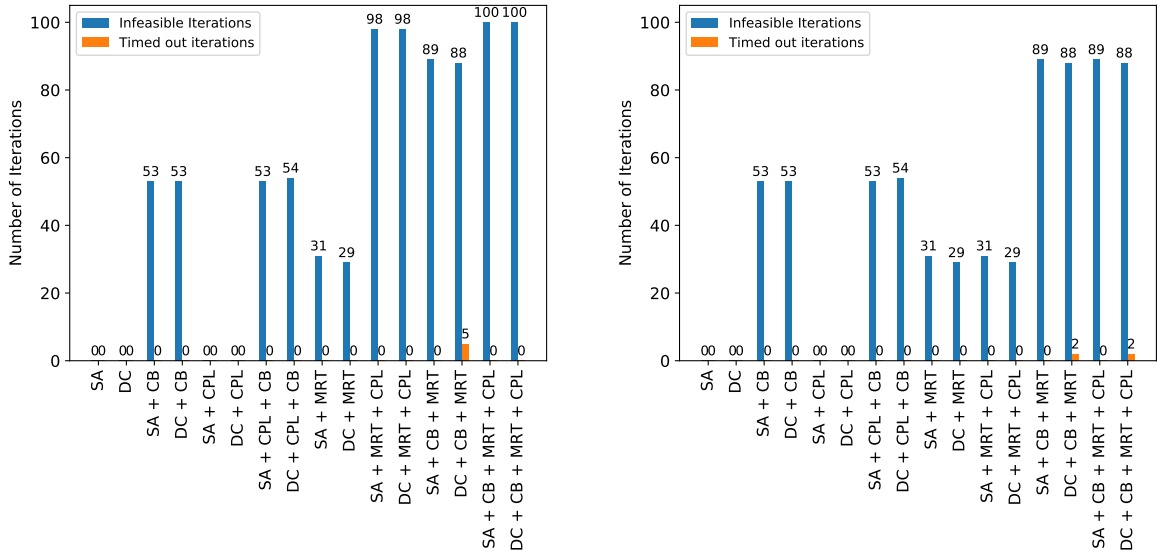
(a) (b)

Figure 6.21: System Fairness Measure for (a) SABEm without Relaxed Backhaul, and (b) SABEm with Relaxed Backhaul scenarios with 275 eMBB users.



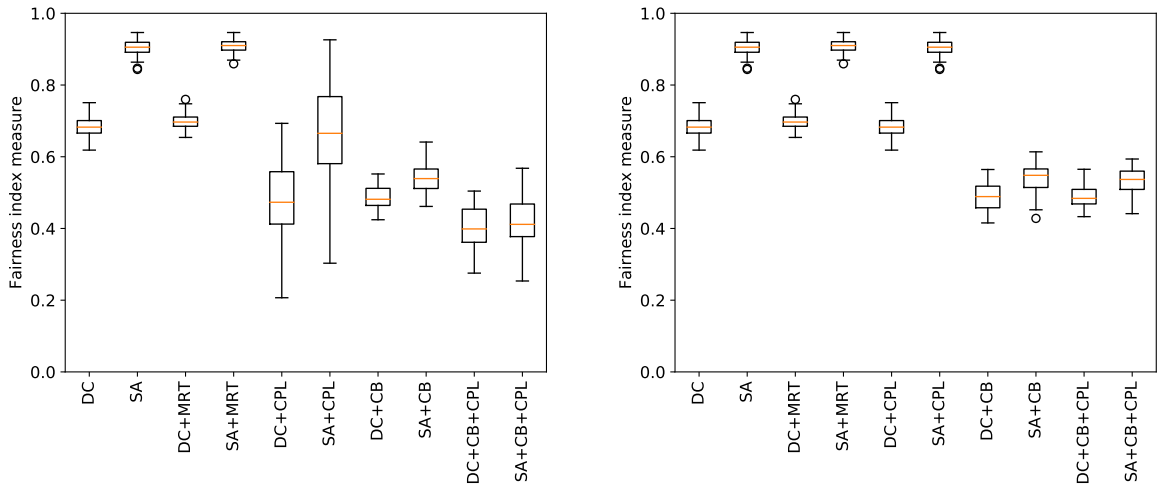
(a) (b)

Figure 6.22: Total Network Throughput for (a) SABEm without Relaxed Backhaul, and (b) SABEm with Relaxed Backhaul scenarios with 275 eMBB users.



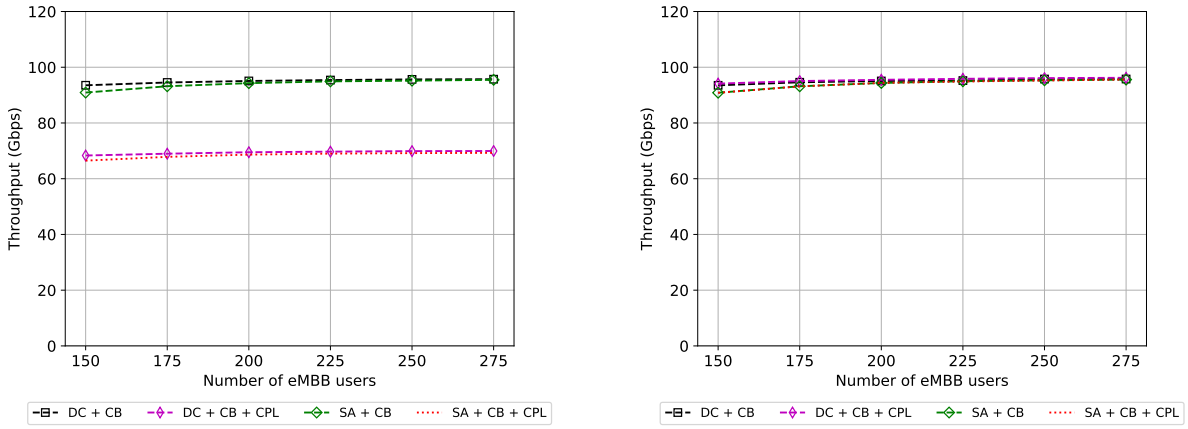
(a) (b)

Figure 6.23: Optimizer Status for (a) SABEm with Relaxed Backhaul and Increased SC density scenario with 275 eMBB users, and (b) SABEm scenario with Relaxed Backhaul, Increased SC density, 5 ms downlink latency requirement and 275 eMBB users.



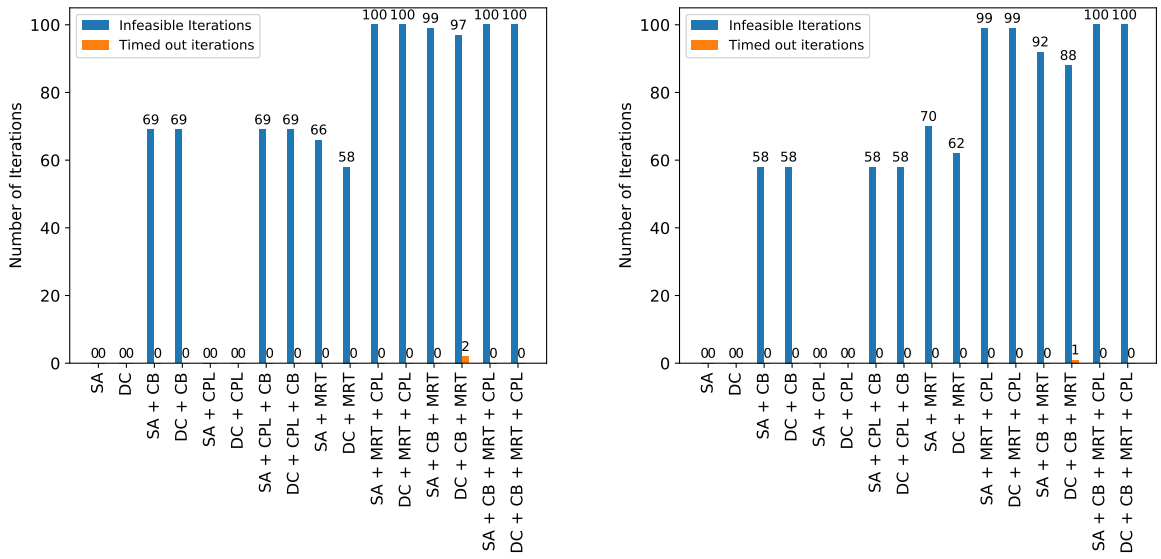
(a) (b)

Figure 6.24: System Fairness Measure for (a) SABEm with Relaxed Backhaul and Increased SC density scenario with 275 eMBB users, and (b) SABEm scenario with Relaxed Backhaul, Increased SC density, 5 ms downlink latency requirement and 275 eMBB users.



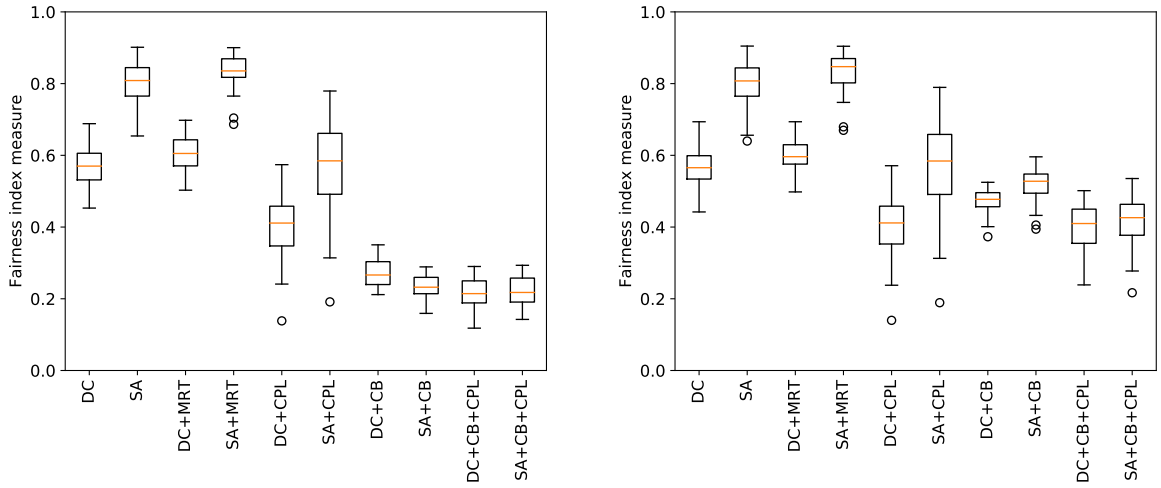
(a) (b)

Figure 6.25: Total Network Throughput for (a) SABEm with Relaxed Backhaul and Increased SC density scenario with 275 eMBB users, and (b) SABEm scenario with Relaxed Backhaul, Increased SC density, 5 ms downlink latency requirement and 275 eMBB users.



(a) (b)

Figure 6.26: Optimizer Status for (a) CABEm without Relaxed Backhaul, and (b) CABEm with Relaxed Backhaul scenarios with 275 eMBB users.



(a) (b)

Figure 6.27: System Fairness Measure for (a) CABEm without Relaxed Backhaul, and (b) CABEm with Relaxed Backhaul scenarios with 275 eMBB users.

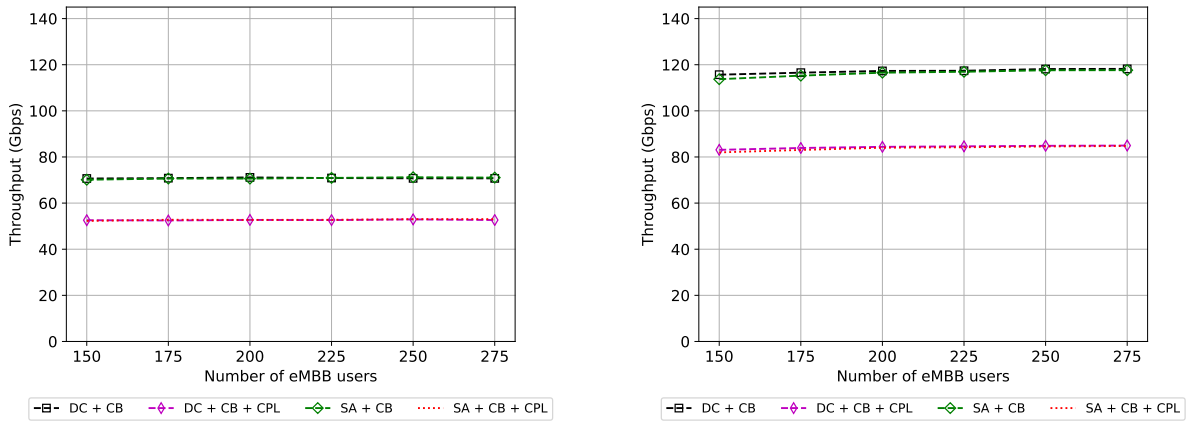
in all the SCs when no re-dimensioning is done for the scenario under study. Next, we increase the backhaul capacity of all the SCs in the system by a percentage of this average consumption. For the percentage increment we utilized four quantized levels, i.e. 30%, 50%, 80% and 100%, and assigned a random and different choice from these values to each

of the SCs. Subsequently, and to account for the worst case scenario, we increment the capacity of the backhaul for each MC by 10 times the aforementioned average SC backhaul utilization. The factor of 10 arises from the fact that in our evaluation framework the maximum number of supported SCs by an MC is also 10. Next, we re-dimension the access network by increasing the average number of SCs per MC in the topology from 6-7 (uniform distribution of 3 to 10 SCs per MC) to 8 (uniform distribution of 6 to 10 SCs per MC). This, automatically provisions more access network resources, in terms of bandwidth, as well as increases the likelihood for a user to find an SC in close proximity. It is important to state here that, we maintain the receive beam angle and beamforming gain. Whilst, these can be exploited to improve the performance of the system further, it might lead to increased capital/operating expenditure for the operator given required infrastructure overhaul, such as antenna replacement, etc. Hence, we leave the discussion on this aspect for a sequel work.

Consequently from the analytical results in Figures 6.20(a)-(b) and 6.23(a), we observe that when the backhaul capacities are enhanced by the methodology explained above, the scenarios where backhaul was a constraint ceases to be so anymore. For example, in Figure 6.20(a), constraint combinations *only* CB, and CB + CPL highlight the fact that the backhaul capacity is a constraint for the scenario under study, i.e. SABEm. Hence, by the re-dimensioning employed as specified in our work, through Figure 6.20(b), we observe that the number of iterations that converge for the *only* CB and CB + CPL constraint combinations increases by 14.47%. Furthermore, when we employ the increased SC density framework to provision more access network resources, the percentage improvements in the number of converged iterations, as can be seen in Figure 6.23(a), for constraint combinations CB, CB + CPL, MRT and CB + MRT are at 30.2%, 30%, 55% and 11%, respectively . As a result, to a great extent the re-dimensioning performed, according to the guidelines specified above, helps in alleviating the bottlenecks that hampered the AURA-5G framework earlier.

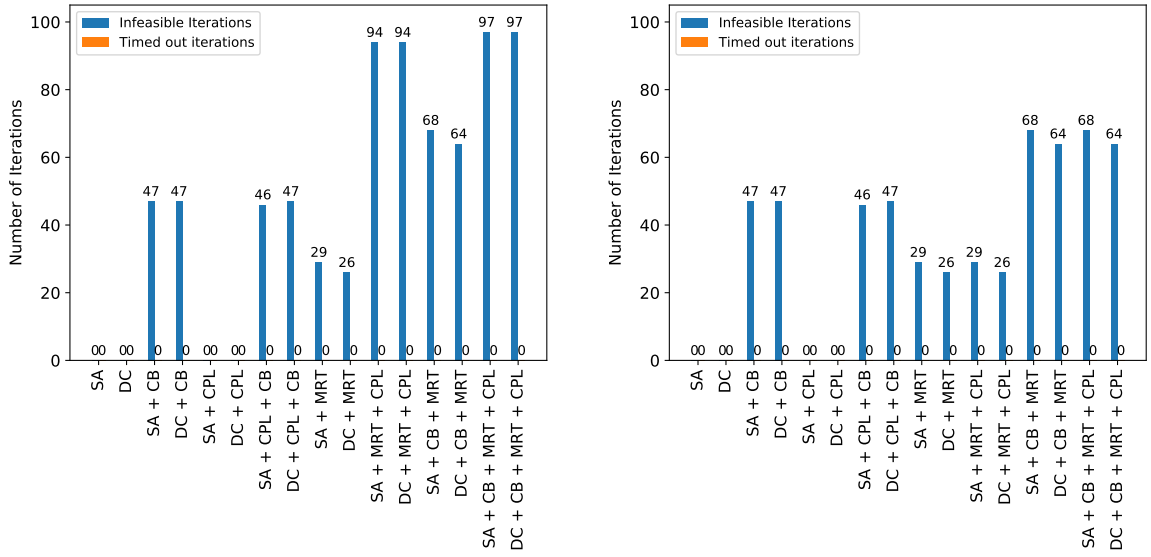
In addition to the solvability analysis, discussed above, we see that the re-dimensioning efforts result in an increase in the system fairness. This is more prominent for constraint combinations CB and CB + CPL, as understood from Figures 6.21(a) and (b), while from Figure 6.24(a) we deduce that a re-dimensioned access network topology leads to an across the board positive effect on the system fairness. The positive effects of network re-dimensioning are also prevalent in the network throughput plots in Figures 6.22(a)-(b) and 6.25(a). From these results we observe that the network re-dimensioning enables approximately 35% and 40% increase in the total network throughput for the CB and CB + CPL constraint combinations, respectively.

However, as can be seen from Figure 6.23(a), path latency still remains a bottleneck constraint in scenarios where MRT+CPL and CB+MRT+CPL constraint combinations are



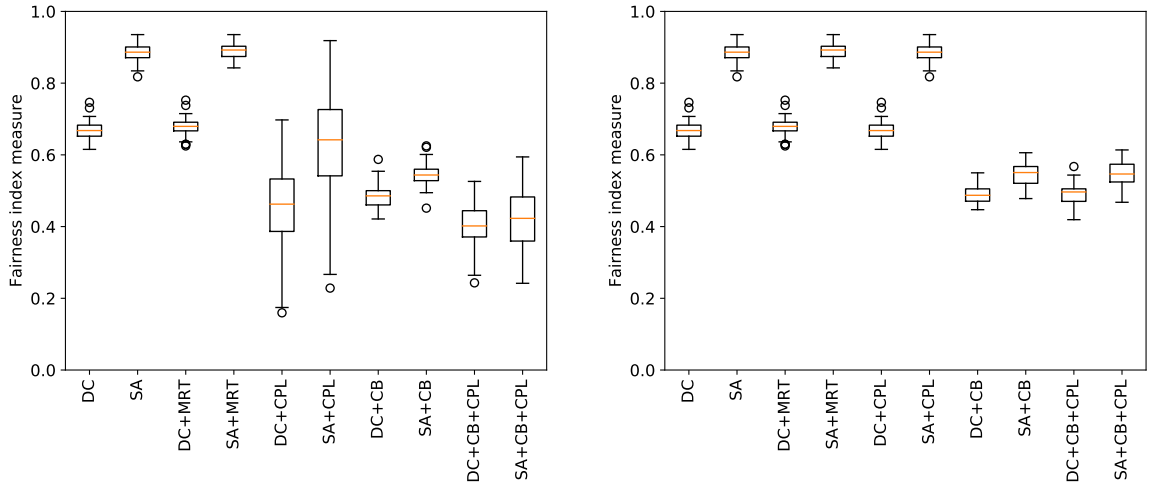
(a) (b)

Figure 6.28: Total Network Throughput for (a) CABEm without Relaxed Backhaul, and (b) CABEm with Relaxed Backhaul scenarios with 275 eMBB users.



(a) (b)

Figure 6.29: Optimizer Status for (a) CABEm with Relaxed Backhaul and Increased SC density scenario with 275 eMBB users, and (b) CABEm scenario with Relaxed Backhaul, Increased SC density, 5 ms downlink latency requirement and 275 eMBB users.



(a) (b)

Figure 6.30: System Fairness Measure for (a) CABEm with Relaxed Backhaul and Increased SC density scenario with 275 eMBB users, and (b) CABEm scenario with Relaxed Backhaul, Increased SC density, 5 ms downlink latency requirement and 275 eMBB users.

imposed. Moreover, when path latency is imposed as the only constraint, our optimizer converges to a solution in each of the 100 Monte Carlo trials. Hence, in a network wherein there can be multiple constraint combinations, such as MRT+CPL and MRT+CPL+CB, the operators should be careful when they sign the service level agreements (SLAs). These SLAs should not be overly restrictive, such as the one we have here where the 3ms downlink latency cannot be guaranteed in most topologies. As a consequence, we present our observations for the case when this downlink latency requirement is relaxed to 5ms. Immediately, through Figure 6.23(b) we observe that the optimizer is able to determine an optimal association in 68.3% more iterations for the CB+MRT constraint scenario, and in 11% more iterations for the CB+MRT+CPL. Further, the fairness and the total network throughput in the presence of MRT+CPL and MRT+CPL+CB constraints are also improved as seen through Figures 6.24(b) and 6.25(b). In addition to the relaxation in the SLAs, edge clouds, through appropriate placement [140, 141], can also provision great improvements in system performance. This is so because, they bring the services closer to the users, which reduces the total round trip time, and hence the downlink delay as well.

Also, from Figures 6.26-6.31, wherein the circular deployment is considered, we observe a similar trend in results as that in Figures 6.20-6.25. Concretely, from Figures 6.26(a) and (b) we notice that the number of iterations that converge to an optimal solution for the CB and CB+CPL constraint combinations increases by 15.9%. For the system fairness, from Figures

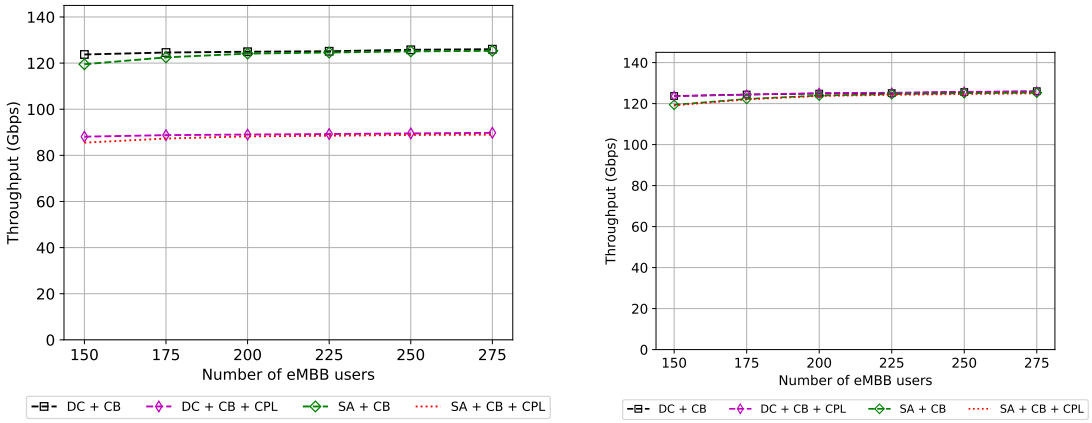


Figure 6.31: Total Network Throughput for (a) CABEm with Relaxed Backhaul and Increased SC density scenario with 275 eMBB users, and (b) CABEm scenario with Relaxed Backhaul, Increased SC density, 5 ms downlink latency requirement and 275 eMBB users.

6.27(a) and (b), it can be observed that for scenarios with CB and CB + CPL, the fairness is also improved. The reason being, the improved backhaul capacity allows the AURA-5G framework to assign resources more equitably to the users in the system. Additionally, as seen from Figures 6.28(a) and (b), the improvement in system throughput is nearly 71.4% and 60% for the CB and CB+CPL scenarios, respectively. Furthermore, from Figures 6.29-6.31, it can be deduced that the increase in the average number of SCs per MC from 6 to 8, as well as having less stricter latency requirements, results in resolving to a great extent the bottleneck nature of the MRT and path latency constraint alongside increasing the system fairness and the total network throughput.

Thus, the above observations highlight an important aspect of the AURA-5G framework, wherein the operators can test their network topology and infer its performance as well as the bottlenecks. And although we randomly increased the various network settings, the operators, through the use of the AURA-5G framework, are empowered with the ability to re-dimension their networks according to the inferences they make and inline with their ultimate objectives.

6.6 Related Work

User association has been an area of major interest ever since wireless networks came into existence. Over the course of these many years, multiple intriguing algorithms and methodologies have been proposed by academia and industry, aiming to resolve the incremental

challenges that user association presents as wireless networks continue to evolve. Concretely, while 2G and Wi-Fi networks utilized an RSSI/SNR based criterion for BS selection, the increasing complexity of the networks has driven the growth of algorithms that utilize TOPSIS [195], Fuzzy logic [196–198], Genetic Algorithms [196], Multi Attribute Decision Making (MADM) [199–204] and Optimization theory [205, 207, 223] among others as tools to accomplish an optimal user association. Moreover, they not only aim to provision an optimal association, based on maximizing sum rate of users for example, but they also try to optimize other system parameters such as interference, bandwidth allocation, interference reduction, etc.

Specifically, the authors in [188] consider an SDN enabled LTE network and utilize the global knowledge of the SDN controller so as to be able to distribute information about the backhaul load of neighboring BSs as well as the bottleneck backhaul link bandwidth for an BS of interest. Such a mechanism consequently allows for a more backhaul and BS load aware load-balancing mechanism. Next, in [189] the authors propose a distributed user association policy wherein they explore multiple dimensions to the user association problem. These dimensions explore the association policy while trying to optimize rate, throughput, minimize file transfer delays and load balancing. They do so by varying a factor, which characterizes each of these dimensions. Further, the distributed nature encompasses the fact that part of the optimal solution search is performed on the UE and a part of it is performed on the BS. Further, authors in [190] utilize the k-nearest neighbors principle and the azimuth angle to determine the next BS to associate to in a mobile environment. They analyze the pattern associated with a given k-value and then utilize pattern recognition methods to determine the optimal BSs to associate to given an ultra-dense topology and severe interference scenarios. Moreover, in [191], the authors utilize Media Independent Handover (MIH) and its corresponding services so as to be able to execute a MADM algorithm. They compare it with other simplistic counterparts, such as Additive Weighting strategies, to project the benefits of MIH based MADM. While these discussed strategies consider multiple constraints and propose effective solutions, they limit the number of BSs that a UE can connect/choose to just one.

As a consequence, authors in [192] proposed a Dual Connectivity (DC) based user asociation solution, wherein they formulate an optimization problem based on maximizing the sum throughput. Since, the problem is ridden with non-linearities and is NP-hard, they propose a tractable solution to achieve the optimal solution. Also, in [193], the authors study the problem of user association with the objective of maximizing the weighted sum rate with user rate (minimum and maximum) constraints as well as with another objective of maximizing the proportional fairness of the system. Note that, in [193], for DC, the authors consider that

a UE can be associated with an MC and an SC that is within the coverage of this selected MC. Next, the authors of [224] consider the problem of user association for the uplink with power allocation optimization being another objective. They consider user minimum rate requirements as well as backhaul limitations at the pico BSs as the constraints and aim to find an association that reduces the overall cost for the network while demands of all the users are satisfied.

Moreover in [20] the authors consider the downlink aspect of dual connectivity as well as the mmWave aided HetNets. Correspondingly, the authors develop a two stage iterative algorithm for user association where their objective is to maximize the total network throughput subject to the fact that the overall access-fairness amongst the users has to be improved. However, for the analysis the authors consider a very small representative scenario, which in essence does not represent the real world densities of SCs, MCs and users. In addition, in [225] the authors propose an opportunistic method for cell selection in a DC scenario, wherein the load characteristics of the network are taken into account. However, according to the network model considered for simulations, their analysis is only limited to the sub-6 GHz scenarios. Another significant study in the direction of user association for 5G networks is in [30]. The authors of this work analyze the impact of wireless backhaul on user association. The optimal association is determined such that it maximizes the overall network throughput whilst not exceeding the backhaul capacity limits.

Lastly, multiple studies such as [205–208] propose a joint optimization approach towards user association. In these studies the objective function is maximized/minimized, depending on the utility, and an optimal association strategy that provisions the same is determined. Additionally, they also incorporate, within their optimization approach, a search for the association that will lead to an optimal system interference/energy consumption regime/spectrum allocation system, etc. Such mechanisms are termed as *joint optimization* approaches, and through the use of binary decision variables they introduce another dimension of non-linearity to an already non-linear optimization problem. However, a relaxation of these decision variables or a decomposition into simpler sub problems is in general possible. And so, in [205–208] such techniques have been utilized and a discussion on the optimal solution obtained has been presented.

6.7 Summary

In this chapter, we have provided a first holistic study in literature with regards to the joint optimization based user association framework for multiple applications in 5G networks. The study entails utilizing a MILP formulation based joint optimization process, which takes into

account the service classes and accordingly assigns the bandwidth and BS to the users in the system. Further, we organized the entire process into a software suite, wherein the location generation, system model specifics, optimization and data analytics are performed. We refer to this softwarized framework as *AURA-5G*, and present it to the academic and industrial community for analyzing the user association and resource allocation behavior in conjunction with each other.

Next, for the optimizer, we developed a MILP framework wherein the objective is to maximize the overall network throughput, and find the appropriate user-BS-bandwidth association that achieves the same. We established the various constraints that help us perform the multi-dimensional study carried out in this chapter. Subsequently, we also establish certain methodologies to resolve the non-linearities introduced by multiplication of binary decision variables. This, assists in reducing the complexity of our optimization problem. Further, we also present a novel discussion on the complexity of SINR calculation and how our computation method, while being a lower bound and sub-optimal, assists in reducing the complexity of the AURA-5G framework.

In addition, as part of this study we presented the performance of the AURA-5G framework for dual connectivity (DC) and single connectivity (SA) modes. For the performance evaluation process, we utilized realistic network scenarios and parameters so as to be able to ensure the efficacy of the AURA-5G framework. Consequently, we showed that the established framework outperforms the baseline scenario in terms of the total network throughput. Further, we presented the performance characteristics of the AURA-5G framework for the DC and SA modes alongside the multiple constraint combinations in terms of system fairness, backhaul utilization, latency compliance, convergence time distribution and solvability. Note that, for DC modes we present a novel analysis for scenarios where the choice of SC does not have to be geo-restricted by the choice of MC and where the user has the opportunity to connect to any two BSs, i.e., SC-SC, SC-MC or MC-MC.

We now summarize some of the important findings from our detailed analysis as follows:

1. For the *total network throughput* metric (Figures 6.3, 6.4, 6.6 and 6.7), scenarios, wherein circular deployment, beamformed, only eMBB services and *AnyDC* setup was considered, showed significant performance gains of upto 17.2 % for dual connectivity as compared to single association. On the other hand, with the *MCSC* setup the gains were not as significant. However, both the *AnyDC* and *MCSC* scenarios registered improvements of upto 529× and 476× over the baseline scenario, respectively. Note that, while these values correspond to the scenario where dual connectivity without any other network or application constraint has been considered, our framework outperforms the baseline framework for all the studied scenarios (Figures 6.3 and 6.4).

Further, the CPL and CB constraints severely impact the overall network throughput for all scenarios. In addition, for the scenarios with the interference limited regime, one can immediately observe a significant reduction in the overall network throughput due to the degradation in SINR. Moreover, for the square deployment scenarios, a gain of nearly 6% in the total network throughput is observed, as compared to the circular deployment scenarios. Finally, for the scenarios with both eMBB and mMTC services, since we consider the mMTC services to operate in the guard band and consume only the backhaul resources, a corresponding reduction in the overall throughput for the eMBB services in the presence of CB constraint was observed.

2. For the *system fairness* metric (Figures 6.8-6.10), scenarios wherein only eMBB services, circular deployment, *AnyDC* setup and beamforming are considered, showed that single association achieved higher fairness than dual connectivity. The Minimum Rate (MRT) constraint however assisted in a slight improvement of system fairness, given the DC setup. Moreover, the CB and CPL constraints resulted in a significant lowering of the overall system fairness. Furthermore, for *MCSC* setup, an overall improvement in system fairness for the DC setup was observed. However, for the scenarios where an interference limited regime was considered, a significant drop in system fairness was noticed, given the SINR degradation and the greedy nature of the objective function (Section 6.1, equation 6.14). Next, the square deployment scenarios showed an overall improvement of 5-6% in system fairness as compared to the circular deployment scenarios. Lastly, we analyzed the scenarios wherein both eMBB and mMTC services co-exist. For these scenarios, we observed that fairness measure is not affected significantly as compared to that noticed for only eMBB scenarios.

3. For the *user throughput distribution* metric (Figures 6.11 and 6.12), we observed that the AURA-5G framework is able to determine user-BS-bandwidth associations such that the rate constraints are satisfied. It provisions an interesting insight into how the user rates are distributed when *AnyDC* and *MCSC* setups are studied in the presence of beamformed and interference limited regimes.

4. For the *backhaul utilization* metric (Figures 6.13 and 6.14), we discern that the AURA-5G framework works exceptionally well in being able to adhere to the strict backhaul capacity constraints imposed by the network. Further, for the scenarios with beamforming we observe that the backhaul capacity is almost completely utilized as compared to that in the scenarios with interference limited regime. Additionally, for scenarios wherein both eMBB and mMTC devices are considered, it was observed that

the overall backhaul utilization by the eMBB devices is lower than that in scenarios where only eMBB devices exist.

5. For the *latency compliance* metric (Figure 6.15), we observed again that the AURA-5G framework is able to determine user-BS-bandwidth associations such that the latency constraints are satisfied. It was observed that, while in *AnyDC* setups the users accessed SCs more than MCs, for the *MCSC* setups, a higher density of users was observed to have access to MCs and thus a reduced latency.

6. Through our novel convergence time distribution and solvability analysis, it was observed that certain constraint combinations are very restrictive and hence, the network requires re-dimensioning. It is imperative to state that such insights will be significantly important for the operators in network planning.

7. We presented, in Section 6.5, an analysis of certain challenging scenarios wherein the network re-dimensioning was carried out on both the access and backhaul network. We showed that, a simple re-dimensioning process, wherein the SC density was increased from 6 to 8 SCs on average per MC, the backhaul capacity was increased and less restrictive SLAs were agreed to, resulted in significant improvement in system performance, thus alleviating the bottleneck constraints concern. Concretely, an improvement in total network throughput performance of upto 75.8% – 77.29%, alongside the illustrated improvements in system fairness and number of solvable iterations for CABEm (Section 6.5), is provisioned by our framework. Additionally, for SABEm, the total network throughput performance improvements by utilizing our framework range from 42.51% – 96.39%. Similar to CABEm, improvements in the system fairness and solvability have been illustrated through our discussions in Section 6.5.

Lastly, the proposed user association and resource allocation methodology, through its ability to utilize application requirements related details, and access and backhaul network related information, provisions a method that satisfies the *cross layer methodology* component of the MM framework defined in Figure 3.1. Moreover, given the fact that the proposed user association strategy facilitates dual connectivity and performs intelligent RAT selection, it also helps cater to the *multi-connectivity* and *RAT selection* solution block components for the MM framework defined in Figure 3.1.

Chapter 7

Conclusions

The dawn of mobile Internet age in the 2000s with the arrival of GPRS and subsequently 3G services, transformed how we have lived our daily lives. This transformation, driven by faster data rates, versatile applications and better QoS, has since grown and evolved at an exponential rate. With the ushering in of the 5G networks, the conventional algorithms and procedures will cease to service the network and user requirements efficiently. The reason being, 5G networks will be more dense in terms of the number of users and BS [7, 9] and will need to support a wider range of applications (eMBB, URLLC and mMTC) with different mobility profiles and QoS requirements [7, 9]. More critically, designing newer methods will be challenging than ever before, given the enhanced requirements for reliability, scalability and flexibility by the 5G networks. Amongst those is the critical mechanism of Mobility Management (MM). As we have already emphasized throughout this thesis, MM is important for users as it grants them seamless mobility and seamless access to services irrespective of their location and movement patterns, thus also cementing its importance for any wireless networks' ubiquity.

Hence, in this thesis we have illuminated the broad field of mobility management, wherein the erstwhile legacy mechanisms devised and developed by multiple standardization bodies, academia and industry have been at first discussed in detail. Through these discussions, which were carried out in Chapter 2, we also highlighted upon the mechanisms that are currently being developed for the 5G networks. A very unique feature of the discussions in Chapter 2 was to also take into cognizance the prevailing works with regards to Beyond 5G networks. With this background, in Chapter 3 we then laid out a framework for a novel qualitative analysis. The main goal of this qualitative analysis was to assess the readiness of existing/developing MM mechanisms for 5G and beyond networks. For this, we first defined the three main pillars of any future mobility management mechanism, i.e., reliability,

scalability and flexibility. Recall that while reliability will ensure seamless connectivity even in the most adverse environments, scalability will guarantee that the MM mechanisms are able to service the exponentially increasing number of users/devices. Additionally, flexibility will ensure that the heterogeneity prevalent in the future networks will be an ally and not an adversary. Following this framework we have provisioned the qualitative analysis for multiple MM mechanisms, which have been widely utilized in the existing wireless networks as well as for those that have been newly developed given the arrival of 5G networks. The analysis determined that the existing and current mechanisms do not fully satisfy the requirements of the 5G and beyond networks. Consequently, we then outlined the persistent challenges towards the development of MM mechanisms for such networks. We have then presented the potential strategies that the research community can explore to circumvent these challenges. With the inferences from the rest of the discussions in Chapters 2 and 3, we finally presented a framework for 5G and beyond MM strategies.

Building on this framework, in Chapter 4 we introduced a novel on-demand MM strategy. This strategy aims at provisioning the required granularity and flexibility for the future MM mechanisms. Concretely, it presents a methodology wherein multiple aspects such as mobility profile, type of flows, network load and predefined policies can be utilized to tailor the MM mechanisms on-demand. In addition, it also aligns the implementational aspect of the MMaaS strategy with the upcoming SDN based framework. And so, through this strategy we provide a MM methodology that is applicable not only in the future softwarized network architectures, but it also provisions scalability and flexibility aspects necessary for the future MM strategies.

Next, in Chapter 5, we have tackled the issue of smart CN signaling, which has also been stated as one of the building blocks for the 5G and beyond MM framework (Chapter 3). Notably, we present a novel handover method and system, wherein we improved the handover preparation and failure phase signaling sequences. We have also proposed a novel handover failure aware handover preparation signaling, which further enhances the performance of the handovers that have been designed for legacy as well as 5G networks. The aforesaid performance is analyzed using the latency, transmission cost, processing cost and message size analysis metric. The analysis has been conducted for the myriad scenarios that have been defined in the 3GPP technical specifications, and subsequently the performance of the proposed mechanism has been compared with the standards. In addition, we also performed a novel network wide analysis, wherein we consider an ensemble of 3 million users and varying percentages of the various handover types studied to assess the benefits that the proposed strategy grants to the network operators in a large deployment scenario. Notably, to realize the gains that the proposed strategy provisions over the standards, we also utilized data from

a Greek and Japanese telecom operator. We showed that the proposed handover sequences in fact can lead to significant reduction in signaling load in real network scenarios. To conclude this work, we then proposed an evolutionary 4G/5G network architecture that would help adapt to the proposed signaling strategy whilst causing minimal disruption through network re-design as well as minimal increase in the CAPEX.

Lastly, in Chapter 6, we have presented a novel user association and resource allocation strategy. As we have already noted in Chapter 3, ensuring context awareness, multi-connectivity, optimal RAT selection and possibility of cross layer strategies will be essential building blocks for 5G and beyond MM framework. Hence, in this chapter we have developed the *AURA-5G* framework, which utilizes a MILP formulation to perform the user association and resource allocation task, with the objective of maximizing total network throughput. We subject the system and topology analyzed to multiple real network constraints and also take into account the application requirements while designing these aforementioned constraints. Through our analysis we observed that the proposed strategy performs better than the baseline scenarios in terms of the total network throughput, which is also the objective function. Further, we analyzed myriad scenarios including dual connectivity scenarios for which, to the best of our knowledge, no studies exist. Our analysis also delves into the fidelity and scalability of our solution through the convergence time, latency compliance, backhaul utilization compliance and user throughput distribution CDF metrics. These will be essential to both industry and academia, as we have developed this tool keeping in mind the requirements of both these communities.

To conclude, our work advances the area of Mobility Management and proposes methods that are of great utility to both industry and academia. This is concretely highlighted by the fact that our research work has been featured in 4 international conferences, 1 Journal (Q1 Journal with IF 4.096) and 1 Patent Application (PCT application with a positive International Search Report), with 2 Journal papers still *under review* (1 in Q1 Journal with IF 3.03 and 1 in Q2 Journal with IF 2.766). However, we believe that more efforts are required to enable the future MM strategies to be capable of satisfying the 5G and beyond network requirements. Thus, in the next chapter we briefly discuss some of our possible future directions.

Chapter 8

Future Work

In this thesis we have endeavored to push the status of the MM strategies from being benign and insufficient towards satisfying the 5G and beyond network requirements, to potent and effective. However, there still remain certain areas that require more research and development efforts to realize this goal.

And so, one of the most important areas that we are already in the phase of addressing is the fact of including URLLC services within the ambit of the *AURA-5G* framework. Recall that, in Chapter 6 we defined the *AURA-5G* framework, which utilizes a joint optimization process to determine an optimal user association and resource allocation strategy. The URLLC services will consequently represent a far more challenging service class for the 5G networks to cater to, given their low latency and ultra-reliability requirements. Hence, we will develop a multi-objective optimization problem wherein we aim to satisfy the QoS requirements of all the prevailing service classes(i.e., eMBB and mMTC), including the URLLC services. In addition to this, we also intend to formally release the code of our simulation tool as an open source toolbox, which can be adapted for research in academia as well as for implementation by the industry.

Next, with the formalization of the Mobile Edge Computing (MEC) paradigm, it will become critical to be able to perform effective service migration and service relocation, given the user mobility and application profile. Particularly, in scenarios where the users move from one PLMN to another, it will be extremely critical to have effective strategies for service migration/relocation in order to ensure seamless mobility. Hence, a future work will be to understand the dynamics of stateless and stateful applications and devise prediction algorithms to understand when and what to migrate to the destination MEC.

Another aspect that has to be actively explored is the implementation and testbed based analysis of the proposed handover strategy in Chapter 5. Note that, through Chapter 5 we

introduced a novel handover preparation and failure phase signaling process, wherein we were able to reduce the amount of handover signaling by upto 50% in certain scenarios. Hence, by utilizing the testbed we intend to elevate the level of our invention to TRL-4, which would then make it viable for introduction to actual field trials and possible adoption as a market product. For this we are already in the process of determining the input costs and accordingly reaching out to relevant industrial partners. Simultaneously, we are also tirelessly pursuing the possibility of including our work into 3GPP specifications. Additionally, we have also been exploring the possibility of utilizing open source platforms, such as Open Air Interface (OAI), for carrying out the testbed analysis. Notably, and in this regard, a training was undertaken in NEC Laboratories Europe GmBH, in 2019, to gain operational experience in utilizing OAI to setup SDN/NFV based testbeds.

Lastly, given the current research trends in Artificial Intelligence and Machine Learning, as well as certain important works in the area of MM utilizing the same [226], we believe it will be prudent to explore these techniques for the future MM strategies as well. Note that, due to the ability of the AI algorithms to learn complex relationships between input and output data, we endeavor to utilize them in determining the user association and resource allocation strategies in an increasingly stringent network scenario. However, it remains to be seen, the scale of performance benefits that can be achieved given the relatively short amount of time MM mechanisms have to determine and orchestrate the best solution. Thus, computational complexity and accuracy of the solution will be an interesting trade off that can be looked at as future work.

Appendix A: Handover Signaling – Other Scenarios

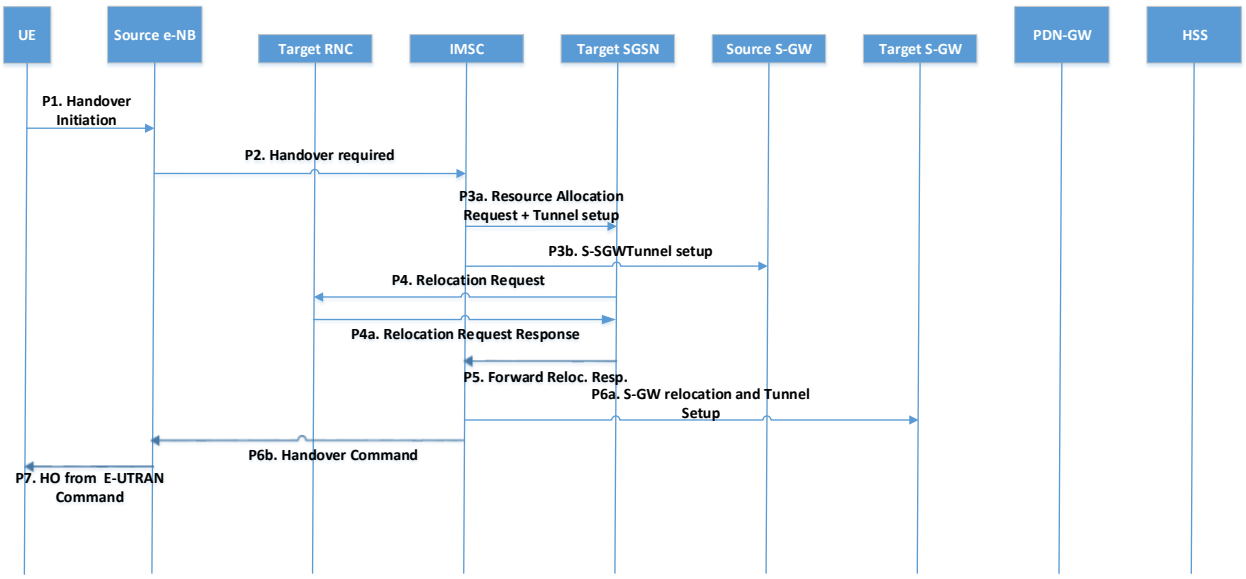


Figure A.1: Proposed Handover Signaling for LTE to 3G/2G Inter-RAT HO with Target S-GW and Direct Tunnel.

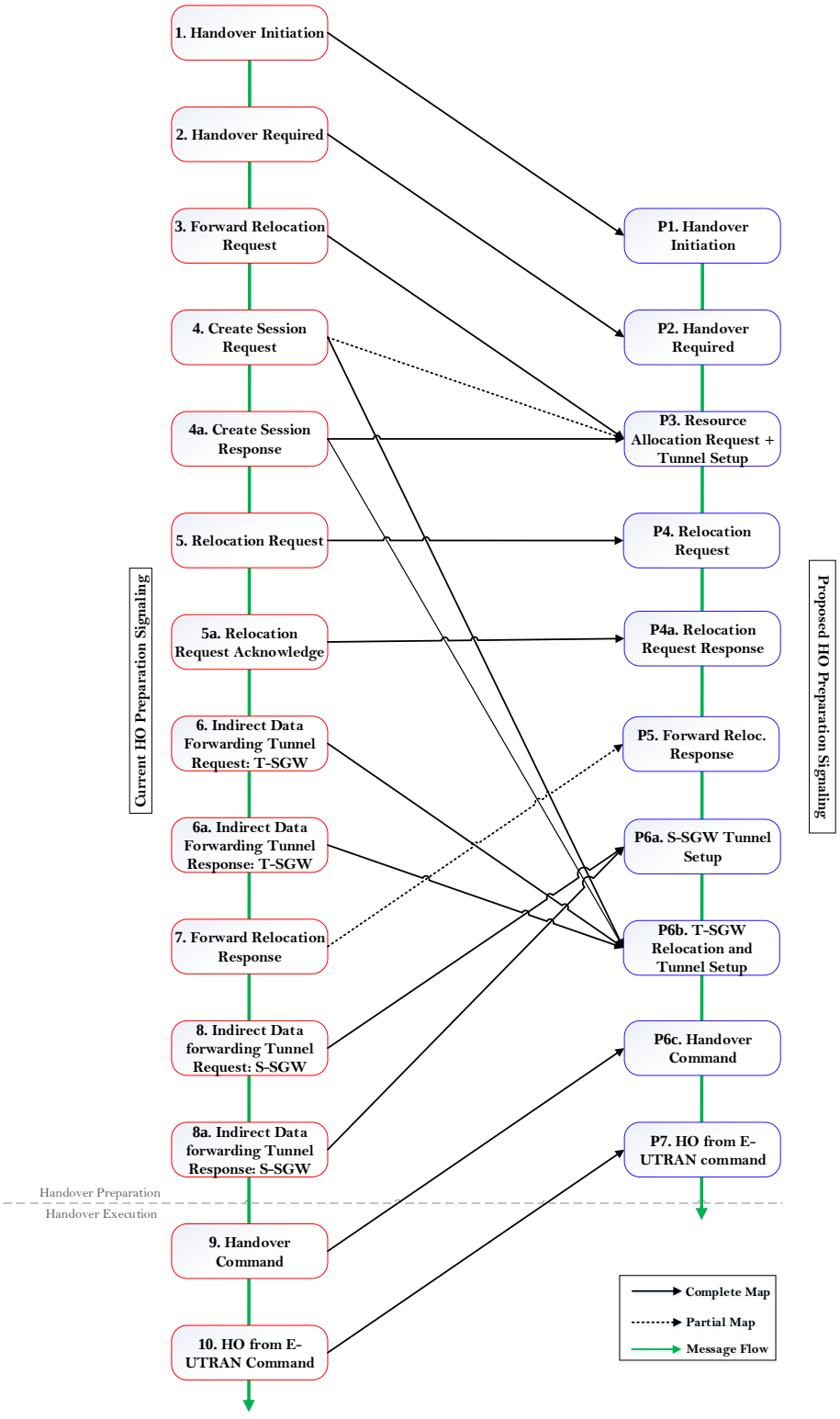


Figure A.2: Proposed Handover Signal mapping for LTE to 3G/2G Inter-RAT HO with Target S-GW and Direct Tunnel.

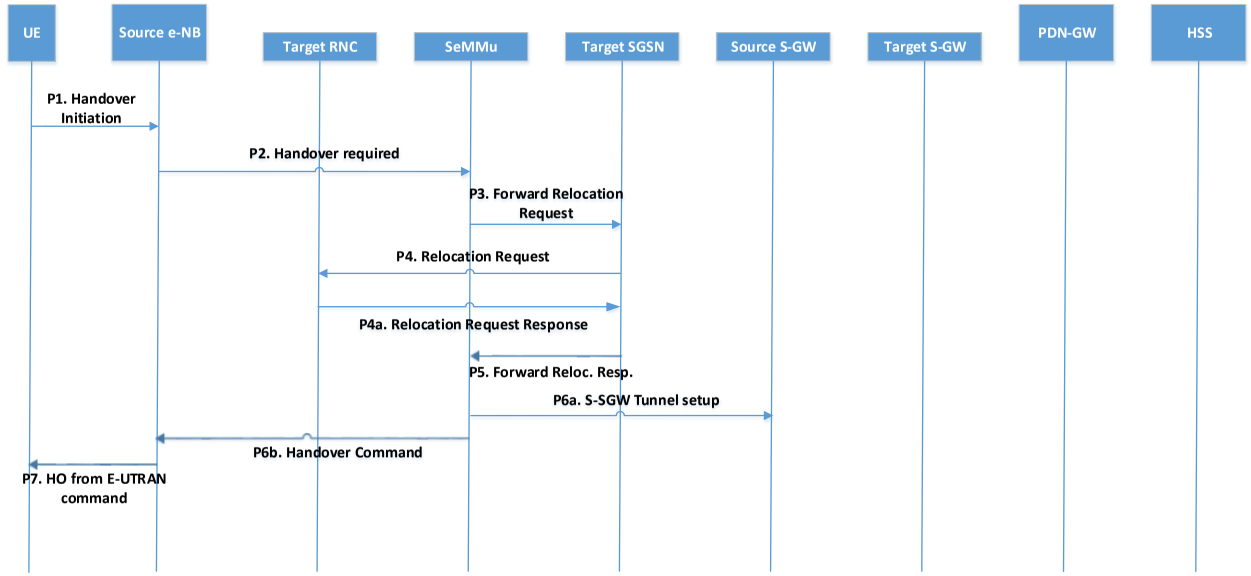


Figure A.3: Proposed Handover Signaling for LTE to 3G/2G Inter-RAT HO without Target S-GW and Direct Tunnel.

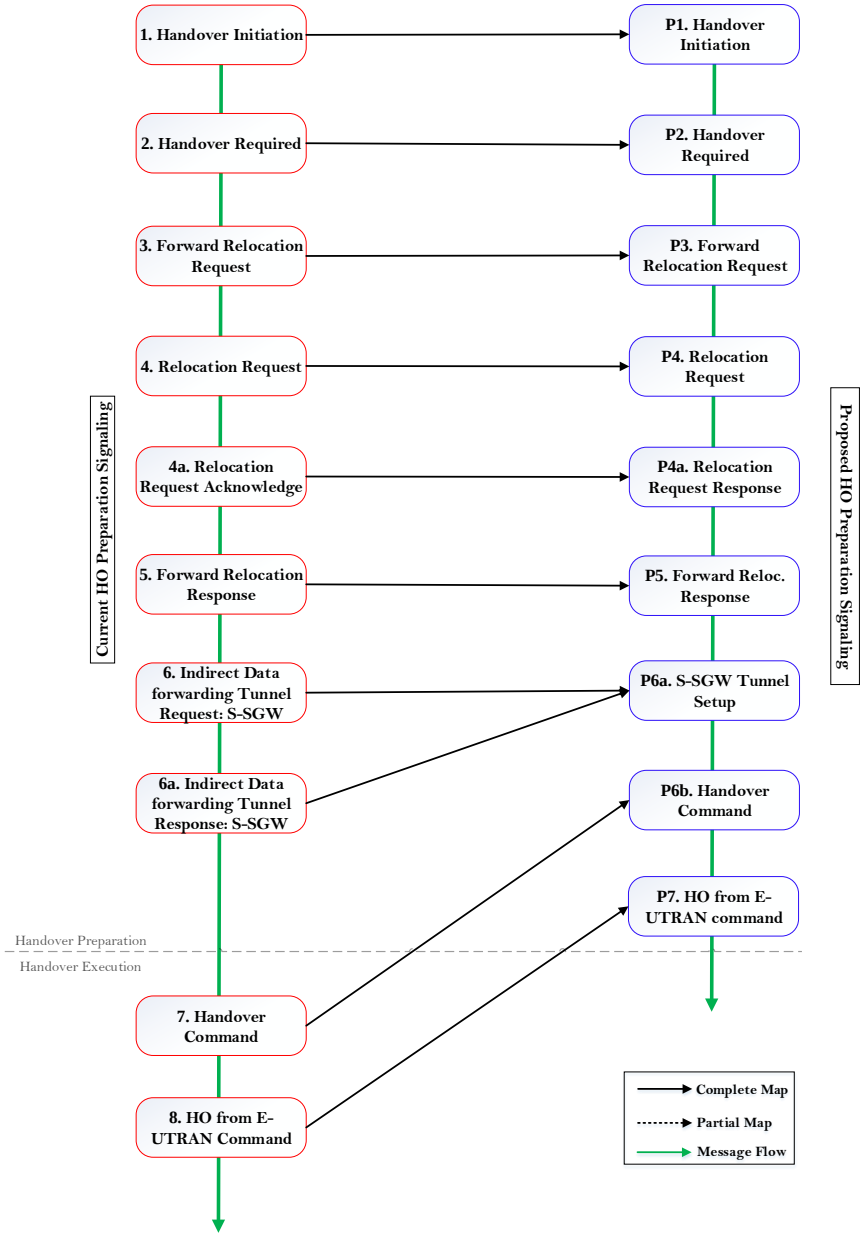


Figure A.4: Proposed Handover Signal mapping for LTE to 3G/2G Inter-RAT HO without Target S-GW and Direct Tunnel.

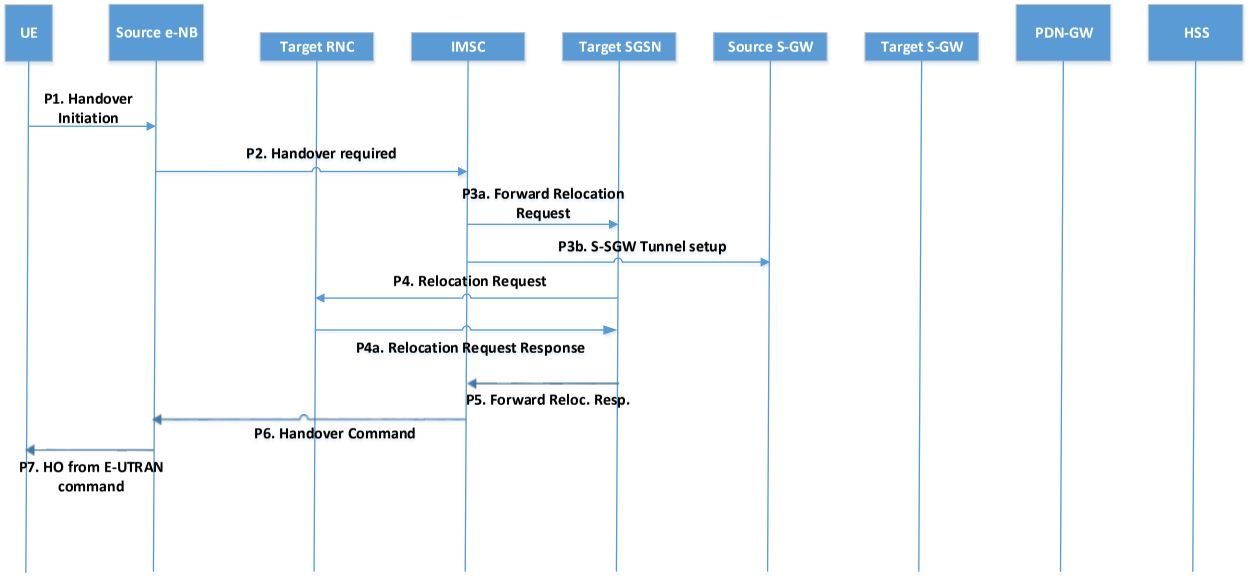


Figure A.5: Proposed Handover Signaling for LTE to 3G/2G Inter-RAT HO without Target S-GW and Indirect Tunnel.

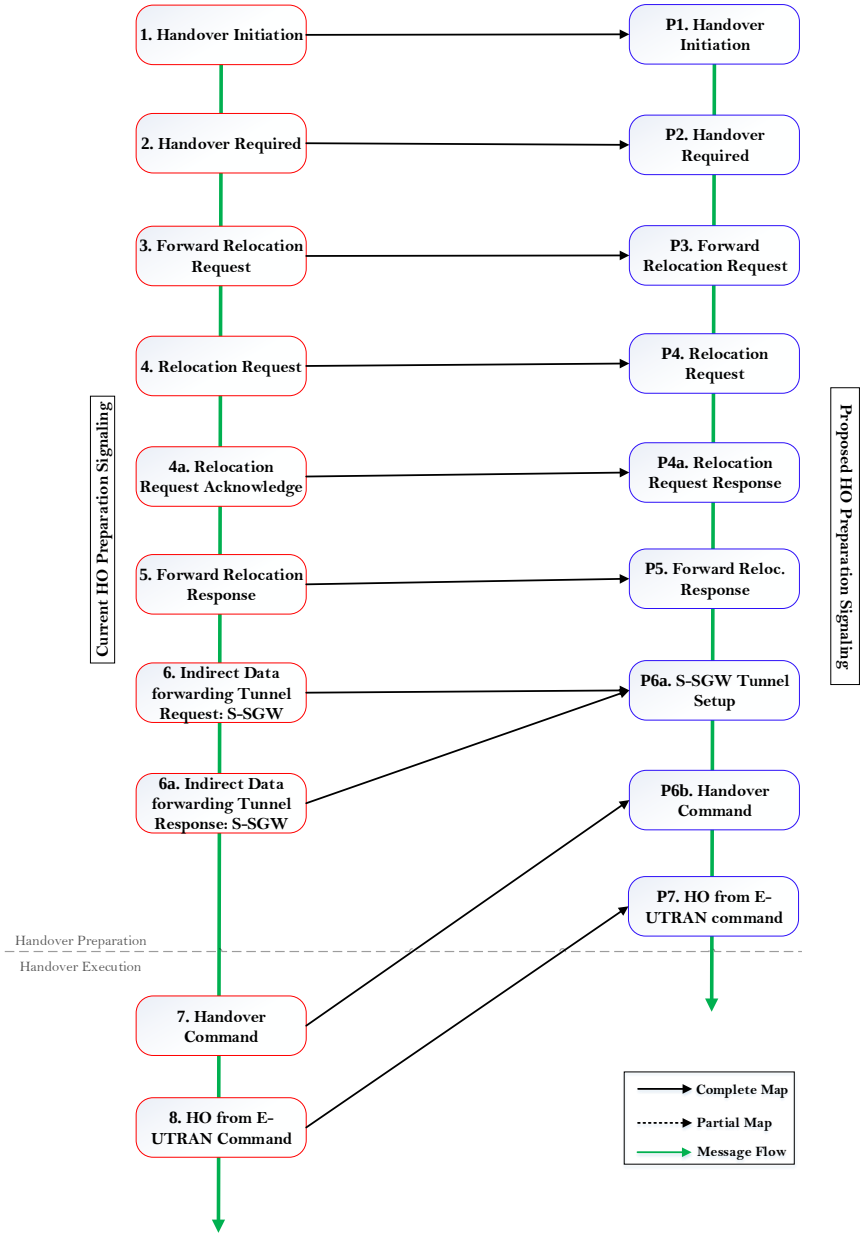


Figure A.6: Proposed Handover Signal mapping for LTE to 3G/2G Inter-RAT HO without Target S-GW and Indirect Tunnel.

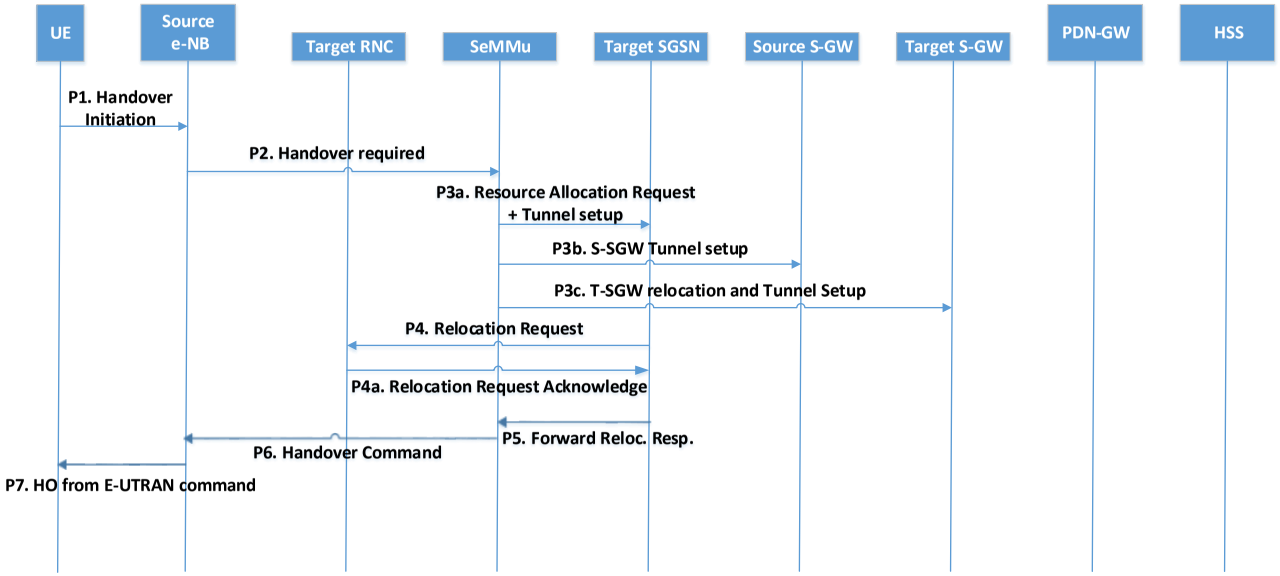


Figure A.7: Proposed Handover Signaling for LTE to 3G/2G Inter-RAT HO with Target S-GW and Indirect Tunnel.

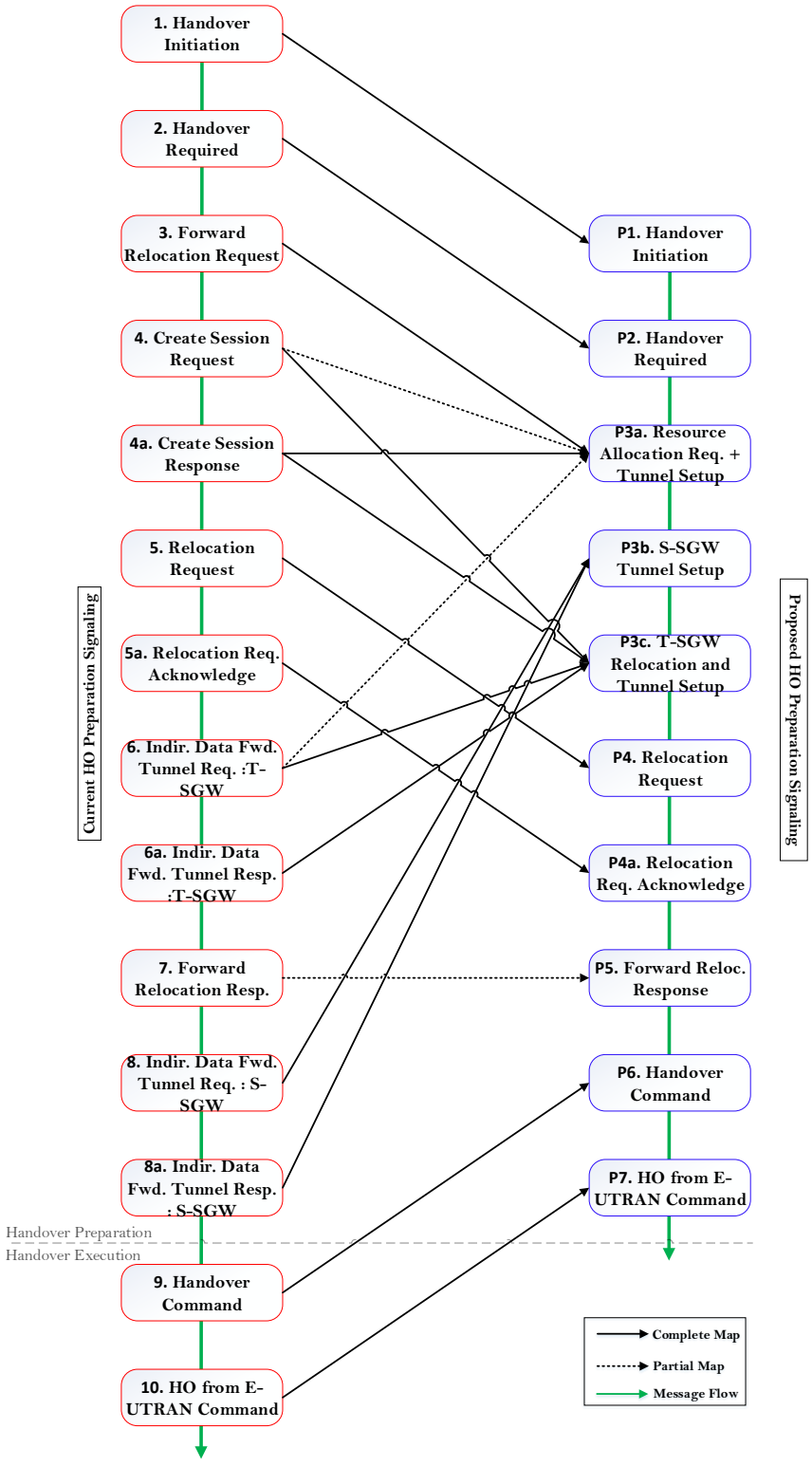


Figure A.8: Proposed Handover Signal mapping for LTE to 3G/2G Inter-RAT HO with Target S-GW and Indirect Tunnel.

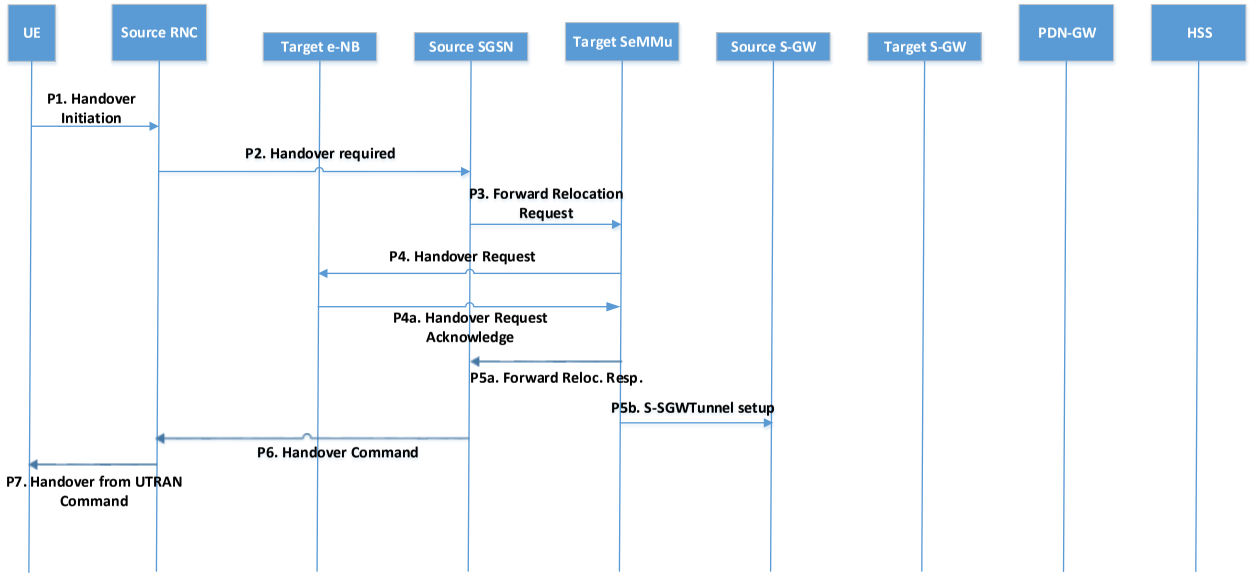


Figure A.9: Proposed Handover Signaling for 3G/2G to LTE Inter-RAT HO without Target S-GW.

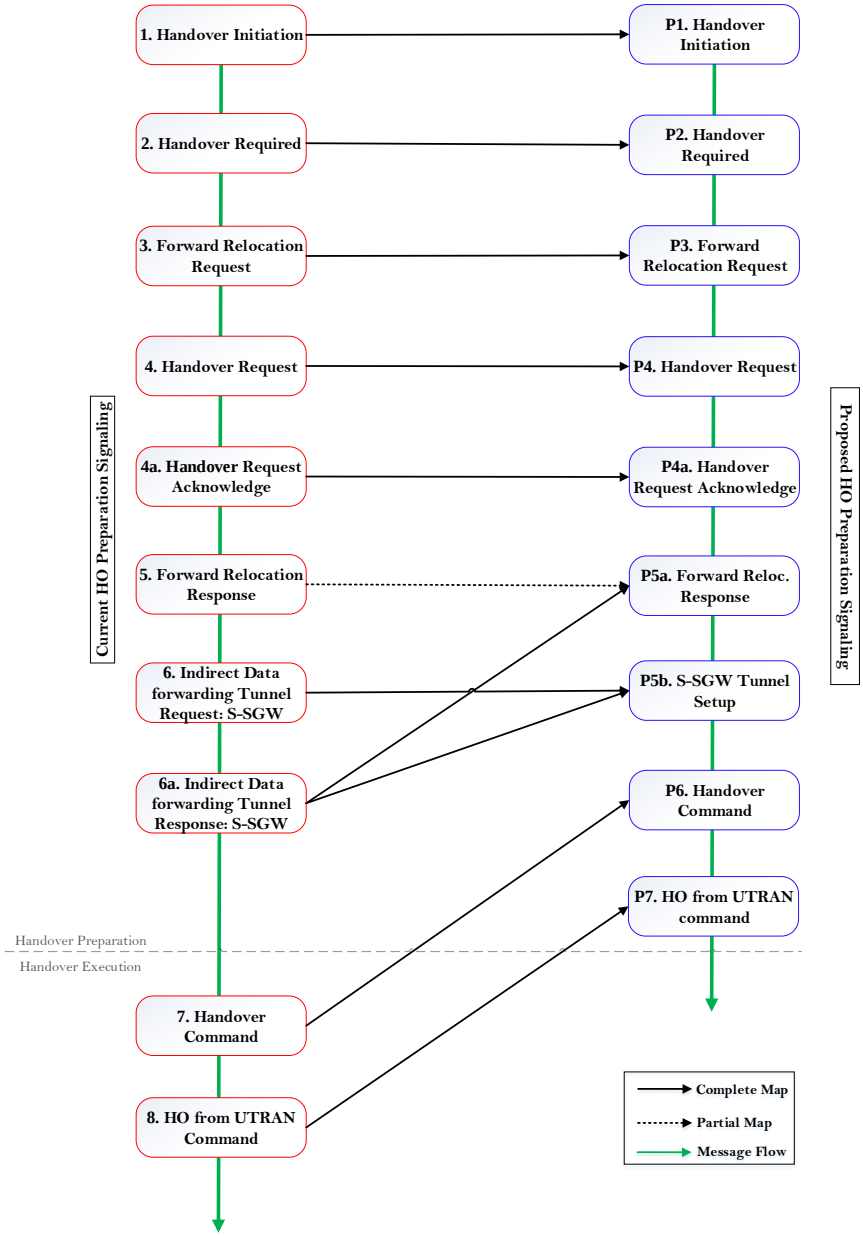


Figure A.10: Proposed Handover Signal mapping for 3G/2G to LTE Inter-RAT HO without Target S-GW.

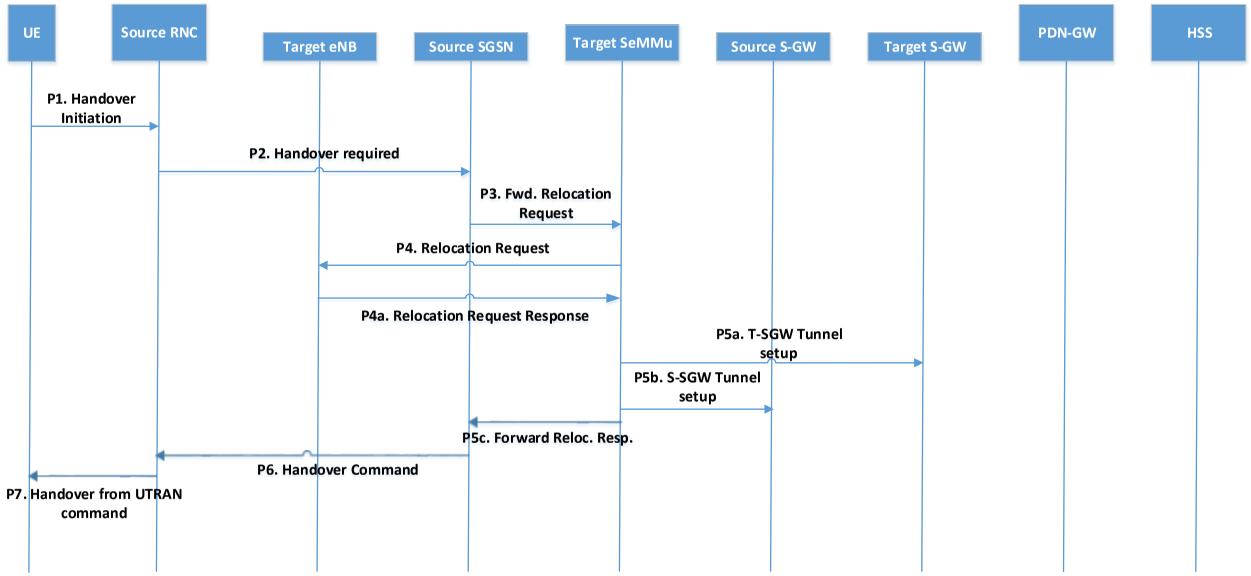


Figure A.11: Proposed Handover Signaling for 3G/2G to LTE Inter-RAT HO with Target S-GW.

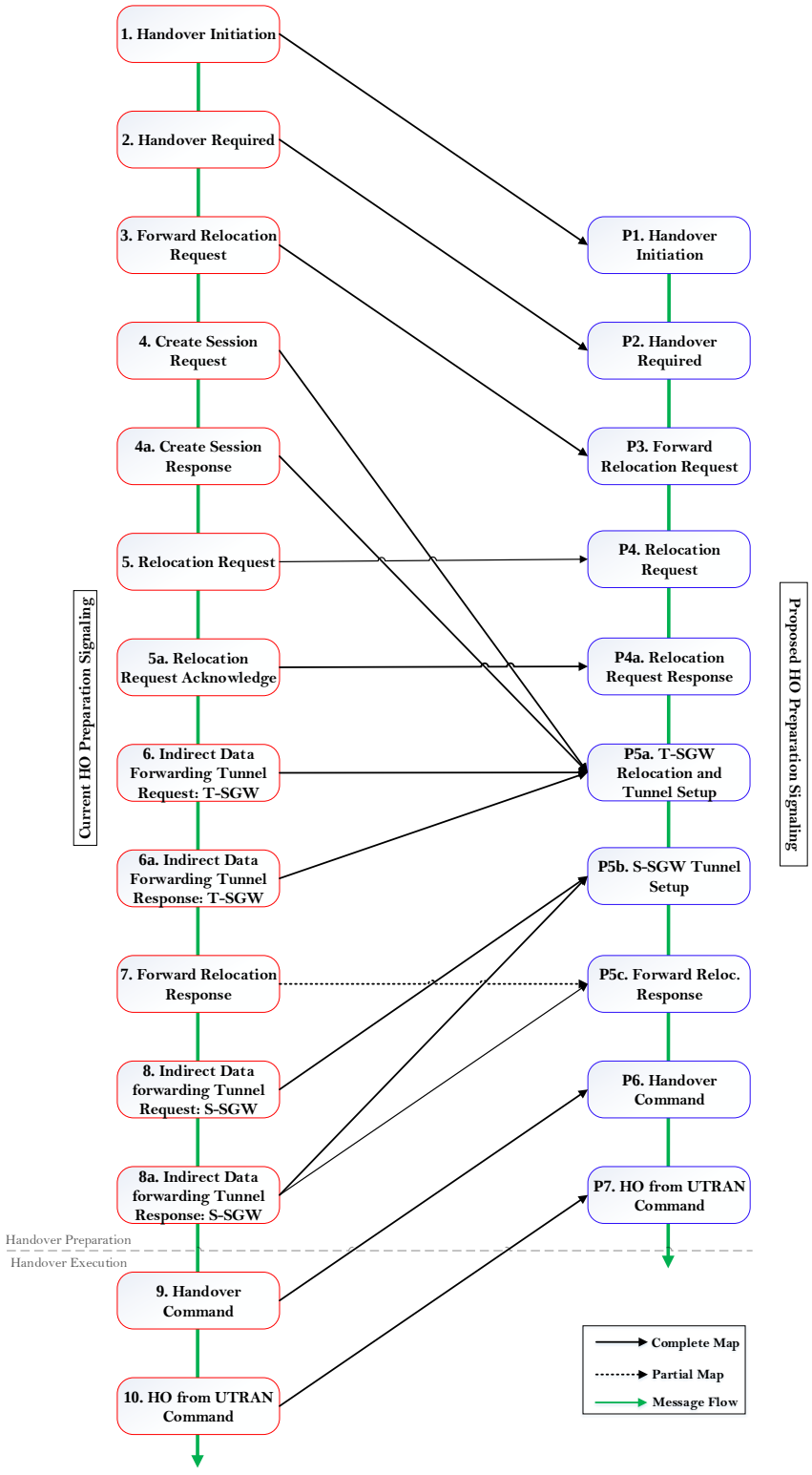


Figure A.12: Proposed Handover Signal mapping for 3G/2G to LTE Inter-RAT HO with Target S-GW..

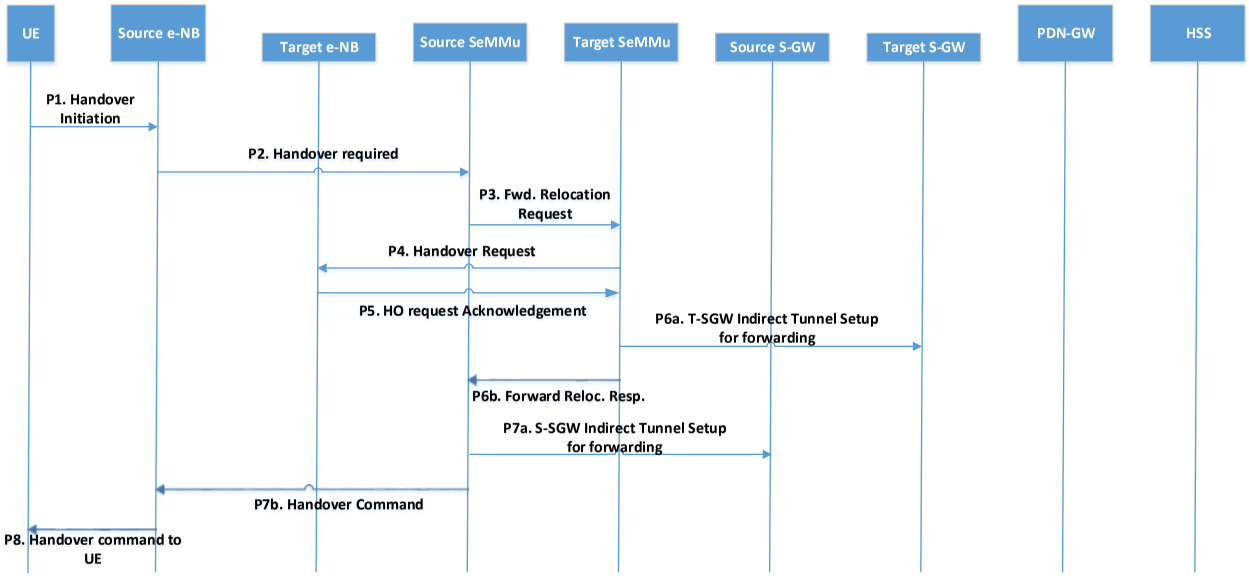


Figure A.13: Proposed Handover Signaling for LTE Intra-RAT HO with Target S-GW and MME.

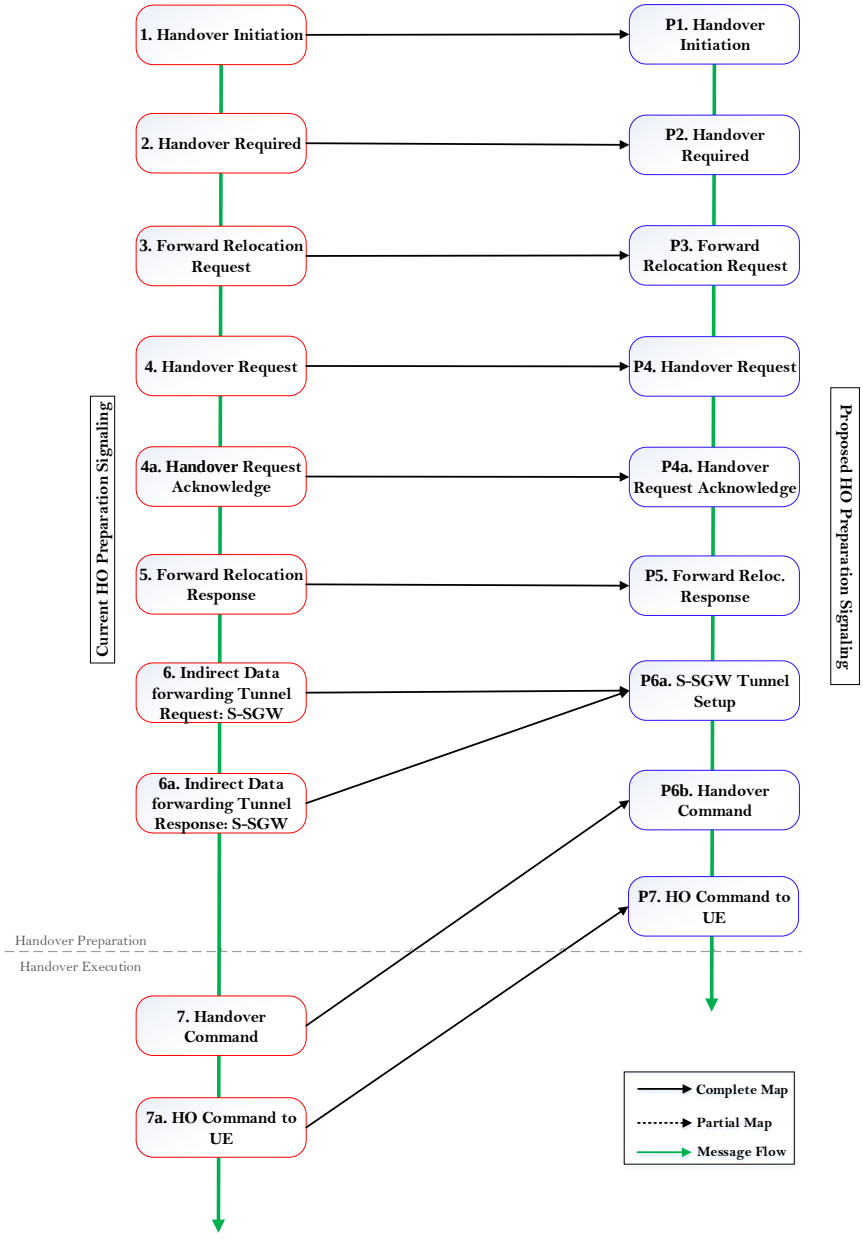


Figure A.14: Proposed Handover Signal mapping for LTE Intra-RAT HO with MME relocation (without S-GW relocation).

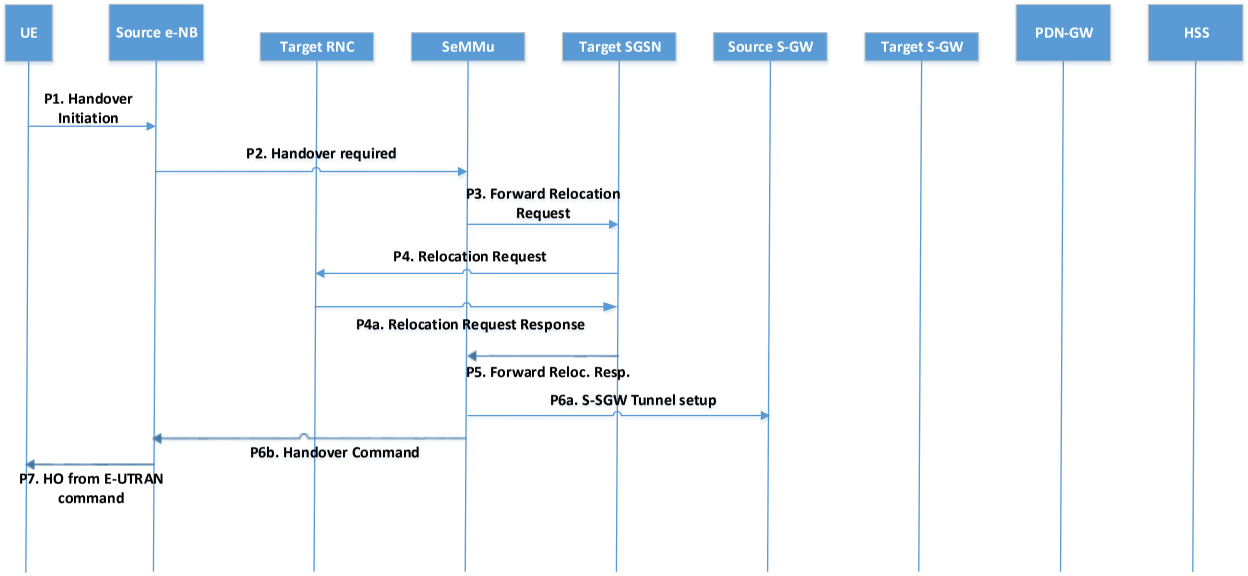


Figure A.15: Proposed Handover Signaling for LTE to 3G/2G Inter-RAT HO without Target SGW and Direct Tunnel.

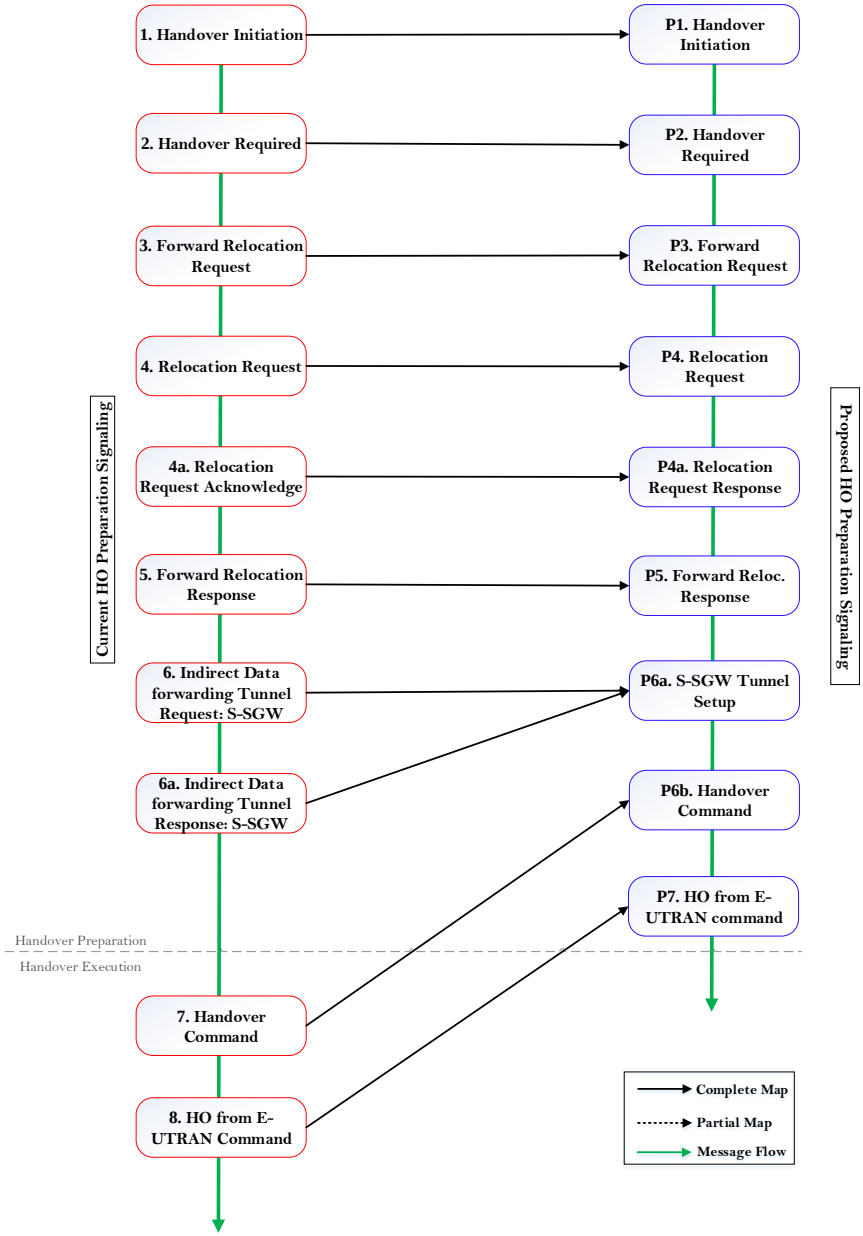


Figure A.16: Proposed Handover Signal mapping for LTE to 3G/2G Inter-RAT HO without Target SGW and Direct Tunnel.

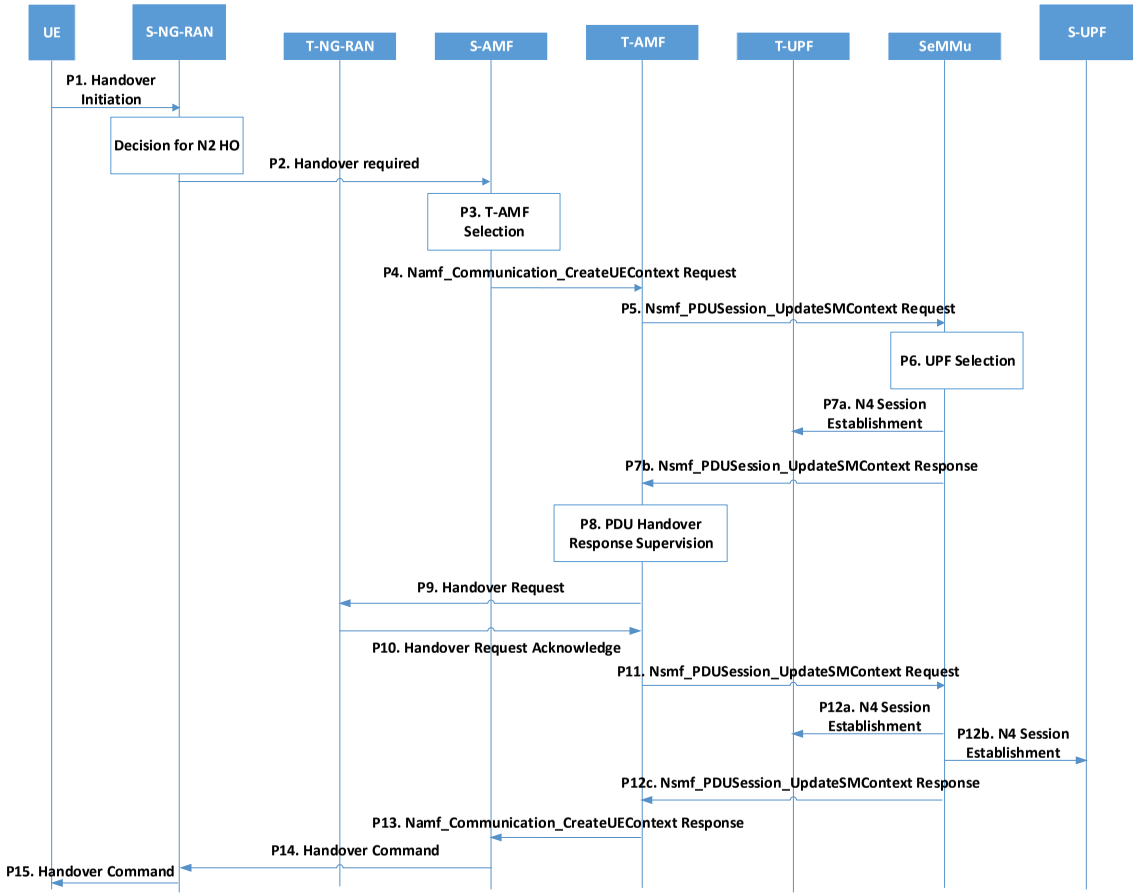


Figure A.17: Proposed Handover Signaling 5G Inter NG-RAN N2 based Handover.

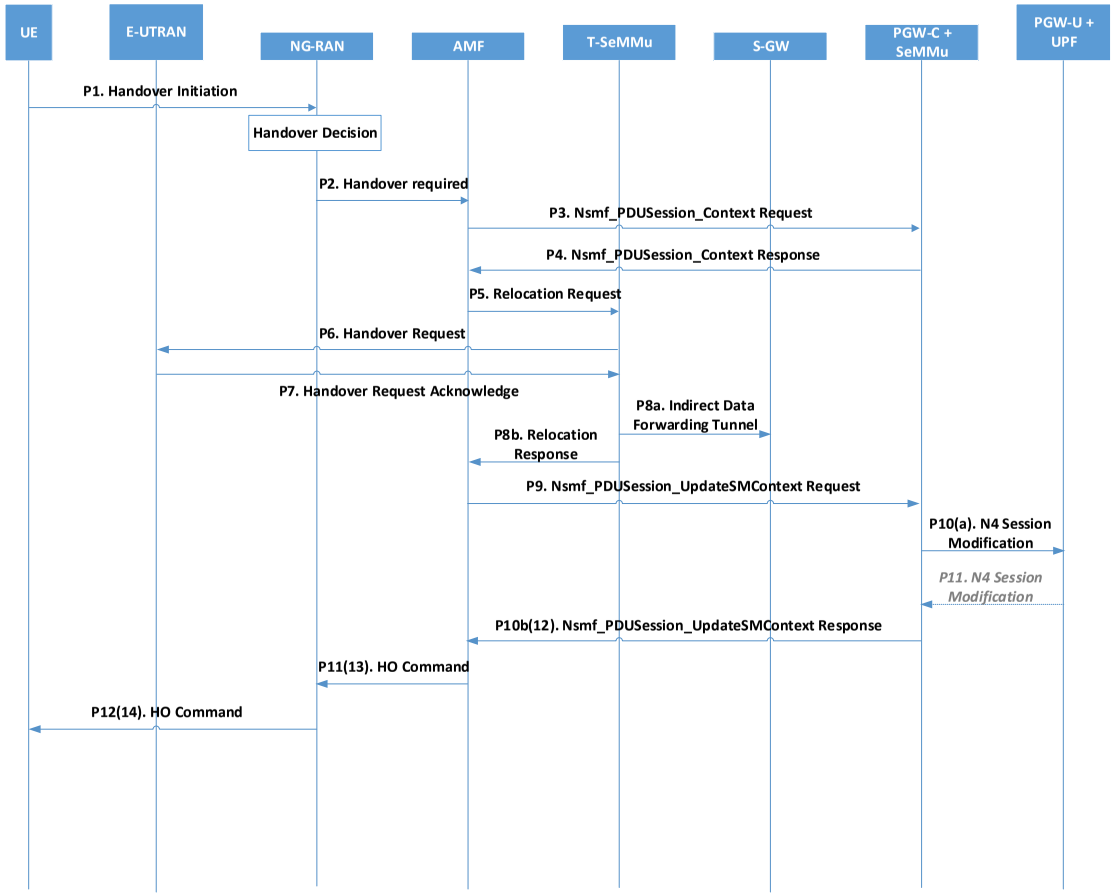


Figure A.18: Proposed Signaling for 5G core to EPS Handover with N26 Interface.

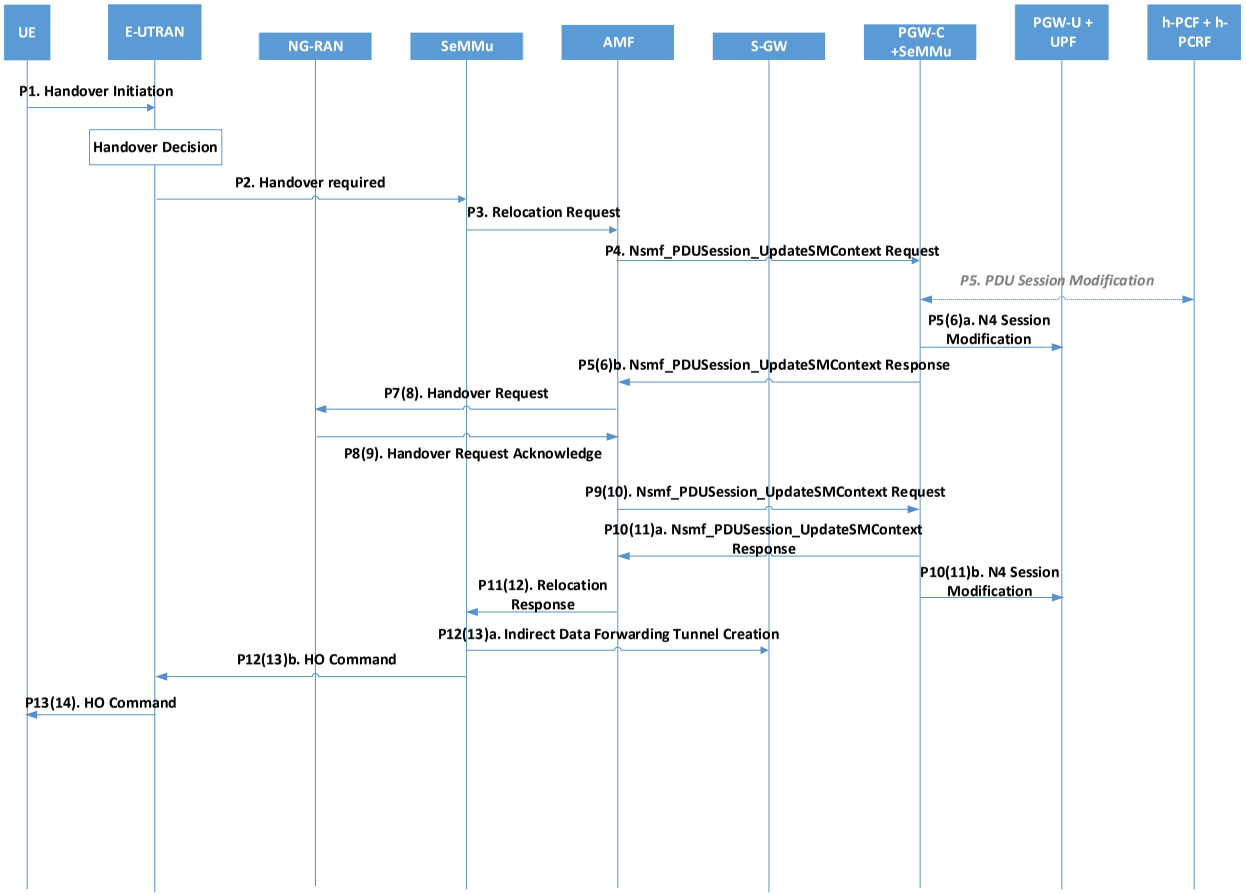


Figure A.19: Proposed Signaling for EPS to 5G Core Handover with N26 Interface.

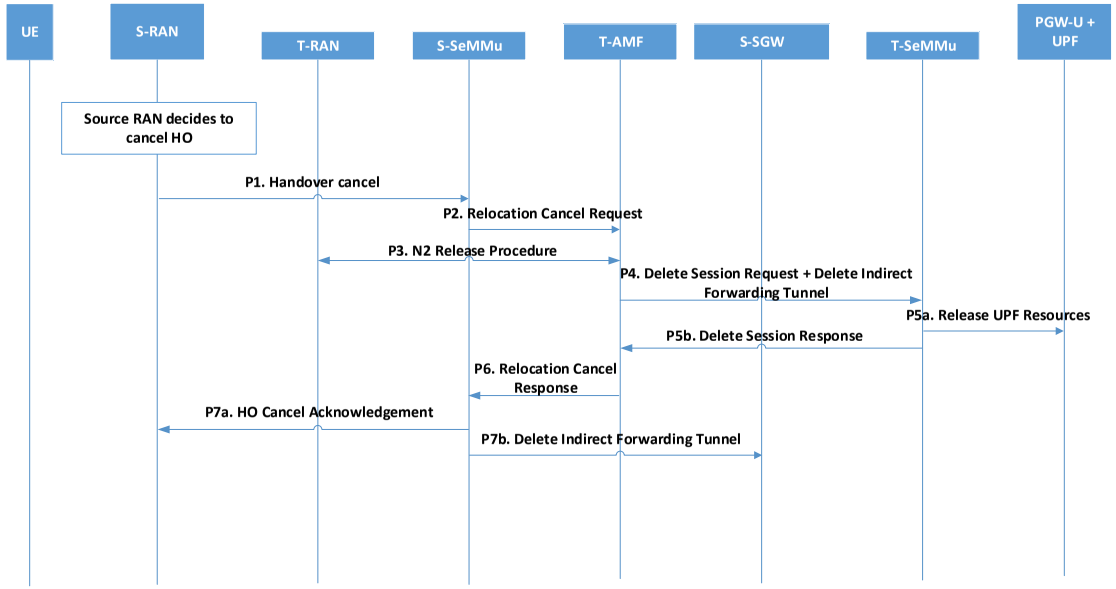


Figure A.20: Proposed Signaling for EPS to 5G Core Handover Cancel.

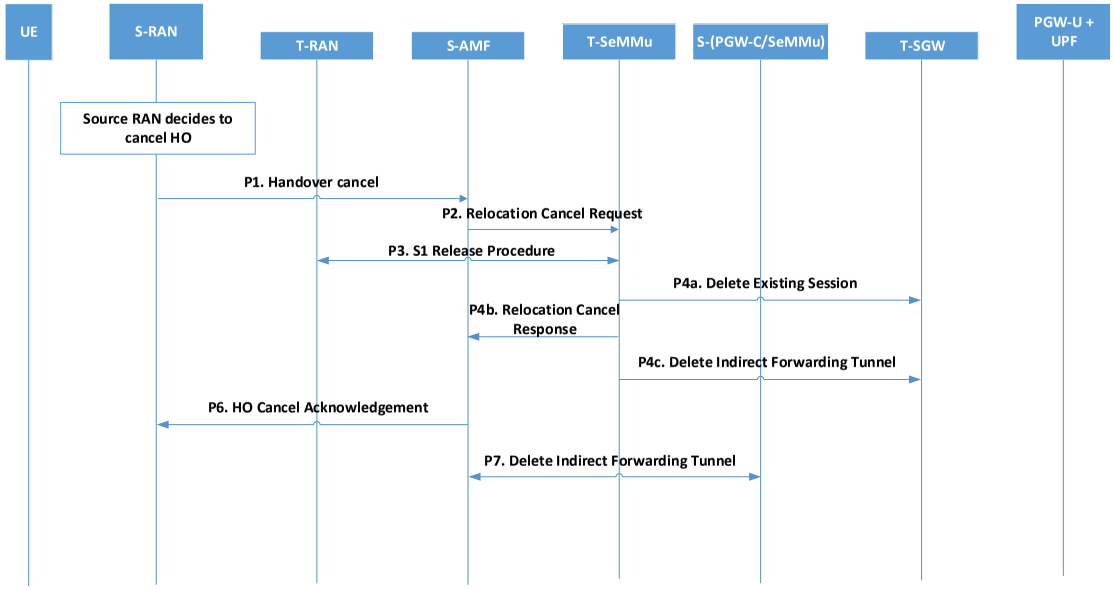


Figure A.21: Proposed Signaling for 5G Core to EPS Handover Cancel.

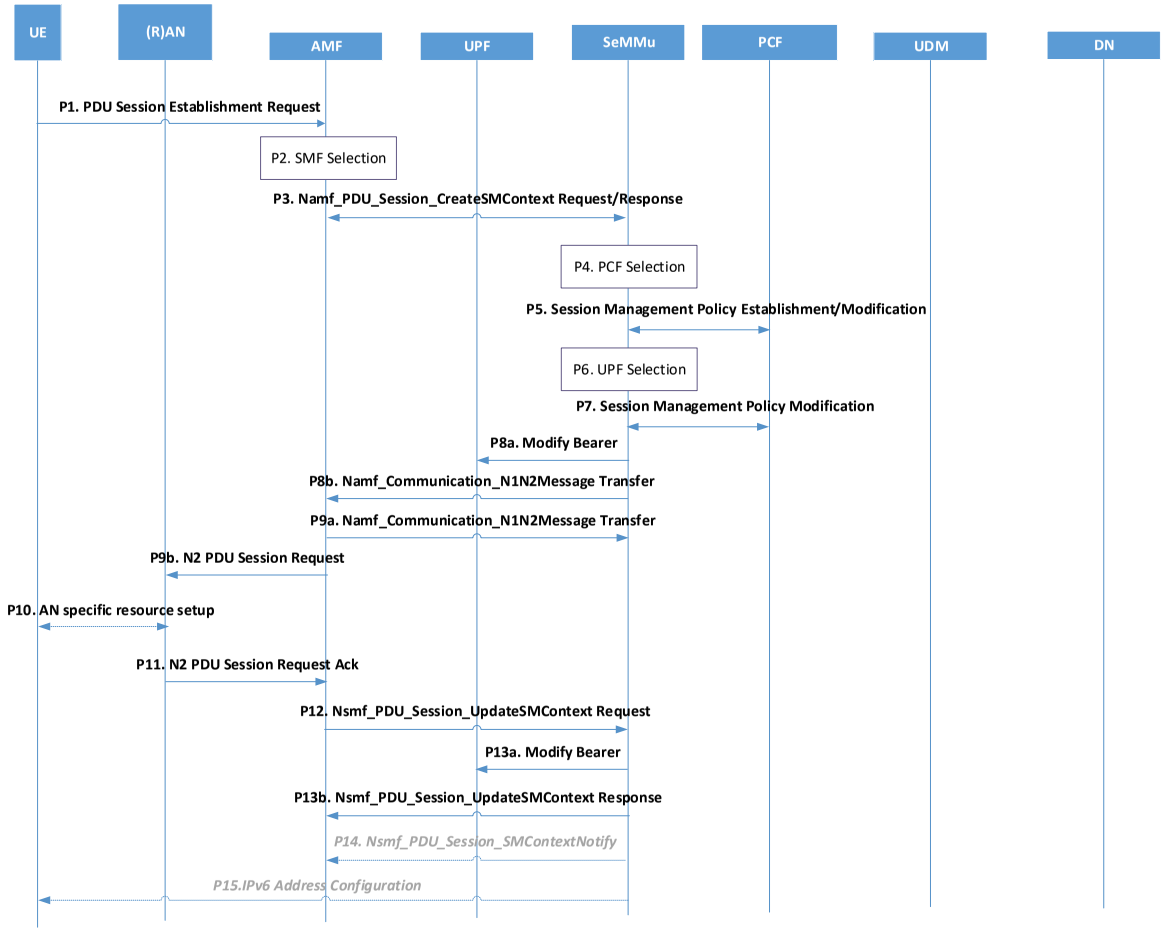


Figure A.22: Proposed Signaling for EPS to 5G Core Handover without N26 interface: PDU establishment.

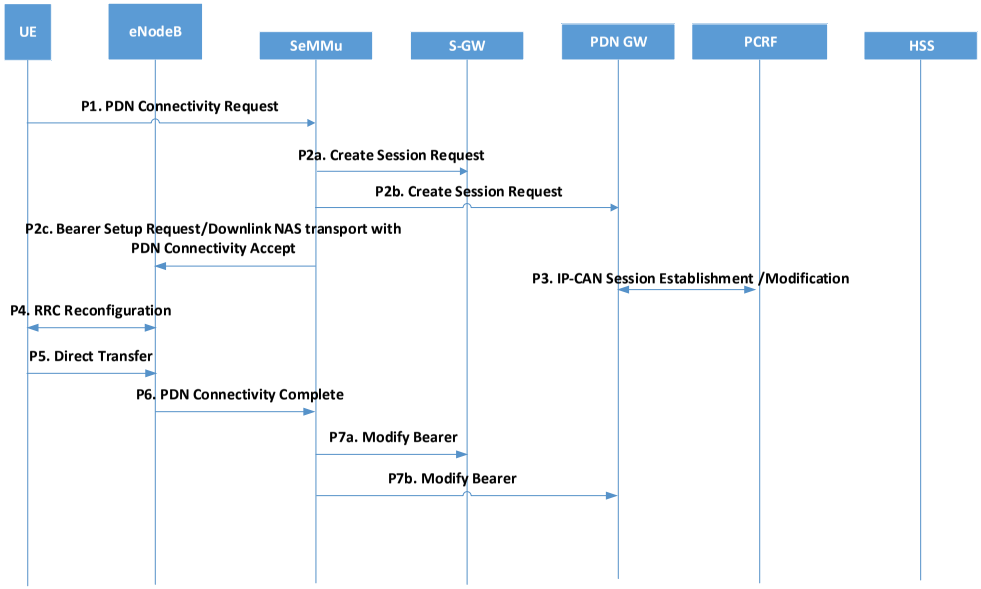


Figure A.23: Proposed Signaling for 5G Core to EPS Handover without N26 interface: UE requested Connectivity.

Bibliography

[1] G. Smail and J. Weijia, "Techno-economic analysis and prediction for the deployment of 5G mobile network," *Proceedings of the 2017 20th Conference on Innovations in Clouds, Internet and Networks, ICIN 2017*, no. 2015, pp. 9–16, 2017.

[2] PWC, "2017 Telecommunications trends," Tech. Rep., 2017. [Online]. Available: https://www.strategyand.pwc.com/media/file/2017-Telecommunications-Trends.pdf

[3] Ericsson, "Ericsson Mobility Report," Tech. Rep. February, 2016.

[4] Ericsson, "Ericsson Mobility Report," Tech. Rep. November, 2018.

[5] Ericsson, "Ericsson Mobility Report," Tech. Rep. June, 2018.

[6] Ericsson, "Ericsson Mobility Report," Tech. Rep. June, 2019.

[7] I. F. Akyildiz *et al.*, "5G roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17–48, 2016.

[8] ITU, "Setting the scene for 5G: Opportunities & Challenges," Tech. Rep., 2018.

[9] ITU, "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond, M Series, Recommendation ITU-R M.2083-0 (09/2015)," vol. 0, 2015. [Online]. Available: https://www.itu.int/dms{_}pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf

[10] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65–75, 2014.

[11] N. Saravanan, N. Sreenivasulu, D. Jayaram, and A. Chockalingam, "Design and Performance Evaluation of an Inter-System Handover Algorithm in UMTS/GSM Networks," in *TENCON 2005*, 11 2005, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/4085225/

[12] K. Sivanesan, J. Zou, S. Vasudevan, and S. Palat, "Mobility performance optimization for 3GPP LTE HetNets," in *Design and Deployment of Small Cell Networks*, A. Anpalagan, M. Bennis, and R. Vannithamby, Eds. Cambridge: Cambridge University Press, 2015, pp. 1–30. [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781107297333A009/type/book_part

[13] W. F. Elsadek and M. N. Mikhail, "Inter-domain Mobility Management Using SDN for Residential/Enterprise Real Time Services," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*. IEEE, 8 2016, pp. 43–50. [Online]. Available: http://ieeexplore.ieee.org/document/7592699/

[14] R. Ratasuk, N. Mangalvedhe, and A. Ghosh, "LTE in unlicensed spectrum using licensed-assisted access," in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 12 2014, pp. 746–751. [Online]. Available: http://ieeexplore.ieee.org/document/7063522/

[15] 3GPP and ETSI, "3GPP TS 36.300 version 13.2.0 Release 13 LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 Terrestrial Radio Access (E-UTRA) Overall description," Tech. Rep., 2013.

[16] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," *IEEE Access*, vol. 7, pp. 70 371–70 421, 2019.

[17] "SDN-based Mobility in iJOIN."

[18] L. Valtulina, M. Karimzadeh, G. Karagiannis, G. Heijenk, and A. Pras, "Performance evaluation of a SDN/OpenFlow-based Distributed Mobility Management (DMM) approach in virtualized LTE systems," in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 12 2014, pp. 18–23.

[19] A. Huang and N. Nikaein, "Demo: LL-MEC A SDN-based MEC Platform," *Mobicom 2017*, pp. 1–3, 2017.

[20] Y. Shi, H. Qu, J. Zhao, and G. Ren, "Downlink Dual Connectivity Approach in mmWave-Aided HetNets With Minimum Rate Requirements," *IEEE Communications Letters*, vol. 22, no. 7, pp. 1470–1473, 2018.

[21] R. Wang, H. Hu, and X. Yang, "Potentials and challenges of C-RAN supporting multi-RATs toward 5G mobile networks," *IEEE Access*, vol. 2, pp. 1200–1208, 2014.

[22] T.-T. Nguyen, C. Bonnet, and J. Harri, "SDN-based distributed mobility management for 5G networks," in *IEEE Wireless Communications and Networking Conference*, 2016, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/7565106/

[23] D. Battulga, J. Ankhzaya, B. Ankhbayar, U. Ganbayar, and S. Sh, "Handover Management for Distributed Mobility Management in SDN-based Mobile Networks," *27th International Telecommunication Networks and Applications Conference (ITNAC)*, 2017.

[24] H. Yang and Y. Kim, "SDN-based distributed mobility management," in *International Conference on Information Networking (ICOIN)*, 2016, pp. 337–342. [Online]. Available: http://ieeexplore.ieee.org/document/7427127/

[25] R. Urgaonkar, S. Wang, T. He, M. Zafer, K. Chan, and K. K. Leung, "Dynamic service migration and workload scheduling in edge-clouds," *Performance Evaluation*, vol. 91, pp. 205–228, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2015.06.013

[26] P. A. Frangoudis and A. Ksentini, "Service migration versus service replication in Multi-access Edge Computing," *14th International Wireless Communications and Mobile Computing Conference, IWCMC 2018*, pp. 124–129, 2018.

[27] R. Akremaddad, D. Leonelcadette Dutra, M. Bagaa, T. Taleb, and H. Flinck, "Towards a Fast Service Migration in 5G," *IEEE Conference on Standards for Communications and Networking, CSCN 2018*, 2018.

[28] 3GPP and ETSI, "TS 22.186 Enhancement of 3GPP support for V2X scenarios," Tech. Rep. Release 16, 2018.

[29] 3GPP, "TS 23.502: Procedures for the 5G System (Stage 2)," Tech. Rep. Release 15, 2017.

[30] A. Sutton, M. A. Imran, R. Tafazolli, A. Tukmanov, F. J. Lopez-Martinez, and M. Jaber, "Wireless Backhaul: Performance Modeling and Impact on User Association for 5G," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3095–3110, 2018.

[31] A. Machen, S. Wang, K. K. Leung, B. J. Ko, and T. Salonidis, "Live Service Migration in Mobile Edge Clouds," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 140–147, 2018.

[32] A. A. A. Boulogeorgos *et al.*, "Terahertz Technologies to Deliver Optical Network Quality of Experience in Wireless Systems beyond 5G," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 144–151, 2018.

[33] M. Z. Chowdhury, M. K. Hasan, M. Shahjalal, M. T. Hossan, and Y. Min Jang, "Optical Wireless Hybrid Networks for 5G and beyond Communications," *9th International Conference on Information and Communication Technology Convergence: ICT Convergence Powered by Smart Intelligence, ICTC 2018*, pp. 709–712, 2018.

[34] Z. Zhang, Y. XIAO, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Vehicular Technology Magazine*, 2019.

[35] E. Basar, M. Di Renzo, J. de Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless Communications Through Reconfigurable Intelligent Surfaces," pp. 1–20, June 2019. [Online]. Available: http://arxiv.org/abs/1906.09490

[36] M. D. Renzo *et al.*, "Smart radio environments empowered by reconfigurable AI metasurfaces: an idea whose time has come," *Eurasip Journal on Wireless Communications and Networking*, no. 1, 2019.

[37] Y. L. Chung, L. J. Jang, and Z. Tsai, "An efficient downlink packet scheduling algorithm in LTE-Advanced systems with Carrier Aggregation," *IEEE Consumer Communications and Networking Conference, CCNC'2011*, pp. 632–636, 2011.

[38] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys and Tutorials*, 2018.

[39] S. Sekander, H. Tabassum, and E. Hossain, "Multi-Tier Drone Architecture for 5G/B5G Cellular Networks: Challenges, Trends, and Prospects," *IEEE Communications Magazine*, vol. 56, no. 3, pp. 104–111, 2018.

[40] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1–18, 7 2015.

[41] S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S.-P. Yeh, and S. Talwar, "Intelligent access network selection in converged multi-radio heterogeneous networks," *IEEE Wireless Communications*, vol. 21, no. 6, pp. 86–96, 12 2014.

[42] P. Fan, J. Zhao, and C.-L. I, "5G high mobility wireless communications: Challenges and solutions," *China Communications*, vol. 13, no. 2, pp. 1–13, 2016.

[43] S. Ferretti, V. Ghini, and F. Panzieri, "A survey on handover management in mobility architectures," *Computer Networks*, 2016.

[44] M. Zekri, B. Jouaber, and D. Zeghlache, "A review on mobility management and vertical handover solutions over heterogeneous wireless networks," *Computer Communications*, no. 17, 2012.

[45] 3GPP, "5G; System architecture for the 5G System (5GS) (3GPP TS 23.501 version 15.8.0 Release 15)," pp. 1–251, 2020. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[46] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine*, no. 5, 2016.

[47] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Computer Networks*, 2016.

[48] S. E. Elayoubi, M. Fallgren, P. Spapis, G. Zimmermann, D. Martin-Sacristan, C. Yang, S. Jeux, P. Agyapong, L. Campoy, Y. Qi, and S. Singh, "5G service requirements and operational use cases: Analysis and METIS II vision," *EUCNC 2016 - European Conference on Networks and Communications*, pp. 158–162, 2016.

[49] W. Khawaja, I. Guvenc, D. W. Matolak, U.-C. Fiebig, and N. Schneckenberger, "A Survey of Air-to-Ground Propagation Channel Modeling for Unmanned Aerial Vehicles," *IEEE Communications Surveys & Tutorials*, 2019.

[50] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2241–2263, 2019.

[51] H. Wymeersch, G. Seco-Granados, G. Destino, D. Dardari, and F. Tufvesson, "5G mmwave positioning for vehicular networks," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 80–86, 2017.

[52] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G Backhaul Challenges and Emerging Research Directions: A Survey," *IEEE Access*, 2016.

[53] D. Liu and H. Chan, "RFC 7429 Distributed Mobility Management: Current Practices and Gap Analysis," pp. 1–34, 2015.

[54] C. Chen, Y.-T. Lin, L.-H. Yen, M.-C. Chan, and C.-C. Tseng, "Mobility management for low-latency handover in SDN-based enterprise networks," in *IEEE Wireless Communications and Networking Conference*, 2016.

[55] S. Wang, J. Xu, N. Zhang, and Y. Liu, "A Survey on Service Migration in Mobile Edge Computing," *IEEE Access*, vol. 6, pp. 23 511–23 528, 2018.

[56] T. Bai and R. W. Heath, "Coverage analysis for millimeter wave cellular networks with blockage effects," *IEEE Global Conference on Signal and Information Processing, GlobalSIP 2013 - Proceedings*, pp. 727–730, 2013.

[57] L. Zanzi and V. Sciancalepore, "On Guaranteeing End-to-End Network Slice Latency Constraints in 5G Networks," *Proceedings of the International Symposium on Wireless Communication Systems*, pp. 1–6, 2018.

[58] R. Molina-Masegosa and J. Gozalvez, "* LTE-V for Sidelink 5G V2X Vehicular Communications," *IEEE Vehicular Technology Magazine*, vol. 12, no. 4, pp. 30–39, 2017.

[59] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*, 1st ed. Elsevier Ltd., 2007.

[60] Ericsson, "Ericsson Mobility Report," Tech. Rep., 2016. [Online]. Available: https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf

[61] Alcatel Lucent, "The LTE Network Architecture," 2009. [Online]. Available: http://www.cse.unt.edu/~rdantu/FALL_2013_WIRELESS_NETWORKS/LTE_Alcatel_White_Paper.pdf

[62] E. Dahlman, S. Parkvall, and J. Sköld, *4G: LTE Advanced Pro and the Road to 5G*, 3rd ed. Academic Press, 2016.

[63] D. P. Ibarra Barreno, "LTE / WIFI AGGREGATION IMPLEMENTATION AND EVALUATION," Ph.D. dissertation, UPC, 2017.

[64] 3GPP, "TS 136 361 - V14.1.0 - LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE-WLAN Radio Level Integration Using Ipsec Tunnel (LWIP) encapsulation; Protocol specification (3GPP TS 36.361 version 14.1.0 Release 14)," pp. 1–12, 2018.

[65] K. Samdanis, T. Taleb, and S. Schmid, "Traffic offload enhancements for eUTRAN," *IEEE Communications Surveys and Tutorials*, vol. 14, no. 3, pp. 884–896, 2012.

[66] C. B. Sankaran, "Data offloading techniques in 3GPP Rel-10 networks: A tutorial," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 46–53, 2012.

[67] ITU-T, "Framework of vertical multihoming in IPv6-based next generation networks," 2011.

[68] R. Irmer, H. Droste, P. Marsch, G. P. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *Commun. Mag.*, no. February, pp. 102–112, 2011. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=5706317

[69] C. Perkins, "RFC 6275 Mobility support in IPv6," pp. 1–169, 2011.

[70] R. Koodli, "RFC 4068 Fast Handovers for Mobile IPv6," pp. 1–42, 2005.

[71] H. Soliman, C. Castelluccia, K. Elmalki, and L. Bellier, "RFC 5380 HMIPv6," pp. 1–25, 2008.

[72] K. Leung, "RFC 5213 Proxy Mobile IPv6," pp. 1–92, 2008.

[73] S. Gundavelli *et al.*, "Proxy Mobile IPv6," *RFC 5213*, pp. 1–92, 2008.

[74] C. Bernados, "Proxy Mobile IPv6 Extensions to Support Flow Mobility," *RFC 7864*, pp. 1–19, 2016.

[75] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Proxy Mobile IPv6 (PMIPv6) based Mobility and Tunnelling protocols; Stage 3 (3GPP TS 29.275 version 8.6.0 Release 8)," pp. 1–73, 2010.

[76] H. N. Nguyen and C. Bonnet, "Scalable proxy mobile IPv6 For heterogeneous wireless networks," *Proceedings of the International Conference on Mobile Technology, Applications, and Systems, Mobility'08*, 2008.

[77] F. Giust, L. Cominardi, and C. Bernardos, "Distributed mobility management for future 5G networks: overview and analysis of existing approaches," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 142–149, 1 2015.

[78] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, "MPTCP RFC 6182," pp. 1–28, 2011.

[79] A. Ford *et al.*, "TCP Extensions for Multipath Operation with Multiple Addresses," *RFC 6824*, pp. 1–64, 2013.

[80] A. Ravanshid *et al.*, "Multi-connectivity functional architectures in 5G," in *IEEE International Conference on Communications Workshops (ICC)*, 2016.

[81] T. Klein, "Enhancements to Improve the Applicability of Multipath TCP to Wireless Access Networks," *IETF (draft)*, no. c, pp. 1–25, 2011.

[82] S. Zannettou, M. Sirivianos, and F. Papadopoulos, "Exploiting path diversity in datacenters using MPTCP-aware SDN," *Proceedings - IEEE Symposium on Computers and Communications*, pp. 539–546, 2016.

[83] C. D. Phung *et al.*, "MPTCP robustness against large-scale man-in-the-middle attacks," *Computer Networks*, vol. 164, p. 106896, 2019. [Online]. Available: https://doi.org/10.1016/j.comnet.2019.106896

[84] Y. Liu, A. Neri, A. Ruggeri, and A. M. Vegni, "A MPTCP-Based Network Architecture for Intelligent Train Control and Traffic Management Operations," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2290–2302, 2017.

[85] P. Natarajan, F. Baker, C. Systems, P. D. Amer, and J. T. Leighton, "SCTP : What , Why , and How," *IEEE Internet Comput.*, vol. 13, no. 5, pp. 81–85, 2009.

[86] C. Raiciu, M. Handly, and D. Wischik, "Coupled Congestion Controol for Multipath Transport Protocols," *IETF RFC6356*, pp. 1–12, 2011.

[87] D. Wischik *et al.*, "Design, implementation and evaluation of congestion control for multipath TCP," *Proceedings of NSDI 2011: 8th USENIX Symposium on Networked Systems Design and Implementation*, pp. 99–112, 2011.

[88] P. Ignaciuk and M. Morawski, "Discrete-time MPTCP flow control for channels with diverse delays and uncertain capacity," *2018 22nd International Conference on System Theory, Control and Computing, ICSTCC 2018 - Proceedings*, pp. 722–727, 2018.

[89] X. Wei, C. Xiong, and E. Lopez, "MPTCP proxy mechanisms," Internet Engineering Task Force, Internet-Draft draft-wei-mptcp-proxy-mechanism-02, Jun. 2015, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-wei-mptcp-proxy-mechanism-02

[90] R. Stewart, "Stream Control Transmission Protocol," *RFC 4960*, pp. 1–152, 2007.

[91] A. De La Oliva, A. Banchs, I. Soto, T. Melia, and A. Vidal, "An overview of IEEE 802.21: Media-independent handover services," pp. 96–103, 2008.

[92] L. Eastwood *et al.*, "Mobility Using IEEE 802.21 in a Heterogeneous IEEE 802.16/802.11-based, IMT-Advanced (4G) Network," *IEEE Wireless Communications*, no. Apr., pp. 26–34, 2008.

[93] IEEE, *IEEE 802.21c-2014: IEEE Standard for Local and metropolitan area networks — Part 21 : Media Independent Handover Services Amendment 3: Optimized Single Radio Handovers*, 2014.

[94] J.-S. Wu, S.-F. Yang, and B.-J. Hwang, "A terminal-controlled vertical handover decision scheme in IEEE 802.21-enabled heterogeneous wireless networks Jung-Shyr," *Int. J. Commun. Syst.*, vol. 22, pp. 819–834, 2009.

[95] R. Qureshi, A. Dadej, and Q. Fu, "Issues in 802.21 mobile node controlled handovers," in *Australasian Telecommunication Networks and Applications Conference*, 2007.

[96] 3GPP, "3gpp TS 36.331 – 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 10) The," no. June, 2011.

[97] D. Xenakis *et al.*, "Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 64–91, 2014.

[98] J. Zhao, "A Survey of Reconfigurable Intelligent Surfaces: Towards 6G Wireless Communication Networks with Massive MIMO 2.0," pp. 1–7, 2019. [Online]. Available: http://arxiv.org/abs/1907.04789

[99] 3GPP, "5G; Procedures for the 5G System (5GS) (3GPP TS 23.502 version 15.8.0 Release 15)," pp. 1–362, 2020.

[100] 3GPP, "5G NR; Overall description; Stage-2 (3GPP TS 38.300 version 15.8.0 Release 15)," vol. 1, pp. 1–102, 2020. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[101] ETSI and 3GPP, "ETSI TS 137 340 v15.5.0," Tech. Rep., 2019.

[102] S. Jung and J. Kim, "A new way of extending network coverage: Relay-assisted D2D communications in 3GPP," *ICT Express*, vol. 2, no. 3, pp. 117–121, 9 2016.

[103] M. Kantor, R. State, T. Engel, and G. Ormazabal, "A policy-based per-flow mobility management system design," in *Proceedings of the Principles, Systems and Applications on IP Telecommunications - IPTComm '15*. New York, New York, USA: ACM Press, 2015, pp. 35–42. [Online]. Available: http://doi.acm.org/10.1145/2843491.2843835http://dl.acm.org/citation.cfm?doid=2843491.2843835

[104] M. Gramaglia, A. Banchs, V. Sciancalepore, Z. Yousaf, C. Mannweiler, L. Yu, B. Sayadi, M.-L. Alberi Morel, R. L. Silva, M. R. Crippa, D. V. Hugo, P. Arnold, V. Frederikos, and I. L. Pavon, "Definition of connectivity and QoE / QoS management mechanisms – 5G Norma deliverable D5.1," 2016.

[105] G. Schütz, "A k-Cover Model for Reliability-Aware Controller Placement in Software-Defined Networks," in *Computational Science – ICCS 2019*. Springer International Publishing, 2019, pp. 604–613.

[106] S. Kuklinski, Y. Li, and K. T. Dinh, "Handover management in SDN-based mobile networks," in *IEEE Globecom Workshops (GC Wkshps)*, 2014.

[107] F. Meneses, C. Guimares, D. Corujo, and R. L. Aguiar, "SDN-based Mobility Management: Handover Performance Impact in Constrained Devices," in *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 2 2018, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/8328716/

[108] T. D. Assefa *et al.*, "SDN-based local mobility management with X2-interface in femtocell networks," *IEEE Int. Work. Comput. Aided Model. Des. Commun. Links Networks, CAMAD*, pp. 3–8, 2017.

[109] S. Basloom and N. Akkari, "Mobility Management in SDN and NFV-based Next-Generation Wireless Networks : An Overview and Qualitative Evaluation," *2018 1st International Conference on Advanced Research in Engineering Sciences (ARES)*, pp. 1–8.

[110] I. Elgendi, K. S. Munasinghe, and A. Jamalipour, "A Three-Tier SDN based distributed mobility management architecture for DenseNets," *2016 IEEE International Conference on Communications, ICC 2016*, 2016.

[111] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "Edge Cloud and Underlay Networks: Empowering 5G Cell-Less Wireless Architecture," in *Proceedings of European Wireless 2014, 20th European Wireless Conference*, 2014.

[112] ETSI, "Mobile Edge Computing (MEC); End to End Mobility Aspects," Tech. Rep., 2017.

[113] A. Mtibaa, R. Tourani, S. Misra, J. Burke, and L. Zhang, "Towards edge computing over named data networking," *Proceedings - 2018 IEEE International Conference on Edge Computing, EDGE 2018 - Part of the 2018 IEEE World Congress on Services*, 2018.

[114] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," 2017. [Online]. Available: http://arxiv.org/abs/1702.05309%0Ahttp://dx.doi.org/10.1109/COMST.2017.2682318

[115] ETSI, "3GPP TS 138 331 - V15.2.1 - 5G; NR; Radio Resource Control (RRC)," vol. 1, 2018.

[116] T. Nakamura, S. Nagata, A. Benjebbour, Y. Kishiyama, T. Hai, S. Xiaodong, Y. Ning, and L. Nan, "Trends in small cell enhancements in LTE advanced," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 98–105, 2 2013.

[117] F. A. A. Emam, M. E. Nasr, and S. E. Kishk, "Coordinated Handover Signaling and Cross-Layer Adaptation in Heterogeneous Wireless Networking," *Mob. Networks Appl.*, vol. 25, pp. 285–299, 2020.

[118] F. A. A. Emam, M. E. Nasr, and S. E. Kishk, "Context-aware parallel handover optimization in heterogeneous wireless networks," *Ann. Telecommun.*, vol. 75, pp. 43–57, 2020.

[119] A. Al-rubaye and J. Seitz, "A Cross-Layer Mobility Management with Multi-Criteria Decision Making," *2016 Eighth Int. Conf. Ubiquitous Futur. Networks*, pp. 821–826, 2016.

[120] N. Nikaein *et al.*, "Demo – Closer to Cloud-RAN : RAN as a Service," *ACM Mobicom*, pp. 193–195, 2015.

[121] A. Outtagarts *et al.*, "When IT meets Telco : RAN as a Service," *2015 IEEE/ACM 8th Int. Conf. Util. Cloud Comput.*, pp. 422–423, 2015.

[122] D. Sabella *et al.*, "RAN as a Service: Challenges of Designing a Flexible RAN Architecture in a Cloud-based Heterogeneous Mobile Network," *2013 Futur. Netw. Mob. Summit*, pp. 1–8, 2013.

[123] L. Liu *et al.*, "Analysis of Handover Performance Improvement in Cloud-RAN Architecture," *7th Int. Conf. Commun. Netw. China*, pp. 850–855, 2012.

[124] V. Passast *et al.*, "Dynamic RAT Selection and Pricing for Efficient Traffic Allocation in 5G HetNets," *IEEE ICC*, pp. 1–6, 2019.

[125] S. Goudarzi *et al.*, "A hybrid intelligent model for network selection in the industrial Internet of Things," *Appl. Soft Comput. J.*, vol. 74, pp. 529–546, 2019.

[126] J. Wang, J. Weitzen, O. Bayat, V. Sevindik, and M. Li, "Interference coordination for millimeter wave communications in 5G networks for performance optimization," *Eurasip Journal on Wireless Communications and Networking*, vol. 2019, no. 1, 2019.

[127] D. Calabuig, S. Barmpounakis, S. Gimenez, A. Kousaridas, T. R. Lakshmana, J. Lorca, P. Lunden, Z. Ren, P. Sroka, E. Ternon, V. Venkatasubramanian, and M. Maternia, "Resource and Mobility Management in the Network Layer of 5G Cellular Ultra-Dense Networks," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 162–169, 2017.

[128] O. N. C. Yilmaz *et al.*, "Smart mobility management for D2D communications in 5G networks," *2014 IEEE Wirel. Commun. Netw. Conf. Work. WCNCW 2014*, pp. 219–223, 2014.

[129] K. Ouali and B. Kervella, "An Efficient D2D Handover Management Scheme for SDN-based 5G networks," *2020 IEEE 17th Annu. Consum. Commun. Netw. Conf.*, pp. 1–6, 2020.

[130] R. Klempous and J. Nikodem, *Smart Innovations in Engineering and Technology*, 2020, vol. 15. [Online]. Available: http://link.springer.com/10.1007/978-3-030-32861-0

[131] S. Barua and R. Braun, "Mobility management of D2D communication for the 5G cellular network system: A study and result," *2017 17th Int. Symp. Commun. Inf. Technol. Isc. 2017*, pp. 1–6, 2017.

[132] S. Barua and R. Braun, "A novel approach of mobility management for the D2D communications in 5G mobile cellular network system," in *2016 18th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 10 2016, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/document/7737272/

[133] 3GPP and ETSI, "ETSI TS 123 401," Tech. Rep., 2015.

[134] S. Oh, B. Ryu, and Y. Shin, "EPC signaling load impact over S1 and X2 handover on LTE-Advanced system," *3rd World Congress on Information and Communication Technologies, WICT 2013*, pp. 183–188, 2013.

[135] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE Release 12 and Beyond David," *Ieee Wirel. Commun. Mag.*, no. July, pp. 154–160, 2013.

[136] W. Sun and J. Liu, "Coordinated multipoint-based uplink transmission in internet of things powered by energy harvesting," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2585–2595, 2018.

[137] J. Lee, Y. Kim, H. Lee, B. Ng, D. Mazzarese, J. Liu, W. Xiao, and Y. Zhou, "Coordinated multipoint transmission and reception in LTE-advanced systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 44–50, 2012.

[138] C. Shen and M. Van Der Schaar, "A learning approach to frequent handover mitigations in 3GPP mobility protocols," *IEEE Wirel. Commun. Netw. Conf. WCNC*, pp. 1–6, 2017.

[139] A. Ahmed, L. M. Boulahia, and D. Gaïti, "Enabling vertical handover decisions in heterogeneous wireless networks: A state-of-the-art and a classification," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 2, pp. 776–811, 2014.

[140] A. Santoyo-González and C. Cervelló-Pastor, "Latency-aware cost optimization of the service infrastructure placement in 5G networks," *Journal of Network and Computer Applications*, vol. 114, no. February, pp. 29–37, 2018. [Online]. Available: https://doi.org/10.1016/j.jnca.2018.04.007

[141] I. Leyva-Pupo, A. Santoyo-González, and C. Cervelló-Pastor, "A framework for the joint placement of edge service infrastructure and user plane functions for 5G," *Sensors (Switzerland)*, vol. 19, no. 18, 2019.

[142] R. Alkhansa, H. Artail, and D. M. Gutierrez-Estevez, "LTE-WiFi carrier aggregation for future 5G systems: A feasibility study and research challenges," *Procedia Computer Science*, vol. 34, pp. 133–140, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.procs.2014.07.068

[143] M. A. Ferrag, L. Maglaras, A. Argyriou, D. Kosmanos, and H. Janicke, "Security for 4G and 5G Cellular Networks: A Survey of Existing Authentication

and Privacy-preserving Schemes," pp. 1–24, 2017. [Online]. Available: http://arxiv.org/abs/1708.04027

[144] M. Jawad Alam and M. Ma, "DC and CoMP Authentication in LTE-Advanced 5G HetNet," *IEEE Global Communications Conference, GLOBECOM 2017 - Proceedings*, pp. 1–6, 2018.

[145] G. Qiao, S. Leng, K. Zhang, and K. Yang, "Joint Deployment and Mobility Management of Energy Harvesting Small Cells in Heterogeneous Networks," *IEEE Access*, vol. 5, pp. 183–196, 2017.

[146] A. Habbal, S. Goudar, and S. Hassan, "Context-aware Radio Access Technology Selection Approach in 5G Ultra Dense Networks," *IEEE Access*, 2017.

[147] 3GPP, "TS22.261: Service requirements for the 5G system (Stage 1)," 2018.

[148] A. Sadeghian, L. Sundaram, D. Z. Wang, W. F. Hamilton, K. Branting, and C. Pfeifer, "Semantic Edge Labeling over Legal Citation Graphs," in *LTDCA*, 2018.

[149] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, and W. Xu, "Use Cases, Requirements, and Design Considerations for 5G V2X," pp. 1–10, 2017. [Online]. Available: http://arxiv.org/abs/1712.01754

[150] G. Report, "Study on MEC Support for V2X Use Cases," vol. 1, pp. 1–19, 2018.

[151] G. A. Zhang, J. Y. Gu, Z. H. Bao, C. Xu, and S. B. Zhang, "Efficient Signal Detection for Cognitive Radio Relay Networks Under Imperfect Channel Estimation," *European Transactions on Telecommunications*, vol. 25, no. 3, pp. 294–307, 2014.

[152] D. Raychaudhuri, K. Nagaraja, N. Brunswick, and A. Venkataramani, "MobilityFirst : A Robust and Trustworthy Mobility- Centric Architecture for the Future Internet," *ACM SIGMobile Mobile Computing and Communication Review (MC2R)*, pp. 1–12, 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.2259&rep=rep1&type=pdf

[153] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 44–53, 2013.

[154] P. Marsch *et al.*, "5G RAN Architecture and Functional Design," 2016. [Online]. Available: https://metis-ii.5g-ppp.eu/documents/white-papers/

[155] A. Kaloxylos, I. Modeas, F. Georgiadis, and N. Passas, "Network Selection Algorithm for Heterogeneous Wireless Networks : from Design to Implementation," *Network*, vol. 1, no. 2, pp. 27–47, 2009.

[156] K. Alexandris, N. Nikaein, R. Knopp, and C. Bonnet, "Analyzing X2 handover in LTE/LTE-A," in *2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2016.

[157] K. Dimou *et al.*, "Handover within 3GPP LTE: Design Principles and Performance," in *2009 IEEE 70th Vehicular Technology Conference Fall*, 2009.

[158] A. F. Almutairi, M. A. Landolsi, and A. O. Al-Hawaj, "Weighting Selection in GRA-based MADM for Vertical Handover in Wireless Networks," in *UKSim-AMSS 18th International Conference on Computer Modelling and Simulation (UKSim)*, 2016.

[159] S. Barmpounakis, A. Kaloxylos, P. Spapis, and N. Alonistioti, "Context-aware, user-driven, network-controlled RAT selection for 5G networks," *Computer Networks*, vol. 113, pp. 124–147, 2 2017.

[160] M. Lahby, L. Cherkaoui, and A. Adib, "An enhanced-TOPSIS based network selection technique for next generation wireless networks," in *ICT 2013*, 2013.

[161] D. Singhal, M. Kunapareddy, V. Chetlapalli, V. B. James, and N. Akhtar, "LTE-advanced: Handover interruption time analysis for IMT-A evaluation," *International Conference on Signal Processing, Communication, Computing and Networking Technologies, ICSCCN*, 2011.

[162] Q. Wang, S. Zhao, and C. Hou, "UE assisted mobility management based on SDN," in *11th International Conference on Computer Science & Education (ICCSE)*, 2016.

[163] L. Wang, Z. Lu, X. Wen, G. Cao, X. Xia, and L. Ma, "An SDN-based seamless convergence approach of WLAN and LTE networks," in *IEEE Information Technology, Networking, Electronic and Automation Control Conference*, 2016.

[164] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; 3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control plane (GTPv2-C); Stage 3 (3GPP TS 29.274 version 12.6.0 Release 12)," pp. 1–316, 2014.

[165] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (3GPP TS 24.301 version 10.3.0 Release 10)," 2011.

[166] 3GPP, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Mobile radio interface Layer 3 specification; Core network protocols; Stage 3 (3GPP TS 24.008 version 8.6.0 Release 8)," 2009.

[167] 3GPP, "Universal Mobile Telecommunications System (UMTS); Radio Resource Control (RRC); Protocol specification (3GPP TS 25.331 version 13.1.0 Release 13)," 2016.

[168] 3GPP, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA);Radio ReResource Control (RRC); Protrotocol specification(3GPP TS 36.3 .331 version 13.0.0 Release 13)," 2016.

[169] 3GPP, "Universal Mobile Telecommunications System (UMTS); UTRAN Iu Interface RANAP Signalling (3GPP TS 25.413 version 3.4.0 Release 1999)," 2000.

[170] 3GPP, "Universal Mobile Telecommunications System (UMTS); LTE; Evolved Packet System (EPS); 3GPP Sv interface (MME to MSC, and SGSN to MSC) for SRVCC (3GPP TS 29.280 version 8.3.0 Release 8)," 2010.

[171] 3GPP, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Numbering, addressing and identification (3GPP TS 23.003 version 10.2.0 Release 10)," 2011.

[172] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Study on Architecture for Next Generation System (Release 14)," Tech. Rep., 2016.

[173] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NR; NR and NG-RAN Overall Description; Stage 2 (Release 15)," 2017.

[174] V. S. Rao and R. Gajula, "Interoperable UE Handovers in LTE," pp. 1–11, 2011.

[175] S. Chourasia and K. M. Sivalingam, "SDN based Evolved Packet Core architecture for efficient user mobility support," in *Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft)*, 2015.

[176] S. Abe, G. Hasegawa, and M. Murata, "Design and performance evaluation of bearer aggregation method in mobile core network with C/U plane separation," in *IFIP Networking Conference (IFIP Networking) and Workshops*, 2017.

[177] Z. Savic, "LTE Design and Deployment Strategies," *Cisco*, pp. 1–79, 2011.

[178] 3GPP, "Domain Name System Procedures TS 29.303," 2008.

[179] 3GPP, "LTE; Evolved Universal T l Terrestrial Radio Access (E- UTRA); Packet Data Converg rgence Protocol (PDCP) spec ecification (3GPP TS 36.3 .323 version 12.4.0 Release 12 12)," pp. 1–35, 2015.

[180] 3GPP, "Digital cellular telecomm mmunications system (Phase e 2+) (GSM); Mobile radio in interface layer 3 specification ion; Radio Resou ource Control (RRC) protocol ol (3GPP TS 44.0 .018 version 13.1.0 Release 13 13) GLOBAL," 2016.

[181] ng4t GmBH, "NG4T project wireshark traces." [Online]. Available: http://www.ng4t.com/wireshark.html

[182] O. Dubuisson, *ASN.1 Communication between Heterogeneous Systems Olivier Dubuisson*, 2000.

[183] J. Costa-Requena, J. Llorente Santos, J. Manner, and R. Kantola, "Enhanced Mobility Management," 2015.

[184] A. Jain, Sadagopan N S, S. K. Lohani, and M. Vutukuru, "A comparison of SDN and NFV for re-designing the LTE Packet Core," in *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2016, pp. 74–80.

[185] A. Mahmoud, A. Abo Naser, M. Abu-Amara, T. Sheltami, and N. Nasser, "Software-defined networking approach for enhanced evolved packet core network," *International Journal of Communication Systems*, vol. 31, no. 1, pp. 1–15, 2018.

[186] 3GPP, "TS22.261: Service requirements for the 5G system (Stage 1)," 2018.

[187] Michał Maternia (Nokia) and S. E. E. A. (Orange), "5G PPP use cases and performance evaluation models," *5G-PPP Initiative*, 2016. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf%0Ahttp://www.5g-ppp.eu/

[188] F. H. Khan and M. Portmann, "Joint QoS-control and handover optimization in backhaul aware SDN-based LTE networks," *Wireless Networks*, vol. 7, 2019. [Online]. Available: https://doi.org/10.1007/s11276-019-02021-7

[189] H. Kim, G. De Veciana, X. Yang, and M. Venkatachalam, "Distributed-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, 2012.

[190] S. Kapoor, D. Grace, and T. Clarke, "A base station selection scheme for handover in a mobility-aware ultra-dense small cell urban vehicular environment," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, vol. 2017-Octob, pp. 1–5, 2018.

[191] M. Q. Khan and S. H. Andresen, "A semi and fully distributed handover algorithm for heterogeneous networks using MIIS," *Proceedings - IEEE Symposium on Computers and Communications*, pp. 000 145–000 150, 2012.

[192] Y. Shi, H. Qu, and J. Zhao, "Dual Connectivity Enabled User Association Approach for Max-Throughput in the Downlink Heterogeneous Network," *Wireless Personal Communications*, vol. 96, no. 1, pp. 529–542, 2017.

[193] N. Prasad and S. Rangarajan, "Exploiting Dual Connectivity in Heterogeneous Cellular Networks."

[194] A. N. Manjeshwar, P. Jha, A. Karandikar, and P. Chaporkar, "Enhanced UE Slice Mobility for 5G Multi-RAT Networks."

[195] S. Kan, H. Chen, Y. Zhu, and W. Li, "Aggregation based Cell Selection Methods for multi-RAT HetNet," in *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, 10 2016, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/7752572/

[196] M. Alkhawlani and A. Ayesh, "Access Network Selection based on Fuzzy Logic and Genetic Algorithms," *Advanced Artificial Intelligence (AAI)*, vol. 1, no. 1, pp. 1–12, 2008. [Online]. Available: http://www.hindawi.com/journals/aai/2008/793058/

[197] K. Radhika and A. V. G. Reddy, "Network Selection in Heterogeneous Wireless Networks Based on Fuzzy Multiple Criteria Decision Making," *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, pp. 136–139, 2011.

[198] M. Mansouri and C. Leghris, "A comparison between fuzzy TOPSIS and fuzzy GRA for the vertical handover decision making," in *2017 Intelligent Systems and Computer Vision (ISCV)*. IEEE, 4 2017, pp. 1–6.

[199] M. Mansouri and C. Leghris, "A battery level aware MADM combination for the vertical handover decision making," *2017 13th International Wireless Communications and Mobile Computing Conference, IWCMC 2017*, pp. 1448–1452, 2017.

[200] A. Bazrafkan and M. R. Pakravan, "An MADM network selection approach for next generation heterogeneous networks," in *2017 Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 5 2017, pp. 1884–1890.

[201] M. Mansouri and C. Leghris, "The use of MADM methods in the vertical handover decision making context," *Proceedings - 2017 International Conference on Wireless Networks and Mobile Communications, WINCOM 2017*, 2017.

[202] S. Radouche, C. Leghris, and A. Adib, "MADM methods based on utility function and reputation for access network selection in a multi-access mobile network environment," in *2017 International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 11 2017, pp. 1–6.

[203] S. Liang, Y. Zhang, B. Fan, and H. Tian, "Multi-Attribute Vertical Handover Decision-Making Algorithm in a Hybrid VLC-Femto System," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1521–1524, 2017.

[204] L. Tie and P. G. Bai, "A New Multi-Attribute Decision-Making Vertical Handover Algorithm," no. 800, pp. 1–4.

[205] Y. Liu, L. Lu, G. Y. Li, Q. Cui, and W. Han, "Joint User Association and Spectrum Allocation for Small Cell Networks With Wireless Backhauls," *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 496–499, 2016.

[206] S. Bayat, R. H. Louie, Z. Han, B. Vucetic, and Y. Li, "Distributed user association and femtocell allocation in heterogeneous wireless networks," *IEEE Transactions on Communications*, vol. 62, no. 8, pp. 3027–3043, 2014.

[207] Z. Su, B. Ai, D. He, G. Ma, K. Guan, N. Wang, and D. Zhang, "User association and backhaul bandwidth allocation for 5G heterogeneous networks in the millimeter-wave band," *2017 IEEE/CIC International Conference on Communications in China, ICCC 2017*, vol. 2018-Janua, pp. 1–6, 2018.

[208] A. Mesodiakaki, F. Adelantado, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Energy and spectrum efficient user association in 5G heterogeneous networks," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, pp. 1–6, 2016.

[209] L. Gurobi Optimization, "Mixed Integer Programming Basics." [Online]. Available: https://www.gurobi.com/resource/mip-basics/

[210] V. Baños-Gonzalez, M. S. Afaqui, E. Lopez-Aguilera, and E. Garcia-Villegas, "IEEE 802.11ah: A technology to face the IoT challenge," *Sensors (Switzerland)*, vol. 16, no. 11, 2016.

[211] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, 2015.

[212] H. Shokri-Ghadikolaei, C. Fischione, and E. Modiano, "On the accuracy of interference models in wireless communications," *2016 IEEE International Conference on Communications, ICC 2016*, pp. 1–6, 2016.

[213] H. Shokri-Ghadikolaei and C. Fischione, "The Transitional Behavior of Interference in Millimeter Wave Networks and Its Impact on Medium Access Control," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 723–740, 2016.

[214] S. Ryoo, J. Jung, and R. Ahn, "Energy efficiency enhancement with RRC connection control for 5G new RAT," *IEEE Wireless Communications and Networking Conference, WCNC*, vol. 2018-April, pp. 1–6, 2018.

[215] O. Koymen, T. A. Thomas, T. S. Rappaport, H. C. Nguyen, I. Rodriguez, A. Ghosh, I. Z. Kovacs, S. Sun, and A. Partyka, "Investigation of Prediction Accuracy, Sensitivity, and Parameter Stability of Large-Scale Propagation Path Loss Models for 5G Wireless Communications," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 2843–2860, 2016.

[216] M. Polese *et al.*, "Improved handover through dual connectivity in 5g mmwave mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 2069–2084, 2016.

[217] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of Millimeter Wave Communications for Fifth-Generation (5G) Wireless

Networks-With a Focus on Propagation Models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213–6230, 2017.

[218] Keysight, "Understanding the 5G NR Physical Layer," Tech. Rep., 2017. [Online]. Available: https://www.keysight.com/upload/cmc_upload/All/Understanding_the_5G_NR_Physical_Layer.pdf

[219] S. Sun, T. S. Rappaport, M. Shafi, P. Tang, J. Zhang, and P. J. Smith, "Propagation Models and Performance Evaluation for 5G Millimeter-Wave Bands," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8422–8439, 2018.

[220] 3GPP, "TR 138 901 - V14.0.0 - 5G; Study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 version 14.0.0 Release 14)," Tech. Rep., 2017. [Online]. Available: http://www.etsi.org/standards-search

[221] T. Report, "TR 138 901 - V15.0.0 - 5G; Study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 version 15.0.0 Release 15)," vol. 0, 2018.

[222] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems," 1998. [Online]. Available: http://arxiv.org/abs/cs/9809099

[223] C. Kim, R. Ford, and S. Rangan, "Joint interference and user association optimization in cellular wireless networks," *Conference Record - Asilomar Conference on Signals, Systems and Computers*, vol. 2015-April, pp. 511–515, 2015.

[224] Y. Liu, M. Derakhshani, and S. Lambotharan, "Dual Connectivity in Backhaul-limited Massive-MIMO HetNets : User Association and Power Allocation."

[225] G. Pocovi, S. Barcos, H. Wang, K. I. Pedersen, and C. Rosa, "Analysis of Heterogeneous Networks with Dual Connectivity in a Realistic Urban Deployment," no. May 2016, 2015.

[226] M. Eisen and A. Ribeiro, "Optimal Wireless Resource Allocation with Random Edge Graph Neural Networks," pp. 1–15, 2019. [Online]. Available: http://arxiv.org/abs/1909.01865