DISSERTATION

# USER-ORIENTED MOBILITY MANAGEMENT IN CELLULAR WIRELESS NETWORKS

Submitted by

Alaa A. R. Alsaeedy

Department of Electrical and Computer Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2020

Doctoral Committee:

Advisor: Edwin Chong

Jade Morton
J. Rockey Luo
Rebecca Atadero

ABSTRACT

USER-ORIENTED MOBILITY MANAGEMENT IN CELLULAR WIRELESS NETWORKS

Mobility Management (MM) in wireless mobile networks is a vital process to keep an individual User Equipment (UE) connected while moving within the network coverage area—this is required to keep the network informed about the UE's mobility (i.e., location changes). The network must identify the exact serving cell of a specific UE for the purpose of data-packet delivery. The two MM procedures that are necessary to localize a specific UE and deliver data packets to that UE are known as *Tracking Area Update (TAU)* and *Paging*, which are burdensome not only to the network resources but also UE's battery—the UE and network always initiate the *TAU* and *Paging*, respectively. These two procedures are used in current Long Term Evolution (LTE) and its next generation (5G) networks despite the drawback that it consumes bandwidth and energy.

Because of potentially very high-volume traffic and increasing density of high-mobility UEs, the *TAU/Paging* procedure incurs significant costs in terms of the signaling overhead and the power consumption in the battery-limited UE. This problem will become even worse in 5G, which is expected to accommodate exceptional services, such as supporting mission-critical systems (close-to-zero latency) and extending battery lifetime (10 times longer). This dissertation examines and discusses a variety of solution schemes for both the *TAU* and *Paging*, emphasizing a new key design to accommodate 5G use cases. However, ongoing efforts are still developing new schemes to provide seamless connections to the ever-increasing density of high-mobility UEs.

In this context and toward achieving 5G use cases, we propose a novel solution to solve the MM issues, named *gNB-based UE Mobility Tracking (gNB-based UeMT)*. This solution has four features aligned with achieving 5G goals. First, the mobile UE will no longer trigger the *TAU* to report their location changes, giving much more power savings with no signaling overhead. Instead, second, the network elements, gNBs, take over the responsibility of *Tracking* and *Locating*

these UE, giving always-known UE locations. Third, our *Paging* procedure is markedly improved over the conventional one, providing very fast UE reachability with no *Paging* messages being sent simultaneously. Fourth, our solution guarantees lightweight signaling overhead with very low *Paging* delay; our simulation studies show that it achieves about 92% reduction in the corresponding signaling overhead. To realize these four features, this solution adds no implementation complexity. Instead, it exploits the already existing LTE/5G communication protocols, functions, and measurement reports.

Our *gNB-based UeMT* solution by design has the potential to deal with mission-critical applications. In this context, we introduce a new approach for mission-critical and public-safety communications. Our approach aims at emergency situations (e.g., natural disasters) in which the mobile wireless network becomes dysfunctional, partially or completely. Specifically, this approach is intended to provide swift network recovery for Search-and-Rescue Operations (SAROs) to search for survivors after large-scale disasters, which we call *UE-based SAROs*. These SAROs are based on the fact that increasingly almost everyone carries wireless mobile devices (UEs), which serve as human-based wireless sensors on the ground. Our *UE-based SAROs* are aimed at accounting for limited UE battery power while providing critical information to first responders, as follows: 1) generate immediate crisis maps for the disaster-impacted areas, 2) provide vital information about where the majority of survivors are clustered/crowded, and 3) prioritize the impacted areas to identify regions that urgently need communication coverage. *UE-based SAROs* offer first responders a vital tool to prioritize and manage SAROs efficiently and effectively in a timely manner.

# ACKNOWLEDGMENTS

I would like to express my sincere appreciation to the Iraqi Ministry of Higher Education and Scientific Research for granting me a scholarship to study abroad, pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering, Colorado State University, USA. My deepest thanks also go to the University of Al-Qadisiyah where I was initially nominated for this scholarship. Thanks so much to the Iraqi Cultural Office in Washington DC and all individuals whose support and encouragement have made my dissertation successful.

Foremost, I would like to express my gratitude and appreciation to my advisor Professor Edwin Chong for his patience, motivation, and immense knowledge. His professional guidance and valuable recommendations helped me to write and publish high-quality technical papers, which are the backbone of this dissertation. Professor Chong's help and encouragement have widened the scope of my research, supporting me to attend conferences and present my work.

My sincere thanks also go to my committee members, Jade Morton, J. Rockey Luo, and Rebecca Atadero, for their insightful comments, supportive suggestions, and serving on my committee.

Last but not least, my deepest thanks go to my parents, wife, and children for their love with care and patience. My special thanks also extend to all my friends and colleagues, especially those who encouraged and supported me while pursuing my Ph.D. degree. I'm very grateful that I have such friendships.

# DEDICATION

*I would like to dedicate this dissertation to my parents and family, with love.*

TABLE OF CONTENTS


ABSTRACT . . . . . ii
ACKNOWLEDGMENTS . . . . . iv
DEDICATION . . . . . v
LIST OF TABLES . . . . . ix
LIST OF FIGURES . . . . . x

Chapter 1 Introduction . . . . . 1
1.1 Background and Preliminaries . . . . . 2
1.2 Open Questions . . . . . 4
1.3 Contributions and Dissertation Outline . . . . . 6

Chapter 2 Mobility Management in LTE Networks . . . . . 7
2.1 Overview . . . . . 7
2.2 Main Focus . . . . . 7
2.3 Mobility Management Entity (MME) in LTE . . . . . 8
2.3.1 *TAU* Procedure . . . . . 12
2.3.2 *Paging* Procedure . . . . . 13
2.3.2.1 *Paging* procedure from the UE's perspective . . . . . 14
2.3.2.2 *Paging* procedure from the MME's perspective . . . . . 15
2.4 *TAU* and *Paging* Overhead . . . . . 17
2.4.1 Modeling of Signaling Cost . . . . . 17
2.4.2 Modeling of UE Battery Power Cost . . . . . 19
2.5 Mobility Management Techniques . . . . . 19
2.5.1 *TAU* Improvement Techniques . . . . . 22
2.5.1.1 Global and static techniques for *TAU* . . . . . 22
2.5.1.2 Local and dynamic techniques for *TAU* . . . . . 23
2.5.2 *Paging* Improvement Techniques . . . . . 26
2.5.3 Combined *TAU* and *Paging* Techniques . . . . . 30
2.5.3.1 Optimization approaches . . . . . 31
2.5.3.2 Information-theoretic approaches . . . . . 33
2.5.3.3 Mobility model-based approaches . . . . . 35
2.6 SDN and Virtualization-based MM in LTE . . . . . 36
2.7 Final Remarks . . . . . 40
2.8 Summary . . . . . 41

Chapter 3 Mobility Management in 5G Networks . . . . . 43
3.1 Overview . . . . . 43
3.1.1 Open Questions . . . . . 43
3.1.2 Our Focus . . . . . 44
3.2 *TAU* and *Paging* Challenges in 5G Networks . . . . . 46
3.2.1 Massive Deployment of UEs . . . . . 47

3.2.2 HetNets Deployment . . . 48
3.2.3 High-Mobility UEs . . . 48
3.3 LTE MM Assessment for 5G Networks . . . 49
3.4 Mobility Models Assessment for 5G Networks . . . 50
3.4.1 Estimation Accuracy . . . 51
3.4.2 Final Notes . . . 52
3.5 Design of MM in 5G Networks . . . 52
3.5.1 5G NFs . . . 53
3.5.2 Mobility States in LTE Systems . . . 55
3.5.3 Mobility States in 5G Systems . . . 56
3.6 NG-RRC States for 5G Systems . . . 57
3.6.1 5G RRC-INACTIVE State . . . 58
3.6.2 NG-RRC Protocols . . . 58
3.7 5G *RNAU* and *Paging* . . . 59
3.7.1 5G *RNAU* . . . 60
3.7.2 5G *Paging* . . . 61
3.7.2.1 DRX cycle specifications for 5G . . . 62
3.7.2.2 DRX cycle periods . . . 63
3.7.2.3 DRX cycle values . . . 63
3.8 MM Improvement Studies for 5G . . . 64
3.9 Performance Improvement for 5G . . . 67
3.10 Summary . . . 69

Chapter 4 5G Mobility Management for Critical and End-User Needs . . . 71
4.1 Overview . . . 71
4.2 Pitfalls of Current Tracking and Locating Procedures . . . 72
4.3 Our Solution Approach . . . 73
4.3.1 The Proposed Solution . . . 74
4.3.2 Preliminaries . . . 76
4.4 Solution Framework and Methodology . . . 77
4.4.1 *gNB-based UeMT* Entity Definitions . . . 77
4.4.2 *gNB-based UeMT* H/V-CT Functions . . . 78
4.5 *gNB-based UeMT* Control Models . . . 82
4.5.1 Basic Workflow of IoT/UE Mobility Tracking . . . 82
4.5.2 Service Switch Procedure (SSP) . . . 83
4.5.3 Home/Visiting-gNBs Mobility Tracking Scheme (H/V-MTS) . . . 83
4.6 Illustrative Scenarios . . . 86
4.6.1 When IoT/UEs Served by Home-gNB . . . 86
4.6.2 When IoT/UEs Served by Visiting-gNB . . . 86
4.7 Cost Functions for Tracking and Locating . . . 88
4.7.1 Power Overhead . . . 89
4.7.2 Signaling Overhead . . . 89
4.8 Simulation Setup and Performance Evaluation . . . 91
4.8.1 IoT/UE Battery Power Consumption . . . 91
4.8.2 Combined Signaling Costs . . . 92

4.8.3 Final Notes . . . 92
4.9 Summary . . . 93

Chapter 5 User-Oriented Mission-Critical Communication . . . 95
5.1 Overview . . . 95
5.1.1 Current MCPSC Systems . . . 95
5.1.2 Failure of Current MCPSC Systems . . . 97
5.2 Network Status Post-Disaster . . . 97
5.2.1 Lack of RFC . . . 98
5.2.2 Isolated gNBs . . . 98
5.2.3 Cell-Edge UEs . . . 98
5.3 Related Studies . . . 99
5.3.1 Deploying Wireless Equipment into RoI . . . 99
5.3.2 Network Recovery Using D2D Communication . . . 100
5.3.3 Network Recovery Using UAV Communication . . . 102
5.4 Solution Approach . . . 103
5.4.1 Issues to be Considered . . . 103
5.4.2 Proposed Approach . . . 103
5.5 *UE-based SARO* System Model . . . 104
5.5.1 Entity Definitions . . . 104
5.5.2 Entity Notation . . . 106
5.6 *UA-gNB* Searching Procedures . . . 108
5.6.1 Procedure for Finding *ref-gNB* . . . 108
5.6.1.1 Cluster centroid-based search (CCBS) . . . 109
5.6.1.2 *UA-gNB* optimal distance . . . 110
5.6.2 Procedure for Finding *X-UEs* . . . 112
5.7 *UA-gNB* Overlapping RFC Issue . . . 113
5.7.1 Beamsteering Antenna . . . 114
5.7.2 Access Control . . . 115
5.8 Mobility Management for *UE-based SAROs* . . . 116
5.9 *UA-gNB* Location Setup for *SZs* . . . 117
5.10 Time Cost for Discovery and Relocation . . . 119
5.10.1 Discovery Time Cost . . . 120
5.10.2 Relocation Time Cost . . . 120
5.10.3 Simulation Setup . . . 121
5.11 Generating Crisis Maps, *UEBCMs* . . . 122
5.11.1 Illustrative Scenario Setup . . . 123
5.11.2 *UEBCM* for UE Densities . . . 123
5.11.3 *UEBCM* for UE RSRP Levels . . . 125
5.11.4 Building *Priority-Driven RFC (PDRFC)* . . . 126
5.12 Summary . . . 127

Chapter 6 Conclusion . . . 128

Bibliography . . . 130

LIST OF TABLES


2.1 Summary of EMM and ECM mobility states . . . . . 11
2.2 Signaling cost of *TAU* and *Paging* in $M$ (adapted from [1]) . . . . . 18
2.3 *TAU* scheme comparisons . . . . . 26
2.4 *Paging* scheme comparisons . . . . . 31
2.5 Trade-off between *TAU* and *Paging* scheme comparisons . . . . . 36

3.1 Cell types in wireless networks . . . . . 48
3.2 NG-RRC protocols and functions . . . . . 60

4.1 *Home-CT* and *Visiting-CT* entries . . . . . 79
4.2 IoT/UE association status (*Resident/Visiting*) . . . . . 80
4.3 Example of *CiPD* index values . . . . . 81
4.4 Signaling load of *TAU* and *Paging* in $M$ (adapted from [1–3]) . . . . . 89

5.1 Equivalence of *gNB-based UeMT* and *UE-based SAROs* entities . . . . . 117
5.2 Information table for each *ref-gNB*$_i$*–UA-gNB*$_i$ pair . . . . . 119
5.3 Association table for each *ref-gNB*$_i$ . . . . . 119
5.4 Information table for the *PDRFC* . . . . . 126

LIST OF FIGURES


1.1 TA and TAL concept in LTE . . . 2
1.2 Distribution of MME signaling events in a U.S. LTE network (adapted from [4]) . . . 3

2.1 EPC and E-UTRAN architecture in LTE networks . . . 9
2.2 EMM and ECM state models between UE and MME . . . 12
2.3 Paging process in LTE networks . . . 16
2.4 Smart *Paging* and *TA* management to reduce signaling load (adapted from [4]) . . . 21
2.5 Classification of *TAU* and *Paging* solution schemes . . . 22
2.6 Comparison of *Paging Success Rate* and *Bandwidth utilization* (adapted from [5]) . . . 28
2.7 Relative cost comparisons for BP, SP, and PP schemes (adapted from [6]) . . . 29

3.1 Key performance requirements for 5G goals (adapted from [7]) . . . 44
3.2 MM 5G promising key features . . . 47
3.3 5G system architecture in NF and reference point representation (adapted from [8]) . . . 54
3.4 Overall architecture of 5G system [9] . . . 54
3.5 Connectivity and UE RRC states for LTE and 5G (adapted from [10]) . . . 57
3.6 UE state model for NG-RAN (adapted from [10, 11]) . . . 59
3.7 DRX cycle [9] . . . 64

4.1 *Tracking* and *Locating* . . . 74
4.2 X2/Xn protocol stack for the UP and CP . . . 77
4.3 Behavior of the *SST* in the *Visiting-CT* . . . 82
4.4 Flow chart of the *gNB-based UeMT* process . . . 84
4.5 Signaling flow of the *SSP* . . . 85
4.6 *Home/Visiting-gNB* and *Home/Visiting-CT* interaction . . . 85
4.7 Shows UEs served by *Home-gNB* . . . 87
4.8 Shows UEs served by *Visiting-gNB* . . . 88
4.9 Draining the IoT/UE battery power while moving . . . 91
4.10 Signaling cost for *TAU/Paging* versus *gNB-based UeMT* . . . 93

5.1 Illustrative example showing the SARO entities . . . 104
5.2 *UA-gNB* searching model . . . 107
5.3 Illustrative example showing *UA-gNBs* centers at cluster centroids, using k-means . . . 109
5.4 Illustrative example showing the LoS and height conditions . . . 110
5.5 Maximum distance according to (5.3) . . . 112
5.6 Avoiding RFC overlap using beamsteering . . . 114
5.7 *UA-gNB*$_i$ screening locations . . . 118
5.8 Discovery and relocation time . . . 122
5.9 Attached UEs corresponding to each *SZ* . . . 124
5.10 *UEBCM*: Density of surviving UEs . . . 125
5.11 *UEBCM*: RSRP levels of the attached UEs . . . 125

# Chapter 1

# Introduction

Over the past few years, wireless mobile networks have expanded significantly, and about 95% of the world's population are now covered by 3rd Generation Partnership Project (3GPP) cellular technologies. One of the fastest-deployed mobile networks is known as Long Term Evolution (LTE) [12]. In addition, 5 billion mobile subscriptions are expected by the end of 2022, and around 98% of the US population lives in areas where LTE technology is available [13]. Moreover, LTE will work in conjunction with its next generation, the $5^{th}$ Generation of mobile networks (5G), to introduce tremendous services beyond current mobile networks (discussed further in Chapter 3). Note that the LTE networks are retuned to use as a basis for design in 5G networks, so our discussion starts from current LTE toward 5G.

Up to this point, LTE Network Operators (LNOs) have been concerned about not only mobile communication coverage but also providing high Quality-of-Service (QoS) to each individual end-user. Henceforth, we will follow the usual convention in LTE, according to the 3GPP technical specification [14], and refer to each end-user cellular device as a UE (or UEs for plural). The acronym UE comes from "user equipment," though UE is often specifically used for an individual cellular device rather than devices in general. To deliver service to huge numbers of UEs, LTE networks need to know the exact location of each UE. In this case, LTE networks (and its successor 5G) have to keep track of all UEs, and because of the high mobility of UEs, it is difficult to determine each UE's location precisely (i.e., serving cell of each UE) across the network coverage area. Therefore, the service being delivered to a specific UE experiences some delays while the network searches for the intended serving cell. However, current LTE networks allow UEs to be connected even with high mobility [1]. For example, in high-speed trains, LTE networks can maintain UEs connected up to speeds of 350 km/h or even 500 km/h in rural areas [15]. Note that the material in this chapter has been published in [16].

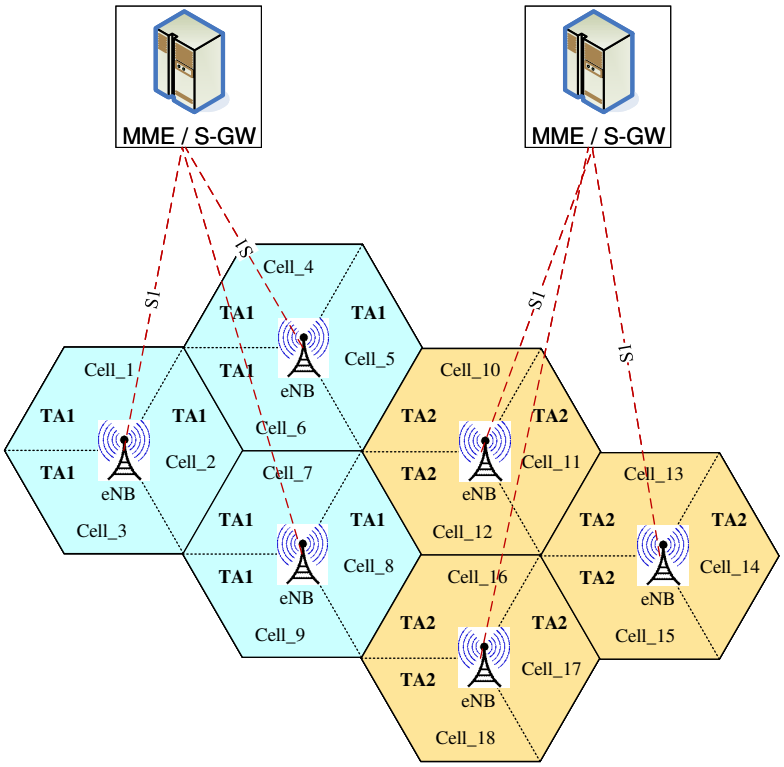



**Figure 1.1:** TA and TAL concept in LTE

## 1.1 Background and Preliminaries

The LTE network component that controls and manages an individual UE's mobility within the network is called the Mobility Management Entity (MME) [2, 17], which handles all mobility-related messages between UEs and the serving network (the MM in LTE and 5G are discussed further in Chapters 2 and 3, respectively). To facilitate this operation, the LTE coverage area is divided into groups of cells (or groups of eNBs, base stations in LTE terminology) called Tracking Areas (TAs), also known as Registration Areas (RAs) [18, 19], and each TA has a unique identity called Tracking Area Identity (TAI). Likewise, these TAs are further grouped into Tracking Area Lists (TALs) [2]. Figure 1.1 shows an illustrative example of how cells and TAs are grouped to form TALs. TA1 consists of Cell_1 to Cell_9. Similarly, TA2 consists of Cell_10 to Cell_18. As described in the 3GPP technical specification [2], once a UE registers with the network, the MME allocates a specific TAL, which comprises a set of TAs in close proximity to the UE's current

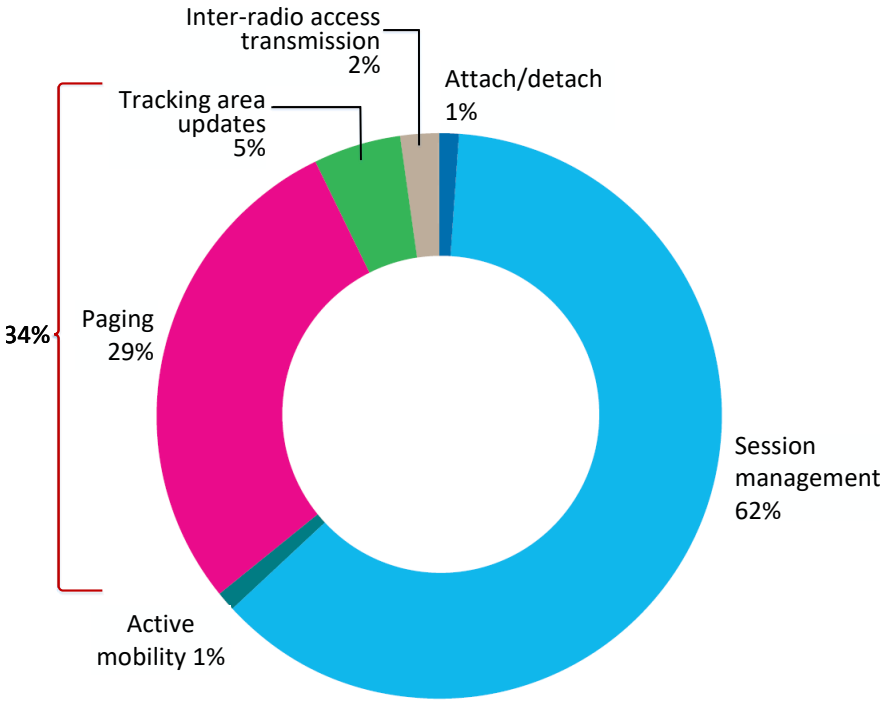



**Figure 1.2:** Distribution of MME signaling events in a U.S. LTE network (adapted from [4])

position. Therefore, the UE will be assigned a new TAL when it moves out of the current TAL by a process called *Tracking Area Update (TAU)*, initiated by the UE through its serving eNB (i.e., the current serving cell).

There is another process initiated by the MME, called *Paging*. The network uses this process to localize a specific UE within the network to forward the incoming data packets. In other words, the MME sends *Paging* messages to determine the exact serving cell of the UE within the network. Therefore, MME is considered the control of LTE networks access. All *TAU* and *Paging* signaling are completely processed by the MME servers, should process all *TAU* and *Paging* requests in an efficient way. However, the MME is burdened by very high signaling loads because of high volume mobility and connection management (connected with thousands of eNBs). This process is considered one of the highest costs of signaling, not only for LTE networks, but also for UE units [20]. During normal busy hours, a MME can handle a signaling load of over 500 to 800 messages/UE and even up to 1500 messages/UE under extreme circumstances [21]. At the same time, this process consumes battery power in UEs and costs over 10 mW of power consumption in

current generation smart-phones per process. A real data set collected from a large metropolitan market in the USA reveals that most signaling loads on the MME is caused by *TAU* and *Paging* procedures, as shown in Figure 1.2; cost about 34% of the total signaling load on the MME [4].

Generally, LNOs use many LTE Key Performance Indicators (KPIs) measurements, as primary indicators, to evaluate and measure the network performance to satisfy the end-user requirements according to Service Level Agreement (SLA) [22]. Basically, the *Paging Success Rate* and *TAU Success Rate* are defined as the two LTE KPIs that measure how well the *TAU* and *Paging* procedures are succeeded [23, 24]. Having low KPI values of *Paging Success Rate* and *TAU Success Rate* can be caused by low UE and/or network performances. Many researchers and practitioners try to mitigate the overall signaling loads on the MME by improving both the *TAU* and *Paging* procedures and the end-user experience to maintain the related KPIs at optimal values.

## 1.2 Open Questions

By taking into account the above discussion and before we proceed further, we give rise, in this context, to the following questions:

1. How to allocate the best TAL to a UE in the network such that the power consumption in the UE is minimized (most UEs are battery-limited)?

2. What is an upper bound on the TAs in that TAL?

3. How many cells (i.e., eNBs) should be in one TA?

4. If a MME has a data packet to deliver to a specific UE, can the UE be reached (i.e., find the serving cell) within the network in a reasonable time of delay (i.e., during an acceptable value of *Paging* latency [25])? This issue tends to be a crucial problem for mission-critical communication.

5. While the MME is burdened by heavy signaling loads, what are the recent solutions to mitigate the overall signaling overhead owing to high mobility connection management?

Basically, all LNOs take into account the above questions when planning a certain network to determine the optimal values for their network parameters (i.e., TA size, assigned TAL size, *Paging* latency value). Currently, LTE networks adopt a dynamic Mobility Management (MM) (also called location management) scheme where each UE has own TAL which comprises up to 16 TAs [1, 18]. The number of cells (i.e., eNBs) in one TA depends on the network topology or whether the coverage is outdoor or indoor. Typically, one TA consists of 1 to 100 eNBs [26]. In the dynamic scheme, each UE is provided with a specific TAL which consists of a number of TAs that in close proximity to the UE's current position [2]. This scheme would reduce the frequency of *TAU* requests that a UE sends when it moves within the network coverage area. When the UE crosses the boundary of its previously allocated TAL, it sends a *TAU* request to inform the network about its location update and acquires a new TAL. Essentially, the *Paging* process will use the new TAL to page the corresponding UE. In other words, if a MME needs to page a specific UE, it should broadcast *Paging* messages to all eNBs in the TAL that are already associated with the intended UE. In this case, a larger number of cells in a TA (i.e., larger area for a TA) would produce more *Paging* messages that burden the MME. On the other hand, a small number of cells in a TA (i.e., smaller area for a TA) would increase the *TAU* requests (i.e., increase power consumption in UEs). Excessive *TAU* requests can reduce the *Paging Success Rate* KPI because some UEs cannot respond to the *Paging* messages while responding to the *TAU* procedures. This problem can be considered as a planning problem. Many researchers have studied this network planning problem by developing algorithms to provide some trade-off solutions.

In this context, this dissertation reviews the existing MM algorithms (in terms of *TAU/Paging* cost) to find out the best trade-off solutions and gives insights to evaluate current LTE algorithms and their role in future wireless networks, especially for 5G networks. As we will see later, LTE systems will be used as a legacy design for 5G systems. However, MM will become a crucial problem for 5G requirements; for example, how to support real-time applications, providing close-to-zero latency for life-critical systems? Because most UEs are battery-limited, how to extend the battery lifetime (10 times longer), supporting a new paradigm of 5G Internet-of-Things (IoT)

devices? More specifically, as many studies and practitioners try to reduce the overall signaling overhead for LTE MM, the 5G MM issue in terms of *TAU* and *Paging* overhead will become more severe than in LTE. All the preceding concerns are the main focus of this dissertation.

## 1.3 Contributions and Dissertation Outline

This dissertation is based on our work in [16, 27–32]. In Chapter 2, we first introduce the MM in current LTE systems, describing the *TAU* and *Paging* procedures and their related UE mobility states. Furthermore, we also discuss and analyze the proposed solutions to deal with the problem of MM offloading. Note that the material in Chapter 2 has been published in [16].

In Chapter 3, we address the challenges that burden 5G networks in terms of *TAU* and *Paging* signaling overhead. Furthermore, we discuss the applicability of current LTE *TAU* and *Paging* schemes and evaluate the new MM improvement studies for 5G use cases, highlighting new aspects of 5G network improvements. Note that the material in Chapter 3 has been published in [27].

In Chapter 4, after studying and analyzing the state-of-the-art MM in terms of *TAU* and *Paging* solutions (which are not sufficient to achieve the 5G critical requirements), we propose a novel approach to achieving these requirements, supporting mission-critical systems and real-time applications. Note that the material in Chapter 4 has been published in part in [28] and in whole in [29].

In Chapter 5, we introduce a new framework for mission-critical and public-safety communications. This is intended to provide swift network recovery for Search-and-Research Operations (SAROs) to search for survivors after large-scale disasters, assuming the wireless network cells are partially operational and exploiting the recent trend of using Unmanned Aerial Vehicles (UAVs) as parts of the network. These SAROs are based on the idea that almost all survivors have their own cellular mobile devices (UEs), which can serve as human-based sensors on the ground. Note that the material in Chapter 5 has been published in part in [30, 31] and in whole in [32].

Finally, in Chapter 6, we highlight and conclude the main points of this dissertation, including future expansions of its underlying topics.

# Chapter 2

# Mobility Management in LTE Networks

## 2.1 Overview

In this chapter, we provide preliminary information about the Mobility Management Entity (MME) in Long Term Evolution (LTE) networks, which is responsible for the Mobility Management (MM), including *TAU* and *Paging* procedures, of all User Equipment (UE) within the network coverage area. Furthermore, because these two procedures are burdensome to both the network resources (very high-volume traffic) and the UE's battery (most UEs are battery-limited), we review the existing MM solutions (in terms of *TAU/Paging* cost) to find out the best trade-off solutions and give insights to evaluate current LTE algorithms and their role in future wireless networks (called the 5$^{th}$ Generation (5G) wireless networks). Next, we list the main focus of this chapter. Note that the material in this chapter has been published in [16].

## 2.2 Main Focus

In the context of the preceding discussion, this chapter focuses attention on the following:

1. While critically discussing the existing LTE solution schemes in terms of *TAU* and *Paging* overhead, we evaluate them with a view toward 5G, which is expected to provide exceptional services beyond current LTE (e.g., 10 times longer UE battery lifetime). Specifically, because LTE systems will be used as a legacy design for 5G systems, this study provides fertile ground for researchers to investigate the current MM solutions (for LTE specifically) to reuse/redesign toward 5G use cases, providing a comprehensive discussion about the implications of the current *TAU* and *Paging* procedures, which impact both the UE experience and network performance.

2. To elaborate on item 1 above, we classify the existing solutions into groups based on the particular approaches taken (as we detail later in Sections 2.5 and 2.6 and illustrate in

Figure 2.5), providing comparison tables and figures (Tables 2.3, 2.4, and 2.5, and Figures 2.4, 2.6, and 2.7). This is to help researchers to retune/reuse a particular solution scheme toward 5G. For more detail about MM issues in 5G specifically, see Chapter 3.

3. We uniquely emphasize the problem of battery power consumption in mobile UEs (most UEs are battery-limited, including IoT devices). All UEs are required to initiate the *TAU* procedure (normal and periodic *TAUs*, detailed in Section 2.3.1), and this gives rise to battery power problems in these battery-limited UEs. In this context, we highlight and discuss quantitative measures for the power overhead (including the accompanied signaling cost), which is the main focus of Section 2.4. These quantitative measures can be used to design new schemes, achieving 5G requirements in terms of reducing power consumption and signaling load—with the current *TAU/Paging* procedure, it will be difficult to realize these 5G use cases.

Apart from *TAU* and *Paging* solution schemes, we also examine new solution trends to mitigate the overall signaling overhead on the MME, such as Software Defined Network (SDN) and Virtualization (SDNV) in LTE [33–35], including Centralized and Distributed MM (CMM and DMM) approaches [36], detailed in Section 2.6. Moreover, because of the rapid increase in UE loads in terms of density, wireless services, and mobility, network KPIs are adversely affected. This issue tends to be a crucial problem for 5G requirements. In addition, we highlight and critically analyze different types of UE mobility models, which are used to evaluate the network performance related to UE mobility, discussed in Section 2.5.3.3.

## 2.3 Mobility Management Entity (MME) in LTE

The core network architecture in LTE, called the Evolved Packet Core (EPC), is responsible for providing complete mobile-broadband services to the underlay Evolved Universal Terrestrial Radio Access Network (E-UTRAN) [37]. According to the 3GPP standards, the EPC consists of several different elements, as briefly shown in Figure 2.1, such as Serving Gateway (S-GW), Packet Data Network Gateway (PDN Gateway, P-GW), Policy and Charging Rules Function (PCRF),

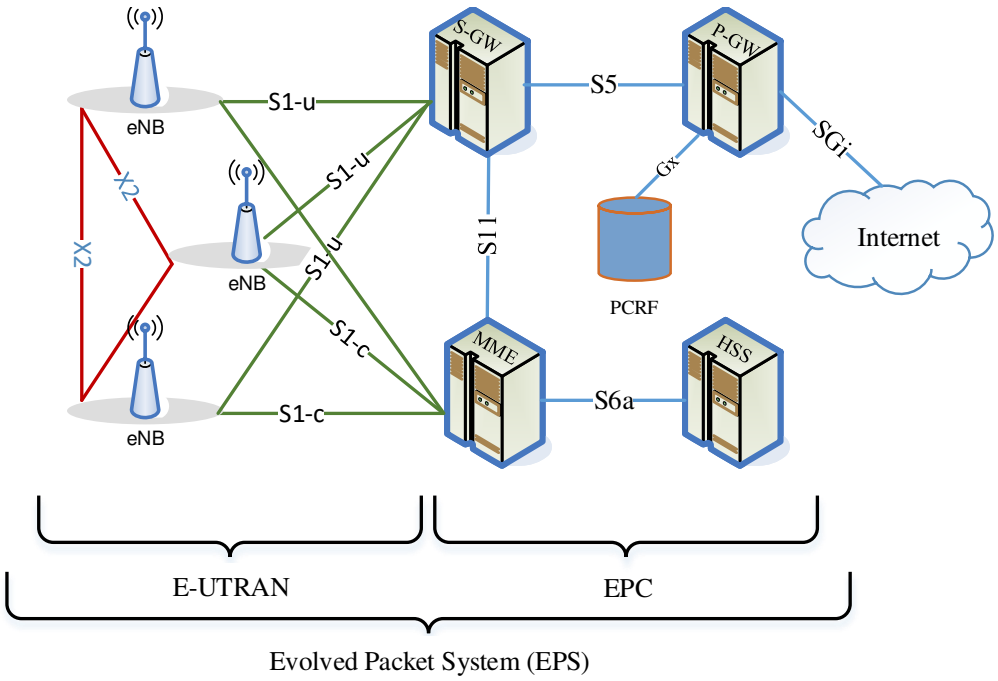



**Figure 2.1:** EPC and E-UTRAN architecture in LTE networks

Home Subscriber Service (HSS), and MME. In this chapter, we focus on the MME, which is in charge of MM, also known as the control-plane of the EPC. Generally, the MME supports many control-plane functions such as authentication, handling of idle to active transitions, handover, TAL management, and paging. (For more details, see [2]). MM is one function within the context of general network management. Network management functions include the following network services: performance management services, configuration management services, and fault supervision services, according to [38]. In this context, we are specifically focus on the *TAU* and *Paging* procedures (managed and controlled by the MME). These two procedures are important network management processes because they are required to track and locate each individual UE while moving across the network coverage area (as stated earlier, these are necessary for the purpose of UE-specific services delivery).

The events between the UE and its serving MME can be described by two main states as below [2, 26]:

1. EPS[1] Mobility Management (EMM) states: Used to describe the results of mobility management procedures such as *Attach*, *Detach*, and *TAU*, and can be defined by the following two EMM sub-states:

   (a) EMM-DEREGISTERED: In this state, a UE is not attached to the network, and no MME has information about the UE's location. More precisely, the UE is unreachable, there is no active context for it, and its recent reported location may be temporarily stored in the last serving MME at a TAL accuracy.

   (b) EMM-REGISTERED: In this state, a UE is attached to the network and its location has been known by the serving MME since the last triggered *TAU* procedure. More precisely, all the attached UEs have location information stored in the MME either within a cell or TAL granularity.

2. EPS Connection Management (ECM) states: Used to indicate whether there is an active signaling connection between the UE and EPC. The ECM states can be defined by the following two ECM sub-states:

   (a) ECM-IDLE: In this state, the serving MME has known the UE's location, the UE is in EMM-REGISTRED state, and there is no active data exchange (i.e., the UE is dormant). the UE's location is known with an accuracy of the assigned TAL (i.e., the UE's location is known within some TAs).

   (b) ECM-CONNECTED: A UE and its serving MME enter this state whenever the UE sends or MME receives any of the following signaling connection messages: *Attach Request, TAU Request, Service Request,* or *Detach Request*. That means that the UE is actively exchanging data packets with the network.

In LTE, the states of the UE connection with the corresponding serving eNB (i.e., radio-access network) is described by what are called the Radio Resources Control (RRC) states [2,27], used to

[1]In LTE, the acronym EPS refers to *evolved packet system*, which comprises the E-UTRAN and EPC [26]; see Figure 2.1.

**Table 2.1:** Summary of EMM and ECM mobility states

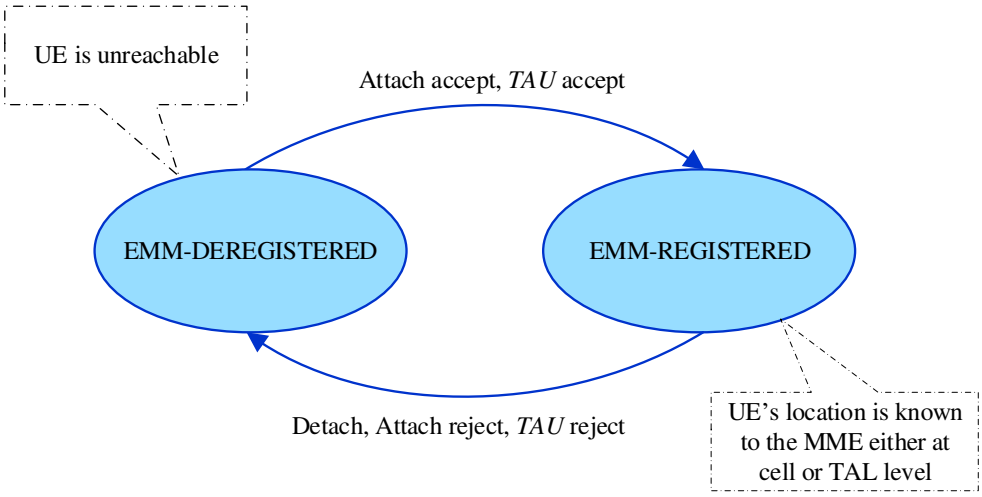
| | | |
|---|---|---|
| EMM-States | DEREGISTERED | -UE is unreachable (*Detached*)<br>-UE's location is unknown to the MME<br>-MME may store the last reported UE's location |
| | REGISTERED | -UE is reachable (*Attached*)<br>-UE's location is known to the MME<br>-MME stores UE's location either at a cell or TAL granularity |
| ECM-States | IDLE | -UE is reachable (*Attached*)<br>-UE's location is known at a TAL granularity<br>-UE has no active data packets to exchange |
| | CONNECTED | -UE is reachable (*Attached*)<br>-UE's location is known at a cell granularity<br>-UE has active data packets to exchange |

identify whether there is a connection established between the UE and its serving eNB—the latter is the control node for the RRC states, unlike the ECM states where the MME is the control node. For that purpose, there are two RRC states:

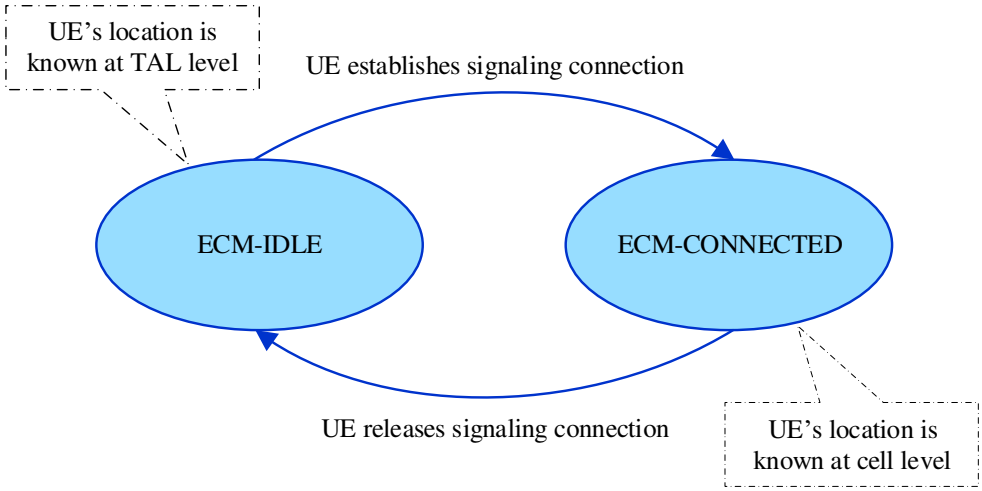
1. RRC-CONNECTED: A UE enters this state when an RRC context is established (there are active signaling messages transferred between the UE and the eNB). This state is intended for the UE-specific data packets to be exchanged with the eNB. When the UE is in the RRC-CONNECTED state, its location is known with cell-level accuracy, similar to the ECM-CONNECTED state.

2. RRC-IDLE: Unlike the RRC-CONNECTED state, when a UE is in the RRC-IDLE state, there is no RRC context established between the UE and its serving eNB; i.e., no data packets are exchanged. The UE's location in this state is known at its allocated TAL, similar to the ECM-IDLE state.

Table 2.1 summaries the EMM and ECM mobility states between a UE and its serving MME. Moreover, Figure 2.2 illustrates the transition conditions for each state. The control node for the above mobility states is the MME, which has many responsibilities including *TAU* and *Paging* tasks (see Figure 1.2). The following subsections describe the *TAU* and *Paging* procedures in more detail.




**(a)** EPS Mobility Management (EMM) states




**(b)** EPS Connection Management (ECM) states

**Figure 2.2:** EMM and ECM state models between UE and MME

### 2.3.1 *TAU* Procedure

A mobile UE always triggers a *TAU* process, sending a *TAU Request* message to its serving MME. Generally, UEs perform *TAU* procedures in a variety of situations [2, 19]. Basically, the UE initiates the *TAU* procedure to report that its location has changed within the network when detecting a new TA that is not listed in current TAL. The *TAU* procedure can be initiated by the two following scenarios:

1. Normal *TAU*: This procedure takes place when a UE detects itself entering a new TA that is not in the assigned TAL. Therefore, it sends a *TAU Request* message to the serving MME. Once received this message, the MME should respond by sending back a *TAU Accept* message and providing a new TAL to the corresponding UE according to the *Central Policy*; see Section 2.5.1.2. At the end of this process, the UE responds by accepting the new TAL, by sending a *TAU Complete* message.

2. Periodic *TAU*: This procedure is periodically triggered by a timer, known as *T3412-timer*, in the UE and is used to notify the network that the UE is still available (i.e., turned-on and under the network coverage area) [39]. The LTE network provides all the registered UEs an initial value for this timer during the EMM-REGISTERED state or during a *TAU Accept* message (the default initial value of *T3412-timer* = 54 min.). When a UE's state is changed from ECM-CONNECTED to ECM-IDLE, the *T3412-timer* resets and starts count-down from its initial value until expires (i.e., *T3412-timer* = 0). When this timer expires, the UE triggers another *TAU* procedure and sets the initial value to this timer again. This behavior continues periodically while the UE stays in the ECM-IDLE state. However, the *T3412-timer* stops counting-down when the UE's state is changed to ECM-CONNECTED or to EMM-DEREGISTERED. This concept is also used by the network controller to stop sending *Paging* messages to a UE that is already turned-off or out of the network coverage area [19].

### 2.3.2 *Paging* Procedure

In LTE networks, the MME always initiates the *Paging* procedure by broadcasting *Paging* messages to inform UEs in the ECM-IDLE and/or ECM-CONNECTED state about the following situations [40–42]:

1. Earthquake and Tsunami Warning System (ETWS): Allow LTE-enabled devices to receive ETWS notification.

2. Commercial Mobile Alert System (CMAS): Allow LTE-enabled devices to receive CMAS alerts during emergencies or natural disasters. The CMAS is also called Wireless Emergency Alerts (WEA).

3. Acquiring system information: Used to trigger all UEs to re-acquire the system information whenever there is an update in the LTE system information.

4. Paging a specific UE in ECM-IDLE state: Used whenever the network needs to locate the exact location of a specific UE on a cell level (i.e., determining the serving cell) to forward the incoming data packets to the intended UE.

The MME may broadcast *Paging* messages to multiple UEs in the ECM-IDLE or ECM-CONNECTED state within a certain area such as in **1**, **2**, and **3** above. In these scenarios, the exact locations of the UEs are not necessary from the network's point of view. However, in occasion **4**, the intended UE's location (i.e., its serving cell) has to be known instead of being located on a TAL level. Once the network identifies the UE on a cell level, the network can deliver the received data packets to their destination (i.e., the UE), and this is about reaching a specific UE that has a static or dynamic location change throughout the network. This is the most challenging scenario in the LTE *Paging* procedure (discussed later in this chapter). In contrast, the location of UEs in the ECM-CONNECTED state have been known on a cell level since establishing an active signaling connection.

To gain insight into the *Paging* procedure for a UE in the ECM-IDLE state, we describe the procedure from both the UE's and MME's perspectives as follows.

#### 2.3.2.1 *Paging* procedure from the UE's perspective

Essentially, in most cases when a UE enters the ECM-IDLE state, the UE monitors the download control channels (i.e., monitoring the *Paging Channel (PCH)*) to extract paging information from the *Paging Control Channel (PCCH)* [37]. Listening to the *PCH* all the time is a power-consuming process from the UE's perspective. Instead, the UE uses a *Discontinuous Reception (DRX)* mechanism to reduce the power consumption—that is, the UE makes multiple sleeping and

waking up at predefined occasions (i.e., battery power saving mechanism). By using the *DRX*, the UE can monitor the *PCH* only on the *Paging Occasions (POs)*, a LTE subframe where the *Paging* message is located. In other words, the UE should listen only to one *PO* per *DRX* cycle, defined as a time interval between monitoring the *POs* for a specific UE [3]. The *DRX* cycle is a configurable parameter in the UE, so the longer the *DRX* cycle, the more battery power is saved. However, using long a *DRX* cycle comes at the cost of adding more time delay to the *Paging* process. This can be considered as a trade-off optimization problem between *Paging* latency and the UE battery power saving.

#### 2.3.2.2 *Paging* procedure from the MME's perspective

Figure 2.3 shows an illustrative example to describe the MME's *Paging* procedure. The location of all UEs in the ECM-IDLE state are known on a TAL level by the serving MME. Also, all eNBs are connected to at least one MME/S-GW (i.e., eNBs are connected to the MME by means of *S1-c* and to the S-GW by means of *S1-u* interfaces, see Figure 2.1) [19]. According to the TAL concept [2], as shown in Figure 2.3, all the UEs within the TA1 boundary have the same TAL (i.e., TAL = {Cell_1, ..., Cell_9}). Likewise, all the UEs under the TA2 boundary have the same TAL (i.e., TAL = {Cell_10, ..., Cell_18}). In this particular example, the serving MME stores the TAL database for each individual UE under TA1 and TA2 coverage areas. When the MME receives downlink data packets designated for UE1 from the S-GW, the MME sends *Paging* messages to all cells in TA1, because the MME already knows that UE1 is within the TA1 coverage area. That means that one cell from the TA1 cell list is acting as a serving cell for UE1, but this cell is not known by the MME yet. While all the UEs in TA1 are monitoring the *PO*, once they receive the *Paging* messages, only the intended UE (i.e., UE1 in this example) will respond by sending back a *Paging* response through its serving eNB, and hence the MME stores the current UE serving cell. Consequently, UE1 should change its state to the ECM-CONNECTED state and establish a signaling connection with the network to exchange the required data packets.

It is worthwhile to mention here that the TAL idea gives rise to another trade-off optimization problem between the *Paging* cost and *Paging* latency [18, 43, 44]. The *Paging* process costs some

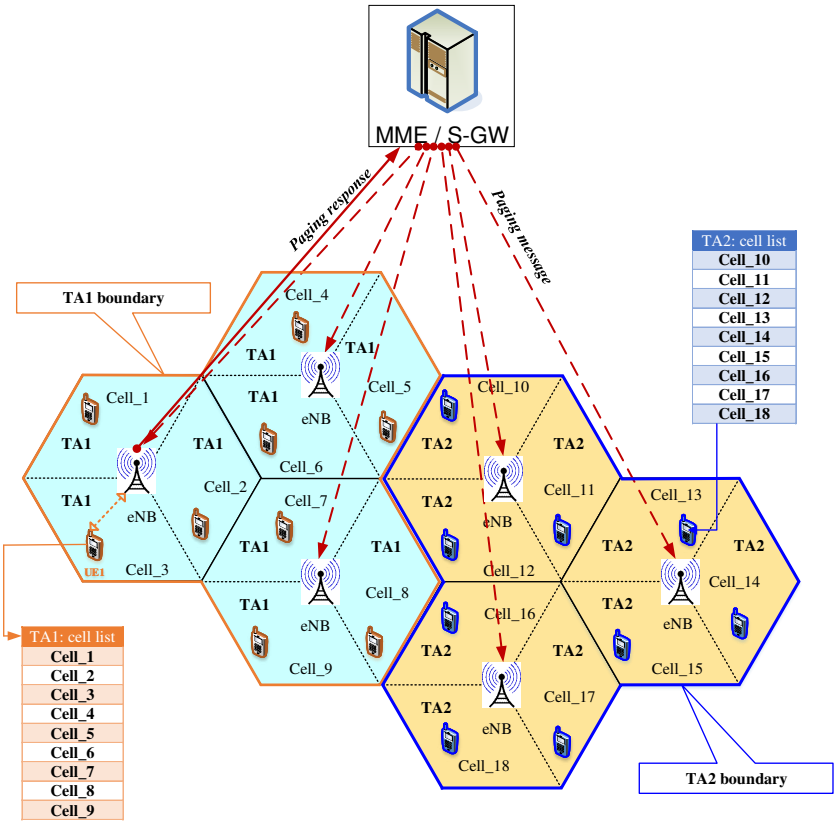



**Figure 2.3:** Paging process in LTE networks

of the available bandwidth (i.e., *Paging* overhead) and the *Paging* overhead is proportional to the number of cells in the TA being paged [45]. In addition, because there is a maximum number of *Paging* attempts in which a network can find the paged UE, the *Paging* attempts add more *Paging* delay to the *Paging* process. However, we could reduce the *Paging* delay at the expense of *Paging* cost [18]. To wit, to reduce the *Paging* delay, we could use more bandwidth to page more TAs at the same time.

If the MME failed to receive the UE's *Paging* response, this will be reported and counted toward the *Paging Failure Rate* KPI. The salient factors that increase the *Paging Failure Rate* KPI are listed as follows [26, 46]:

1. The UE is faulty because of a hardware problem but is still attached to the network—it cannot respond to the *Paging* messages. After the *POs* have reached to the maximum value, this will be reported as a *Paging* failure.

2. The UE is being located in a poor radio coverage area—it cannot receive *Paging* messages with sufficient signal strength to resolve the PCCH information. This is also reported as a *Paging* failure.

3. While a UE in the ECM-IDLE state is performing a cell (re)selection or a *TAU* procedure, the MME cannot receive a *Paging* response from that UE. Hence, the MME will report this case as a *Paging* failure.

Obviously, improving the *TAU* procedure (i.e., optimizing the size of the TAs) would save the UE's battery and reduce the relevant signaling cost. Keeping the *Paging* delay and *Paging* cost at lower values can reduce not only the overall signaling overhead but also the bandwidth usage. The following sections review some solutions to reduce the *TAU* and *Paging* signaling overhead and/or improve the UE experience (i.e., optimizing power consumption in the UEs).

## 2.4 *TAU* and *Paging* Overhead

As we have seen in Section 1.1, the *TAU* and *Paging* procedures contribute over 34% of the total signaling overhead on the MME (illustrated in Figure 1.2). In addition, the *TAU* procedure consumes over 10 mW of the UE battery power—the UE frequently initiates *TAU* while moving. In this context, we detail the multiple impacts of these procedures in the following subsections.

### 2.4.1 Modeling of Signaling Cost

Based on the 3GPP LTE specifications [2, 3] and the illustration of [1], we introduce Table 2.2 to show the signaling overhead of *TAU* and *Paging* corresponding each involved network entity (network entities are shown in Figure 2.1), measured in the number of required messages $M$ for each network element. For example, for *TAU*, when a mobile UE needs to trigger this process,

**Table 2.2:** Signaling cost of *TAU* and *Paging* in *M* (adapted from [1])

| **Network element** | *TAU* (*M*) | *Paging (M)* |
|---|---|---|
| UE | 9 | 5 |
| eNB | 14 | 8 |
| MME | 7 | 8 |
| S-GW | 4 | 6 |
| P-GW | 4 | 4 |
| PCRF | 2 | 2 |

reporting its location change, the total cost will be 40 *M* per *TAU*. For *Paging*, when a network needs to trigger this process, delivering UE-specific data packets, the total cost will be 33 *M* per *Paging*.

As we have discussed earlier, the *TAU* overhead is a function of how a UE moves throughout the network coverage area. In addition, the *TAU* overhead is inversely proportional to the size of the corresponding TAL—the larger the TAL size, the smaller the number of *TAUs*. Moreover, the *Paging* overhead is proportional to the rate of incoming data packets, which are designated to specific UEs. Furthermore, as we have stated before, these two procedures give rise to a trade-off optimization problem (because of the dependency). In this context, we can calculate the total cost (detailed later in Section 4.7), denoted by $C_{\text{tot}}$, for the combined *TAU* and *Paging* overhead, using the following expression (adapted from formula (4.6)):

$$C_{\text{tot}} = C_{\text{tau}} \cdot \lambda + (1 + \frac{\alpha}{\sigma}) \cdot C_{\text{pag}} \cdot N_{\text{TAL}} \cdot \sigma, \tag{2.1}$$

where $C_{\text{tau}}$ and $C_{\text{pag}}$ are the relevant message overhead of *TAU* and *Paging*, respectively (detailed in Table 2.2), $N_{\text{TAL}}$ is the TAL size (i.e., number of eNBs in the relevant TAL, detailed in Section 1.1), and $\lambda$ is the rate of triggering *TAU*. As detailed in Section 2.3.2, the MME always triggers the required *Paging* procedure whenever there are incoming data packets intended for a certain UE. This process is captured by $\sigma$, which refers to the rate of triggering the *Paging* procedure. After trigging the *Paging* procedure (which is intended to localize a certain UE), the MME waits to receive a *Paging* response from the intended UE (via its serving eNB) within a predefined time

limit (specified by the network operators). If the UE cannot respond in time to the first incoming *Paging* message (for the reasons described in items **1**, **2**, and **3** of Section 2.3.2.2), the MME starts sending multiple *Paging* attempts. These are captured by $\alpha$, which refers to the rate of the *Paging* attempts. The parameter $\alpha$ also influences the *Paging* delay—the more the *Paging* attempts, the higher the *Paging* delay. If the UE responds to the first incoming *Paging* message, there is no need for subsequent *Paging* attempts. Typically, network operator sets a maximum number of allowable *Paging* attempts. If the intended UE does not respond even after this maximum number of *Paging* attempts is reached, this will be counted toward the *Paging Failure Rate* KPI.

### 2.4.2 Modeling of UE Battery Power Cost

From the preceding discussion, as long as UEs move throughout the network coverage area, each UE must report its location change, initiating the required *TAU* procedure. This consumes not only network resources but also battery power in these UEs, which we detail here. Each time the UE starts the *TAU* procedure, about 10 mW will be consumed from the UE's battery. Moreover, as stated before, the triggering rate of *TAU* depends on how a specific UE moves via the network coverage area—that is, the more frequent the initiations of *TAU*, the more the battery power drains. In this context, we introduce the following expression to model the total battery consumption, denoted by $Apwr_{\text{tau}}$ (adapted from formula (4.1)):

$$Apwr_{\text{tau}} = Pwr_{\text{tau}} \cdot \lambda, \tag{2.2}$$

where $Pwr_{\text{tau}} = 10$ mW (the average battery consumption per *TAU*, according to [4, 47]) and $\lambda$ is the rate of triggering *TAU*. From this formula, it is clear that if the UEs are very mobile, more battery power is consumed in these UEs—this adversely affects the battery lifetime of mobile UEs.

## 2.5 Mobility Management Techniques

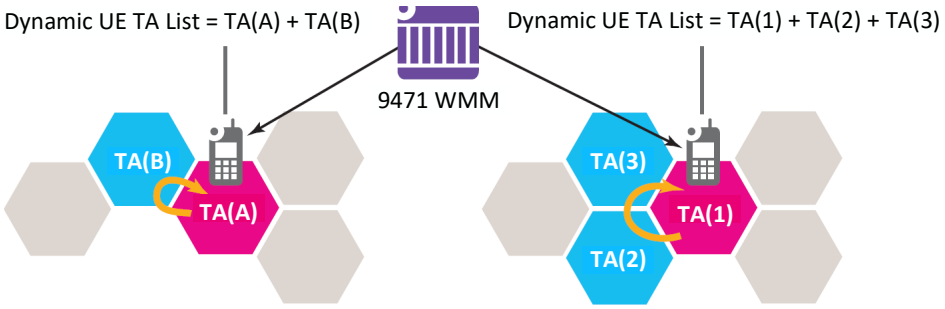
As mentioned earlier, the MME functions (i.e., the MM in LTE networks) enable the network to keep track of all UEs under the E-UTRAN coverage area to deliver data packets and maintain

signaling links between UEs and the E-UTRAN. *TAU* and *Paging* are the two procedures that MME uses to track the locations of all UEs and locate the corresponding serving cells, respectively.

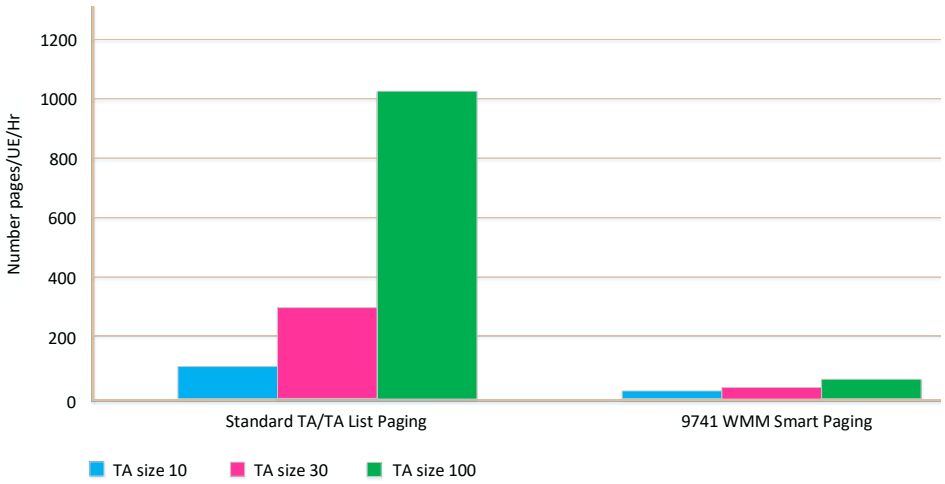
The controversial issue here is that the *TAU* and *Paging* message overhead leads to a trade-off optimization problem. Making a TA contain a large number of cells reduces the signaling load associated with the *TAU* procedure. Hence, the power consumption in the UEs is reduced. On the other hand, because the UE is paged through all the cells in the assigned TAL, the *Paging* overhead increases accordingly. In addition, making a TA contain a small number of cells reduces the *Paging* load, but increases both the *TAU* load and battery power consumption in the UEs [48]. In other words, a small size of the TAs causes an individual UE to trigger *TAU* requests more frequently. This process, as mentioned earlier, drains about 10 mW of battery power in current-generation smart-phones [4, 47]. When the UE crosses the boarder between the TAs that do not belong to its TAL, the *TAU* load becomes extremely high, producing excessive signaling of the *TAU* requests. This is known as "toggling" or "ping-pong" effect [26, 49]. To make that clear, we consider the following relevant example:

Alcatel-Lucent introduced a smart *Paging* and dynamic TAL management scheme to significantly reduce the signaling load on the MME [47]. This technique is called *"Alcatel-Lucent 9471 Wireless Mobility Manager (WMM)"* and is intended to help LNOs minimize the signaling load on the MME, as shown in Figure 2.4. Two different scenarios are shown in Figure 2.4(a), in which when a UE moves in a circular pattern between two TAs (e.g., TA(A) and TA(B)) or between three different TAs (e.g., TA(1), TA(2), and TA(3)). In the first scenario, the UE is currently located within TA(A) but before has been in TA(B). In the second scenario, the UE is currently registered in TA(1) but the *"9471 WMM"* has detected that the UE has a circular movement pattern between TA(1), TA(2), and TA(3). Once observing such a movement pattern, the *"9471 WMM"* will send a new TAL to the corresponding UE, comprising all TAs involved in the circular pattern.

The *Paging* load increases significantly when the TA size is large in the standard *TAL* (one TAL comprises up to 16 TAs, and one TA comprises up to 100 eNBs). By using this technique, *"9471 WMM,"* the *Paging* load can be reduced to a lower rate, as shown in Figure 2.4(b). At the same




**(a)** Dynamic TA management




**(b)** Smart signaling paging benefits

**Figure 2.4:** Smart *Paging* and *TA* management to reduce signaling load (adapted from [4])

time, the "toggling" issue can also be reduced by using dynamic *TAL*, as illustrated in Figure 2.4(a), where the TAL is assigned dynamically to each individual UE according to its movement between the last visited TAs.

Recently, many researchers have proposed schemes for MM planning to find reasonable solutions to mitigate signaling loads on the MME. Generally, some of these schemes focus on *TAU* and the others focus on *Paging*. These techniques are highly related to each other because if the UE can report its location precisely (i.e., its serving cell), the MME can page that UE (i.e., its reachability) directly in one *Paging* attempt (i.e., very low *Paging* cost). Therefore, careful TAL planing would significantly reduce the consequent *Paging* procedure because the latter uses the last updated TAL to page a specific UE (as described in Section 2.3.2.2). In this context, we classify these solution

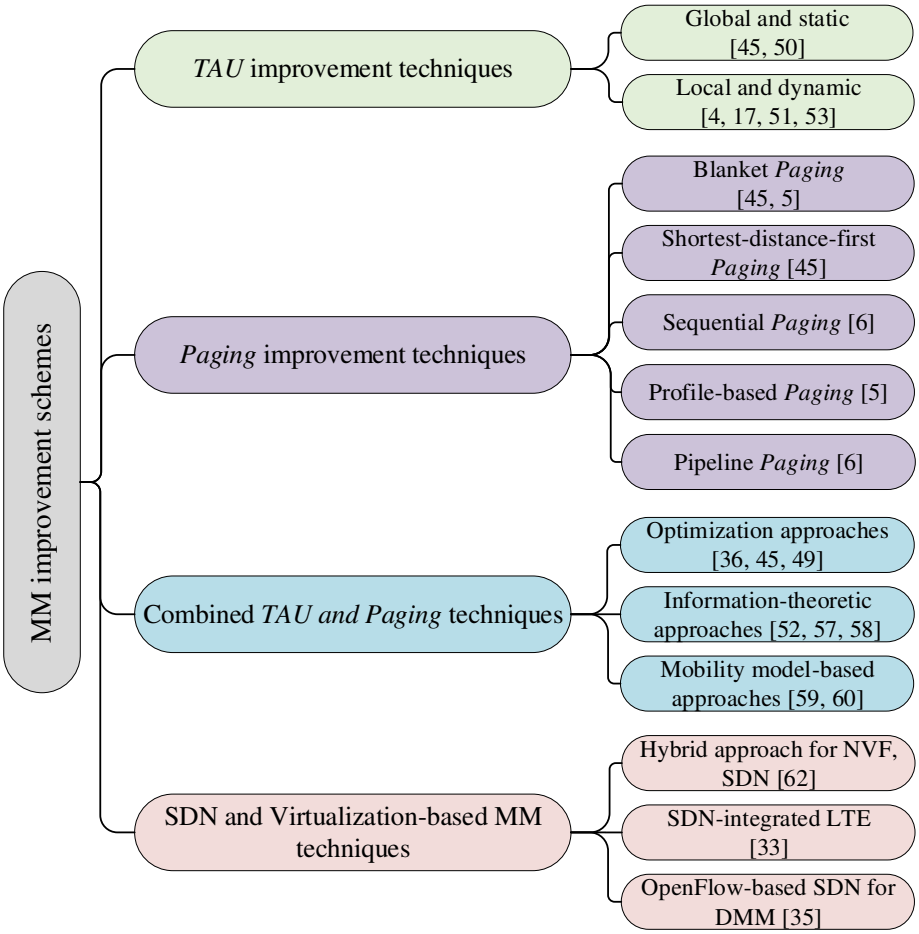



**Figure 2.5:** Classification of *TAU* and *Paging* solution schemes

schemes, which this chapter discusses, as shown in Figure 2.5. The following subsections describe the proposed *TAU* and *Paging* schemes.

### 2.5.1 *TAU* Improvement Techniques

These techniques can be classified according to whether or not they account for individual UE behavior in terms of its movement pattern or traffic characteristic, as follows.

#### 2.5.1.1 Global and static techniques for *TAU*

In these techniques, the TA size is fixed and configured at the time of network planning. In other words, all UEs within a certain area have the same TAL, and hence have the same *Paging*

Area (PA). Some common schemes that can be classified with this category are: *Reporting cell*, *Never update*, and *Always update* [45, 50]. Practically, such techniques are no longer used in current LTE networks because of the following pitfalls [18]:

1. These schemes are costly because they do not take into account the individual UE behavior. They produce excessive *TAU*, *Paging* messages, or even both.

2. These schemes have no way to reduce the "toggling" effect, especially when a group of UEs cross the TA boarder back and forth at the same time.

3. These schemes could generate uneven signaling distribution when a massive number of UEs cross the boarder of a TA that is not in their TAL. This process triggers a huge number of *TAU* messages causing signaling congestion in that area.

#### 2.5.1.2 Local and dynamic techniques for *TAU*

In these techniques, the TA is not fixed beforehand. Instead, they account for individual UE moving patterns and traffic characteristics. As mentioned earlier, LTE adopts a dynamic *TAU* in which the center of the assigned TAL is close to the UE's current location (i.e., neighboring eNBs), also called the *Central Policy* [2]. In this policy, instead of all UEs having the same TAL, each individual UE is assigned a specific TAL based on its location in the network. Thus, the overall *TAU* and *Paging* loads can be reduced, as depicted in the example of Figure 2.4. The dynamic schemes can mitigate the drawbacks of the global and static schemes. However, the *Central Policy* can generate a negative impact on the network if the UEs have been allocated TALs which are irrelevant to their mobility and traffic characteristics [18]. Therefore, many algorithms have been proposed to reduce the signaling loads on the MME, which we can classify according to how the *TAU* is triggered:

1. Each UE performs *TAU* process independently: Each individual UE triggers a *TAU* procedure after a specific threshold. The authors of [51] proposed three strategies in which UEs can update their location change while moving throughout the network (i.e., trigger the *TAU*

procedure). These strategies are: time-based, movements-based, and distance-based. For a performance analysis, they assume a memoryless and a Markovian motion model, and consider the cellular topology as a ring. According to the authors of [51], the distance-based strategy performs better than the other two when assuming a memoryless movement pattern. When using a Markovian model, the authors show that for some movement patterns the time-based strategies outperforms the movement-based. So, such strategies are mostly dependent on the UE mobility patterns (we highlight this issue in Section 2.5.3.3).

Such techniques, as expected, are difficult to apply to future wireless networks (applied to limited network topologies). They are cost ineffective because of the rapid increase in both the number of UEs and their mobility (e.g., cost much more bandwidth). Furthermore, these schemes are basically heuristic and are often far from optimal [52]. In some cases, the network cell topology should be available to the UEs (i.e., the UEs should store information about how eNBs are distributed), which is impractical in real wireless networks [17].

2. A single *TAU* process is performed for a group of UEs: The *TAU* procedure is initiated based on a group behavior of mobility—that is, if there is a similarity in the mobility patterns among a group of UEs in the same geographic area, one of the UEs can trigger a *TAU* message instead of many messages being sent simultaneously (from multiple UEs). The authors of [53] introduced a Group MM (GMM) in which UEs are grouped according to their mobility correlations; one UE initiates the *TAU* procedure on behalf of others (i.e., in the same group) in a certain area, aiming at mitigating the corresponding signaling overhead. In this method, a history of UE paths is stored in what is called a Location Database (LDB) for the purpose of finding the correlated mobility among UEs. Once the correlations are determined, one UE is defined as a leader for each mobility group, which is responsible for triggering *TAU* on behalf of all UE members in the corresponding group. For the purpose of simulation, the authors have used three mobility models for the UEs: *random walk, bioinspired mobility, and transportation mobility models*.

According to this study, the performance trend of the random-walk and bioinspired model are similar while the transportation mobility model performs better by saving most of the signaling cost, reducing the *TAU* load that comes from UE location updates. Also, this work can support Internet-of-Things (IoT) applications and M2M communications because these are likely have a correlated mobility, which can significantly reduce the signaling cost from the simultaneous *TAUs*. The GMM can support increases in UE density and save network resources (e.g., save bandwidth) for UEs that have correlated mobility patterns. However, the GMM requires maintaining a history for each individual UE to identify which group of UEs have a correlated mobility, and this process would increase the computation/storage overhead to identify such similar UE groups. Also, when it comes to the *Paging* procedure and because the latter depends on the *TAU* (i.e., *TAL*), the GMM may increase the *Paging Failure Rate* KPI. In other words, when some UEs in the group have different mobility patterns than the leader UE, the stored LDB (movement paths) for these UEs will be uncorrelated, resulting in UE location error and *Paging* failure.

3. Forming TALs adaptively: Unlike the solution in [53], the authors of [1] introduced a solution in which the TAL is allocated adaptively for each UE—that is, each UE initiates a *TAU* adaptively according to its allocated TAL. This scheme is intended to support automotive users or connected cars while mitigating the corresponding signaling overhead (because of the high mobility, these UEs can generate high volume of *TAU* signaling). According to [1], TALs are formed adaptively as rings and sectors of different sizes (for urban and rural areas) that depend on the most probable movement angle; two different cell sizes are deployed based on known geographic area, position, speed, acceleration, and heading information.

   Based on a Markov model for mobility prediction and variable TAL forms, a reduction of 33% of signaling overhead for MME as compared with *TAU* based on movements and distance strategies, as proposed by [51]. The used Markov model allows allocation of TAs in more tailed shapes as sectors of rings instead of full rings. Although this approach achieves significant savings in the MM signaling overhead, these savings come at expense of higher

**Table 2.3:** *TAU* scheme comparisons

| | **Bandwidth/ Signaling** | ***Paging* success rate KPI** | **Toggling effect** | **UE battery power saving** | **High-mobility UEs** | **Network topology** | **Mobility prediction** | **Future IoT/5G** |
|---|---|---|---|---|---|---|---|---|
| **UEs trigger *TAU* independently [51, 52]** | no bandwidth conservation; need dedicated resources for each UE | cannot support *Paging*; the exact serving cell is unknown | provide no toggling effect | no UE battery power saving | increase signaling overhead at high-mobility UEs | no variety of network support. Cell topology may be needed | no mobility prediction considered | cannot support future network; much more bandwidth needed |
| **UEs trigger *TAU* per groups [53]** | better bandwidth utilization; each UE group needs one *TAU* | improve *Paging*; UEs are known at TAL group level | improve toggling effect | improve UE battery power saving | support vehicular and M2M comm. | limited to some types of networks | mobility models considered | support IoT and M2M comm. |
| **UEs trigger *TAU* adaptively [1,51]** | 33% signaling reduction than independent *TAU*, bandwidth saving | improve *Paging*; UEs are known at TAL group level | improve toggling effect | improve UE battery power saving | support vehicular and M2M comm. | support vehicular and M2M comm. | mobility models considered | support IoT and M2M comm. |

complexities. Also, the method used for mobility prediction gives accurate prediction only for about 70% of the cases. In this case, however, the assigned TAL may not coincide with the UE movements (i.e., mobility pattern), and hence this produces adverse effects by increasing the related TAU signaling overhead.

At this point, we highlight some important performance comparisons for the above *TAU* schemes in Table 2.3.

### 2.5.2 *Paging* Improvement Techniques

The *Paging* process is to search for a specific UE in the network and is related to how the UE reports its location while moving throughout the network. Basically, The *Paging* uses the last-updated TAL (for a specific UE) to trigger *Paging* messages throughout this TAL (the MME always initiates *Paging* messages). Common schemes can be generally classified as follows:

1. Blanket *Paging* (BP): This is also called broadcast or simultaneous *Paging* [5, 45]. In the BP scheme, all TAs in a UE's TAL are paged simultaneously. Although this scheme is widely used, the bandwidth utilization is ineffective; BP needs very high bandwidth because multiple cells are being paged at the same time. This may decrease the *Paging Success Rate* KPI due to peak hours of traffic [5].

2. Shortest-Distance-First (SDF) *Paging* (SDFP): This scheme starts by sending a *Paging* message to the last serving cell where the UE has triggered the *TAU* message and then tries the other cells based on the SDF order [45]. This scheme is difficult to implement in practice because the neighboring cells are difficult to determine dynamically (when the coverage area changes, the neighboring cells may also change [17]). Also, in commercial LTE system, the network has no information about the last serving cell [2].

3. Sequential *Paging* (SP): This scheme starts by sending *Paging* messages to a group of cells (i.e., PA) where the probability of locating a specific UE in that PA is expected to be high. These PAs are paged sequentially in descending order according to their probabilities. Although this scheme reduces the network congestion as compared with the BP, it suffers from a high implementation cost, requiring storage of *Paging* profiles for all UEs, and results in increased *Paging* delay [6]—that is, because the PAs are being paged one by one until the aimed UE is located. Hence, this scheme cannot handle increases in traffic load and may lead to a lower *Paging Success Rate* KPI.

4. Profile-based *Paging* (PBP): According to [5], this scheme is intended to improve both the *Paging Success Rate* KPI and bandwidth utilization, taking into account both the probability of locating a specific UE in a certain PA (as in the SP scheme) and the mobility pattern of the UE. This scheme extracts the UE movements and improves the *Paging* process. As a result, the PBP achieves 9% more *Paging Success Rate* than the BP scheme and 3% more than SP. Moreover, based on the results shown by the authors of [5], we produce Figure 2.6 to show that this scheme has the lowest bandwidth usage compared to BP and SP. Improving both *Paging Success Rate* and bandwidth utilization is crucial to allowing networks to support the tremendous increase in UE density per unit area. Also, the PBP scheme works well with UEs that exhibit periodicity in their movements. However, mobility patterns of UEs need to be captured for the purpose of location estimation, which increases the computation/storage overhead.

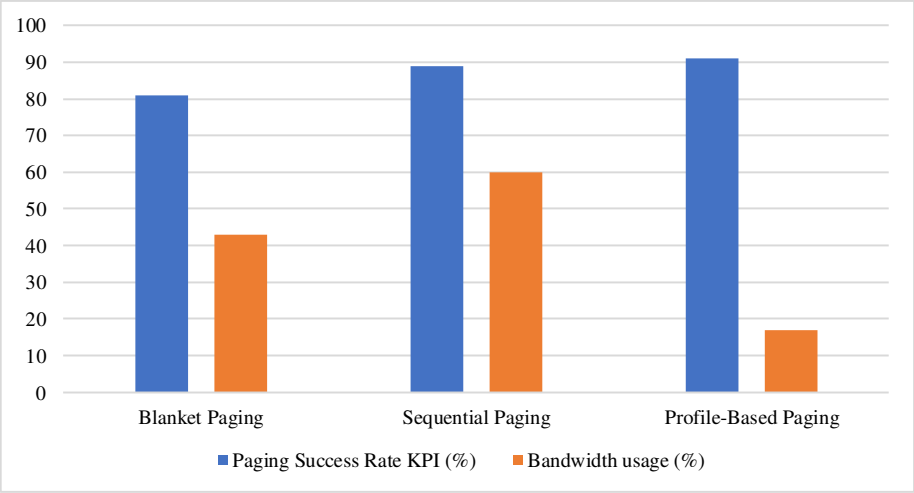



**Figure 2.6:** Comparison of *Paging Success Rate* and *Bandwidth utilization* (adapted from [5])

5. Pipeline *Paging* (PP): Unlike the SP scheme, the authors of [6] introduced a PP scheme, assuming no prior knowledge about the probability of presence of a specific UE in a certain PA. Instead, multiple fixed-size PAs are paged for all UEs in a pipeline manner (the TAs are divided into fixed PAs). According to [6], PP performs better than SP and BP in terms of different metrics. The *Paging* delay/cost is reduced because multiple UEs can be paged in a parallel way in the same cell or PA. Based on the results shown by the authors of [6], we produce Figure 2.7 to show that the PP and SP schemes have the same behavior in their *Paging* cost, which decreases as the *Paging* delay bound increases and is lower than the *Paging* cost for the BP scheme. In addition, the PP scheme can handle more *Paging* requests than the SP scheme.

   Although the PP scheme has some advantages over BP and SP in terms of *Paging* delay/cost and bandwidth utilization, PP still has some implementation costs when applied to large-scale networks (could increase the *Paging* load when the traffic load is high). Furthermore, because the *Paging* delay and cost are as functions of traffic load, this scheme cannot handle the expected exponential growth of mobile UEs especially at peak hours.

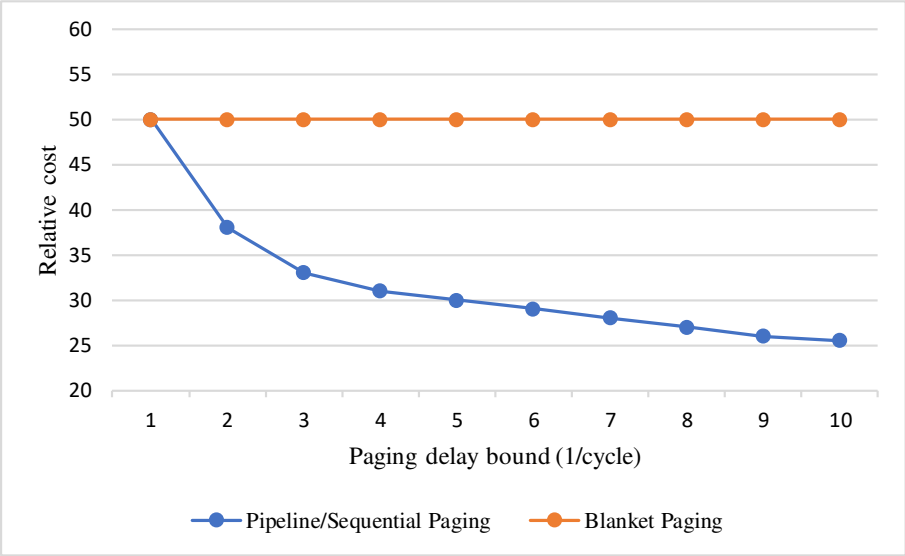



**Figure 2.7:** Relative cost comparisons for BP, SP, and PP schemes (adapted from [6])

6. PA-based *Paging* (PAP): To overcome the drawbacks in the PP scheme (which considers fixed-size PAs for *Paging* messages), the authors of [54] proposed variable-size PAs, aiming to provide lightweight *Paging* signaling load than the ordinary *Paging*, in which the PAs are pre-defined. In this scheme, the MME, which is responsible for initiating *Paging* messages, selects PAs based on the Elapsed Time (ET) of the last location update (i.e., last *TAU*) of UEs (assuming that when the ET value is high, the UE would be farther away from the last *TAU*).

   The MME stores a time stamp of the last triggered *TAU* procedure (last location update) for each UEs and the *Paging* log where the *Paging* succeeds—that is, the last serving cell (including eNB, TA, and TAL) is saved. Accordingly, three sizes of PAs are defined: last-eNB, last-TA, and last-TAL (small PA, medium PA, and large PA). The MME selects the appropriate size for PA according to the ET value. For example, when the ET value is low (as compared with a pre-defined threshold), the MME chooses the small PA to page the corresponding UE. If this PA reports a *Paging* failure, the larger PA will be used for a second attempt of *Paging*. In [54], the authors used a random way-point model to simulate

the UE mobility. They show that the *Paging* messages are lower by 63% than the ordinary one. However, because this scheme depends mainly on the last serving cell, it adds some implementation difficulties. According to [2], in commercial LTE systems, the network has no information about the last serving cell. Moreover, the authors of [54] calculated the ET threshold based on the UE movements, which are governed by a single mobility model. This would not reflect the mobility behaviors of UEs in practical cases.

7. Call Data Record-based *Paging* (CDRP): This scheme is quite similar to PBP scheme, but adds some modifications. The authors of [55] proposed the CDRP scheme aiming to reduce the *Paging* resources. This scheme extracts the history of UE mobility behavior from the past knowledge of UE movement integrating with sets of CDR for UEs. Thus, according to [55], this adds more predication accuracy to locate the intended UEs while reducing the corresponding *Paging* resources. For this purposes, large-scale CDRs and location information of UEs are collected on a daily basis, developing a database to profile each UE. Then, the aimed UE is paged using profile-based paging (similar to PBP). The authors show 4 to 5 times better performance than the conventional one. However, this scheme is limited in terms of UE density, and adds some computational/storage overhead (requiring maintenance of a history for each individual UE).

In sum, we have summarized the most important metrics about these *Paging* solutions in Table 2.4.

### 2.5.3 Combined *TAU* and *Paging* Techniques

The preceding solution schemes (in terms of *TAU* and *Paging* as in Section 2.5.1 and Section 2.5.2, respectively) are typically studied separately. However, some studies have proposed combined solutions for both the *TAU* and *Paging* problems. In this context, the combined solutions in the literature can be classified according to approaches that are used to solve this problem, which we describe below.

**Table 2.4:** *Paging* scheme comparisons

| | **Bandwidth/ Signaling** | ***Paging* delay/ attempt** | **Mobility prediction** | **Future IoT/5G** |
|---|---|---|---|---|
| **Blanket *Paging* (BP) [5,45]** | low bandwidth utilization, high signaling overhead | high *Paging* delay, low *Paging Success Rate* at peak hours | no UE mobility prediction | cannot support future networks; need much more bandwidth, no support for real-time applications |
| **Shortest-Distance-First *Paging* (SDFP) [45]** | low bandwidth utilization, high signaling overhead | high *Paging* delay/attempt | UE mobility prediction is needed | may be used, but need information about last UE serving cell |
| **Sequential *Paging* (SP) [6]** | improve bandwidth utilization | high *Paging* delay, needs sequential *Paging* | UE mobility prediction is needed | cannot support future networks; high *Paging* delay, no support for real-time applications |
| **Profile-based *Paging* (PBP) [5]** | improve bandwidth utilization | improve *Paging* delay and *Paging Success Rate* than SP and BP | UE mobility prediction is needed | may be used, especially when UEs show periodic mobility behavior |
| **Pipeline *Paging* (PP) [6]** | no bandwidth saving; multiple PAs are paged simultaneously | improve *Paging* delay/attempt than SP and BP | no UE mobility prediction | cannot support future networks; may increase *Paging* load at peak hours |
| **PA-based *Paging* (PAP) [54]** | lightweight *Paging* than ordinary *Paging* (no fixed PA size as in PP) | improve *Paging* delay/attempt than ordinary *Paging* | UE mobility prediction is needed | may be used, but need information about last UE serving cell |
| **Call Data Record-based *Paging* (CDRP) [55]** | improve bandwidth utilization (quite similar to PBP) | improve *Paging* delay/attempt than ordinary *Paging* | UE mobility prediction is needed and large-scale of CDR | cannot support future networks; cannot handle high-density UEs |

#### 2.5.3.1 Optimization approaches

These approaches consider *TAU* and *Paging* overhead as a multi-objective optimization problem to find feasible solutions to the trade-off between the *TAU* and *Paging* costs [36,45,49], which are discussed as follows:

1. In [45], the author formulated this problem as an integer programming problem to provide a set of Pareto-optimal solutions. Basically, this solution conducts TAL re-optimization. That means that the TAL size is adapted according to a budget constraint. In this case, some cells will be allocated to different TAL (i.e., the given TALs may be divided or combined) such that the *TAU* and *Paging* signaling overhead is minimized. According to the simulation results, the overall signaling cost of the designed TAL is 49–56% better than the optimal standard TAL, resulting in reducing the *Paging* cost between 67–73% compared to the values

obtaining from the standard TAL setup. The integer programming model may have to be run many times, depending on the number of these solutions, and hence this process is time-consuming and sometimes leads to infeasible solutions for large-scale networks. Therefore, the author uses a genetic algorithm embedded with local search, but still this process is also time-consuming.

Despite the significant improvement in the signaling overhead, it is difficult to apply this approach to large-scale networks (i.e., the solutions vary according to the network topology and may increase the infeasible solutions). Moreover, although the author of [45] proposed different algorithms to solve this optimization problem, what is lacking is a comparison between his proposed solution and other related studies in this context.

2. In [49], the authors proposed an Evolutionary Multi-objective Optimization (EMO) algorithm based on a Population Decomposition Strategy (PDS) (see [56] for more detail about the PDS). This study has two different contributions. First, the authors build a multi-objective TA planning model by integrating the geographic information that can directly affect an individual UE's mobility. For example, areas with many roads experience more user mobility than areas blocked by high mountains or barriers. This model can provide a set of trade-off solutions for TA planning (according to the authors, multiple trade-off solutions can be obtained in a single run), and thus give more options to the decision makers. Second, the authors design a new PDS-based EMO algorithm for the proposed TA planning model. This trend tends to merge several small TAs into a big one and split a big TA into several small ones. The PDS-based EMO algorithm is designed to make better use of the information about not only the feasible but also infeasible solutions.

   This approach potentially reduces both *TAU* and *Paging* costs than in a single-objective TA planning model. The initialization process of the EMO uses fuzzy clustering based on the geographic information, and hence the computational complexity is still significantly high (as detailed in [49]).

3. The optimization schemes above introduce *TAU/Paging* trade-off solutions to find better cell allocations to the TAs/TALs (i.e., cell mapping) such that the *TAU/Paging* overhead is minimized, not including the MME itself in the optimization problem. However, the authors of [36] included not only cell-TAs/TALs but also TAs/TALs-MME mappings. In this scheme, the authors exploit the concept of MME pooling that is already introduced in LTE to ensure equally loaded MMEs within an MME pool area (the MME load balancing functionality is not discussed further here; see [2] for more detail). In [36], the authors introduced two MME pooling schemes, which are as follows Centralized/Distributed MME (C/DMME). The CMME scheme, according to the authors, allows one MME to control one TAL while the DMME scheme allows one MME to control one TA, assuming the TA comprises one cell (one eNB) and the TAL comprises a set of TAs. The authors of [36] showed that the CMME scheme outperforms the DMME because of the additional costs resulting from multiple TA relocation in DMME scheme. Also, in this study, the *TAU/Paging* signaling overhead is minimized in the CMME scheme.

   However, in the above study, the authors drew their conclusions based on certain simulation assumptions (e.g., assuming 10 cells, UEs uniformly distributed with an average of 100, and TALs vary between 3 and 4), which might be insufficient to apply their results to large-scale networks. Moreover, it would be beneficial to consider a mobility model for moving UEs to show how the *TAU/Paging* overhead changes while UEs are moving. Also, this study should consider a traffic load (i.e., cell load) to make it more practical.

#### 2.5.3.2 Information-theoretic approaches

Some schemes rely on information-theoretic frameworks (using Shannon's entropy) to deal with the trade-off between the combination of *TAU/Paging* costs and storage/computational overhead [52, 57, 58], which are discussed as follows:

1. In [52], the authors proposed two entropy-based frameworks: Bayesian-based *TAU* and entropy-coding based *TAU*. In these frameworks, UE mobility patterns are collected online to perform profile-based *Paging* and improve the *TAU* overhead. The first one reduces

the *Paging* overhead and has less storage and computational cost. The second one minimizes *TAU/Paging* overhead, but needs higher storage and computational overhead than the first one. Hence, these approaches have some trade-off between both the *TAU/Paging* and storage/computational costs. The authors of [52] showed that the Bayesian-based and entropy-coding based *TAUs* reduce signaling load by about 60% and 80%, respectively, than the existing *TAU* in LTE. Hence, the entropy-coding based *TAU* performs better than the first one in terms of reducing update overhead. But the entropy-coding based *TAU* exhibits much more computational load (almost 50%) than the Bayesian-based *TAU*. Moreover, the two frameworks give rise to additional computational overhead to UEs themselves, which are limited in terms of battery power and processing capability.

2. In [57], the authors proposed spatial and temporal quantizations for UE tracking schemes, to deal with the trade-off between *TAU* and *Paging* overhead, assuming no prior knowledge about UE mobility patterns (mobility-model independent). The authors showed that the *TAU* cost can be reduced at the expense of increasing the corresponding *Paging* cost, but maintains low computational and storage overhead as compared with the solution in [52]. The proposed schemes can reduce *TAU* frequency to 3–4 updates/day.

3. In [58], the authors used Shannon's entropy to predict an individual UE mobility path by designing an adaptive on-line algorithm for tracking the UE's movement to reduce the *Paging* cost. However, such scheme is prone to a high computational/storage cost owing to building and maintaining a dictionary for all UE mobility patterns. The above Shannon's entropy solutions are not taking into account the *Paging* latency and still have either high computational or storage overhead.

To make these solutions more realistic, it would be preferable to minimize *Paging* latency, *TAU/Paging* overhead, and storage/computational cost, optimizing all these parameters at the same time.

#### 2.5.3.3 Mobility model-based approaches

Mobility models are built to predict a UE movement pattern, providing information about UE location changes such that the *TAU* and *Paging* signaling overhead can be reduced [59, 60], which are discussed as follows:

1. In [59], the authors proposed a User Mobility Pattern (UMP) scheme for *TAU* and *Paging* operations based on a User Mobility History (UMH). During *TAU*, a UE derives the expected UMP from its UMH and registers this information to a database in the network. In this case, the UE does not need to trigger the *TAU* while moving in its registered UMP. In addition, when data packets arrive for a specific UE, the cells are paged sequentially starting from the cell where the UE is expected to be located based on the registered UMP.

2. Unlike the above solution, the authors of [60] presented a framework to predict a UE's mobility (traveling trajectory and destination) by analyzing the UE contextual information (i.e., to estimate the UE's location) without taking into account the UE mobility history. The primary goal behind the above mobility prediction techniques is to reduce the *Paging* cost by assigning the best TAL that is consistent with the UE's movement.

In sum, we have summarized the combined schemes for *TAU* and *Paging* in Table 2.5. Although these schemes have been applied in some cases, they are practically ineffective because of being very costly in terms of storage capacity and implementation complexity. Because of the tremendous increase in the number of mobile UEs, it becomes more difficult to apply such solutions in large-scale networks. Moreover, when it comes to 5G use cases, a huge number of highly mobile UEs would trigger control messages simultaneously that current LTE networks fail to handle (the network becomes more congested). Hence, it will be very challenging to apply such solution techniques in 5G (we highlight these issues in Chapter 3).

**Table 2.5:** Trade-off between *TAU* and *Paging* scheme comparisons

| | Trade-off method used | Pros | Cons |
|---|---|---|---|
| **Optimization approaches [36, 45, 49]** | -Pareto-optimal-based integral programming | -reduce *TAU/Paging* signaling cost.<br>-help to optimize TAL sizes. | -not applied to large-scale networks.<br>-limited to some types of networks.<br>-time consuming.<br>-no UE mobility prediction. |
| | -PSD-based EMO algorithm | -reduce *TAU/Paging* signaling cost.<br>-integrating the geographic information of the network.<br>-give more planning solutions to choose from. | -computational complexity is significantly high.<br>-no UE mobility prediction. |
| | -C/DMME pooling | -reduce the *TAU/Paging* signaling cost.<br>-balancing TAL between MME pools.<br>-offer load distribution. | -simulation results are insufficient.<br>-not applied to large-scale networks.<br>-no UE mobility prediction. |
| **Information-theoretic approaches [52, 57, 58]** | -Bayesian-based | -improve *Paging* overhead.<br>-reduce update cost by 60% as compared to exist *TAU*.<br>-less computational and storage costs. | -the proposed scheme increases the complexity of UEs. |
| | -Entropy-coding | -minimize *TAU/Paging* overhead.<br>-reduce update cost by 80% as compared to exist *TAU*. | -the proposed scheme increases the complexity of UEs.<br>-add more computational than Bayesian-based.<br>-higher computational and storage costs. |
| | -Spatial and temporal quantizations | -reduce *TAU* cost by 3-4 update/day.<br>-use real data for UE traces.<br>-low computational and storage costs. | -no UE mobility prediction.<br>-increase *Paging* cost. |
| **Mobility model-based approaches [59, 60]** | -UE mobility prediction based on: history (i.e., UMP) or contextual information | -improve *Paging* cost compared to SP and BP.<br>-create less *TAU* signaling compared with other schemes. | -need to store UE mobility history over time.<br>-high computational and storage costs. |

## 2.6 SDN and Virtualization-based MM in LTE

Apart from the previous solution schemes that try to mitigate (or manage) the load on the MME which mostly comes from *TAU/Paging* signaling overhead (Section 2.5), a new network paradigm has been introduced to control and manage problems of the heavy traffic load in the whole network, including the MME load, which is known as SDN and Virtualization (SDNV) [33]. Here we will pay some attention to such techniques, which aim to deal with the overall network load, and describe briefly this trend (see [33] for more detail). In other words, for example, the

SDNV is proposed not to solve the trade-off of the *TAU/Paging* issues or to allocate a best TAL to UEs; instead, it is introduced to meet the significant growth in mobile traffic, mitigating the overall signaling overhead in the network. In terms of traffic management, for example, traffic offloading and load balancing (by using the SDNV) have been introduced as an efficient scheme to optimize network resources, mitigate network congestion, and handle the rapid increase in traffic demand, which in turn maximizes UE experiences.

According to [33], the SDN idea is intended to separate the control and data planes, giving rise to a programmable network, and virtualization enables network infrastructure sharing and the "softwarization" of the network functions. More specifically, SDN facilitates network configuration and management by shifting the signaling loads to a SDN controller (i.e., a centralized controller). Virtualization comprises two technologies, which are Network Virtualization (NV) and Network Function Virtualization (NFV). The NV allows many different virtual networks to be served by the same network infrastructure. In the NVF, network functions are implemented as software running on general purpose computing/storage platform (for more detail about SDN and NFV architecture, see [33]).

SDNV will be a major trend in 5G systems. For example, the authors of [61] reviewed the main 5G trends in terms of SDN and NFV architecture design. Specifically, 5G will exploit the SDN and NFV principles to provide ever flexible/scalable network management, which is intended to approach close-to-zero latency, accommodating the rapid increase in high-mobility UEs and supporting both mission-critical and real-time applications (the SDNV architecture for 5G is beyond the scope of this dissertation). In this context, we describe some of the SDNV solutions as follows:

1. To take advantage of the SDNV, the authors of [62] studied the effects of integrating this technology on LTE systems and propose a hybrid approach to select whether to apply NFV or SDN technology, formulating the selection decision (i.e., SDN decomposition/NFV Virtualization) as an optimization problem such that the overall network loads are minimized subject to a set of constraints: number of active datacenters, the population of the area under consideration, packet delay budget, and traffic volume. This is intended to offload the LTE

gateways (such as S-GW and P-GW; see Figure 2.1) by steering data traffic of the gateways to datacenters, which are virtualized by SDNV. This solution responds to the dynamic state of the network—at each time slot, it decides whether to apply SDN or NFV on each gateway (S-GW/P-GW), depending on the state of the network. According to [62], the SDN decomposition reduces the network delay while increasing the total network load. On the contrary, the NFV gateway does not increase the network load because there is no additional control layer at the expense of increasing the traffic delay.

2. As small cells are now broadly accepted (addressed in Chapter 3; see [4, 47] for more detail), this causes more signaling overhead between the small cells and the network backhaul (via the MME). To deal with this issue, the authors of [34] introduced a framework for MM in SDN-integrated LTE, in which the backhaul between the S-GW and small cells is implemented as a SDN with QoS differentiation support. In this design, the SDN controller can receive mobility events directly from the MME. Also, this solution proposes a dynamic localized forwarding scheme to support UEs while moving within the small cells for data-packet exchange (i.e., deliver ongoing data traffic). In other words, by using this SDN controller, the data-packet of an ongoing session can be directly exchanged between the source and destination cells without switching the whole forwarding path—that is, the path-switch signaling overhead can be significantly reduced for the SDN-integrated LTE. The authors of [34] showed that the signaling overhead is reduced by 50% relative to traditional path switching (e.g., can mitigate the handover signaling cost). Although this SDN can achieve significant reduction in overall signaling overhead, there is some degradation on the data delivery.

3. Some MM solutions are based on what is called centralized MM, as we have seen in Section 2.5.3.1, item 3. However, such central solutions are prone to some performance pitfalls, such as low scalability, suboptimal routing, and a potential single point of failure. Therefore,

the authors of [35] proposed a SDN/OpenFlow[2] based DMM scheme that can be applied in virtualized LTE networks. Two DMM solutions can be applied known as Full OpenFlow and Partial OpenFlow, which are provided by the OpenFlow protocol. As stated by the authors, choosing a more suitable approach depends on the LNO's investment plan. At this point, in [35], the authors only considered the Partial OpenFlow scheme for simulation and evaluations. Basically, this solution is intended to re-forward the traffic to the current UE serving P-GW such that the average delay of downlink data packets is minimized. For example, according to the simulation results, X2 handover with P-GW relocation shows that the DMM traffic re-forwarding (of the mobile UEs) outperforms the X2 path in terms of downlink data-packet latency. This is to support the continuity of sessions in a seamless manner in case of inter P-GW handover, which shows a significant reduction in average handover latency.

Generally, when it comes to the MME offloading, the above solutions (in terms of SDNV scheme) will help to mitigate/manage the loads on the MME by shifting the signaling load to a SDN controller, which can support the critical requirement of 5G use cases (in this context, see [64] for more detail). In some cases, the current LTE-based SDNV schemes might be used for 5G use cases, but need some modifications to match the new design of 5G systems. To the best of our knowledge, however, the ongoing 5G solution schemes in the literature (in terms of *TAU* and *Paging* management) are still limited (more detail in Chapter 3).

Although these schemes increase the network performance, they do not directly reduce the trade-off signaling overhead that mostly comes from *TAU/Paging* procedures, which directly affect both the user experience and network resources. For example, while moving throughout the network coverage area, UEs are still always triggering the *TAU* procedures more frequently (it is required to report UE location change), which increase battery power consumption in the battery-limited UEs in addition to dedicated resources (e.g., bandwidth) (as stated earlier, each *TAU* procedure consumes over 10 mW of battery power in current-generation smart-phones). This will

[2]OpenFlow is the most common communication protocol used in SDN; see [63] for more detail.

become a crucial issue in 5G because most IoT devices are battery powered—that is, it is desirable to minimize the power consumption on these devices to extend the battery lifetime, making IoT devices rarely (or never) trigger *TAU* procedures while moving.

Furthermore, other network KPIs (such as *Paging* delay and *Paging Success Rate*) are also impacted by the *TAU* because of the fact that while these UEs are busy in responding to the *TAU* procedures, they cannot respond to the incoming *Paging* messages.

## 2.7 Final Remarks

As we have seen from preceding discussion, the state-of-the-art MM (*TAU* and *Paging*) solutions that have been proposed for LTE networks will not be sufficient to achieve the 5G critical requirements, and hence they need to be redesigned accordingly or new solutions need to be developed. In other words, the current *TAU* and *Paging* solution schemes give rise to some challenges when applied to 5G networks. Let's take the following case:

The proposed *TAU* schemes discussed in Sections 2.5.1.1 and 2.5.1.2 for LTE networks are not applicable to 5G networks. The global and static *TAU* schemes have the same drawbacks mentioned in Section 2.5.1.1 and are even worse when applied to 5G use cases. In addition, the local and dynamic *TAU* schemes (Section 2.5.1.2) are more suitable for LTE networks. However, they are still not suitable for 5G networks in terms of computational cost, complexity, latency (from the UE point of view, with limited battery and processing capabilities). Moreover, as stated earlier, the mobile UEs always initiate *TAU* procedures, which are burdensome to both the UE's battery and network resources.

To this end, some of the proposed MM solutions would work for LTE networks, especially the dynamic ones (see Section 2.5.1.2), but these solutions might not work well for 5G use cases. To accommodate the exceptional requirements for 5G, the standard LTE system parameters need to be retuned accordingly to match 5G requirements. For example, LTE networks have used the *DRX* and *T3412-timer* technologies to optimize UE experience and network resources, as mentioned earlier, in terms of UE battery savings and the frequency of *TAU* messages, respectively. This will

also give rise to a trade-off between UE battery power saving and access latency [11]. In other words, the *DRX* cycles prevent a UE from monitoring the *Paging* signals frequently, to save the UE's battery. On the other hand, the *DRX* cycles decrease the UE's reachability (i.e., increase the *Paging* attempts). As a result, many studies have been conducted to achieve the 5G goals. In this context, Chapter 3 extensively addresses this issue, specifically for 5G systems.

## 2.8 Summary

The *TAU* and *Paging* procedures in LTE networks are essential to keep track of all UE units (enabling UE-specific data packets to be exchanged) throughout the network, which are completely controlled by the MME. The tremendous increase in the number of high-mobility UEs will adversely affect the MME performance. This will also impact the related network KPIs and end-user experiences because of the limited network bandwidth and UE battery capacity. To mitigate the MM overhead (in terms of *TAU* and *Paging*), a variety of solution schemes have been proposed. We have examined these solutions in terms of complexity, latency, and computational cost. Because of the trade-off between *TAU* and *Paging* overhead, some studies have considered this trade-off as a multi-objective optimization problem while other studies try to minimize either the *TAU* or *Paging* costs (trade-off between UE battery power consumption, computation cost, network resources, or *Paging* latency). On the other hand, most of the current LTE MM schemes are designed according to UE movements in different scenarios taking into account different mobility models, and hence the network-performance evaluation is highly influenced by these mobility models.

Apart from minimizing the *TAU/Paging* signaling overhead, which this chapter has addressed in detail, we have brought attention to another solution schemes for LTE MM, known as SDNV. These schemes have been proposed to mitigate and manage the overall signaling load on the network, but not intended to solve the optimization problem of the *TAU/Paging* (i.e., the trade-off between *TAU* and *Paging*); we have discussed such solutions in terms of MM.

Moreover, we have investigated applying current LTE LM solutions to 5G use cases. The current solutions have to be modified to meet the expected 5G goals, supporting life-critical systems and real-time applications (close-to-zero latency, on the order of 1 millisecond).

As we have seen throughout this chapter, the two vital MM procedures, *TAU* and *Paging*, are required to locate and track all UEs while moving within the network coverage area, which are still used in current LTE and 5G. These procedures are prone to failure (e.g., might not locate a target UE within a reasonable time of delay and result a congested network in dense area). This raises concerns about how to achieve 5G goals. First, we are concerned with how to provide a very fast way to locate UEs, achieving the order of 1 millisecond latency. Second, we are concerned with how to minimize the power consumption in these devices, especially because most of them are battery powered, supporting IoT devices. Moreover, because of the tremendous increase in high-mobility UEs, the consequent signaling control (UEs always initiate *TAU*) can impact not only the network performance (i.e., cost more network resources and even lead to a congested network) but also the UE experience (i.e., increase power consumption in UEs). This will also impact the *Paging* performance in terms of increase delay, attempts, and even failure.

# Chapter 3

# Mobility Management in 5G Networks

## 3.1 Overview

The 5th Generation of mobile networks (5G) is coming soon (AT&T announces the first 19 US cities to be covered with 5G technology throughout 2019 [65]) to provide exceptional services beyond current cellular systems. To achieve this goal, however, ongoing studies are still developing new schemes to provide seamless connections to the ever-increasing density of high-mobility User Equipment (UE). As stated before, 5G systems will work in conjunction with current Long Term Evolution (LTE) systems and the latter is retuned to use as a base design for 5G. Because of the exceptional requirements of 5G, the 5G Mobility Management (MM) should face tremendous challenges to achieve its uses cases, which are the main focus of this chapter. Note that the material in this chapter has been published in [27].

### 3.1.1 Open Questions

Before proceeding further, it is worthwhile to address the following questions:

1. Will current LTE MM schemes work for 5G use cases?

2. What are the effective solutions to deal with the trade-off between *TAU* and *Paging* overhead in 5G networks?

3. How do we provide services with close-to-zero latency (i.e., making *Paging* latency extremely low relative to LTE networks to support real-time 5G applications)?

4. Because most of the devices on the Internet-of-Things (IoT) are battery powered [66], how do we minimize the power consumption on these devices to extend the battery lifetime?

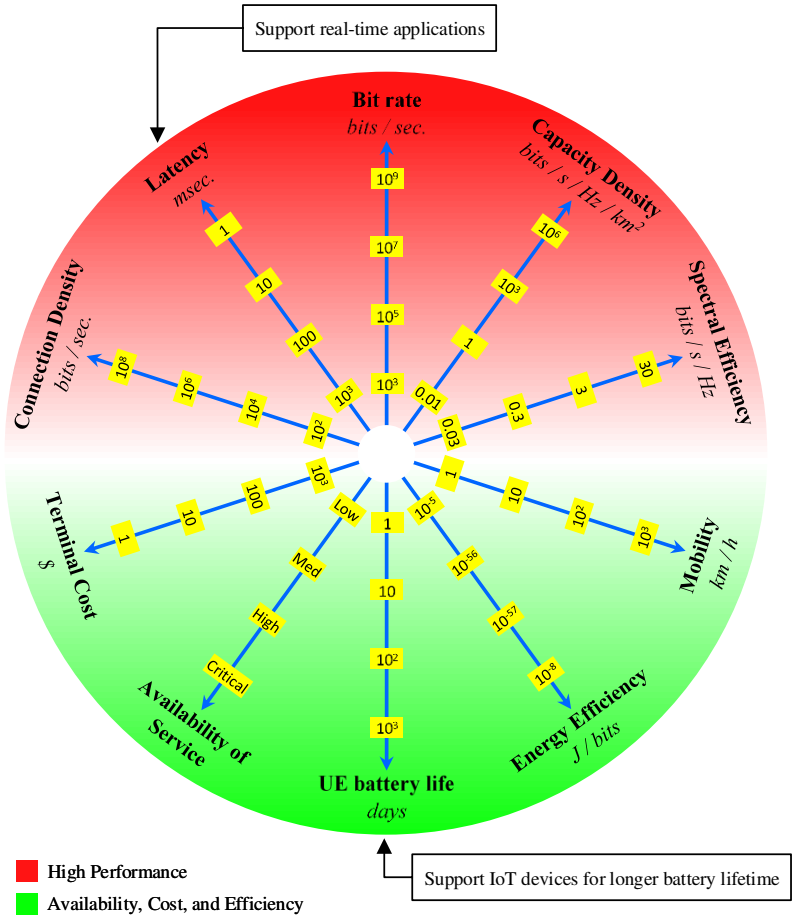



**Figure 3.1:** Key performance requirements for 5G goals (adapted from [7])

Throughout this chapter, we answer the above questions and investigate state-of-the-art MM schemes to meet the expected 5G goals and support the IoT. Figure 3.1 shows the target values that network operators and researches would achieve for 5G.

### 3.1.2 Our Focus

In this chapter, we raise attention to two important MM procedures, *TAU* and *Paging*, in 5G networks. We now argue why it is important to address *TAU* and *Paging* problems.

As we can see later throughout the chapter, these procedures are two related components in location tracking of mobile UEs in current LTE and 5G systems. The two procedures aim to keep

the network better informed about the UE's mobility, in which the network must identify the exact serving cell of a specific UE for the purpose of data-packet delivery; otherwise, the network fails to guarantee this service. To address why *TAU* and *Paging* are important, we provide the following relevant data (recalled from Section 1.1):

> Currently, a LTE MME can process a signaling load of over 500 to 800 messages/UE and even up to 1500 messages/UE under extreme circumstances [21]. These message loads become extremely high when the UEs move across the network and even more when the number of the connected UEs increases (e.g., 5 billion UE units are expected by the end of 2022 [12]).

Therefore, there will be a huge number of control signals that are associated with *TAU* and *Paging* procedures and even worse in 5G use cases (e.g., IoT). The consequent loads can negatively affect both the network performance (i.e., cost more network resources and even lead to a congested network) and the end-user experience (i.e., drain more battery power in UEs); each *TAU* procedure drains about 10 mW of battery power in current-generation smart-phones [4]. Moreover, the *Paging* delay and *Paging* failure are also impacted.

In this context, we mostly focus on the following critical issues in 5G use cases (see Figure 3.1):

1. Achieving high performance in terms of significant reduction in the latency to support real time applications (e.g., life-critical systems).

2. Achieving high availability in terms of significant increase in UE battery lifetime (e.g., supporting IoT devices; see [66] for more details).

Furthermore, since network resources and battery power in UEs are mostly wasted by these two procedure, and because 5G networks are promising to be green in terms of Energy Efficiency (EE), this will become a critical goal to achieve in 5G under the current *TAU* and *Paging*. However, many studies have been introduced to mitigate this issue. The author of [67] introduces a novel resource allocation scheme called hybrid resource management scheme to maximize the network

performance in terms of EE. In [68], the authors present solutions for what is called "smart cities", supporting IoT devices to optimize UE experiences (i.e., maximize the system delivery capacity and optimize the network resource sharing). The authors of [69] have designed an optimal rate allocation and description distribution for high performance video streaming, involving device-to-device (D2D) communication in 5G systems.

In terms of green communication, Energy Harvesting (EH) has become a promising approach in wireless networks to provide what are called self-powered devices (to power up sensors and battery-limited UEs) [70]. According to [71], this can be achieved by converting the received radio frequency signal into a direct current, which is used for charging low power devices such as sensors. Enabling such potential techniques will be very beneficial to maximize EE in 5G systems, enabling what is called EH Relay (EHR) in 5G. However, the EHR system can be susceptible to security attacks since the energy and information are transmitted simultaneously. Hence, the authors of [71] have investigated this security issue and proposed new schemes to improve both the security and EH in the system. All the above improvement solutions are addressed in detail later in Section 3.9.

## 3.2 *TAU* and *Paging* Challenges in 5G Networks

It is expected that 5G mobile networks will support services with zero latency, on the order of 1 millisecond, to accommodate demanding services such as life-critical systems and real-time applications [72,73]. Further, it is envisioned that 5G networks will handle connectivity to as many as 300,000 devices within a single cell and reduce UE power consumption significantly (giving 10 times longer UE battery lifetime) [74,75], supporting IoT devices. Furthermore, 5G networks will shift toward dense heterogeneous network deployments to provide very efficient networks [76]. To satisfy these promising key features, there are several challenges to be faced, and the 5G network KPIs will be impacted in terms of MM load. Because of the above critical requirements, it is clear that the loads on the MME will be extremely high as compared with LTE networks (most of the

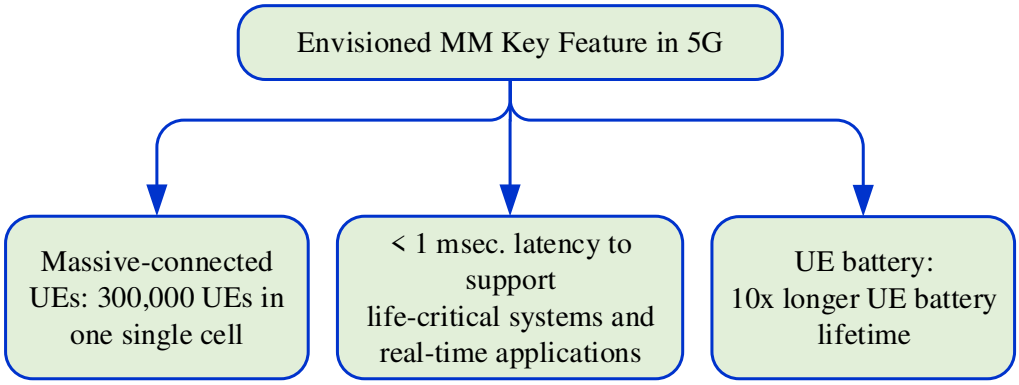



**Figure 3.2:** MM 5G promising key features

loads come from the *TAU* and *Paging* signaling). In this context, the following three subsections summarize the key challenges in terms of *TAU* and *Paging* overhead, as shown in Figure 3.2.

### 3.2.1 Massive Deployment of UEs

As mentioned earlier, there will be numerous UEs connected to 5G networks and could be new version of UEs such as embedded sensors in human body (or clothing), sophisticated equipment for monitoring vital signs (i.e., clinical measurements), or even connected cars [77,78]. Consequently, this scenario will produce a series of negative impacts to the related KPIs. First, if these UEs need to update their location because of their mobility, the triggered *TAU* signaling will grow rapidly. As a result, the *TAU Success Rate* drops to a lower value because of limited bandwidth. Also, the network becomes highly congested if these UEs trigger the *TAU* messages simultaneously, and that could generate adverse effects on the other KPIs such as *Handover Success Rate*. Second, the *Paging Success Rate* will also drop to a lower value because some UEs, as mentioned earlier, would not respond to the incoming *Paging* messages while responding to the *TAU* procedure. This will become a crucial problem to solve; however, it can be avoided trivially at the expense of increasing network resources (e.g., bandwidth), which is unrealistic for 5G.

**Table 3.1:** Cell types in wireless networks

| Cell type | Coverage range (meter) | Capacity |
|---|---|---|
| Femtocell | $10 - 20$ | A few UEs |
| Picocell | 200 | $20 - 40$ UEs |
| Microcell | 2000 | $> 100$ UEs |
| Macrocell | $(30 - 35) \times 10^3$ | Many UEs |

### 3.2.2 HetNets Deployment

5G will shift toward ultra-dense small-cell HetNets, containing multiple layers of different cell sizes: macrocell, microcell, picocell, and femtocell (see Table 3.1 [72]). HetNets are required to improve the network coverage and increase the network capacity while maintaining the energy consumption as low as possible for both the network elements and the connected UEs [74]. To achieve these requirements, [78] presents some 5G design solutions to accommodate the evolution of communication types, UE behavior, and technology. Such solutions, however, would initiate a series of control messages that are necessary for the *TAU* procedures when the UEs cross the ever-small TAs (i.e., femtocell). In addition, when the UEs move along the boarder between the TAs that are not in its TAL, the *TAU* overhead becomes extremely high (relative to a larger TAL) because of the excessive *TAU* signaling (this is also known as the "toggling" or "ping-pong" effect [49]). Moreover, the ultra-dense small-cell deployment increases the *Paging* attempts [79].

Obviously, multiple *Paging* attempts will increase the *Paging* latency or produce *Paging* failure if the UEs cannot respond to that messages within a reasonable time delay. Therefore, the loads on current LTE MME becomes extremely high because of the significant increase in both the "toggling" effect and multiple *Paging* attempts. In this case, life-critical systems and real-time applications will fail (because of the latency) in the 5G.

### 3.2.3 High-Mobility UEs

The other critical requirement for 5G systems is to accommodate the rapid increase in the density of high-mobility UEs. In other words, it is envisioned that 5G systems will provide mobile service for moving UEs in speeds up to 500 km/h such as UEs in ground vehicles, subways, or high-

speed trains [80]. In this case, the very high-mobility UEs will generate excessive control messages that are needed for both *TAU* and *Paging* signaling. In other words, triggering simultaneous control messages from massive number of UEs can lead to a low performance and congested network (the LTE MME cannot handle all the triggered control messages). For example, the *Paging Success Rate* varies in the range of 67–79.3% for the highly mobile UEs [5].

In sum, the current LTE MM solution in terms of *TAU* and *Paging* overhead will not be sufficient to accommodated the above requirements for 5G use cases. The next section discusses whether it suffices to reapply the state-of-the-art LTE MM solutions (*TAU* and *Paging* schemes) for 5G MM.

## 3.3 LTE MM Assessment for 5G Networks

As we have seen in Chapter 2, Section 2.5, numerous solution schemes have been proposed to improve the LTE MM in terms of *TAU* and *Paging* overhead not only to enhance the network performance but also optimize power consumption in the UEs. When it comes to the 5G future, many challenges lie ahead especially when applying the LTE MM solutions to 5G networks. Some of the proposed solutions try to improve *TAU* or *Paging* procedures independently while the others focus on both procedures jointly to deal with the trade-off between them. In this context, we evaluate the receding LTE MM solutions for 5G MM, as follows:

1. As stated before, the schemes in Section 2.5.1.1,*"Global and static techniques for TAU,"* are not used in current LTE MM because of their drawbacks [18], which are as follows cost ineffective, initiate excessive *TAU* and *Paging* messages, do not take into account the "ping-pong" effect, and could generate uneven signaling distribution. Obviously, these schemes are not suitable for the 5G use cases.

2. The schemes in Section 2.5.1.2,*"Local and dynamic techniques for TAU,"* are more practical for LTE networks than schemes in point 1 above. But, they can degrade the network KPIs [18]. Hence, the authors of [1, 5] have proposed solutions to deal with this problem (for LTE), which are no longer applied, according to [81], to future wireless networks (5G).

Moreover, these schemes are heuristic and might be far from optimal [52]. Furthermore, sometimes the UEs need to store the network cell topology, which is unrealistic in real wireless networks [17]. Also, the solution in [53] is cost ineffective for the 5G future because of the rapid increase in the number of high-mobility UEs, which requires extremely high storage capacity to store all the UE mobility data.

3. As we have discussed the solution schemes in Section 2.5.2,*"Paging Improvement Techniques,"* such solutions still have limitations in current LTE systems, and hence when it comes to 5G systems, these schemes cannot satisfy the exceptional requirements for 5G use cases.

4. As illustrated before, despite the fact that the solution techniques in Section 2.5.3, *"Joint Solutions for both TAU and Paging,"* have been applied in some cases, they are still far from realistic because developing such models are very costly in terms of implementation complexity and storage capacity. Moreover, it becomes extremely difficult to apply such solution schemes to large-scale networks (ultra-dense small-cell HetNets 5G; see Section 3.2.2) because the rapid increase in the density of mobile UEs would trigger a huge number of control messages that current wireless networks (LTE) fail to handle.

In sum, the current LTE MM solution in terms of *TAU* and *Paging* overhead will not be sufficient to accommodated the exceptional requirements for 5G use cases.

Because the recent research on designing UE mobility models have attracted the attention of network operators, especially for 5G MM [82], the next section analyzes using these mobility models and studies their impact on the network performance in the purpose of network evaluation (or network simulation).

## 3.4 Mobility Models Assessment for 5G Networks

Many different mobility models have been proposed to predict not only UE locations but also how their velocity and acceleration change over time while moving throughout the network cov-

erage area. Some of these mobility models are fully stochastic and independent with past locations, such as random walk models, random waypoint models, and fluid flow models [82]. Others are built based on a priori knowledge of the UE movement pattern and/or traffic characteristic [52, 57–60], which we discuss here. Basically, wireless network designers have used these mobility models to evaluate the network KPIs that interact with UE movements (i.e., mobility pattern of a UE). Hence, these models can play important roles in MM design (i.e., TAL planning) as well as performance analysis of network performance to mimic real-life networks.

When it comes to the ultra-dense small-cell HetNets (such as in 5G), UE mobility will also impact the 5G performance as UEs move throughout the ever-small cell coverage area (e.g., this increases *handover rate*). Furthermore, the *Paging* solution schemes have mostly used a variety of mobility models to estimate UE locations.

### 3.4.1 Estimation Accuracy

Based on the preceding discussion, it is very important here to address this question: Because these mobility models are designed to estimate UE positions over time, what is the degree of the estimation accuracy that models can provide? As stated in [60], the estimation accuracy of the used mobility models decreases as the randomness in UE movement increases. In other words, unsurprisingly, when UEs possess a high degree of movement randomness, it is difficult to predict UE positions and build mobility models that reflect the actual movement of UEs [43]. In this context, we have proceeded further to address the following vignette:

> The authors of [83] conduct a case study to track the movement paths of users by using different types of estimation schemes. The authors show that most estimation schemes give an estimation accuracy ratio in the range of 50 to 70%. In [84], another case study shows estimation accuracies in the range of 80 to 90%.

In the following subsection, the problem of the prediction accuracy of the mobility models are highlighted.

### 3.4.2 Final Notes

From the above discussion, we find some skepticism about the predictability degree of the these mobility models, which we elaborate on as follows:

1. Such models are constructed based on collected data from UE movement habits over time. To build a very precise mobility model entails assembling a very long-term history of movement data, which is often impractical in terms of memory requirement and computation overhead.

2. According to [83], if the UE enters new places where there is no mobility history data available, history-based estimation might fail to estimate a UE's location, and hence this will impact the related KPI such as *Paging Success Rate*.

3. Because a UE's mobility history can be logged for a long period of time, any unexpected change in the UE movement habits may not modify the overall probability distribution of the UE's location.

Therefore, for the majority of UEs, such mobility models might not be realistic. However, they could give a good accuracy when the UE exhibits a high degree of periodicity in its movement. Users in offices, campuses, or malls can show periodic behaviors (i.e., mobility patterns) that can easily be extracted and modeled [5, 85]. In general, the accuracy of predictability of a given mobility model is highly impacted by the nature of the UE mobility behavior over time.

## 3.5 Design of MM in 5G Networks

Based on legacy LTE system architecture and 3GPP specification for new 5G systems [9], the Next Generation (NextGen) design is based on Network Function (NF) rather than Network Entity (NE) as in LTE. In other words, the Evolved Packet Core (EPC), called the Core Network (CN) in LTE, defines for each network entity (e.g., Serving Gateway (S-GW) and MME) the required network protocols and interfaces between these entities—see [2] for more details—while in 5G Core Network (5GC), the network protocols and interfaces are defined for each NF. The following subsection discusses the NF in more detail in terms of MM.

### 3.5.1 5G NFs

According to the 3GPP specification for NextGen [86], the NF is a processing function in NextGen networks, which has functional behavior and interfaces defined. Basically, the NF can be implemented in three different ways: 1) as a network element on dedicated hardware, 2) as a software instance running on dedicated hardware, or 3) as a virtualised function instantiated on an appropriate platform, e.g., on a cloud infrastructure. Figure 3.3 briefly shows these NFs and their interface connections, also called reference points (e.g., NG1, Next Generation (NG)1, carries signaling between UE and AMF) [8]. In this architecture, the overall 5GC comprises two different planes: User Plane (UP) and Control Plane (CP). The NF that serves the UP (support UE traffic) is called the User Plane Function (UPF). The CP (support UE signaling) is served by six NFs: Access Mobility Function (AMF), Session Management Function (SMF), Policy Control Function (PCF), Application Function (AF), Authentication Server Function (AUSF), and User Data Management (UDM).

The key idea behind this design is to separate the NFs and reduce the latency relative to current LTE systems. For example, as mentioned earlier in Chapter 2, the LTE MME is in charge of all the corresponding mobility management (i.e., all control-plane functions), including *TAU* and *Paging* procedures, and it suffers from the heavy loads that mostly come from the high-mobility UE signaling. Therefore, the 5GC design aims at splitting up the UP and CP to guarantee each plane resources scale independently and allow deploying UPFs in a distributed fashion. This will shorten the Round Trip Time (RTT) between UEs and the data network to achieve the 5G goals (e.g., to support real-time applications) [87].

Figure 3.4 shows the overall NG Radio Access Network (NG-RAN) 5G system architecture. This design supports not only 5G technology but also LTE to provide a variety of services through different types of NG-RAN nodes, which are defined as follows:

- ng-eNB: New version of LTE BS that provides services (UP and CP) to the connected UEs (whether current version of UEs or NG-UEs).
- gNB: 5G BS that provides services (UP and CP) to the connected NG-UEs.

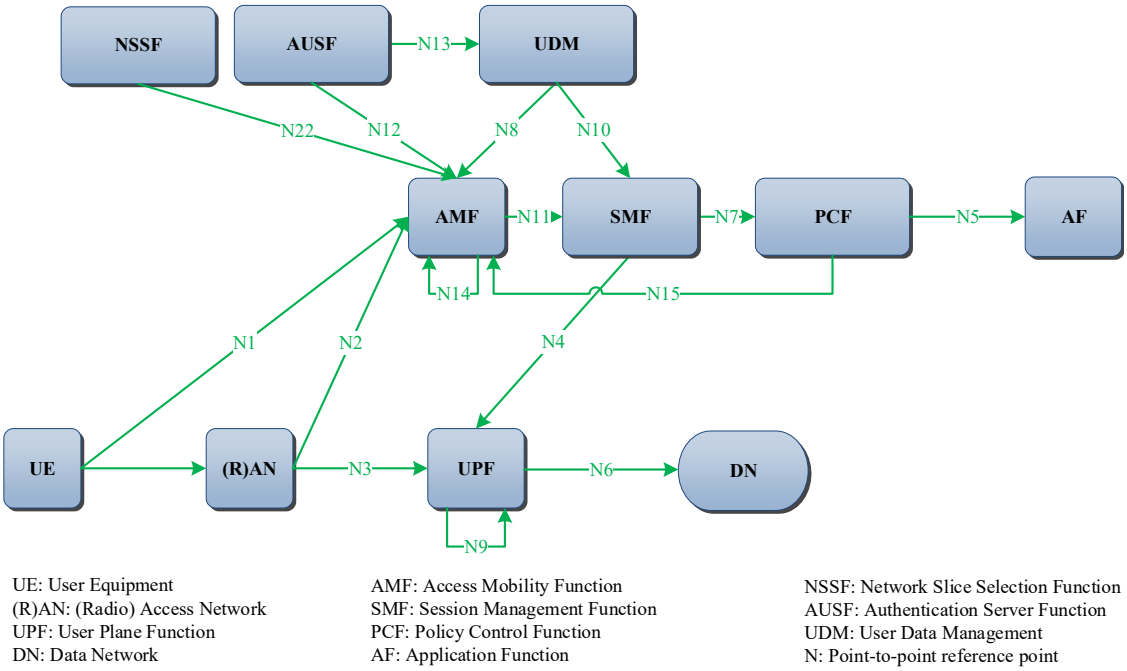



**Figure 3.3:** 5G system architecture in NF and reference point representation (adapted from [8])

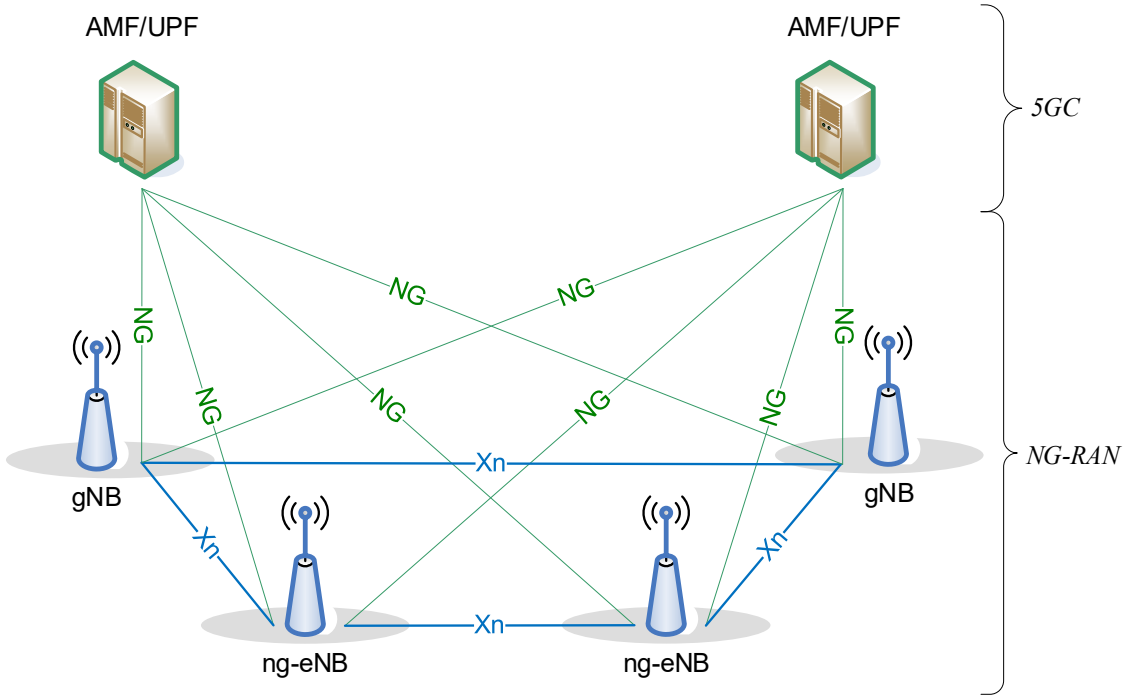



**Figure 3.4:** Overall architecture of 5G system [9]

In 5G, the MM is controlled by two separated NFs, which are AMF and SMF, that can be briefly described as follows [8, 9]:

- AMF: Provide many different functions such as access authentication, authorization, and mobility management control.

- SMF: Provide many different functions, including session management, IP address allocations to UEs.

As we can see that AMF and SMF work independently (i.e., AMF and SMF are represented by different NFs) to allow the CP supporting many services in flexible way. Generally, the events (i.e., states) between UEs and their serving network can be described by different states. The following subsections describe these states for both LTE and 5G systems (for comparative purposes).

### 3.5.2 Mobility States in LTE Systems

In LTE systems and according to the 3GPP specifications [2], three types of states can describe the UE status in the network as shown below:

1. EPS[3] Mobility Management state (EMM): Used to represent if a UE is registered in the EPC or not. This state is managed by the CN.

2. EPS Connection Management state (ECM): Used to represent if a Non-Access Stratum (NAS)[4] is active between the UE and EPC. It is also controlled by the CN.

3. Radio Resource Control (RRC): Used to represent if there is a connection signaling established between UE and its serving RAN (i.e., eNB). The RAN manages (i.e., scheduling and/or resource allocations) the RRC status (RRC-IDLE and RRC-CONNECTED).

The UE can enter different types of sub-states that describes its status with respect to each network entity (i.e., CN and RAN). These sub-states can be a combination of different sub-state such as *EMM-REGISTERED, EMM-DEREGISTERED, RRC-CONNECTED,* and *RRC-IDLE*. Figure 3.5(a) briefly shows these states, describing the different UE mobility states.

[3]Both LTE RAN and the EPC are referred to as the Evolved Packet System (EPS) [19].

[4]Non-Access Stratum (NAS) is a set of protocols that are used to convey non-radio signaling between the UE and its serving MME for an LTE/E-UTRAN access.

### 3.5.3 Mobility States in 5G Systems

In 5G systems and according to the 3GPP specifications [8], the mobility states are introduced based on current LTE mobility states, but add some modifications. Figure 3.5(b) shows these states in comparison with what we have for LTE system. We can classify these states as follows:

1. Registration Management (RM): Mainly used to register or deregister a UE with the network and establish the UE context in the network, to receive services that require registration (UE authentication and access authorization). In RM, the UE's state can be either RM-DEREGISTERED or RM-REGISTERED. Upon registration, the UDM will store the registration information regarding the UE and its serving AMF, including current TALs to enable the AMF to page the UE.

2. Connection Management (CM): Mainly used to establish and release a signaling connection between the UE and its serving AMF over N1 connection (see Figure 3.3). In other words, to enable the NAS signaling exchange between the UE and the CN (i.e., also called 5GC; see Figure 3.4), the N1 connection is used to establish the signaling between the UE and its serving NG-RAN (gNB or ng-eNB). The UE N2 connection is also established via the serving NG-RAN toward the corresponding AMF; see Figure 3.3. The CM uses two states to reflect the NAS activity between the UE and AMF, which are CM-IDLE and CM-CONNECTED.

3. Radio Resource Control (RRC): Unlike the RRC states in LTE which comprises only two states, RRC-IDLE and RRC-CONNECTED, to reflect whether the UE exchanges data packets with its serving RAN, 5G has designed three UE NG-RAN states: IDLE, CONNECTED, and INACTIVE (Figure 3.5(b) does not show the RRC-IDLE state, which is specifically intended for fault recovery or link failure). Because of the importance of this new NG-RRC, we describe it in more detail in the following section.

| | | | | | |
|---|---|---|---|---|---|
| **UE status** | Off | Attaching | Idle/Registered | Connecting to EPC | Active |
| **EMM** | Deregistered | | Registered | | |
| **ECM** | Idle | | | | Connected |
| **RRC** | Idle | Connected | Idle | Connected | |
| **Mobility control** | - | UE based | UE based | NW based | |

**(a)** LTE

| | | | | | |
|---|---|---|---|---|---|
| **UE status** | Off | Attaching | Connected/**Inactive** | | Connected/ Active |
| **RM** | Deregistered | | Registered | | |
| **CM** | Idle | Connected | | | |
| **RRC** | - | Connected | **Inactive** | | Connected |
| **Mobility control** | - | UE based | UE based, NW based | | NW based |

**(b)** 5G

**Figure 3.5:** Connectivity and UE RRC states for LTE and 5G (adapted from [10])

## 3.6 NG-RRC States for 5G Systems

Generally, the NG-RAN is a new 5G Radio Access Technology (RAT) and provides NG-RRC protocols via the air interface for a UE to access the network and exchange the required data packets, providing very fast system access. Typically, the registered UE is in the RRC-IDLE state when there is no active data packet to be exchanged with the network. The UE enters the RRC-CONNECTED state when it needs to monitor downlink control channels (e.g., for *Paging* or system information) and performs signal measurement (e.g., channel status estimation). To perform such operations, the UE needs to switch between RRC-IDLE and RRC-CONNECTED more frequently, controlled by what is called Discontinuous Reception (DRX) mechanism to en-

able UE power saving by allowing the UE to sleep and wake up at predefined occasions [9]. These RRC state transitions are costly in terms of the required signaling overhead, especially when a huge number of UEs wake up to transmit or receive data simultaneously. In addition, most of the RRC-CONNECTED states send a small data size (i.e., less than 1 Kbyte) and then switch back to RR-IDLE state, and hence that will also add more delay to the UE access (or UE reachability) because of the DRX. In LTE, the inactivity timers (i.e., DRX cycles) are configured to be quite short, in the range of 10–60 seconds, which produces a high amount of triggered states from RRC-IDLE to RRC-CONNECTED [11].

### 3.6.1 5G RRC-INACTIVE State

The new RRC-INACTIVE state has been introduced to meet the 5G requirements in terms of signaling overhead, access latency, and mobility management optimization (see Figure 3.5(b)). The NG-RRC states (including RRC-INACTIVE) are ruled by a state machine as in Figure 3.6. Mainly, this new design aims at reducing the CP latency and achieve a seamless RRC state transition. In other words, the RRC-INACTIVE state will keep the UE connected from the 5GC perspective [48]. That means that the UE Access Stratum (AS) (set of protocols between UE and its serving RAN) context is stored in both the UE and serving RAN. In this case, the RRC-INACTIVE state has shown to have several advantages for *TAU* and *Paging* procedures as compared with LTE. One of the important features that is introduced for 5G is the way that the *Paging* procedure can be initiated. This procedure is triggered not only from the CN (such as in LTE) but also from the NG-RAN. That would significantly reduce the *Paging* latency to meet the 5G requirements (close-to-zero latency).

### 3.6.2 NG-RRC Protocols

The NG-RRC protocols support a variety of functions that can be summarized in Table 3.2. In this context, the NG-RRC controls the MM of the UE (i.e., in terms of *Paging* and *TAU*). Unlike the *Paging* procedure in LTE which can be initiated only by the CN (specifically by the LTE MME), the *Paging* process in 5G can be triggered by either the 5GC (5GC-based *Paging*, specifically from

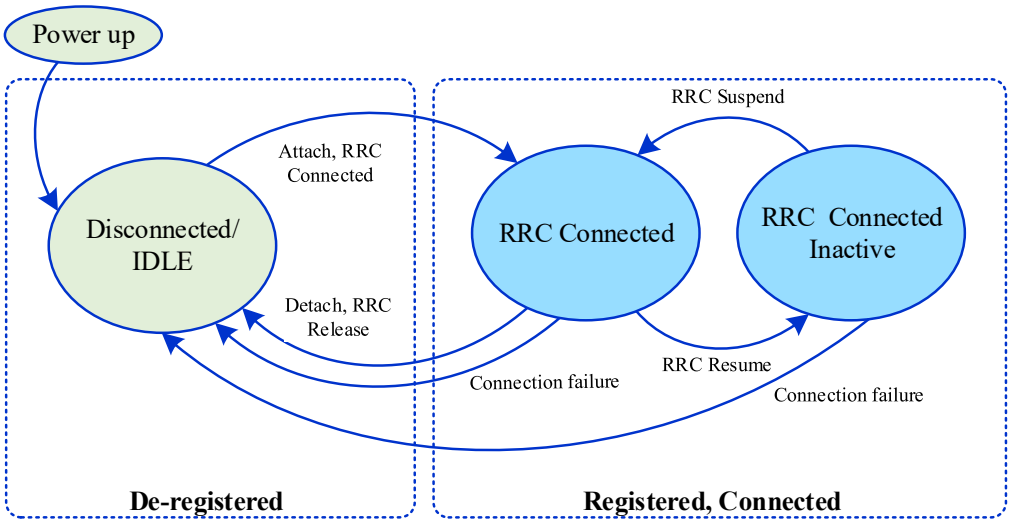



**Figure 3.6:** UE state model for NG-RAN (adapted from [10, 11])

the AMF) or NG-RAN (NG-RAN-based *Paging*, specifically from the ng-eNB or gNB) depending on the NG-RRC status.

As mentioned earlier for LTE systems, the MME is responsible for assigning TALs to each registered UE (i.e., RAs). In 5G, however, this list is assigned by the NG-RAN, called RAN-based Notification Area (RNA) [9]. When the UE moves out of the RNA cell list, it needs to report its location change similar to the *TAU* concept in LTE, which is called RAN-based Notification Area Update (*RNAU*) (triggered via the serving ng-eNB or gNB). In this context, the MM in terms of *Paging* and *RNAU* procedures for 5G have some improvement over the MM for LTE to meet the 5G use cases. The following section describes the 5G MM in more detail.

## 3.7 5G *RNAU* and *Paging*

In LTE, the location tracking of a UE is controlled by the MME on different location levels (i.e., on cell and TAL granularities), depending on the UE RRC states. If the UE is in RRC-CONNECTED state, its location will be determine at the cell level. Otherwise, its location level will be known to be in the assigned TAL (UE is in RRC-IDLE state). In 5G, the 5GC can track the

**Table 3.2:** NG-RRC protocols and functions

| RRC Protocols | RRC States | | |
|---|---|---|---|
| | IDLE | INACTIVE | CONNECTED |
| Network selection/registration | ✓ | | |
| Cell re-selection | ✓ | ✓ | |
| Broadcast system information | ✓ | ✓ | |
| 5GC based Paging | ✓ | | |
| NG-RAN based Paging | | ✓ | |
| DRX configuration for 5GC Paging | ✓ | | |
| DRX configuration for NG-RAN Paging | | ✓ | |
| UE AS context stored in NG-RAN and UE | | ✓ | ✓ |
| NG-RAN manages the UE RNA | | ✓ | |
| NG-RAN knows UE RNA | | ✓ | |
| NG-RAN knows UE serving cell | | | ✓ |
| Keep 5GC/NG-RAN connections for UE (both UP, CP) | | ✓ | ✓ |

UE's location at two levels. First, the 5GC knows that a UE location level is known at its allocated TAL (TAL: an area assigned at the time of the UE registration) when the UE's state in the 5GC is CM-IDLE. Second, when the UE's state in the 5GC is CM-CONNECTED, it can be tracked at the serving NG-RAN level. When the UE state is in RRC-INACTIVE, the network can track its location at cell level also to provide faster UE *Paging*. For that reason, two levels of *Paging* can be applied for UE reachability depending on its NG-RRC state: 5GC and NG-RAN based *Paging* (see Table 3.2). The following subsections describe both the *RNAU* and *Paging* in 5G.

### 3.7.1 5G *RNAU*

The 5GC has introduced a new mechanism for *TAU*, also called *RNAU*, for a UE in RRC-INACTIVE state to track its location more precisely with low signaling overhead. Basically, the RRC-INACTIVE state is introduced to reduce the *Paging* delay and enable lightweight transitions between this new state and RRC-CONNECTED. To facilitate this operation, as mentioned in Table 3.2, the context information of the last UE connection (i.e., during the RRC-CONNECTED state) is kept in both the UE and last serving NG-RAN (i.e., ng-eNB or gNB). However, in commercial LTE systems, the network has no context information stored for the last serving RAN [2].

Therefore, the RRC-INACTIVE state reduces the overall signaling overhead, including *Paging* delay (providing faster and lightweight transitions from the inactive to active) and power consumption in the UEs. Also, RRC-INACTIVE can provide an efficient way to serve the UE applications that send small data packets more frequently, providing lightweight signaling overhead (from RRC-INACTIVE to RRC-CONNECTED) relative to RRC transitions in LTE.

According to the 5G 3GPP standard [9], the AMF assigns to the NG-RAN a RRC-INACTIVE Assistant Information (RIAI), such as the corresponding UE registration area, the UE-specific DRX (see Table 3.2), and Periodic Registration Update (PRU) timer, to assist the serving NG-RAN to decide whether the UE should be sent to the RRC-INACTIVE state. If the NG-RAN does so, the UE remains in the CM-CONNECTED state from the 5GC perspective. The RNA is assigned to a UE by its serving NG-RAN based on the RIAI (i.e., UE registration area) and can cover a single or multiple cells (can be a subset of the 5GC TA). That means that the UE can move freely within its allocated RNA without notifying the NG-RAN (set of ng-eNB or gNB). Otherwise, when it moves into an area that does not belong to its current RNA, it initiates RNAU (for more details about the RNAU procedure, see Section 9.2.2.5 in [9]). Once the serving cell (ng-eNB or gNB) receives the RNAU request from the UE, it may send the UE to one of the following RRC states: RRC-INACTIVE, RRC-CONNECTED, or RRC-IDLE (essentially, the RRC-IDLE state is needed for system maintenance such as recovery from radio link failure; see Figure 3.6). Once a UE enters the RRC-INACTIVE state, the serving NG-RAN may send a periodic RNAU timer to the UE, used to notify the network that the UE is still active. The value of the RNAU time is assigned based on the RIAI (i.e., based on PRU). Also, the NG-RAN uses the UE-specific DRX for *Paging* messages, which we discuss below.

### 3.7.2 5G *Paging*

Unlike the *Paging* messages in LTE systems, which are initiated exclusively by the MME to reach a specific UE in the RRC-IDLE state (or ECM-IDLE from the CN point of view), two types

of *Paging* messages (UE reachability) can be used in 5G systems (as mentioned in Table 3.2), which we explain as follows:

1. **5GC-based *Paging*:** Used as a default *Paging* procedure (similar to LTE MME *Paging*) that 5GC can trigger (i.e., AMF) to locate the serving cell of a specific UE in RRC-IDLE (or CM-IDLE) state for the purpose of data-packet delivery (its current location is known at the 5GC TA level).

2. **NG-RAN-based *Paging*:** Introduced to serve the new UE RRC state, RRC-INACTIVE. As mentioned earlier, when a UE is in the RRC-INACTIVE state, the 5GC considers this UE to be connected (in CM-CONNECTED from the 5GC perspective). That means that the 5GC (i.e., AMF) can simply deliver the UE-specific incoming data packets to the serving NG-RAN. In this case, the NG-RAN should broadcast the corresponding *Paging* messages to find the exact serving cell of the intended UE, with assistance of the RIAI.

#### 3.7.2.1 DRX cycle specifications for 5G

Typically for 5G systems, while UEs are in the RRC-IDLE or RRC-INACTIVE states, they may use the DRX mechanism to save the power consumption in the UE. Because 5G has two types of *Paging* occasions, 5GC and NG-RAN based *Paging*, the UEs should be configured with two DRX cycles to monitor the corresponding *Paging* occasions. The following points summarize these DRX cycles (see Table 3.2):

1. **DRX cycle for 5GC:** The UE receives the DRX cycle length when receiving the System Information (SI) configurations. That means that a UE-specific DRX cycle is transmitted via UE dedicated signaling (this information is assigned by the RIAI, see Section 3.7.1) to monitor the incoming *Paging* messages from the 5GC.

2. **DRX cycle for NG-RAN:** The UE receives this DRX cycle configuration from the serving NG-RAN to monitor the incoming *Paging* messages from the NG-RAN with assistance of the AMF (i.e., RIAI).

#### 3.7.2.2 DRX cycle periods

The UEs, whether in RRC-IDLE or RRC-INACTIVE state, always monitor the PDCCH (Physical Downlink Control Channel) activity within a predefined period of time. This activity is governed by the DRX cycles. As indicated in [9], the DRX cycles are mainly defined by four periods—see Figure 3.7—which are briefly described by the following:

1. **on-duration:** In this period of time, the UE wakes up and monitors the downlink channel (PDCCH). After UE successfully decodes the PDCCH, it stays active and starts an **inactivity-timer**.

2. **inactivity-timer:** In this period of time, the UE waits to decode the received PDCCH. After decoding the PDCCH, the UE restarts the inactivity-timer for the first transmission only. If the UE fails to decode the PDCCH, it can return to sleep.

3. **retransmission-timer:** A period of time in which a retransmission can be expected.

4. **cycle:** This period of time specifies the periodic repetition of the **on-duration** time followed by a possible period of **inactivity-timer**.

#### 3.7.2.3 DRX cycle values

Basically, current LTE system supports only two configurable values of DRX cycles: long and short DRX cycles [11]). However, in 5G systems, the DRX value is negotiated between the UE and the serving AMF (unlike in LTE), which is applied to the UE in the CM-IDLE state. The UE may provide a DRX value (that the UE wants to use) to the serving AMF, which in turns returns back the DRX value to be used by the UE. Otherwise, the NG-RAN broadcasts the DRX values (during the registration) to be applied to all registered UEs (e.g., DRX cycles in the range of milliseconds to hours [11]). This will allow UEs to support different requirements and improve their experiences in terms of *Paging* delay and battery lifetime (see Figure 3.1).

In addition, to achieve high value of *Paging Success Rate* with low signaling overhead, the NG-RAN sends *Paging* assistance information to help the AMF in *Paging* a specific UE. This information will help to identify the NG-RAN nodes and the recommended cells of these NG-RAN

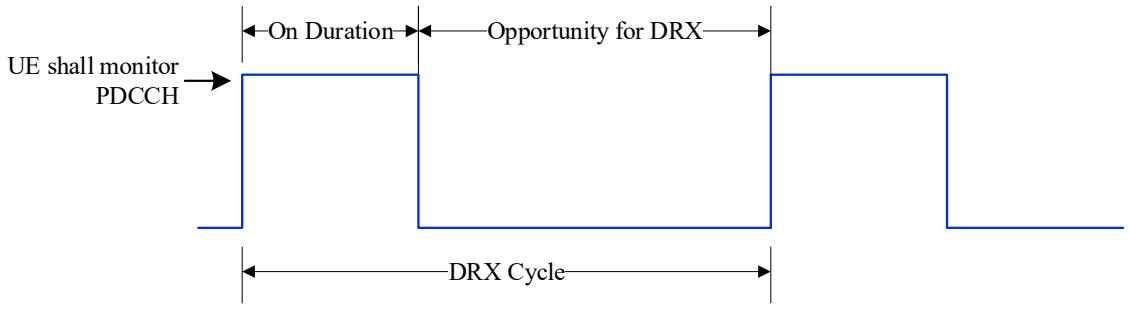



**Figure 3.7:** DRX cycle [9]

to reach the intended UE [8]. But still the ongoing studies focus on these issues, minimizing the *Paging* delay, optimizing UE power consumption, and managing the TAs to meet the 5G requirements (e.g., providing longer UE battery lifetime and close-to-zero latency). The following section discusses these studies in details.

## 3.8 MM Improvement Studies for 5G

To this end, despite the fact that the ongoing 3GPP technical solutions have been proposed to achieve the 5G use cases, especially in terms of *RNAU* and *Paging* signaling overhead, including reducing *Paging* latency and improving the power consumption in the UEs, the network operators are still working to introduce requirement improvements. For example, as mentioned earlier, the 3GPP standards introduce the new NG-RRC state, RRC-INACTIVE, which has benefits in terms of reducing of the *Paging* latency. In this case, when a UE is sent to the RRC-INACTIVE state, the signaling link between the NG-RAN and 5GC is kept. But this mechanism is not always useful, especially for the highly mobile UEs. In this context we have discussed the most recent solutions as follows:

1. Authors of [48] propose hybrid *Paging* and a location tracking scheme for UEs in the RRC-INACTIVE state to control the NG-RAN and 5GC initiated *Pagnig*. According to the author discussions, the *Paging* is initiated based on the mobility status of UEs. A slow-moving UE (i.e., quasi-stationary) is configured with a RNA such that the PA is limited to small number of cells to reduce the related signaling overhead and *Paging* latency as well. A fast-moving

UE should be configured with a RNA such that the PA comprises a large number of cells to reduce the *RNAU* signaling overhead, which is beneficial because the fast moving UEs do not need to trigger the *RNAU* while moving within a large RNA. In other words, for low-mobility UEs, the RAN-initiated *Paging* has lower signaling overhead than the 5GC-initiated *Paging* while the latter has better performance than the RAN-initiated *Paging* for high-mobility UEs. This study shows a significant reduction in the signaling overhead that comes from the *Paging* and *RNAU*.

However, in this study, the authors make an assumption that UEs move with a straight line trajectory with specific speed values (e.g., 3, 30, 60, 90, and 120 km/h), which is not applicable to all the 5G use cases, and does not take into account a variety of UE mobility patterns.

2. Another study introduces a framework to find optimal distributions of TAs for TALs and then allocate these TALs to the moving UEs with minimizing the overall signaling overhead from both *Paging* and *TAU* [88]. The authors of [88] propose two parts to achieve their goal. The first part is responsible for assigning TAs to TALs, and the second part is responsible for distributing the resultant TALs to the moving UEs such that the *Paging* cost is minimized, starting with an inefficient solution and then converges to the best one (through iterations). This study tries to find better trade-off solutions between the two conflicting variables, the *Paging* and *TAU* signaling overhead, formulating this problem as a linear programming problem. As mentioned earlier, the mobility models play an important role in evaluating the network performances (i.e., network KPIs) that interact with the UE movements.

   However, the authors generate their results without identifying which UE mobility model is used to justify their results. Moreover, this work does not take into account the UE mobility status (NG-RRC states), especially the RRC-INACTIVE state, which is mainly introduced to reduce the corresponding *Paging* and *RNAU* signaling overhead. For that reason, this study would be more applicable to current LTE networks. Also, the authors do not consider the different types of 5G *Paging* strategies, whether 5GC or NG-RAN initiates *Paging* messages.

3. Apart from current centralized MM solutions, a study in [89] proposes Distributed Mobility Management (DMM) solutions. The DMM is intended to deploy distributed mobility anchor points close to terminal locations. The authors presented four fundamental designs for DMM solutions. The authors illustrate throughout a comprehensive analysis that their proposal overcoming the limitations from current centralized mobility solutions, including workload distribution, optimizing packets routing, reduce packet delivery latency, and improving handover performance.

   However, such a solution would increase both computation and implementation costs because of the need to deploy multiple anchoring nodes across the network. Also, this solution does not take into account the *Paging* and *TAU* signaling overhead, which is a crucial issue in 5G systems. This solution might adversely affect the network KPIs because it increases both the *TAU* and *Paging* signaling overhead when UEs moves between multiple anchoring nodes.

4. Unlike the solution in [89], the solution in [90] introduces an Autonomous Distributed MME (ADMME) solution, in which many distributed MMEs are responsible for UE mobility management. The ADMME solution distributes the MME load, reduces the CP latency, and gives rise to load balancing. Also, the authors introduce a control node to monitor and control the network performance, including ADMME switching and selection decisions. In this scheme, each ADMME periodically collects the load status of the control nodes. This scheme would reduce the CP latency (for UEs) by choosing an optimal ADMME (e.g., one closer to the UE) and achieve load balancing.

   From the perspective of UEs, however, the proposal still suffers from triggering *TAU* procedures frequently while moving throughout the network, which is burdensome to both the UE's battery and network resources. Moreover, this solution does not mitigate the *Paging* cost or solve the *TAU/Paging* trade-off issue.

## 3.9 Performance Improvement for 5G

In the preceding section, we have discussed different solutions to mitigate the MM signaling overhead and improve the UE experience, but these solutions may not be sufficient to satisfy the 5G use cases. However, and apart from *TAU* and *Paging* solutions, a new trend of solutions have emerged to achieve the 5G goals in terms of energy saving for both the network and battery-limited UEs, including resource allocation and security [67–69, 71]. In this context, we discuss these solutions as below:

1. Since resource allocation affects directly the EE of wireless networks, [67] provides and analyzes hybrid resource allocation approaches to maximize EE performance in 5G. The author provides extensive discussions/comparisons on different 5G potential use cases, including HetNets, massive Multiple-Input Multiple-Output (MIMO), small cell, and cell-free scenarios, introducing some new solutions. For example, the same author with others in [91] have proposed a novel quadratic program (under three objectives: EE, Quality of Service (QoS), and service loading) optimization, resulting in improving EE, guaranteeing QoS, and service loading is optimized. Such solutions are aimed at maximizing the whole system performance to achieve 5G use cases.

2. It is expected from 5G to accommodate exceptional services, supporting a huge number of IoT devices in smart cities. Achieving this goal is a challenge because of limited network sources and battery power of IoT devices [66], resulting in a congested network and low data-packet delivery. To mitigate this issue, the authors of [68] propose a joint Caching and downlink resource Sharing optimization Framework (CSF). According to the authors, this smart solution integrates Wireless Multimedia Sensor Networks (WMSNs) into 5G to efficiently deliver multimedia content to the UEs (to maximize the system delivery capacity). The CSF comprises two optimization problems. The first one is called the Number of Replicas Optimization (NRO) problem, which is then solved for the optimal number of replicas to maximize the average number of replicas, providing high hit rate. Finding the optimal set

of femto BSs and UEs to cache the replicas with high hit rate maximizes the system delivery capacity, which is solved by where to Cache and with whom to Share Optimization (CSO) problem. The authors show through simulations that CSF achieves a high hit rate and system delivery capacity, increasing the system performance.

Solving the NRO and CSO can be time consuming and costly, especially when the search space for these two problems is relatively large (e.g., 5G ultra-dense HetNets). For example, the authors have solved the CSO by using exhaustive binary matrix search, which gives rise to computation delay/complexity even with dividing the searching space into multiple sub-search spaces (as the authors suggest). Also, this problem can be difficult to solve when considering the mobility of UEs. A more effective solution is desirable; they might need a strategy to reduce the search space (e.g., search space reduction; see [92] for more detail).

3. To achieve high performance video streaming in dense 5G networks, the authors of [69] have proposed a joint encoding Rate allocation and Description distribution Optimization (RDO), involving D2D communications and BSs to satisfy high Quality of Experience (QoE) to UEs (also called cellular users in this solution). This solution exploits storage and energy resources available in the D2D helpers, in which the requested videos are already cached. The authors have achieved low interference effect from D2D communications on the UEs and low energy consumption. They also show that the solution can adaptively change the energy constraint taking into account the energy status of the D2D helpers and BS to achieve high playback quality. As the authors indicate, solving the RDO problem is very complicated; instead, a heuristic search algorithm is preferred, such as using Genetic Algorithms (GAs) to find approximate optimal solutions. But, such solution methods are still inefficient in terms of time delay and/or implementation cost (i.e., in mission-critical systems, time and cost are crucial), especially when applying such solutions for 5G ultra-dense HetNets (i.e., very large searching space)—that is, it may not be able to achieve close-to-zero latency in 5G.

4. As stated in Section 3.1.2, although the EHR system is a promising technique (harvest energy from multi-antenna beacons) to approach the green communication prospect of 5G networks, the EHR can be susceptible to security vulnerabilities. To secure the 5G communication and take advantage of the EHR at the same time, the authors of [71] have investigated the security performance of the EHR and proposed new schemes to secure the system and maximize EH. Two Relay Selection (RS) schemes and two EH scenarios are introduced. To secure the EHR, two RS schemes are proposed, called Optimal/Partial Relay Selection (O/PRS). To maximize the EH, two antenna selection scenarios are considered, called Maximizing EH at both the Source (MEHS) and selected Relay (MEHR).

   As the authors illustrate, the ORS scheme outperforms the PRS scheme (in terms of security performance) whether MEHS or MEHR scenario is used. In other words, both the MEHS and MEHR scenarios have equal performance in the ORS scheme. However, according to the authors, the MEHR scenario shows a better enhancement (in terms of security performance) when used with the PRS scheme than MEHS scenario.

   In this solution, the authors have succeeded to improve the security performance of the EHR systems, but at the expense of adding some implementation and/or computation overhead to the system. A solution that adds essentially no overhead would of course be preferable.

## 3.10 Summary

This chapter has presented two of the most vital processes of the MM in mobile networks, *TAU* and *Paging*, which exhibit very high signaling overhead owing to high volume mobility of the connected UEs. Hence, we explore many different solution schemes to mitigate the signaling overhead from the *TAU* and *Paging* control messages, which are adapted to current LTE networks. Yet these solution schemes still have some drawbacks (as explained earlier) such as the trade-off issues, minimizing *TAU* overhead at the expense of maximizing the corresponding *Paging* cost, and vice versa. Based on our evaluations, we examine the ability of applying such solution schemes to 5G use cases. As a result, these schemes might fail to achieve 5G requirements because of the rapid

increase in the density of high-mobility UEs. In this case, the above LTE solution schemes will not satisfy the 5G use cases because of their limitations owing to high implementation complexity, high latency, and high computation cost (e.g., do not maintain close-to-zero latency).

Most of the current MM solution schemes try to localize a UE within the network by building mobility models to predict the UE's location. Also, these models are used to evaluate a network performance that interacts with the UE mobility patterns. Therefore, we have investigated these mobility models in terms of prediction accuracy and implementation cost. To this point, we have come up with the fact that the prediction accuracy of a given mobility model is highly influenced by the nature of a UE mobility behavior over time. That means that, for the majority of UEs, these prediction models might not be realistic.

To meet 5G requirements, a new 5G system architecture is developed (based on legacy LTE system), which is mainly based on the NF rather than NE to produce efficiency. Specifically, this new design aims to reduce not only the *TAU* but also *Paging* signaling overhead and maintains the *Paging* latency to be extremely low (e.g., $<$ 1 millisecond) relative to current LTE systems. In this chapter, many new aspects in terms of 5G MM have been discussed, which include the NG-RAN, NG-RRC, RNA, *RNAU*, and the *Paging* DRX cycle configurations (5GC/NG-RAN-based *Paging*). According to these new key design, the envisioned 5G use cases would be achieved for not only the network performance but also UE experience. However, network operators and many research groups are still developing more MM solutions to satisfy 5G goals, which are examined at the end of this chapter.

# Chapter 4

# 5G Mobility Management for Critical and End-User Needs

## 4.1 Overview

The 5th Generation (5G) wireless networks aim to accommodate extraordinary use cases beyond current networks, Long Term Evolution (LTE), handling very high density of mobile User Equipment (UE), supporting life-critical systems (or real-time applications with close-to-zero latency), and achieving 10 times longer UE battery lifetime. Furthermore, 5G is considered as a basic platform to run emerging of what is called Internet-of-Things (IoT) technology; see [66, 93] for more detail about IoT scenarios. Basically, the concept of IoT is one of the promising 5G technologies, which enables large numbers of "things" to be connected to the Internet and to communicate via 5G networks. For example, these devices can range from smart-phones to novel devices: embedded sensors in the human body (or clothing), wearable devices, equipment for monitoring biometrics, or even autonomous cars (e.g., also called V2X communications [77, 94]).

As expected, there will be tremendous increase in density of connected "things" in 5G, including high-mobility IoT/UEs. According to [95], more than 24 billion IoT devices will be connected by 2020; approximately four devices for every human being on the Earth. To this end, however, achieving this goal (i.e., supporting the massive number of IoT/UEs) will become a crucial problem for 5G requirements. This problem becomes even worse with the existence of high-mobility IoT/UEs. Ongoing research continues to introduce new schemes/algorithms to meet the needs of 5G use cases, taking the existing schemes/algorithms for current LTE systems as a basis for design. At this point, 5G networks need very efficient algorithms for Mobility Management (MM) to control and manage the highly mobile IoT/UE devices. More precisely, the network needs to identify the exact serving cell (i.e., eNB/gNB; LTE/5G base stations, also called Radio Access Network

(RAN)) for each connected device with very low latency, providing very fast IoT/UE reachability (i.e., achieving close-to-zero latency). Note that the material in this chapter has been published in part in [28] and in whole in [29].

Next, we explain the pitfalls of current *Tracking* and *Locating* (i.e., *TAU/RNAU* and *Paging*) toward achieving the critical requirements for 5G IoT/UEs (in terms of IoT/UEs power saving, signaling overhead, and close-to-zero latency).

## 4.2 Pitfalls of Current Tracking and Locating Procedures

As we have seen from the preceding discussion/chapter, the *TAU/RNAU* and *Paging* are vital procedures to keep the CN/5GC informed about the IoT/UE location changes—that is, the CN/5GC must determine the serving eNB/gNB of each IoT/UE, enabling IoT/UE-specific data packets to be exchanged with the serving network. However, these localization procedures are prone to multiple failures. This gives rise to the following questions:

1. Can the *Tracking* and *Locating* procedures fail to identify the exact serving eNB/gNB of an intended IoT/UE?

2. How fast can the network identify an intended IoT/UE?

3. Can these procedures impact the network performance or the served IoT/UEs experience? How?

We have discussed the above concerns in [27] (also addressed in Chapter 3) for both LTE and 5G. In the following points, we briefly summarize some of these concerns:

1. The two localization procedures give rise to a trade-off optimization problem and the associated signaling overhead. This problem impacts not only the network but also the IoT/UEs because the latter are limited in battery power and processing capabilities.

2. As mentioned earlier, because of the increasing density of high-mobility IoT/UEs, there will be a huge number of control signals associated with *Tracking* and *Locating* procedures,

becoming even worse in 5G use cases. For example, according to [21], a LTE MME can process a signaling load up to 1500 messages per UE. The rapid increase in the number of connected IoT/UEs will produce extreme signaling loads on the network. And even worse, with high mobility, the network gets congested because of limited resources.

3. In 5G use cases, the resultant loads can adversely affect both the network performance and the end-user experience; each *TAU* procedure drains about 10 mW of battery power in current-generation smart-phones [4]. Moreover, the *Paging* delay and *Paging* failure Key Performance Indicators (KPIs) are also impacted.

The current *Tracking* and *Locating* procedures will prevent achieving the 5G use cases in terms of extending battery lifetime of IoT/UEs, minimizing the *Paging* attempts/latency, and reducing the corresponding signaling overhead. In the next section, we introduce our approach to solve this problem.

## 4.3 Our Solution Approach

Based on our previous discussion and before proceeding further, we want to answer the following question about the two MM procedures, *TAU/RNAU* and *Paging*:

- Which procedure is the more important of the two, and why?

TAL and the corresponding *Tracking* concepts aim to help the network find the exact serving cell; TAL is mainly used by the *Paging* procedure (see Figure 4.1). While the *Locating* procedure involves all the cells in the intended TAL, TAL is the key control for both *Paging* and *TAU*. Also, *TAU* and TAL burden not only the network but also the battery-limited IoT/UE because the latter always initiates the *TAU* procedure. We can now answer the above question based on the following fact: *if the serving cell (eNB/gNB) of a UE can be identified precisely, the CN/5GC can page the intended UE directly with extremely low latency*. In other words, it suffices for the CN/5GC to send only *one Paging* message to the exact serving cell of the intended IoT/UE instead of sending multiple *Paging* messages to all cells in the TAL. In addition, the *Paging* overhead is directly

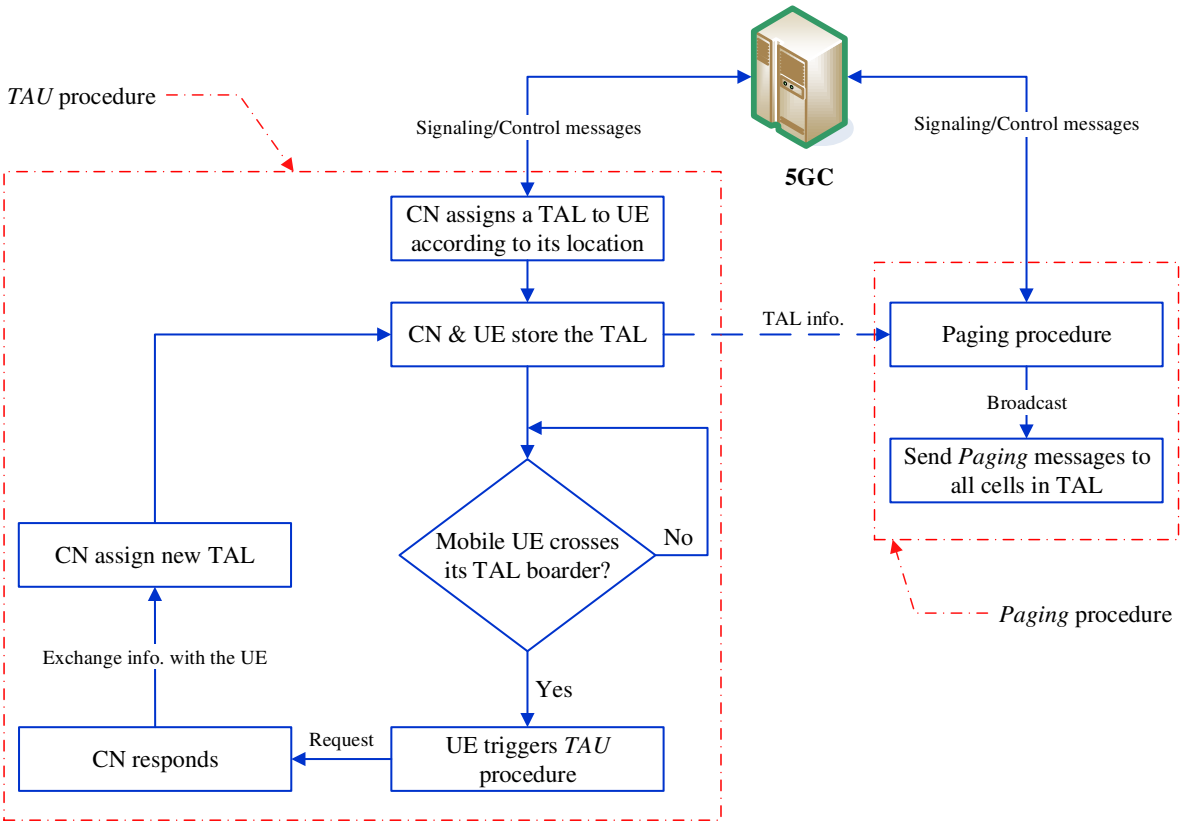



**Figure 4.1:** *Tracking* and *Locating*

proportional to the TAL size, and the latter impacts both the network and UEs. Hence, it is apparent that the current TAL and its IoT/UE-based *TAU/RNAU* are key to solving the problems of *Tracking* and *Locating* in 5G.

### 4.3.1 The Proposed Solution

We propose a novel solution in which the TAL and its corresponding IoT/UE-based *Tracking, TAU/RNAU*, are avoided. Also, our new solution provides sufficient information to the AMF such that only *one Paging* message is sent directly to the serving cell where the intended IoT/UE is located. We highlight some interesting features of this solution as follows:

1. UEs will no longer need to be assigned a TAL/RNA and initiate *TAU/RNAU* procedures while moving throughout the network—that is, the IoT/UEs will be freed from triggering

frequent *TAU/RNAU* procedures, which is beneficial to the IoT/UEs and serving network. First, the battery-limited IoT/UEs will no longer consume 10 mW for each *TAU* procedure, thus extending battery lifetimes. Second, the accompanied *TAU/RNAU* signaling overhead will be avoided, saving network resources when supporting massive number of connected IoT/UEs. Because all the IoT/UEs are no longer involved in responding to the *TAU/RNAU* procedures, the corresponding KPIs, such as *Paging Success Rate*, *Paging Failure Rate*, and *Paging delay/attempt Rate*, will significantly improve. Moreover, the chance of a network being congested will reduce; there is no massive simultaneous requests of *TAU/RNAU*, supporting the rapid increase in the density of IoT/UEs.

2. LTE and 5G use the same TAL and *TAU/RNAU* concepts, which involve effort to optimize the TAL. Specifically, the *TAU* statistical data should be monitored frequently to reduce the corresponding signaling overhead (e.g., add/remove cells to/from TALs to mitigate the corresponding *TAU/Paging* signaling costs). Typically, the TALs are planned by taking into account the surrounding geographic area where the NG-RAN nodes are distributed. This geographic area may change over time (e.g., because of urban development, geographic growth, new roads, and new buildings), adding more complexities to TAL allocations. In our solution, all the above planning efforts are no longer needed; the TAL and *TAU* concepts are not used anymore.

3. There is no need to develop new communication protocols to achieve the proposed solution, adding no additional implementation costs/complexities. Instead, we use the already existing RRC measurement reports, protocols, and interfaces, such as Received Signal Reference Power (RSRP) report [72,96], anchor NG-RAN (ng-eNB or gNB) for 5G system, and X2/Xn (X2 and Xn terminologies are used in LTE and 5G, respectively, covering the same functions) interface/protocol [2, 96, 97]. To describe our design, we first briefly define the used protocols/interfaces in the following section.

### 4.3.2 Preliminaries

Because our solution relies on the existing protocols/interface, we briefly describe some them here:

1. RSRP measurement report: Each IoT/UE monitors the RSRP of its current serving and neighboring cells; they are used for cell (re)selection, handover, and path loss calculation (e.g., for power control). When the RSRP drops below a specific threshold, the UE performs a cell (re)selection or handover (depending on the UE RRC state) by sending measurement reports to their serving NG-RAN nodes, including RSRP report; see [26,40] for more detail.

2. X2/Xn interface and protocol stack: The X2/Xn interface is used to link NG-RAN nodes and exchange control signals to support different functions, such as handover and load management [96, 97]. X2/Xn supports both User Plane (UP) and Control Plane (CP) protocols. Figure 4.2 shows the C/UP protocol stack on the X2/Xn interface. The Xn-UP is used to tunnel IoT/UE data packets between eNBs/gNBs (for data transfer and flow control function). The Xn-CP comprises procedures to manage/control the Xn interface and IoT/UE mobility (support RRC-CONNECTED or RRC-INACTIVE state), such as NG-RAN *Paging* and Retrieve/Release UE Context. The Xn-AP (Xn Application Protocol) has many different functions, which are important to our design functionality in terms of mobility and load management [98]. MM allows the serving eNB/gNB to move the responsibility of a specific IoT/UE to another eNB/gNB, exchange traffic-related and radio quality measurement reports between the eNBs/gNBs, and transfer the IoT/UE status. Load management allows eNBs/gNBs to share information such as resource status, overload, and traffic load.

3. Anchor-gNB concept: In 5G (but not in LTE), the context information[5] of the last UE connection (during the RRC-CONNECTED state) is kept in the UE and its last serving NG-RAN, which is known as anchor-gNB. When the UE enters the RRC-INACTIVE state, the

[5]Includes UE-specific configuration parameters, such as AS security context and most recent TAL/RNA; see [2] for more detail.

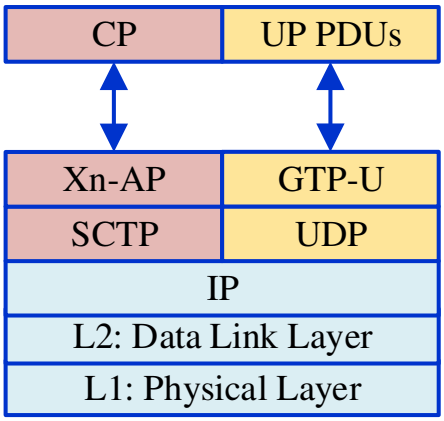



**Figure 4.2:** X2/Xn protocol stack for the UP and CP

5GC can page that UE directly via its anchor-gNB because the latter stores the association information of its UEs. As discussed in Sections 3.6.1 and 3.7.1, this is to provide very fast and lightweight signaling for a UE to transit from RRC-INACTIVE to RRC-CONNECTED. On receiving the 5GC *Paging* message, the anchor-gNB sends *Paging* messages to the RNA cells of the corresponding UE unless the UE is still served by its anchor-gNB.

## 4.4 Solution Framework and Methodology

In our solution, unlike in LTE, the NG-RAN is responsible to track and locate mobile IoT/UE: these devices are no longer involved in reporting their location changes to the network. We call our solution ***gNB-based UE Mobility Tracking (gNB-based UeMT)***, which is defined by two essential parts as follows.

### 4.4.1 *gNB-based UeMT* Entity Definitions

We classify the UEs into two types, according to their serving gNB: *Home-UE* and *Visiting-UE*. Likewise, each of serving gNB is classified as either *Home-gNB* or *Visiting-gNB*. We define their behaviors/functions as follows:

1. *Home-gNB*: Acts as a serving gNB (anchor-gNB) to a group of connected IoT/UEs. The *Home-gNB* registers these devices as *Home-UEs*. So, normally, the *Home-gNB* will be re-

sponsible for exchanging data packets between its *Home-UEs* and the 5GC. In this case, when the AMF needs to page a certain IoT/UE within this group, it simply forwards the incoming data packets to the corresponding *Home-gNB*, which then sends the *Paging* message to the intended IoT/UE.

2. *Visiting-gNB*: Provides services to two types of UEs (called *Home/Visiting-UEs*) in two different ways. First, it serves its own *Home-UEs* such that all the corresponding data packets are exchanged with the AMF, which is the normal function of the gNB. Second, it acts as a temporary serving gNB to *Visiting-UEs* such that all the corresponding data packets are exchanged with their original *Home-gNBs* where these UEs came from (using the Xn interface).

3. *Home-UE*: An IoT/UE that is served by its *Home-gNB* and registered in a control table, which we describe next, in the *Home-gNB*. The gNB uses this table for *Tracking* and *Locating* the registered *Home-UEs*.

4. *Visiting-UE*: An IoT/UE that has camped on a new gNB[6] rather than its current *Home-gNB*; it becomes a *Visiting-UE* with respect to the new gNB. Likewise, this gNB becomes a *Visiting-gNB* to the new IoT/UE (as described in point 2 above). This device is registered in a control table in the gNB, which is used to exchange data packets with the original *Home-gNB*.

As mentioned above, the gNBs use a control table, called *Home/Visiting-Control Table (H/V-CT)*, as shown in Table 4.1. The *H/V-CT* structure and functions are described next.

### 4.4.2 *gNB-based UeMT* H/V-CT Functions

We have designed the *H/V-CT* for the purpose of IoT/UE mobility tracking, controlling their data flow, and guaranteeing low overhead/latency mobility management. To elaborate on the *H/V-CT*, we define the following:

[6]Basically, the IoT/UE (re)selects a new serving cell (gNB) either because of its mobility or when its received RSRP drops below a certain threshold.

**Table 4.1:** *Home-CT* and *Visiting-CT* entries

**(a)** *Home-CT* entry

| **Home-CT** | | |
|---|---|---|
| *Home-UE ID* | *Resident-flag (h/v)* | *Visiting-gNB CID* |
| ue1 | h | 0 |
| ue2 | h | 0 |
| ue3 | v | cid1 |
| . | . | . |

**(b)** *Visiting-CT* entry

| **Visiting-CT** | | | |
|---|---|---|---|
| *Visiting-UE ID* | *Home-gNB CID* | *CiPD index* | *SST (sec.)* |
| ue9 | cid7 | CiPD_1 | 0 |
| ue6 | cid6 | CiPD_0 | 0 |
| ue7 | cid8 | CiPD_4 | 0 |
| . | . | . | . |

1. *Home-CT*: Each gNB uses this control table as a registration table for its own *Home-UEs*. The table structure is shown in Table 4.1(a), comprising three entries: *Home-UE ID*, *Resident-flag*, and *Visiting-gNB CID*. The first entity is the list of registered UE IDs currently associated with a gNB as *Home-gNB* (depending on the gNB capacity/bandwidth), which informs the corresponding AMF that these UE IDs are associated with a specific gNB ID, called *Home-gNB* Cell ID (CID). For simplicity, the list of *Home-UEs* IDs are assigned a specific group ID, which is associated only with the corresponding *Home-gNB* CID. In other words, each *Home-gNB* CID has a unique group ID through which the 5GC can reach these UEs. The second entity indicates whether the corresponding UE is still associated with its *Home-gNB* or is camped on a neighboring cell (*Visiting-gNB*). The third entity refers to a CID of the *Visiting-gNB* in which the *Home-UE* is currently camped. Notice that the number of gNBs in this field defines the potential number of neighboring cells that are connected via the Xn interface. The *Resident-flag* and *Visiting-gNB CID* take values according to Table 4.2. The *Home-CT* is used to page each of its IoT/UEs in two different ways based on whether the device is served by its *Home-gNB* or a *Visiting-gNB*. When the *Home-UE* is still camped

**Table 4.2:** IoT/UE association status (*Resident/Visiting*)

| ***Home-UE* association** | ***Resident-flag*** | ***Visiting-gNB*** |
|---|---|---|
| with its Home-gNB | h | 0 |
| with a Visiting-gNB | v | cid |

on its *Home-gNB* (*Resident-flag* = *h*), a direct-*Paging* is used. Otherwise, an inter-*Paging* is used when the *Home-UE* is camped on a neighboring *Visiting-gNB*, whence *Resident-flag* = *v* for that UE (e.g., ue3 in Table 4.1(a)). The *Home-gNB* forwards the corresponding *Paging* message (via Xn) to the *Visiting-gNB* (e.g., cid1) to reach the aimed IoT/UE (e.g., ue3). As a result, the *Home-gNB* knows the exact location of their registered *Home-UEs* at all times (unless the *Home-UE* is turned off or out of the network coverage area).

2. *Visiting-CT*: This control table is similar to the *Home-CT*, but adds some control functions to guarantee efficient MM in terms of maintaining low signaling overhead and *Paging* delay. The corresponding table entries are defined as below (and illustrated Table 4.1(b)):

   (a) *Visiting-UE ID*: A list of temporarily connected *Visiting-UEs* that have been handed off from their *Home-gNB* to camp on a neighboring *Visiting-gNB*. These devices are listed in this table as *Visiting-UEs* along with their *Home-gNB* CIDs.

   (b) *Home-gNB CID*: A list of *Home-gNB* CIDs that relate each registered *Visiting-UE* in this table. For example, in Table 4.1(b), ue9 is currently camped on a neighboring *Visiting-gNB*, but its *Home-gNB* ID is cid7.

   (c) *Calculated inter-Paging Delay (CiPD) index*: To guarantee that the inter-*Paging* delay between *Home-gNB* and *Visiting-gNB* does not exceed a predefined value, we propose an index value, called the *CiPD*. This value is determined by the network requirements or prioritized according to application requirements, such as in mission-critical scenarios. For illustration, we introduce Table 4.3 to show an example of predefined values for the inter-*Paging* delay (we will explain *CiPD* index values later).

**Table 4.3:** Example of *CiPD* index values

| CiPD Table index | |
|---|---|
| CiPD (msec.) | index value |
| CiPD < 1 | CiPD_0 |
| 5 > CiPD ⩾ 1 | CiPD_1 |
| 9 > CiPD ⩾ 5<br>13 > CiPD ⩾ 9<br>. | CiPD_2<br>CiPD_3<br>. |

(d) *Service Switch Timer (SST)*: This is another control parameter, used to manage the mobility of the *Visiting-UEs* to guarantee that the cumulative time overhead is below a specific value. For ease of presentation, we propose an ascending timer, called the *SST*, with default initial value of $SST_{ini} = 0$ *sec*. and a maximum value of $SST_{max} = t_{max}$ *sec*., specified by the network requirements (e.g., $SST_{max} = 3600$ *sec*.). The behavior of the *SST* is shown Figure 4.3. As we see in this figure, when $SST > t_{max}$, a new procedure is triggered, called the *Service Switch Procedure (SSP)*. This is to transfer the UE context information to the new *Home-gNB* and report this change to the AMF (described next).

To illustrate how the *SST* controls the signaling overhead for a *Visiting-UE*, consider the following example:

When the *Visiting-UE* is registered in the *Visiting-CT*, all the corresponding UE-specific data packets are transfered between the current *Visiting-gNB* and the AMF via the original *Home-gNB* of the UE. If this UE resides in the *Visiting-CT* for a relatively long time (per the *SST*), the accumulated overhead becomes relatively high. So, it is better to transfer the UE-specific data packets between the *Visiting-gNB* and the 5GC directly instead of being transfered via the *Home-gNB*. As such, the relevant signaling overhead is minimized.

In addition, if a *Visiting-UE* is removed from its *Visiting-CT* (when it (re)selects another cell or returns to its *Home-gNB*) before the corresponding *SST* reaches its maximum value (e.g., $SST_{max} =$ 3600 *sec*.), the timer is reset to zero. Figure 4.3 shows how the *SST* acts in the *Visiting-CT*.

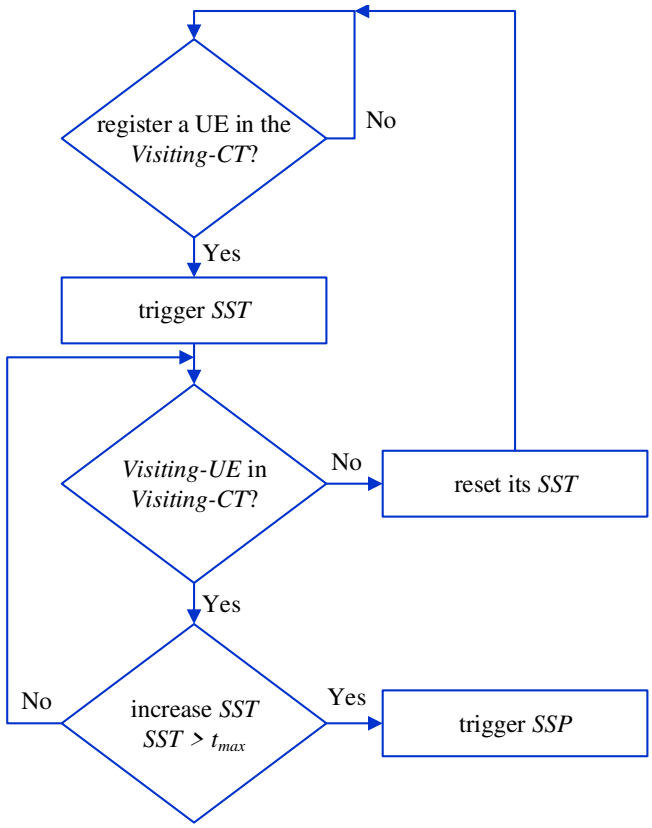



**Figure 4.3:** Behavior of the *SST* in the *Visiting-CT*

In *gNB-based UeMT*, the *CiPD* and *SST* are the two key controls to guarantee low signaling overhead and *Paging* delay. These two control parameters are used by our next proposed procedure, called *Service Switch Procedure (SSP)*.

## 4.5 *gNB-based UeMT* Control Models

We now take a step further to show how all the previous components and control functions of our design are related to each other. To help explain the solution, we divide the *gNB-based UeMT* into three main parts as below.

### 4.5.1 Basic Workflow of IoT/UE Mobility Tracking

Figure 4.4 shows the basic process flow of the *gNB-based UeMT*. This process starts either when UEs begin initial access to the network or when they end up in the RRC-INACTIVE state (the

new UE RRC state in 5G, as stated earlier). Following the normal procedure of cell (re)selection, these UEs make their (re)selection based on the RSRP measurement report[7]. As a result, they cluster around gNBs; each gNB acts as an anchor-gNB to a group of UEs. In other words, the *Home-gNBs* register their UEs as *Home-UEs* in the *Home-CT* (detailed in Table 4.1(a)). Likewise, some of UEs are registered as *Visiting-UEs* by the *Visiting-gNB* in the corresponding *Visiting-CT* (detailed in Table 4.1(b)). Each gNB manages and controls the mobility of UEs by using the *H/V-CT*; this ensures seamless mobility and lightweight signaling overhead (the UEs are not involved in the above managing/controlling process; no *TAU* is needed).

### 4.5.2 Service Switch Procedure (SSP)

The *SSP* is a conditional procedure, as Figure 4.5 shows, that is controlled by either the *SST* or *CiPD* control parameters (see Figures 4.3 and 4.4). The *SSP* is triggered to guarantee that the corresponding signaling overhead/delay is relatively low (according to the setting of *SST* and *CiPD*; see Table 4.3).

### 4.5.3 Home/Visiting-gNBs Mobility Tracking Scheme (H/V-MTS)

While IoT/UEs move throughout the network coverage area, many cell (re)selections occur. To keep track of these mobile UEs, the serving/neighboring gNBs interact (via Xn) to manage and control the mobility tracking of these UEs (specifically, gNB functions in the *gNB-based UeMT*, including regular mobility management [9]). The procedure in Figure 4.6 shows the necessary messages of the H/V-MTS, in which the *Home/Visiting-gNBs* are talking to each other to track and locate their IoT/UEs. All these messages are transported over the Xn interface.

As we see throughout this procedure, mobile UEs use the regular measurement reports (RSRP) to camp on the best serving cell. Also, while moving, UEs do not use the conventional *TAU/RNAU* procedure to update their location changes. Instead, the *Home/Visiting-gNB* locates and tracks these UEs without their intervention. Also, Figure 4.6 shows how the corresponding *Home-CT*

[7]When in RRC-IDLE/INACTIVE, the UE does cell (re)selection; otherwise, the UE undergoes handover.

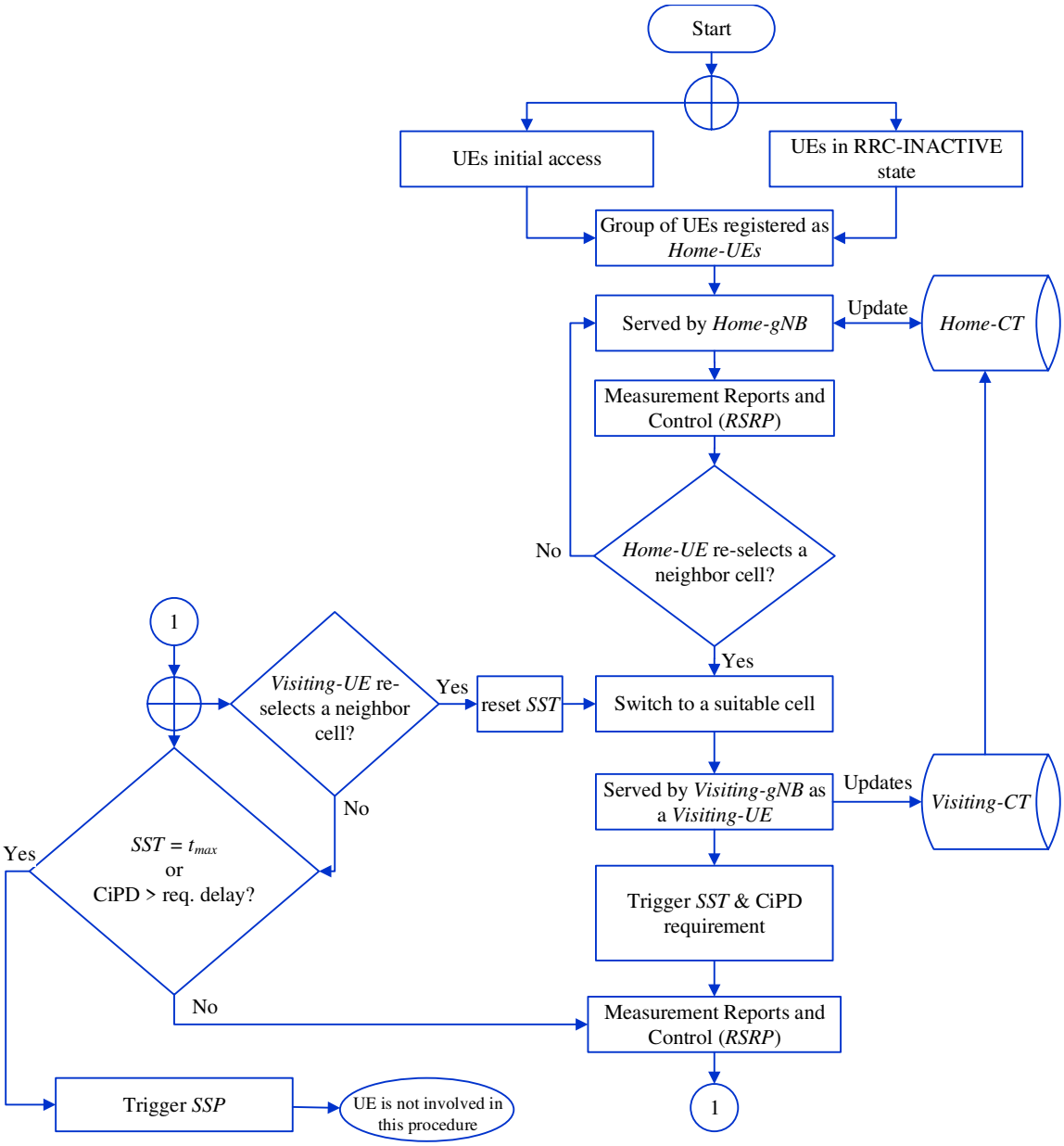



**Figure 4.4:** Flow chart of the *gNB-based UeMT* process

and *Visiting-CT* track the mobility of UEs during a handover, maintaining service continuity and session management. In our design, we use the already existing (3GPP) control messages and measurement reports such that no implementation complexity is added.

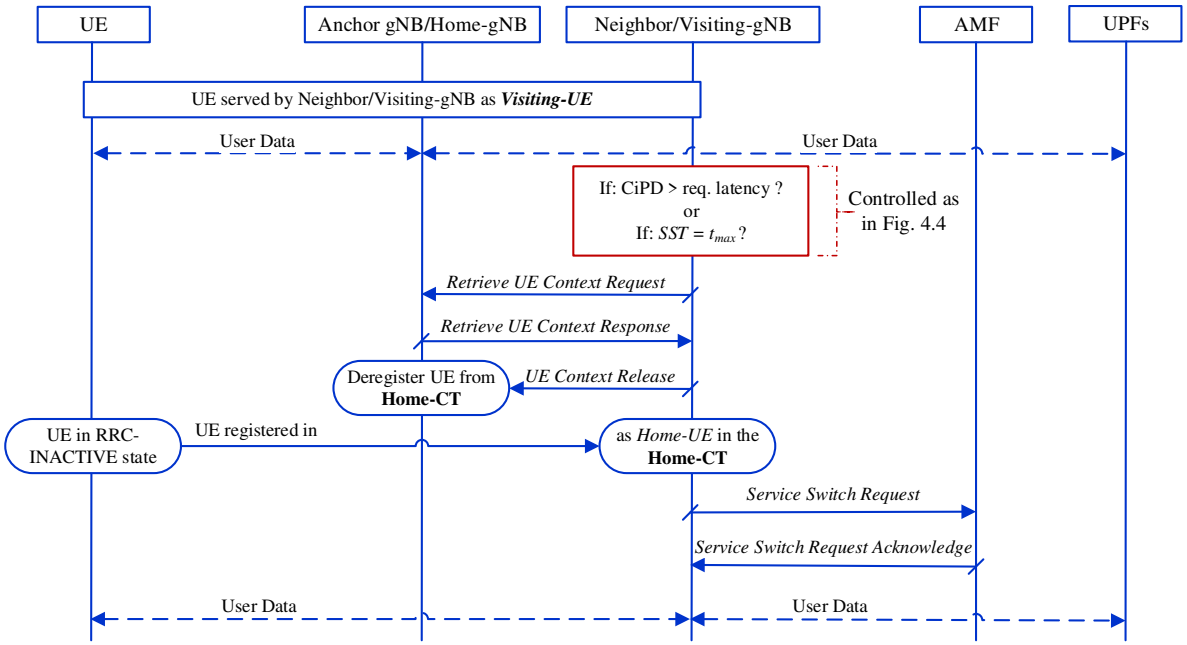



**Figure 4.5:** Signaling flow of the *SSP*

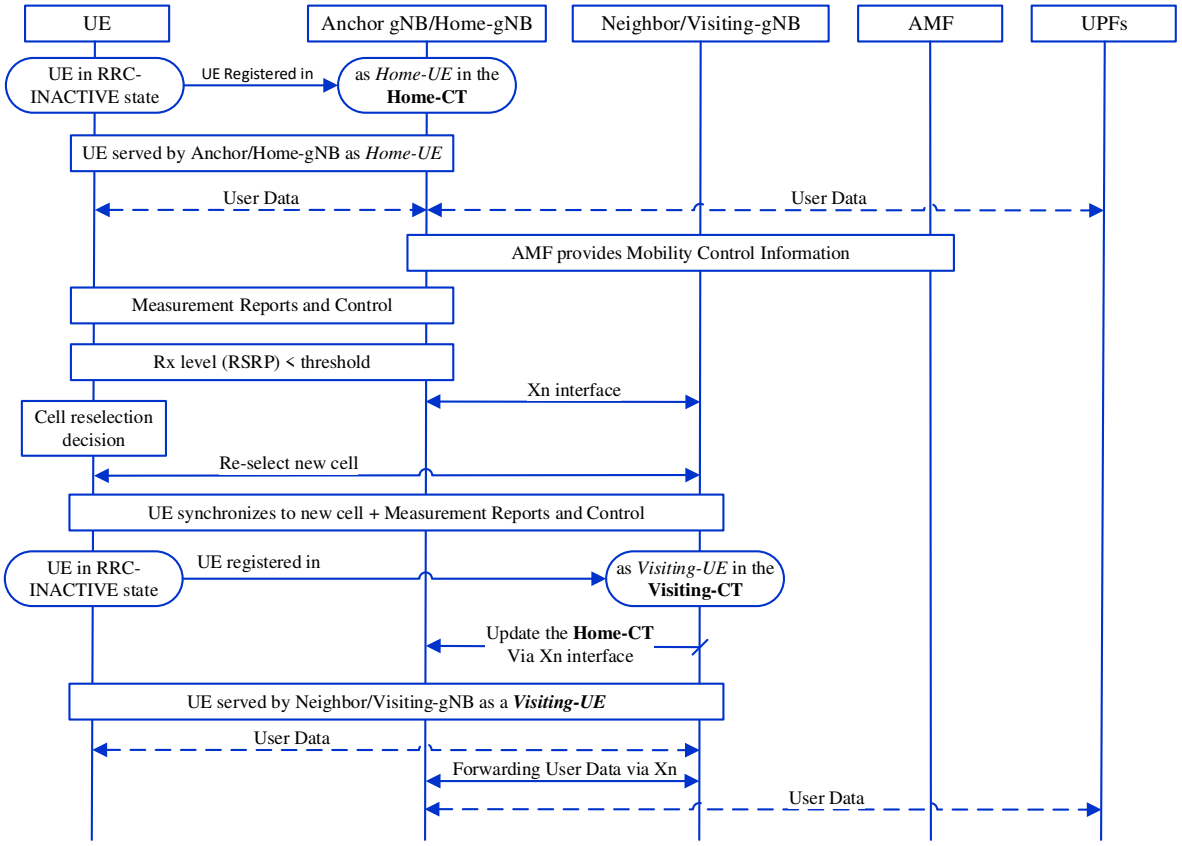



**Figure 4.6:** *Home/Visiting-gNB* and *Home/Visiting-CT* interaction

## 4.6 Illustrative Scenarios

For illustrative purposes, we present some examples to show how the proposed solution, *gNB-based UeMT*, works. In Figure 4.7, IoT/UEs are camped on gNBs based on the RSRP level. In this example, [ue1, ue2, ue3, ue4] are associated with $gNB_1$, and the latter stores their context information, serving as anchor-gNB. These UEs are registered in the *Home-CT* as *Home-UEs* (Figure 4.7 uses colors for clarity). Also, this group of UEs are known to the 5GC under *groupID#*1, for example. Likewise, [ue5, ue6, ue7, ue8] are associated with $gNB_2$, being served as *Home-UEs*, and are known under *groupID#*2. All other parameter are assigned accordingly (*Resident-flag, Visiting-gNB, Visiting-UE, Home-gNB, CiPD, SST*). For presentation purposes, we assume all these IoT/UEs are in the NG-RRC INACTIVE state[8]. To elaborate, we consider the following use cases for *Tracking* and *Locating*.

### 4.6.1 When IoT/UEs Served by Home-gNB

In Figure 4.7, when the AMF needs to reach any UE within *groupID#1* or *groupID#*2, it simply forwards the incoming *Paging* messages to the corresponding *Home-gNB*. Then, the *Home-gNB* sends the *Paging* message to the intended UE (using its *Home-CT*), providing very fast reachability because the *Home-gNB* knows exactly where their UEs are. There is no need to page multiple cells simultaneously as in the conventional *Paging* procedure. Here, one *Paging* message is sent directly to the intended UE.

### 4.6.2 When IoT/UEs Served by Visiting-gNB

Providing services for the *Visiting-UEs*, [ue1, ue4], come with some signaling overhead and delay (caused by packet forwarding and inter-*Paging*). To mitigate this overhead/delay, the control parameters (*CiPD and SST*) of the serving cell, $gNB_2$, are monitored by the *SSP* (see Figures 4.3 and 4.5). After specific thresholds, the *SSP* is initiated (as in Figure 4.5) aimed at transferring the responsibility of these *Visiting-UEs*, [ue1, ue4], from their old *Home-gNB*, $gNB_1$, to be

[8]In 5G, the NG-RRC IDLE state is intended for system maintenance (e.g., link failure) [10]. No *Paging* is needed for the NG-RRC CONNECTED state.

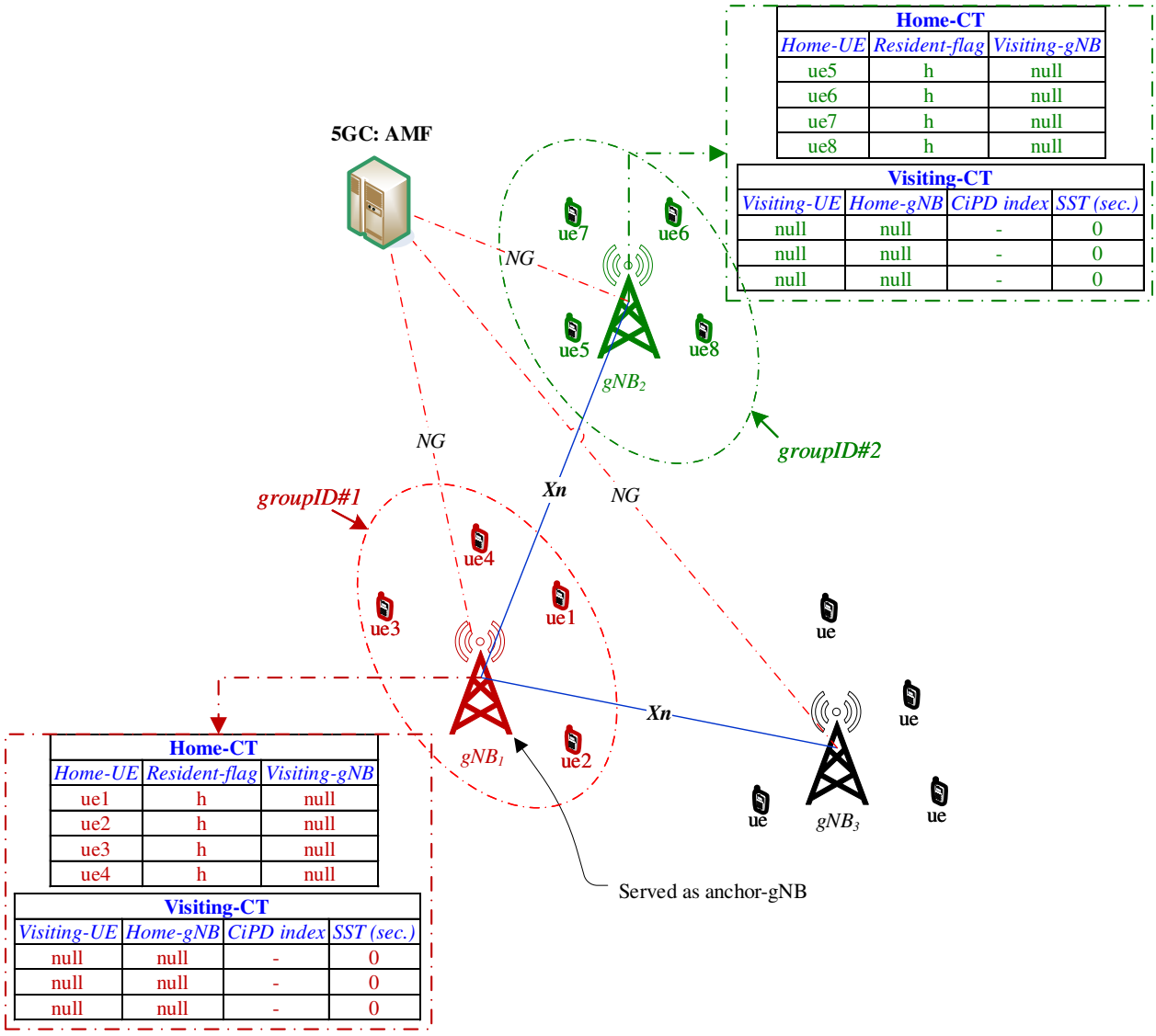



**Figure 4.7:** Shows UEs served by *Home-gNB*

served by the new *Home-gNB*, $gNB_2$, which was serving as *Visiting-gNB* before initiating the *SSP*. As a result, the corresponding *Home-CTs* for both the old and new gNBs are updated, as shown in Figure 4.8. At this time, the new gNB (new *Home-gNB*, $gNB_2$), not the UE, will notify the AMF of this change—the UEs are freed from reporting their location change in our *gNB-based UeMT* solution.

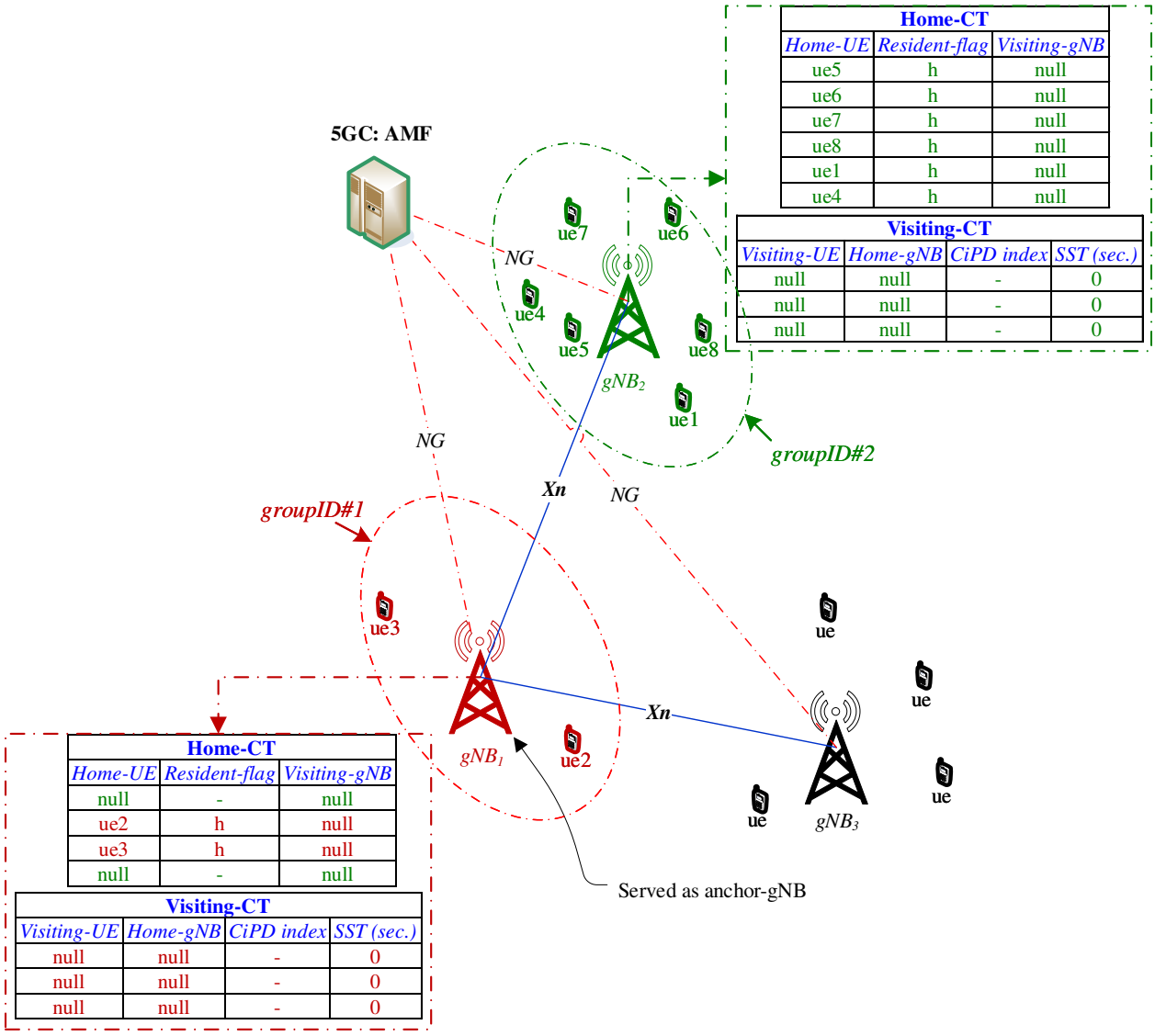



**Figure 4.8:** Shows UEs served by *Visiting-gNB*

## 4.7 Cost Functions for Tracking and Locating

For the purpose of evaluations, we derive and calculate the corresponding signaling overhead and power consumption next. As we have seen throughout this dissertation, both the battery-limited IoT/UEs and network resources are adversely affected by *Tracking* and *Locating* (i.e., *TAU/RNAU* and *Paging*) because of the combined high-volume signaling messages. In this context, we elaborate on the various impacts of these procedures below.

**Table 4.4:** Signaling load of *TAU* and *Paging* in $M$ (adapted from [1–3])

| Network element | TAU ($M$) | Paging ($M$) |
|---|---|---|
| IoT/UE | 9 | 5 |
| eNB/gNB | 14 | 8 |
| MME/AMF | 7 | 8 |

### 4.7.1 Power Overhead

In this section, we calculate how the frequent *TAU/RNAU* impacts the battery power of IoT/UEs while moving. Note that how frequent the mobile IoT/UEs trigger the required *TAU/RNAU* is a function of different parameters, such as IoT/UE mobility patterns/behaviors, network planning (topology), and TAL sizes (as detailed in Section 1.1; see [27] for more detail). To capture these dependencies, we assume that each mobile IoT/UE initiates its *TAU/RNAU* with rate $\lambda_i$ per a given time interval $i$ (e.g., with duration 1 *hour* ($h$)). Accordingly, we write the following formula:

$$Apwr_{\text{tau}} = \sum_{i=1}^{T} Pwr_{\text{tau}} \cdot \lambda_i \tag{4.1}$$

where $Apwr_{\text{tau}}$ is the total accumulated battery consumption during a time duration of $T$ (e.g., $T = 24\ h$) per IoT/UE unit and $Pwr_{\text{tau}}$ is the average battery consumption per *TAU/RNAU* procedure.

### 4.7.2 Signaling Overhead

We now calculate the signaling overhead of the two procedures, measured in the number of the corresponding messages, denoted by $M$. In this context and to elaborate on these message loads ($M$), we have considered the 3GPP technical specification (detailed in [2] and [3])—this is also adopted by [1] to calculate $M$ for each procedure, as detailed in Table 4.4. Note that in our calculations, we consider the signaling loads ($M$) of the following network elements: IoT/UE, eNB/gNB, and MME/AMF—these account for the majority of network elements involved in the two procedures (*Tracking* and *Locating*) [1].

To quantify the message loads and because the two procedures depend on each other (Figure 4.1), we calculate the total accumulated load jointly (per IoT/UE), denoted by $C_{\text{tot}}$ and mea-

sured in $M$ per time interval $T$. We write the following formulas for the accumulated costs of the *TAU* ($Acost_{\text{tau}}$), *Paging* ($Acost_{\text{pag}}$), and *Paging attempts* ($Acost_{\text{att}}$):

$$Acost_{\text{tau}} = \sum_{i=1}^{T} C_{\text{tau}} \cdot \lambda_i \tag{4.2}$$

$$Acost_{\text{pag}} = \sum_{i=1}^{T} C_{\text{pag}} \cdot N_{\text{TAL}} \cdot \sigma_i \tag{4.3}$$

$$Acost_{\text{att}} = \sum_{i=1}^{T} C_{\text{pag}} \cdot N_{\text{TAL}} \cdot P_i \tag{4.4}$$

where $C_{\text{tau}}$ and $C_{\text{pag}}$ are the corresponding message costs of the *TAU/RNAU* and *Paging* (as Table 4.4 shows), respectively, $N_{\text{TAL}}$ is the total number of eNBs/gNBs (i.e., TAL size, as defined in Section 1.1), $\sigma_i$ is the rate of triggering the *Paging* procedure during the time interval $i$ (e.g., with duration 1 $h$), and $P_i$ is the rate of the *Paging attempts* during time interval $i$. Thus, the total accumulated load becomes:

$$C_{\text{tot}} = Acost_{\text{tau}} + Acost_{\text{pag}} + Acost_{\text{att}}. \tag{4.5}$$

By (4.2)–(4.5), we write $C_{\text{tot}}$ in the following form:

$$C_{\text{tot}} = \sum_{i=1}^{T} \left[ C_{\text{tau}} \cdot \lambda_i + (1 + \frac{P_i}{\sigma_i}) \cdot C_{\text{pag}} \cdot N_{\text{TAL}} \cdot \sigma_i \right]. \tag{4.6}$$

Note that we have included $P_i$ to take into account the impact of the multi-*Paging attempts* in which the IoT/UE cannot respond to the first incoming *Paging* message because they might be involved in responding to the *TAU/RNAU* procedure, causing the network to send successive *Paging attempts*. This also increases the *Paging* latency [27]—if the IoT/UEs exhibit high-mobility behavior, this would increase the *TAU/RNAU* occurrence, and hence the *Paging attempts* increase accordingly. In the worst case, *Paging* failure occurs.

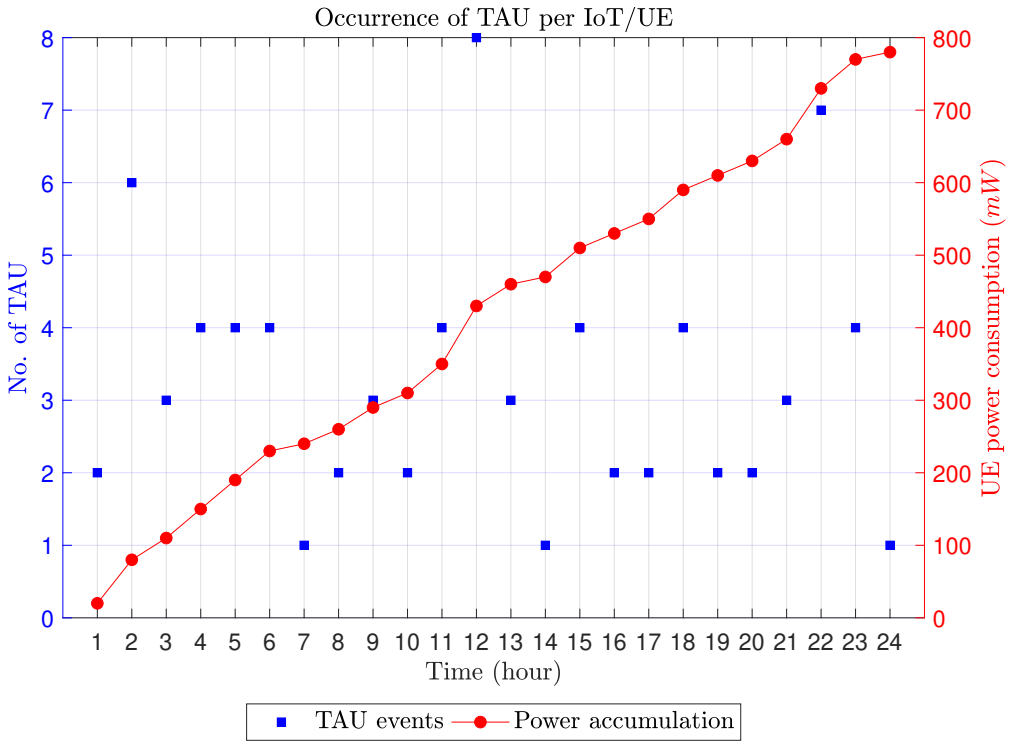



**Figure 4.9:** Draining the IoT/UE battery power while moving

## 4.8 Simulation Setup and Performance Evaluation

### 4.8.1 IoT/UE Battery Power Consumption

Based on (4.1), we examine how the frequent *TAU/RNAU* influences the battery power of a mobile IoT/UE. Assume that the IoT/UE triggers the conventional *TAU/RNAU* according to a Poisson distribution with a rate $\lambda_i$ of 3 tau/h over a time interval of $T = 24$ h. By using (4.1) and setting $Pwr_{\text{tau}} = 10$ mW (recall from Section 4.2), we produce Figure 4.9 to illustrate how the accumulated power consumption increases while the IoT/UE moves through the network over $T = 24$ h. Note that even when the IoT/UEs stay still, they trigger the *TAU/RNAU* periodically about every 60 min—this is called a periodic *TAU/RNAU* (also denoted by *T3412-timer*, defined by the network operator), used to notify the network that the IoT/UEs are still available [39]. So, in Figure 4.9, we do not see occurrences of *TAU/RNAU* below one per hour; see blue boxes. Indeed, such a procedure drains the IoT/UE battery, which is detrimental to the 5G use cases—this becomes even worse when the IoT/UEs exhibit high-mobility behavior. However, one of the *gNB-based UeMT*

features is that the mobile IoT/UEs will no longer use the conventional *TAU/RNAU* to report their location changes; instead, the serving eNBs/gNBs take over this responsibility (*Tracking*), saving IoT/UE battery power.

### 4.8.2 Combined Signaling Costs

Based on (4.6), we show how the conventional *Tracking* and *Locating* procedures impact the network resources, comparing this with the proposed solution, *gNB-based UeMT*. Recall $\lambda_i$ from above for the *TAU/RNAU* rate, and assume that $\sigma_i$ possesses the same characteristic as $\lambda_i$ but at a different rate, e.g., $\sigma_i = 5$ pag/h; this depends on the incoming traffic volume (assume $\sigma_i > \lambda_i$), and $P_i$ has an exponential distribution with mean of 2 att/h. Note that based on the technical specification in [99], $P_i$ should range roughly between 0 (no attempt) and 5. Figure 4.10 shows how the corresponding messaging overhead (from the *Tracking* and *Locating*) burdens the network, comparing the conventional techniques with the proposed scheme. Clearly, Figure 4.10 shows that the signaling overhead is significantly lower in *gNB-based UeMT*. For example, at the end of the time interval $T$ (24 h), the conventional procedures have a dramatically higher signaling overhead ($C_{\text{tot}} = 34,470\ M/T$) than the proposed solution ($C_{\text{tot}} = 2,499\ M/T$)—the *gNB-based UeMT* achieves a reduction of 92% in the signaling cost, including the battery power saving in the IoT/UEs.

### 4.8.3 Final Notes

As shown above, the *gNB-based UeMT* solves the conventional *TAU/RNAU* and *Paging* problems, which are still the default in current LTE and 5G systems. Up to the time of writing this dissertation, all existing solutions involve a trade-off between the *TAU* and *Paging* signaling overhead, but the optimization problem has not been solved efficiently (because of the dependency; see Figure 4.1) until the solution proposed here. On top of that, UEs themselves play an important role in these other solutions [27]. Some solutions rely on collecting frequent information about the mobility patterns of the UEs and then will use this information for UE location predictions, which is costly in terms of computation and implementation complexity [27].

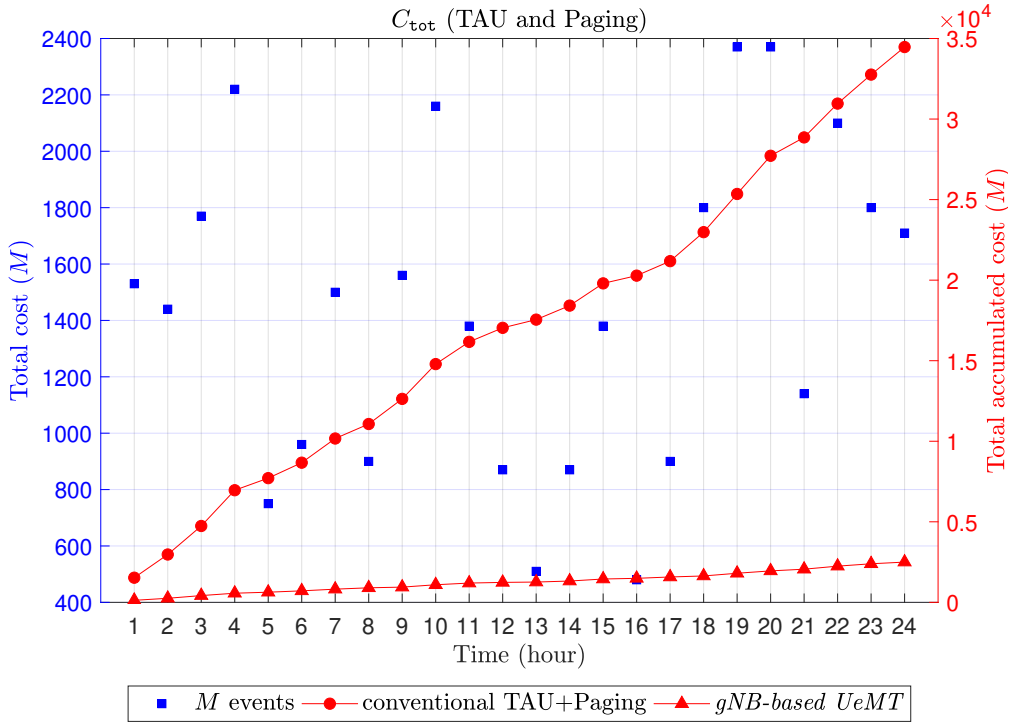



**Figure 4.10:** Signaling cost for *TAU/Paging* versus *gNB-based UeMT*

Finally, it is very important here to emphasize that the current *TAU/RNAU* procedure has high signaling overhead, significantly higher than for *Paging*—this is because when IoT/UEs need to report their location changes, they establish more signaling messages/channels than the *Paging* procedure. In other words, *Tracking* is more expensive than *Locating*, as quantified by Table 4.4 and illustrated by Figure 4.1—this decisively answers the question posed in Section 4.3: *"Which procedure is the more important of the two, and why?"*

## 4.9 Summary

The *gNB-based UeMT* solution achieves the following outcomes: It completely bypasses the *TAU/RNAU* procedure and the accompanying signaling cost, significantly improves the *Paging* procedure, and provides always-known IoT/UE locations. Also, this solution monitors the delay and the signaling cost to guarantee that the overhead does not exceed a specified threshold, providing lightweight signaling overhead.

Furthermore, avoiding the *TAU/RNAU* procedure will remove the relevant signaling overhead and add more power saving to the battery-limited IoT/UEs, a critical requirement for 5G—the signaling overhead is reduced by about 92%. The *gNB-based UeMT* scheme adds no implementation complexity or computation cost. Instead, it exploits the already existing protocols/functions, such as RSRP report, anchor-gNB for 5G, and Xn interface protocol. Also, this solution provides very fast IoT/UE reachability compared with the conventional *Paging* procedure. While the IoT/UEs move throughout the network, *gNB-based UeMT* provides always-known UE locations. This is to provide a very fast way to reach each mobile IoT/UE with lightweight signaling cost, supporting mission-critical applications in 5G.

Furthermore, by leveraging the key features of the *gNB-based UeMT* solution, we apply this solution to a new vision of life-critical missions, aiming to recover the mobile wireless network after large-scale disasters (for search-and-rescue operations), which is the main focus of the next chapter.

# Chapter 5

# User-Oriented Mission-Critical Communication

## 5.1 Overview

In this chapter, we introduce a new approach for Search-and-Rescue Operations (SAROs) to search for survivors after large-scale disasters, assuming the wireless communication network cells are partially operational and exploiting the recent trend of using Unmanned Aerial Vehicles (UAVs) as a part of the network. These SAROs are based on the idea that almost all survivors have their own wireless mobile devices, called User Equipment (UE), which serve as human-based sensors on the ground. Our approach is aimed at accounting for limited UE battery power while providing critical information to first responders: 1) generate immediate crisis maps for the disaster-impacted areas, 2) provide vital information about where the majority of survivors are clustered/crowded, and 3) prioritize the impacted areas to identify regions that urgently need communication coverage. Note that the material in this chapter has been published in part in [30, 31] and in whole in [32].

Mission-Critical and Public-Safety Communications (MCPSCs) are intended to provide vital mobile wireless communication services for first responder entities, such as police and firefighters, enabling them to exchange information during emergency situations. In the following subsection, we discuss the main trends in MCPSCs. Following that, we describe a potential point of failure in current MCPSC systems. Details of our approach, based on UAVs as network elements, begin in Section 5.4, after discussing post-hazard issues in Section 5.2 and reviewing the relevant literature in Section 5.3.

### 5.1.1 Current MCPSC Systems

Many conventional communication systems have been deployed to support MCPSCs. Since the 1930s, Public Safety Agencies (PSAs) have considered Land Mobile Radio (LMR)[9] systems

[9]Basically, LMR systems are terrestrially-based networks of portable/mobile radios, base stations, and repeaters.

as the primary means to support MCPSCs for voice communication among emergency responders [100]. LMR systems are limited to voice and low-speed data communication. Other MCPSC systems, notably Terrestrial Trunked Radio (TETRA) and Project 25 (P25), are still currently in service, although they are inefficient in terms of spectral utilization, data rate, and cost [101, 102]. Thus, many PSAs have migrated from conventional LMR systems to more advanced mobile broadband systems. TETRA and Critical Communications Association (TCCA) have asserted that the commercial Long Term Evolution (LTE) and its next generation (5G) are the most promising technologies for MCPSCs [101, 103, 104]. For this reason, in 2012, the US developed a nationwide MCPSC system called FirstNet, which uses the current LTE network as a basic platform; the US has spent $7 billion and reserved the use of the 700 MHz band for FirstNet communication. A major recent milestone along these lines is that AT&T announced that it will spend $40 billion toward developing FirstNet as a global wireless network dedicated to the US first responders, according to the First Responder Network Authority-AT&T 2018 contract [105].

Concurrently, the 3rd Generation Partnership Project (3GPP) has developed a specific set of mission-critical standards not only for LTE but also its successor 5G to support MCPSC functionalities. These 3GPP standards comprise Proximity-Service (ProSe) [106], Mission Critical Push to Talk (MCPTT) [107], Group Communication System Enabler (GCSE) [108], and network enablers for critical communications [109].

It is clear that there is a pressing need for reliable, extremely efficient and effective, and quick access networks for PSAs to handle life-critical missions. This interoperability between PSAs and existing commercial wireless networks will be extremely vital for MCPSC missions because the latter covers almost all the living population. For example, around 98% of the US population live in areas covered by LTE technology [13]—in this case, the PSAs can communicate even without TETRA/P25 Radio Frequency Coverage (RFC). Moreover, the mobile wireless communication over LTE/5G are beneficial not only to the PSAs but also to the people in need; they can use their smart phones for video streaming, making texts/calls, and even location sharing of UEs wherever they are located in an area of consideration, called the Region of Interest (RoI). This allows PSAs

to be better informed about the emergency status and hence prioritize their operations to save lives and manage the available resources. But MCPSC systems are susceptible to challenges that are unavoidable, which we address next.

### 5.1.2 Failure of Current MCPSC Systems

As we have seen above, there are numerous communication technologies dedicated to PSAs, the most prominent being the interworking between LTE/5G and FirstNet. For example, Los Angeles deployed about 231 sites as a first step toward the FirstNet project in March 2014 [102]. This is needed to keep the PSAs connected—anytime, almost anywhere, and in any emergency situation. At the same time, communication between PSAs and other persons (e.g., potential victims) is also crucial for life-saving purposes. However, LTE/5G and FirstNet technologies can be dysfunctional temporarily after a hazard—the network infrastructure can be devastated partially or completely by natural disasters (e.g., earthquake, hurricane, or tsunami) or even by attacks. In the worst case, the communication between the PSAs and disaster victims becomes impossible. Specifically, Search-and-Rescue Operations (SAROs), mostly location-based missions dedicated to life-saving, become extremely difficult. In such cases, it is important for the PSAs to have some awareness of where the disaster victims are mostly located or clustered, so that the PSAs can conduct SAROs in a timely and more effective manner. But how do we obtain sufficient information on disaster victim locations without the ability to communicate? This is the main focus of this chapter.

## 5.2 Network Status Post-Disaster

After a disaster, some of the wireless base stations may not survive (henceforth, we will use the abbreviation gNB for such base stations, as this is the abbreviation used in 5G). For example, after Hurricane Maria hit on September 21, 2017, 95.6% and 76.6% of the cellular sites were dysfunctional in Puerto Rico and the US Virgin Islands, respectively [110]. Accordingly, the serving

network and its users in the RoI are adversely impacted in various ways, which we highlight as follows.

### 5.2.1 Lack of RFC

The surviving gNBs provide limited RFC only to UEs in close proximity. But not all UEs can exchange information with the surviving gNBs because the latter can serve only a limited number of UEs. Other UEs in the same area might receive a good level of Reference Signal Received Power (RSRP), but cannot access the available network. More specifically, these UEs try to associate with these gNBs by sending multi-access requests simultaneously without success, thus producing congested gNBs in that area.

### 5.2.2 Isolated gNBs

Potentially, the surviving gNBs are unable to communicate—the necessary links (called X2 or Xn in LTE and 5G, respectively [97]) between them are disconnected, leaving these gNBs isolated from each other. Furthermore, as long as the surviving gNBs are scattered across the RoI (and isolated), it is difficult for the PSAs to reach these UEs by wireless communication. In such cases, the SAROs are crucial—victims might be trapped or isolated and risk not being found and rescued.

### 5.2.3 Cell-Edge UEs

At the cell edges of a surviving gNB, UEs might struggle to associate with the gNB because of low RSRP levels. Moreover, parts of the RoI might have no RFC at all. In this case, the UEs start searching for a suitable cell (gNB) to camp on, initiating what is called the Cell Search Procedure (CSP) [111, 112]. Typically, as long as no suitable serving cell is found, the UEs continue to perform the CSP, attempting to find one. This gives rise to a power consumption problem for the UEs—most UEs are battery-limited, and hence conducting CSPs continually without success drains the battery power in these UEs. Eventually, they will be out of service and unreachable, remaining lost even when the RFC is restored.

In this context, this chapter deals with large-scale disasters in which the RFC area is limited or nonexistent, leaving the surviving UEs struggling to get connected.

## 5.3 Related Studies

Many different solutions have been introduced to address the problem of lack of wireless communication between the PSAs and victims in emergency situations. Here, we classify the existing solutions for network restoration into three main groups based on the particular approaches taken, as follows.

### 5.3.1 Deploying Wireless Equipment into RoI

Early solutions have been proposed for emergency managements and triaging patients, allowing first aid teams to prioritize their efforts, named "ARTEMIS" and "CodeBlue," as in [113] and [114], respectively. These two systems are similar in design but differ in data transmission protocols. The two systems deploy wireless-based sensors (for monitoring victims' vital signs) into the RoI. For the data transmission, in [113], medics (with hand-held devices) can move within a deployed ad-hoc wireless network, in which data from multiple devices can be transmitted to remote high-level medical personnel. In [114], the authors propose to create a dedicated wireless sensor network throughout the RoI, comprising multi-purpose sensors (e.g., location and biomedical sensors), used for data transmission.

For military and battlefield assistance, the author of [115] develops a system to track and identify causalities in severe environments, called the Tactical Medical Coordination System (TacMedCS). This system comprises a set of hand-held devices to collect vital signs of victims (including their locations and IDs), providing near real-time awareness of casualty status and allowing medics to respond quickly. In the absence of wireless communication, TacMedCS uses satellite phones for data transmission.

Recently, unlike the solutions in [113–115], Nokia introduced man-portable and vehicle-mounted LTE eNBs to provide temporary LTE RFC for the RoI [116], recovering the network and enabling the PSAs to communicate with disaster victims.

The solutions in [113–116] are effective in dealing with small-scale emergency situations, where the PSAs or vehicle-mounted eNBs can move freely into the RoI. However, in large-scale disasters (e.g., earthquakes), such solutions can fail because of the difficulty in moving into the disaster area quickly (e.g., because of ground rupture and landslides).

### 5.3.2 Network Recovery Using D2D Communication

A well-known technique has emerged to enhance the overall performance of current LTE networks, called Device-to-Device (D2D) communication (also called the 3GPP ProSe feature in LTE) [117]. This is to enable UEs in close proximity to communicate through direct links without passing through eNBs. Exploiting this feature, the authors of [118] introduce a D2D communication scheme and a clustering procedure for network recovery. This study addresses the energy efficiency and battery lifetime of UEs, but it requires special devices to be deployed that have high transmission power, long battery lifetime, and the ability to control radio resources. These are critical requirements because such devices are not widely used or available to the end-users. Furthermore, the PSAs cannot easily deploy such devices in large-scale disasters (as detailed in Section 5.3.1).

Similarly, the authors of [119] propose an efficient network architecture using D2D communication for disaster situations when the network infrastructure is partially unavailable. The authors of [119] use multi-hop concepts of D2D to extend the network coverage of functional eNBs to regions where the coverage is unavailable. This is done by using Relay Nodes (RNs) that route wireless coverage toward uncovered areas. Although this work shows some benefits of using multi-hop D2D for extending network coverage and reducing transmission power, its availability would be limited—the RNs are mostly typical UEs, which are limited in battery power and processing

capabilities. Moreover, if these RNs move, the system would select new suitable RNs; this process impacts the system complexity and stability (because of frequent association and dissociation).

To address the power constraint in the solution of [119], the authors of [120] introduce a Wireless Energy Harvesting (WEH) scheme, exploiting the ability to convert the received RF into energy. In this scheme, the RNs are able to transmit data and energy to UEs via RF. Although this work has shown that WEH can reduce the power consumption of UEs, UEs need to be equipped with RF energy-harvesting circuitry, which is not available in common UEs. Also, when RNs move, it would impact the system stability.

Like in [120], the authors of [121] introduce an energy-efficient UE discovery scheme under interference constraints, called D2D Discovery Maximization (D2D-DM), providing a switching capability for discovery modes (half-duplex and in-band full-duplex). Specifically, when the signal-to-interference-noise ratio of a D2D link drops below a specified threshold, the discovery mode switches from half-duplex to in-band full-duplex. In addition, for battery-limited UEs, the authors adopt an open-loop power control scheme to reduce power consumption. According to [121], the D2D-DM scheme shows a significant improvement in the number of discovered UEs as compared to static resource allocation and the random backoff scheme.

Recently, in the context of WEH, the authors of [122] propose a D2D-based framework with energy-efficient clustering and routing for disaster communication relief. This solution shows a significant improvement in terms of power consumption and end-to-end transmission delay compared to the solutions in [119, 120]. The authors use a Particle Swarm Optimization (PSO) algorithm for routing and clustering. PSO is a time consuming algorithm, especially when the search space is large and it does not guarantee finding an optimal solution [123]. The use of PSO adds to the overhead and complexity, which compromises the success in dealing with large-scale disasters. Solutions that add no overhead are naturally preferable.

### 5.3.3 Network Recovery Using UAV Communication

Recently, some studies have proposed the use of Unmanned Aerial Vehicles (UAVs) as mobile gNBs for rapid network recovery. These UAVs are useful in a variety of applications, especially in wireless mobile communication [124]. More recently, this attention has brought current LTE closer to supporting UAV communication [125]. The authors of [126] provide a comprehensive survey of using UAVs in public-safety communication, highlighting power consumption issues. In addition, the authors propose a multi-layered architecture for emergency situations, providing alternative paths for emergency communication. The study in [126] highlights a variety of issues that are related to UAV placement, UAV communication links, and UAV trajectory plans.

Exploiting this trend, the authors of [127] propose an optimization scheme to place the UAVs, improving the 5th percentile capacity of LTE networks. This UAV placement scheme aims to improve the network throughput. However, in [127], the UAV locations are optimized using brute-force search, which can be prohibitively time consuming. Moreover, the placement scheme of [127] depends on how the surviving gNBs are distributed. Doing so does not account for where the deployed UAVs are mostly needed; it ignores where the majority of UEs are located.

To provide full coverage to all users in the RoI, the authors of [128] propose a solution scheme in which a very large number of UAVs are deployed to cover all UEs. Although the scheme provides full RFC, it is inefficient and possibly impractical; the complexity is proportional to the size of the RoI, and the scheme might provide unnecessary RFC.

While such developments are beneficial to emergency communication, they give rise to the concomitant issue of cybersecurity in UAV communication. This issue is important because UAV communication often involves critical or sensitive data. The cybersecurity vulnerabilities in UAV communication are beyond the scope of this dissertation; see [129] for more detail.

In sum, the preceding studies have proposed many solution schemes for wireless network restoration after hazards. However, they do not focus on where the majority of UEs are located to provide the necessary coverage in a timely manner. Furthermore, at the time of writing this disser-

tation, no solution provides the corresponding PSAs with information on surviving UE locations for the purpose of SAROs. In this chapter, we address these issues.

## 5.4 Solution Approach

### 5.4.1 Issues to be Considered

The issues highlighted in Section 5.2 are the main focus of this chapter. As we have seen from the recent studies, the focus has been on how to provide RFC in the RoI instead of finding the UEs. Specifically, it is crucial not only to provide RFC but also to locate these UEs—where and how the UEs are clustered. After disasters, the location distribution of surviving UEs is likely to be nonuniform. Estimating this distribution provides critical situational awareness to PSAs and helps to focus the use of scarce resources. In addition, it is important to locate these UEs without their assistance—these UEs might be unable to telecommunicate because of the lack of wireless service, network congestion, injuries, unconsciousness, or even unresponsiveness. Awareness of victim locations is a very critical requirement for the MCPSCs, often racing against time.

### 5.4.2 Proposed Approach

Considering the above issues, we propose a new method for SAROs to find potential survivors by finding their UEs without their assistance, even in the absence of RFC, while providing temporary RFC based on certain prespecified priorities (e.g., number of survivors in each sub-area of the RoI; later, we call this sub-area the *Searching Zone (SZ)*). Conducting SAROs by searching for surviving UEs to locate individuals is effective especially because these UEs have become more ubiquitous—each individual (likely) is equipped with at least one of the following: smart phones, tablets, smart watches, or even embedded sensors in the human body or clothing; most of these devices are embedded with RF equipment. This is a quick way to provide vital information to the PSAs even before they arrive at the scene.

In this chapter, the SAROs are based on the idea that each individual has its own UE and potentially is still alive and willing to be found, and hence we call our solution ***UE-based SAROs***.

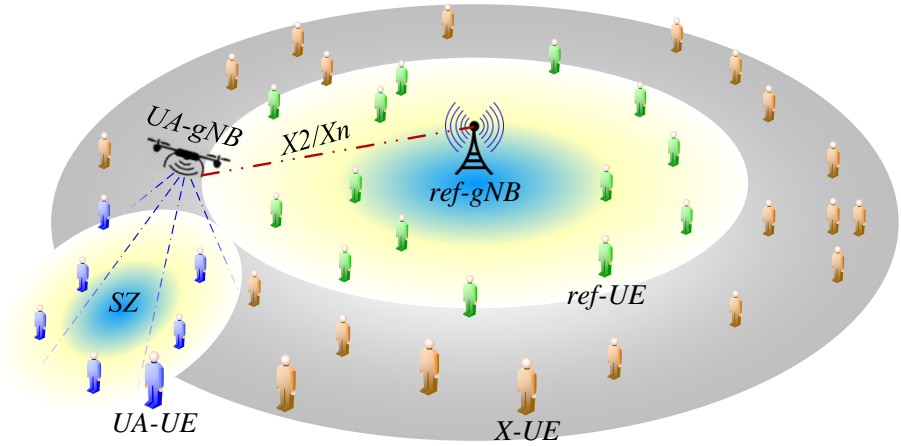



**Figure 5.1:** Illustrative example showing the SARO entities

Specifically, this work is mainly intended to (among other functions, as we will see later) generate immediate crisis maps[10], providing information to the corresponding PSAs to prioritize their operations in disaster-affected regions. Essential entities of *UE-based SAROs* are discussed next.

## 5.5 *UE-based SARO* System Model

In our solution, we use UAVs as mobile gNBs, called *UA-gNBs*, as a part of the network infrastructure in the RoI. We assume that the impacted network is only *partially* dysfunctional; some of gNBs are still able to broadcast and exchange signaling. Before proceeding further, we define the essential entities in our solution (illustrated in Figure 5.1 in color for clarity), as below.

### 5.5.1 Entity Definitions

1. *ref-gNBs*: These are the surviving gNBs, called reference gNBs (*ref-gNBs*); see Figure 5.1. These *ref-gNBs* provide Radio Resources Management (RRM) functionalities, such as resource allocation, scheduling, and mobility control [9, 131]. The deployment of a *UA-gNB* does not need its own RRM. Instead, the *ref-gNBs* provide the necessary RRM to the corre-

[10]Google Crisis Map [130] is a well-known example of a crisis map, but its availability needs Internet connectivity and does not provide immediate information about how and where individuals are distributed.

sponding *UA-gNBs*. This is to minimize the load on these *UA-gNBs*, to address limitations in battery power and processing capabilities.

2. *UA-gNBs*: These are the deployed UAVs, as mentioned above. We use them to provide mobile picocells with range expansion capabilities, with a cell radius of 100–300 $m$ and transmit power of 24–33 $dBm$ [132]. Each *UA-gNB* has five main functions: 1) search for a *ref-gNB* to associate with, establishing X2/Xn interfaces[11]; 2) search for surviving UEs that are actively seeking a serving cell to camp on, around the detected *ref-gNB* (based on a screening procedure we define later), broadcasting UE-specific control messages; 3) feed back the screening results to the corresponding *ref-gNB*, via X2/Xn, for further processing/analysis; 4) provide the necessary RFC according to where the UEs are in need (decisions made by the *ref-gNB*); 5) while conducting the screening procedure, the *UA-gNBs* and *ref-gNBs* broadcast paging messages to the corresponding UEs, including emergency alert messages, using the already existing public warning system in LTE, known as Commercial Mobile Alert System [40]. For example, the *UA-gNBs* may broadcast the following message: *"If your location is safe, stay; otherwise go the nearest safe location and remain there. Refrain from using your mobile phone to conserve battery; we will reach you by phone."*

3. *ref-UEs*: These are surviving UEs that have been associated and registered with *ref-gNBs* (e.g., because of their close proximity) as shown in Figure 5.1 (in green).

4. *UA-UEs*: These are surviving UEs that have been discovered and associated with *UA-gNBs* after the screening procedure. As mentioned above, the association information will be sent to the corresponding *ref-gNB* for further processing. The *UA-UEs* are shown in Figure 5.1 (in blue).

5. *X-UEs*: These are surviving UEs but not associated to any gNB. While the RFC is not available (or the received RSRP is too low), the *X-UEs* continually execute the CSP, which

[11] Essential interfaces between gNBs in LTE/5G networks for exchanging necessary control signaling [97].

consumes the battery power of the UEs. These *X-UEs* need to be found as quickly as possible; otherwise, as stated before, they might go out of service. Figure 5.1 shows these *X-UEs* (in brown).

6. *Searching Zones (SZs)*: To facilitate the screening and searching for *X-UEs*, the area around each *ref-gNB* is partitioned into a set of sub-areas, called *SZs*. Each *UA-gNB* performs the screening procedure in its own assigned set of *SZs*. These *SZs* are known to the *ref-gNBs*, as detailed later in Section 5.9. The *SZs* are also used to generate crisis maps called *UEBCMs* (defined below).

7. *Priority-Driven RFC (PDRFC)*: After the screening procedure, decisions will be made by each *ref-gNB* to identify the areas that are in need of immediate RFC based on UE clustering, which we call *PDRFC*. The resulting *PDRFC* includes a set of *SZs* arranged in priority orders.

8. *UE-based Crisis Map (UEBCM)*: Based on the collected information (e.g., *PDRFC*), each *ref-gNB* generates its own crisis maps, called *UEBCMs*. These maps will contain all the necessary information (*ref-gNB* locations, surviving UE locations, and the corresponding RSRP measurement reports). These maps will be accessible to the PSAs later.

### 5.5.2 Entity Notation

For ease of presentation, we introduce some precise notation for the entities defined in the last section:

1. *ref-gNBs*: We denote the set of their locations by $\mathcal{R} = \{(x_i^{\text{ref}}, y_i^{\text{ref}}) : i = 1, 2, \ldots, R\}$, where $(x_i^{\text{ref}}, y_i^{\text{ref}})$ is the location of *ref-gNB*$_i$ and $R$ is the total number of the detected *ref-gNBs*. These form a sub-set of all gNBs, whose locations are denoted by the set $\mathcal{A}$ (i.e., $\mathcal{R} \subset \mathcal{A}$).

2. *UA-gNBs*: We denote the set of associated *UA-gNB* locations relevant to *ref-gNB*$_i$ by $\mathcal{L}_i = \{L_{i,j} : j = 0, 1, \ldots, J_i - 1\}$, where $j$ is the index of the underlying *SZ* and $J_i$ is the total number of the *SZs* around *ref-gNB*$_i$ (defined in item 5 below), as detailed in Figure 5.2 ($L_{i,j}$ and $J_i$

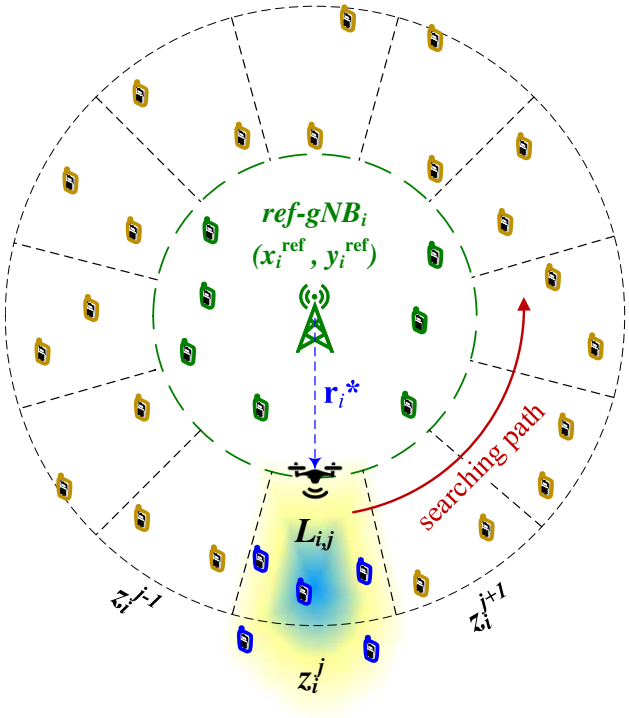



**Figure 5.2:** *UA-gNB* searching model

are calculated in Section 5.9). Here, we assume that for each detected *ref-gNB*$_i$ there is one associated *UA-gNB*$_i$ (they have the same index $i$, as defined in item 1 above).

3. *ref-UEs*: We denote the set of *ref-UEs* associated with *ref-gNB*$_i$ by $\mathcal{G}_i = \{ue_{i,q}^{\text{ref}} : q = 1, 2, \ldots, Q_i\}$, where $Q_i$ is the total number of associated UEs. For example, if *ref-gNB*$_3$ has a total of 100 associated UEs, then $\mathcal{G}_3 = \{ue_{3,1}^{\text{ref}}, ue_{3,2}^{\text{ref}}, \ldots, ue_{3,100}^{\text{ref}}\}$.

4. *UA-UEs*: We denote the set of *UA-UEs* associated with *UA-gNB*$_i$ within *SZ* $j$ by $\mathcal{S}_i^j = \{ue_{i,p}^j : p = 1, 2, \ldots, P_i^j\}$, where $P_i^j$ is the total number of associated UEs within zone index $j$. For example, suppose that *UA-gNB*$_5$ has screened the zone with index 2 (located at $L_{5,2}$) and detected 20 UEs. Then, the set of these UEs is identified as $\mathcal{S}_5^2 = \{ue_{5,1}^2, ue_{5,2}^2, \ldots, ue_{5,20}^2\}$.

5. *SZs*: We denote the *SZs* surrounding *ref-gNB*$_i$ by $\mathcal{Z}_i = \{z_i^j : j = 0, 1, \ldots, J_i - 1\}$, where $J_i$ is defined in item 2 above. For example, if *ref-gNB*$_2$ has 12 surrounding *SZs*, then $\mathcal{Z}_2 = \{z_2^0, z_2^1, \ldots, z_2^{11}\}$. The set $\mathcal{Z}_i$ surrounds the cell edge of the corresponding *ref-gNB*$_i$.

## 5.6 *UA-gNB* Searching Procedures

The *UA-gNB* performs two essential searching procedures. The first (detailed in Section 5.6.1) is to find a *ref-gNB* to obtain the required RRM. Second (detailed in Section 5.6.2), and after association with a *ref-gNB*, the *UA-gNB* starts searching for surviving UEs (*X-UEs*), which likely are starving for a serving cell. Because the *UA-gNBs* and UEs are battery-power limited, the following concerns arise: 1) the *UA-gNB* should find a *ref-gNB* as quickly as possible to save its battery capacity; 2) the *X-UEs* (in brown in Figure 5.1) should be found and located within a reasonable time, by the searching *UA-gNB*, before many of these UEs run out of battery power. These concerns involve time-critical requirements. Thus, the search schemes must be time efficient, as described next.

### 5.6.1 Procedure for Finding *ref-gNB*

To locate a *ref-gNB*, the *UA-gNB* can scan the whole RoI or use some prediction algorithms, but under the above requirements, such searching schemes are too inefficient. Our procedure involves two essential schemes, by which the searching *UA-gNB* finds its best candidate location and optimal distance from the corresponding *ref-gNB*, as we will describe in Sections 5.6.1.1 and 5.6.1.2, respectively.

To expedite the searching efforts and develop time-efficient strategies, we consider the following factor. Typically, all the gNBs at locations in set $\mathcal{A}$ are already deployed according to a predefined plan—the locations in $\mathcal{A}$ are distributed based on where the RFC is most needed (e.g., hot spots and crowded areas). Hence, the locations in $\mathcal{A}$ (also defined by their Physical Cell Identifications (PCIs)) are known to the *UA-gNBs*. But initially (after hazards), the *UA-gNBs* do not know $\mathcal{R}$; i.e., where the potential *ref-gNBs* are located. In other words, the *UA-gNBs* need to find and locate operational potential *ref-gNBs*—this is necessary for association purposes (for RRM; see Section 5.5.1). By exploiting the known $\mathcal{A}$, we introduce the following scheme for *UA-gNBs* to find *ref-gNBs* to associate with.

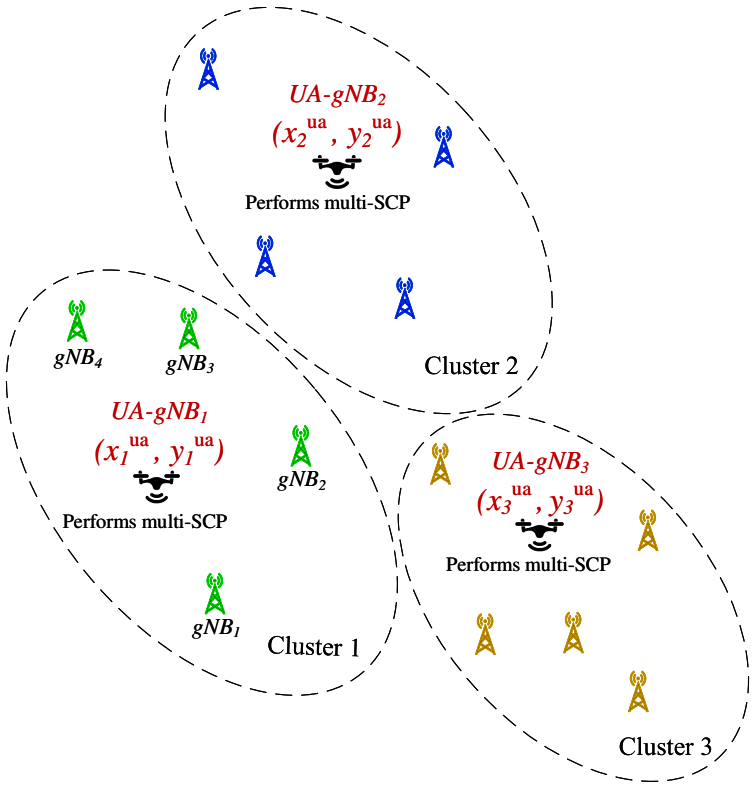



**Figure 5.3:** Illustrative example showing *UA-gNBs* centers at cluster centroids, using k-means

#### 5.6.1.1 Cluster centroid-based search (CCBS)

In this scheme, the *UA-gNBs* localize all the *ref-gNBs* ($\mathcal{R}$) simultaneously with low computation overhead, which we detail in the following steps:

1. The RoI is partitioned, based on the location set $\mathcal{A}$, into $K$ groups using some well-known partitioning algorithm, such as *k-means++* [133]. This is a simple and fast way to find points that serve as centroids for each partition subset of $\mathcal{A}$, to serve as initial searching points for *UA-gNBs*. These initial points are defined by the set $\mathcal{K} = \{(x_k^{\text{ua}}, y_k^{\text{ua}}) : k = 1, 2, 3, \ldots, K\}$, where $K$ is the total number of cluster centroids. If we assume that for each defined cluster $k$ there is exactly one searching *UA-gNB*, $K$ will be equal to the total number of the searching *UA-gNBs*, and hence (based on the previous assumption in Section 5.5.2, item 2), there will be one *ref-gNB* for each cluster $k$, as illustrated in Figure 5.3.

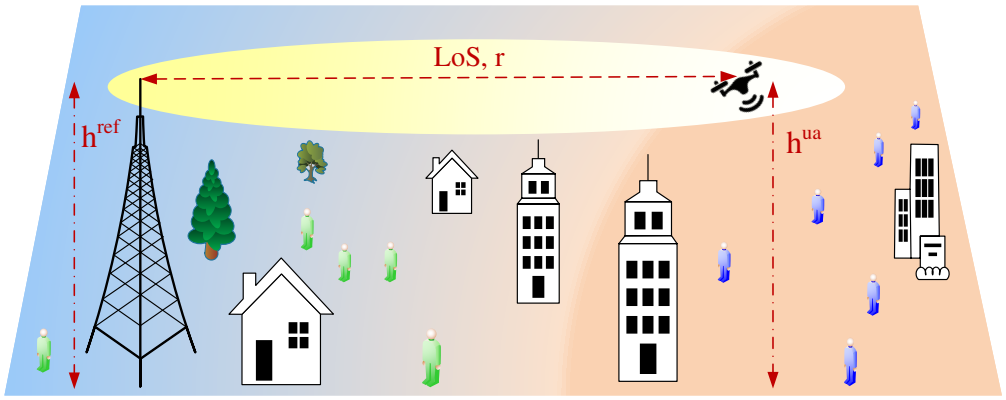



**Figure 5.4:** Illustrative example showing the LoS and height conditions

2. In the *CCBS* scheme, for each cluster centroid, there is one *UA-gNB* located initially at $(x_k^{\text{ua}}, y_k^{\text{ua}})$; $K$ *UA-gNBs* are assigned, one each, to all $K$ cluster centroids in $\mathcal{K}$. Because *UA-gNBs* are deployed at some altitude, it is likely for them to receive Line-of-Sight (LoS) signaling from multiple potential *ref-gNBs*. Furthermore, deployment of *UA-gNBs* at appropriate altitudes ensures that the ground UEs would receive good levels of RSRP while getting screened by these *UA-gNBs*, as we will see in Section 5.6.2. The LoS distance is denoted by **r**, as illustrated in Figure 5.4.

3. While flying around its initial location (from the set $\mathcal{K}$), the *UA-gNB* performs multi-cell search (using the conventional SCP). Once the *UA-gNB* detects a serving cell, the PCI of the decoded cell is identified (a *ref-gNB* found), and hence its location, $(x_i^{\text{ref}}, y_i^{\text{ref}})$, becomes known. In this case, the *UA-gNB* associates with this *ref-gNB* for exchanging the necessary information, detailed further in Section 5.10.1.

#### 5.6.1.2 *UA-gNB* optimal distance

The associated *UA-gNB*$_i$ has to search for *X-UEs*, which are mostly located at the cell edge of *ref-gNB*$_i$ (detailed in Section 5.2.3 and shown in Figures 5.1 and 5.2), assuming the *ref-UEs*, those in $\mathcal{G}_i$, have already associated with *ref-gNB*$_i$. We describe the procedure to search for *X-UEs* in the following steps:

1. *UA-gNB$_i$* sets its initial distance ($\mathbf{r}_i$) from *ref-gNB$_i$* such that it can search its *SZs*, $\mathcal{Z}_i$, circulating around the cell edge. The initial distance is calculated as follows:

$$\mathbf{r}_i = \sqrt{(x_i^{\text{ref}} - x_k^{\text{ua}})^2 + (y_i^{\text{ref}} - y_k^{\text{ua}})^2}, \tag{5.1}$$

where $(x_i^{\text{ref}}, y_i^{\text{ref}})$ and $(x_k^{\text{ua}}, y_k^{\text{ua}})$ are defined in $\mathcal{R}$ and $\mathcal{K}$, respectively. To locate the *ref-gNB$_i$* cell edge, we use the Path Loss (**PL**) to calculate a maximum distance, called $\mathbf{r}_i^*$, subject to the constraint that **PL** does not exceed a predefined value $\mathbf{PL}_{\text{threshold}}$. This constraint is required to maintain the communication link between them (Xn). In this context, we consider the following **PL** formula, which is widely used in the literature (for urban and suburban areas) for system-level simulations [134]:

$$\begin{aligned}\mathbf{PL}(\mathbf{r}_i) = 40 \cdot (1 - 4 \cdot 10^{-3} \cdot \mathbf{h}_i^{\text{ref}}) \cdot \log_{10}(\mathbf{r}_i) \\ - 18 \cdot \log_{10}(\mathbf{h}_i^{\text{ref}}) + 21 \cdot \log_{10}(f) + 80\ dB,\end{aligned} \tag{5.2}$$

where $\mathbf{r}_i$ is the distance (in kilometers) between the *ref-gNB$_i$* and *UA-gNB$_i$*, $f$ is the carrier frequency in MHz, and $\mathbf{h}_i^{\text{ref}}$ is the *ref-gNB$_i$* antenna height (in meters), measured from the average rooftop level.

2. Now we calculate the maximum distance $\mathbf{r}_i^*$ (the radius of the *ref-gNB$_i$* cell edge), which is the solution to the following optimization problem:

$$\begin{aligned}\mathbf{r}_i^* = \arg\max_{\mathbf{r}_i} \mathbf{PL}(\mathbf{r}_i) \\ \text{subject to } \mathbf{PL}(\mathbf{r}_i) \leq \mathbf{PL}_{\text{threshold}}.\end{aligned} \tag{5.3}$$

After solving (5.3), *UA-gNB$_i$* will be placed at distance $\mathbf{r}_i^*$ from its *ref-gNB$_i$*, screening around the cell edge, as shown in Figure 5.5. This will help the uncovered *X-UEs* to associate with *UA-gNB$_i$* while screening its *SZs*. But this placement gives rise to an issue involving overlapping RFCs, which we highlight in Section 5.7.

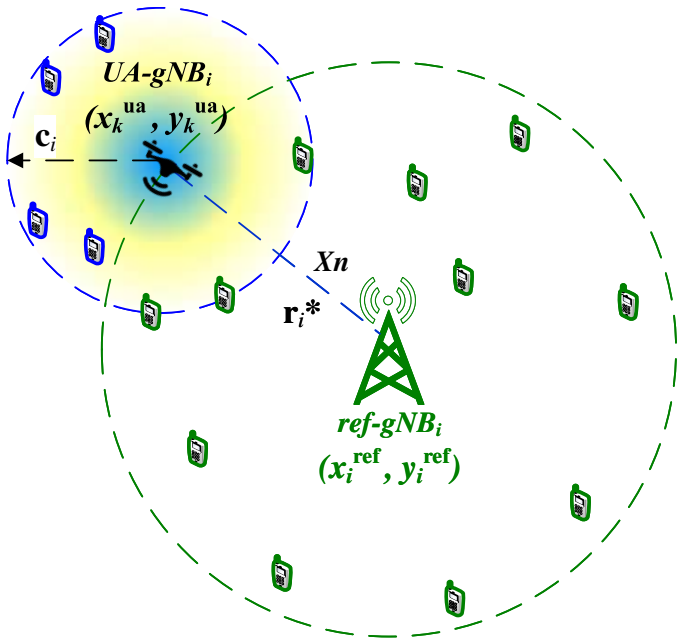



**Figure 5.5:** Maximum distance according to (5.3)

### 5.6.2 Procedure for Finding *X-UEs*

After finding the optimal placement $\mathbf{r}_i^*$, the associated *UA-gNB*$_i$ is ready to surveil the corresponding *ref-gNB*$_i$ cell edge, as described in Section 5.5.1. Specifically, *UA-gNB*$_i$ follows a path around the cell border at a distance of $\mathbf{r}_i^*$ from *ref-gNB*$_i$. Here and for ease of presentation, we assume that the cell edge is a circular boundary; the searching model is detailed in Figure 5.2. As defined in Section 5.5.2, item 5, there are *SZs* around *ref-gNB*$_i$, called $z_i^j$ (introduced in Section 5.9), in which *UA-gNB*$_i$ searches for *X-UEs*. Specifically, *ref-gNB*$_i$ assigns its *SZs*, $z_i^j$, to the associated *UA-gNB*$_i$. The following steps elaborate on the screening procedure:

1. In each $z_i^j$, when *X-UEs* are exposed to the RFC of *UA-gNB*$_i$ (at location $L_{i,j}$), they normally initiate the CSP. As a result, *UA-gNB*$_i$ becomes the serving cell for all the detected UEs in the underlaying zone $j$.

2. All *X-UEs* in $z_i^j$ that are found (by *UA-gNB*$_i$) will be registered as *UA-UEs* (to differentiate them from those that are already associated with the corresponding *ref-gNB*$_i$). As stated in

Section 5.5.2, item 4, the number of these *UA-UEs* is defined by $\mathcal{S}_i^j$—that is, for each $z_i^j$, there is a corresponding $\mathcal{S}_i^j$.

3. The context information[12] of the *UA-UEs* in $\mathcal{S}_i^j$ will be stored in *UA-gNB*$_i$. To detect as many *X-UEs* as possible, *UA-gNB*$_i$ may circulate around the cell edge multiple times. In this case, every time *UA-gNB*$_i$ makes a new round, it does not need to re-register the already detected *UA-UEs*. Moreover, after association, *UA-gNB*$_i$ broadcasts control messages to the corresponding *UA-UEs* to change their RRM status to the INACTIVE state (as described in [27], Section 6.1). This will save more power in the battery-limited UEs and provide very fast network access with lightweight signaling overhead when *UA-UEs* (in the INACTIVE state) are exposed multiple times to the RFC of *UA-gNB*$_i$.

In this context, we will discuss a mobility management issue in our *UE-based SAROs* to meet its critical requirements (signaling overhead and battery power consumption) in Section 5.8.

## 5.7 *UA-gNB* Overlapping RFC Issue

As discussed in Section 5.6.2, the associated *UA-gNB*$_i$ searches for *X-UEs*, mostly located close to the cell edge of the corresponding *ref-gNB*$_i$. This *UA-gNB*$_i$ does so at the optimal distance $\mathbf{r}_i^*$ (calculated in (5.3)) from its *ref-gNB*$_i$. But this placement gives rise to the following issue.

As we see in Figure 5.5, the RFCs of *UA-gNB*$_i$ and *ref-gNB*$_i$ are overlapping. In this case, some of the UEs in $\mathcal{G}_i$ (more precisely, those in the intersected area, $\mathcal{G}_i \cap \mathcal{S}_i^j$, of the current *SZ*) would initiate cell (re)selection[13] procedures each time they are exposed to the RFC of *UA-gNB*$_i$. This might occur when these UEs receive higher levels of RSRP from *UA-gNB*$_i$ than from *ref-gNB*$_i$. More specifically, these UEs (those in $\mathcal{G}_i \cap \mathcal{S}_i^j$) may switch back and forth between these two cells (*UA-gNB*$_i$ and *ref-gNB*$_i$), resulting in what is often called the toggling effect. This increases the signaling load on the both serving cells and their associated UEs. Moreover, initiating multi-cell

[12] Includes UE-specific configuration parameters [2].

[13] Assuming these UEs are in idle mode; otherwise, they would undergo handover.

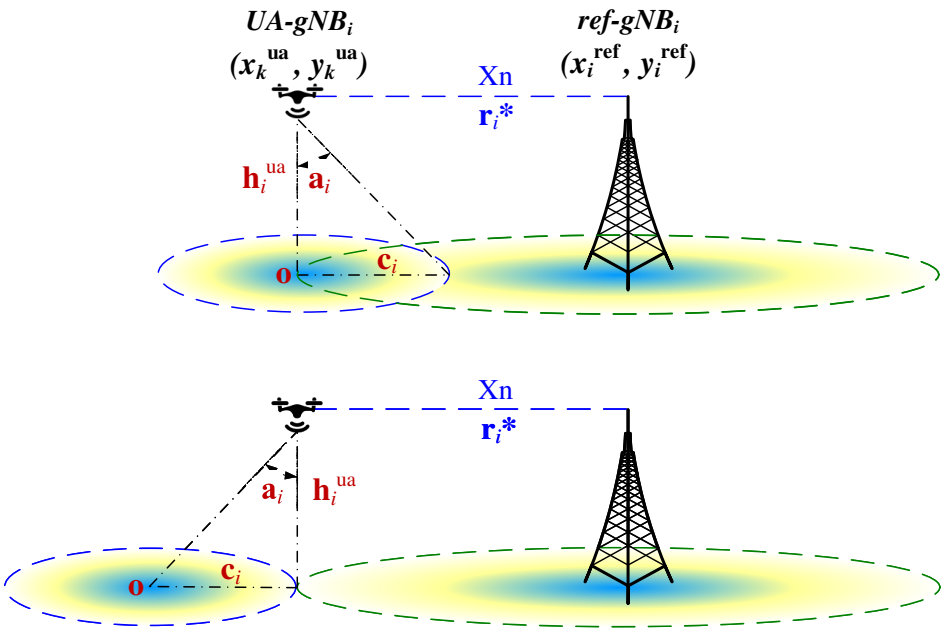



**Figure 5.6:** Avoiding RFC overlap using beamsteering

(re)selection drains battery power in these UEs. Furthermore, as we will see later, this will impact the accuracy of the generated *UEBCM*. Specifically, each $z_i^j$ has its own associated UEs and this is required because UE locations are defined by their serving cell (whether *UA-gNB*$_i$ or *ref-gNB*$_i$). In addition, all UEs within the overlapped RFC will receive relatively high intercell interference.

It would appear that this problem can be solved easily by increasing $\mathbf{r}_i^*$ such that the RFC of *UA-gNB*$_i$ lies outside the RFC of *ref-gNB*$_i$ (to achieve no overlap, $\mathbf{r}_i^*$ must be increased by $\mathbf{c}_i$). But by doing so, the necessary connection (Xn) will be lost, and hence this is not a feasible solution. To deal with this issue, we introduce two different techniques (which can be used separately or together), involving no additional computation cost. In both techniques, we need to keep the distance $\mathbf{r}_i^*$ unchanged, but ensure that the RFCs of *UA-gNB*$_i$ and *ref-gNB*$_i$ are nonoverlapping.

### 5.7.1 Beamsteering Antenna

Beamsteering antennas are widely used in LTE and are expected to be used in upcoming 5G networks [135, 136]. In this technique, the antenna radiation pattern can be electrically steered to a desire direction without physically moving the antenna [137]. In our case, the antenna radiation

pattern of the *UA-gNB*$_i$ should be steered in such a way that its RFC lies outside the RFC of *ref-gNB*$_i$, as we illustrate in Figure 5.6. Specifically, the center of the coverage area (labeled **o** in Figure 5.6) is shifted to the left by distance $\mathbf{c}_i$. Accordingly, the main beam (as shown in the top illustration in Figure 5.6) is shifted by angle $\mathbf{a}_i$, as shown in the bottom illustration in Figure 5.6. For that purpose, the angle $\mathbf{a}_i$ is calculated from the following formula: $\mathbf{a}_i = \tan^{-1} \mathbf{c}_i/\mathbf{h}_i^{\text{ua}}$, where $\mathbf{c}_i$ is the radius of the cell coverage and $\mathbf{h}_i^{\text{ua}}$ is the height for the corresponding *UA-gNB*$_i$.

By using this technique, the overlapping issue is addressed (to avoid multi-cell (re)selections). Moreover, because the center of the *UA-gNB* coverage area is shifted away from the *ref-gNB* cell edge, it will cover more *X-UEs* beyond the cell edge.

### 5.7.2 Access Control

In this technique, unlike the above one, the overlapping is allowed (as in the top illustration in Figure 5.6). But the multi-cell (re)selection (i.e., toggling effect) in the overlapping area is avoided using what is called the "barred cell" access control [3]. Under extreme circumstances (e.g., emergency situations), there is likely to be a huge number of UE access attempts triggered simultaneously—this leads to service degradation and lack of radio resources. To deal with this issue, when it is appropriate, network operators apply the "barred cell" mechanism to prevent many UEs from initiating simultaneous access attempts toward a certain set of gNBs, preventing the network from being overloaded. In this case, the gNBs broadcast cell access restrictions via system information messages to their associated UEs. In doing so, the corresponding UEs are prohibited from triggering multi-access attempts.

Taking advantage of this mechanism, the UEs in the overlapping area ($\mathcal{G}_i \cap \mathcal{S}_i^j$) would no longer switch back and forth between the two RFCs. In other words, the UEs in $\mathcal{G}_i$ are not allowed to make access attempts toward *UA-gNB*$_i$. Likewise, the UEs in $\mathcal{S}_i^j$ are prohibited from accessing *ref-gNB*$_i$.

Note that the two techniques above prevent the toggling effect, and hence will save more power in the battery-limited UEs and provide lightweight signaling overhead. But the beamsteering

method provides a larger RFC area than the "barred cell" access control method, as shown in Figure 5.6—this is because the former shifts the center of its RFC away from the cell edge while the latter still results in overlapping RFC.

## 5.8 Mobility Management for *UE-based SAROs*

In 5G networks, each gNB manages and controls its own associated UEs, including providing Mobility Management (MM) (see [27], Section 5.5.1). This MM imposes signaling overhead that is unsuitable for *UA-gNBs* because the latter are limited in battery power and processing capabilities. Instead, MM tasks, including RRM, will be handled by the corresponding *ref-gNB*, which is much more powerful than the *UA-gNB*.

Typically, in 5G, two essential MM procedures are involved to track and locate mobile UEs within the network, called *Tracking Area Update (TAU)* and *Paging* [27], which burden not only the serving network but also the associated battery-limited UEs. In this context, we propose in [29] efficient MM schemes that can be applied to mission-critical applications—that is, it can meet the critical requirements imposed by *UE-based SAROs*. We call this MM solution ***gNB-based UE Mobility Tracking (gNB-based UeMT)***, in which *TAU* is avoided and *Paging* delay is improved, enhancing the overall network performance, including power consumption in UEs and lightweight signaling overhead.

As we have seen in *UE-based SAROs*, two types of UEs are defined: *ref-UEs* and *UA-UEs* based on their associated gNBs (*ref-gNB* and *UA-gNB*, respectively). These two base stations interact together to handle UE mobility. As we have described in our *gNB-based UeMT* solution, a gNB takes over the responsibility of the MM—this gNB is called *anchor-gNB* (or *Home-gNB*), as defined in [29], Section IV-A. To apply *gNB-based UeMT* for *UE-based SAROs*, we now define the equivalent entities for both systems, as in Table 5.1.

It is worth mentioning here that our MM solution, *gNB-based UeMT*, by design has the potential to deal with mission-critical applications. This can be achieved by choosing the relevant control system parameters appropriately (refer to Section IV-B in [29]). For example, to guaran-

**Table 5.1:** Equivalence of *gNB-based UeMT* and *UE-based SAROs* entities

| *gNB-based UeMT* | *UE-based SAROs* |
|---|---|
| *anchor-gNB* | *ref-gNB* |
| *Home-UE* | *ref-UE* |
| *Visiting-UE* | *UA-UE* |
| *Visiting-gNB* | *UA-gNB* |

tee that the inter-*Paging* delay between the *ref-gNB* and its associated *UA-gNB* does not exceed a predefined value, the relevant system parameter, called *Calculated inter-Paging Delay (CiPD)* index should take a value index equal to *CiPD_0*—this is to maintain the *CiPD* at below 1 *msec*. Applying *gNB-based UeMT* and according to its features, the *ref-gNB* will take over the responsibility of MM tasks (to reduce the load on the UEs and *UA-gNBs*), offer lightweight signaling overhead (it achieves about 92% reduction in the relevant load [29]), and provide always-known UE locations—these are important for *UE-based SAROs*.

## 5.9 *UA-gNB* Location Setup for *SZs*

As we have stated earlier (see Figure 5.2), after calculating $\mathbf{r}_i^*$, *UA-gNB*$_i$ starts screening around the corresponding cell edge. But so far, it is not clear that how *UA-gNB*$_i$ moves around the cell edge within the radius $\mathbf{r}_i^*$. Recall the set $\mathcal{L}_i = \{L_{i,j} : j = 0, 1, \ldots, J_i - 1\}$. As mentioned before, $L_{i,j}$ represents the *UA-gNB*$_i$ location relevant to its *ref-gNB*$_i$ cell edge while in *SZ* $j$. In other words, each value of $j$ corresponds to a specific location of *UA-gNB*$_i$ on the cell edge path. So, we need to find these locations for each value of $j$, where *UA-gNB*$_i$ is located. We explain how to find these coordinates for each value of $j$. When screening *SZ* $j$, the *UA-gNB*$_i$ should be located at $L_{i,j}$, which is defined (in polar) as:

$$L_{i,j} = \mathbf{r}_i^* \angle \underline{j \cdot \mathbf{b}_i} \quad , \; j = 0, 1, \ldots, J_i - 1, \tag{5.4}$$

where $L_{i,0} = \mathbf{r}_i^* \angle \underline{\mathbf{0}}$ is the initial location and $J_i$ is the total number of the *SZs* around *ref-gNB*$_i$ (calculated below). Now, we calculate the angle $\mathbf{b}_i$ as follows. After completing the screening process, *UA-gNB*$_i$ moves to the next *SZ* at $L_{i,j+1}$ such that the distance between its previous and next positions, denoted by $d(L_{i,j}, L_{i,j+1})$, is approximately equal to its RFC radius, the quantity labeled

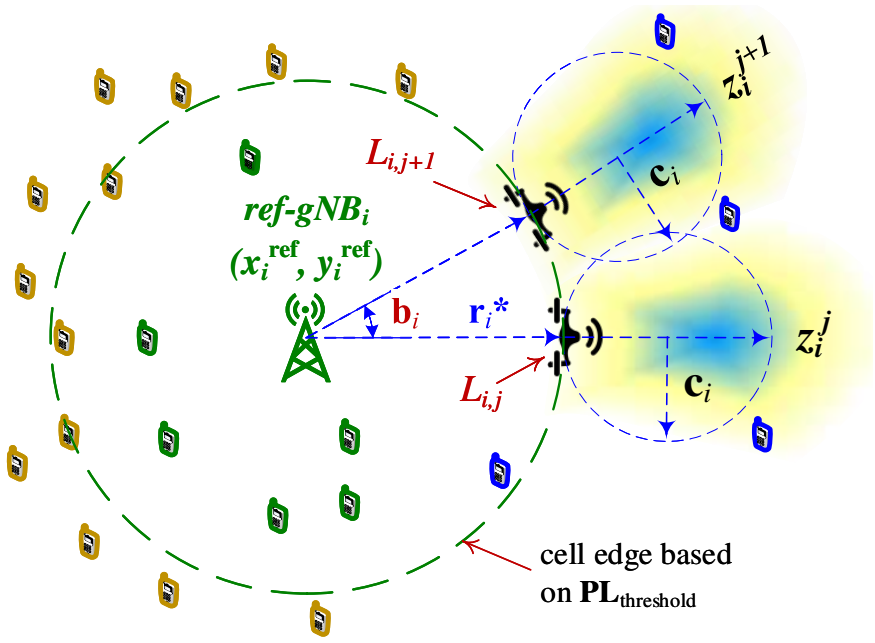



**Figure 5.7:** *UA-gNB$_i$* screening locations

$\mathbf{c}_i$ in Figure 5.7 (the illustration in Figure 5.7 assumes beamsteering, as detailed in Section 5.7.1). This is to keep the *SZs* separated and minimize the overlap between their corresponding RFC, as Figure 5.7 illustrates. Note that even when a swap occurs, there is no confusion among the detected UEs (to which *SZ* they belong); this is because *UA-gNB$_i$* registers the detected *UA-UEs* such that each *SZ* has its own UEs, which means that $\mathcal{S}_i^j \cap \mathcal{S}_i^{j+1} = \emptyset$.

From above, we have $d(L_{i,j}, L_{i,j+1}) \approx \mathbf{c}_i$ (provided $\mathbf{c}_i \not\ll \mathbf{r}_i^*$, as we can see in Figure 5.7). Hence,

$$\mathbf{b}_i \approx 360 \cdot \mathbf{c}_i/(2\pi \cdot \mathbf{r}_i^*) \quad \text{degrees.} \tag{5.5}$$

So, the resulting total number of *SZs* is $J_i = \lceil 2\pi \cdot \mathbf{r}_i^*/\mathbf{c}_i \rceil$, and hence $j$ should range between 0 and $J_i - 1$. It should now be clear how *UA-gNB$_i$* circulates around *ref-gNB$_i$*, searching for *X-UEs*.

To implement the scheme above in practice, *UA-gNB$_i$* can use the Discontinuous Transmission (DTX) technique [138] while moving along the path of the cell edge. In doing so, it does not transmit its RF signal while moving from one *SZ* to another. For example, while moving from position $L_{i,j}$ to $L_{i,j+1}$ (shown in Figure 5.7), *UA-gNB$_i$* refrains from broadcasting its RF signal. More specifically, RF transmission is necessary only when *UA-gNB$_i$* is positioned at location $L_{i,j}$,

**Table 5.2:** Information table for each *ref-gNB*$_i$–*UA-gNB*$_i$ pair

| $\mathcal{Z}_i$ | *UA-gNB locations (polar)* | *UE densities* | *Average RSRP (dBm)* |
|---|---|---|---|
| $z_i^0$ | $L_{i,0}$ | $\mathcal{S}_i^0$ | RxLev#0 |
| $z_i^1$ | $L_{i,1}$ | $\mathcal{S}_i^1$ | RxLev#1 |
| . | . | . | . |
| . | . | . | . |
| $z_i^{J_i-1}$ | $L_{i,J_i-1}$ | $\mathcal{S}_i^{J_i-1}$ | RxLev#$J_i - 1$ |

**Table 5.3:** Association table for each *ref-gNB*$_i$

| $\mathcal{R}$ | *ref-gNB*$_i$ | $\mathcal{G}_i$, *set of ref-UEs* | *Average RSRP (dBm)* |
|---|---|---|---|
| $(x_1^{\text{ref}}, y_1^{\text{ref}})$ | 1 | $\mathcal{G}_1$ | RxLev#1 |
| $(x_2^{\text{ref}}, y_2^{\text{ref}})$ | 2 | $\mathcal{G}_2$ | RxLev#2 |
| . | . | . | . |
| . | . | . | . |
| $(x_R^{\text{ref}}, y_R^{\text{ref}})$ | $R$ | $\mathcal{G}_R$ | RxLev#$R$ |

screening the corresponding *SZ*; this will save battery power in *UA-gNB*$_i$ and increase its functional lifetime.

Once the whole screening process is complete, each *UA-gNB*$_i$ generates an information table, as detailed in Table 5.2. This table will be shared with the corresponding *ref-gNB*; the latter will process the incoming information in conjunction with its own association table (see Table 5.3) to generate the necessary *UEBCM* and *PDRFC*—that is, each *ref-gNB* will have its own *UEBCM* and *PDRFC*.

## 5.10 Time Cost for Discovery and Relocation

Before starting the screening process (Section 5.6.2), each *UA-gNB* needs to discover a *ref-gNB* for acquiring system information. Once completing this process, the *UA-gNB* relocates to the *ref-gNB* cell edge (according to (5.3)). These two processes consume time, which we consider below.

### 5.10.1 Discovery Time Cost

We now calculate the time required for a *UA-gNB* (the *i*th, say) to discover a *ref-gNB*. As detailed in Section 5.6.1.1, *UA-gNB*$_i$ starts at its initial location, $(x_k^{\text{ua}}, y_k^{\text{ua}})$, according to the *CCBS*. At this point, *UA-gNB*$_i$ will need to find and synchronize with *ref-gNB*$_i$. To do so, it initiates the CSP, receiving and decoding what is called *Cell System Information (CSI)* [96]. Two necessary signals, called *Primary Synchronization Signal (PSS)* and *Secondary Synchronization Signal (SSS)*, must be obtained to get the PCI and frame timing of the detected *ref-gNB*$_i$.

The total discovery time, denoted by $T\text{dis}_i$, required for *UA-gNB*$_i$ to discover *ref-gNB*$_i$ consists of two parts. The first is the time required to obtain the *CSI*, which we call $T_{\text{csi}}$. The second is the time required to process the *CSI*. We introduce a symbol $\eta_i > 0$ such that the second part, the processing time, is $\eta_i T_{\text{csi}}$. So $\eta_i$ is a measure of how fast the *CSI* can be decoded, normalized by $T_{\text{csi}}$. The higher the processing capability of *UA-gNB*$_i$, the smaller the value of $\eta_i$. We can now write the following expression:

$$T\text{dis}_i = T_{\text{csi}} \cdot (1 + \eta_i). \tag{5.6}$$

The value of $T_{\text{csi}}$ is in turn given by

$$T_{\text{csi}} = 10 \cdot TTI \cdot SFN, \tag{5.7}$$

where *TTI* is the *Transmission Time Interval (TTI)* relevant to one subframe, and *SFN* is the *System Frame Number (SFN)*, used to define different system frame cycles [96]. In LTE, a single radio frame comprises 10 subframes; hence the factor of 10 in (5.7).

### 5.10.2 Relocation Time Cost

As detailed in Section 5.6.1.2, the location of *UA-gNB*$_i$ should be at the cell edge of *ref-gNB*$_i$. Specifically, the *UA-gNB*$_i$ should move from its initial location, $(x_k^{\text{ua}}, y_k^{\text{ua}})$, toward the cell edge of *ref-gNB*$_i$ by a distance equal to $(\mathbf{r}_i^* - \mathbf{r}_i)$, where $\mathbf{r}_i^*$ and $\mathbf{r}_i$ are defined in (5.1) and (5.3), respectively $(\mathbf{r}_i \leq \mathbf{r}_i^*)$. From here on, we assume that all the deployed *UA-gNBs* move in their horizontal plane

with equal radial velocity, denoted by $v_r$. To calculate the required time, denoted by $T\text{rel}_i$, for the *UA-gNB*$_i$ to relocate its initial location (to reach the cell edge), we have the following formula:

$$T\text{rel}_i = \frac{\mathbf{r}_i^* - \mathbf{r}_i}{v_r}. \tag{5.8}$$

By (5.6)–(5.8), the total time (for discovery and relocating) is

$$T\text{tot}_i = 10 \cdot TTI \cdot SFN \cdot (1 + \eta_i) + \frac{\mathbf{r}_i^* - \mathbf{r}_i}{v_r}. \tag{5.9}$$

It should be clear now that the average time, denoted by $T_{\text{ave}}$, required for the all detected *ref-gNB*$_i$ (i.e., $R$) can be written in the following form using (5.6)–(5.9):

$$T_{\text{ave}} = T_{\text{csi}} + \frac{T_{\text{csi}}}{R} \sum_{i=1}^{R} \left[ \eta_i + \frac{\mathbf{r}_i^* - \mathbf{r}_i}{T_{\text{csi}} \cdot v_r} \right]. \tag{5.10}$$

As we see from the last formula, many factors can impact the total average time. In the next section, for the purpose of simulation, we assume that $R$, $v_r$, and $\mathbf{r}_i^*$ are predefined.

### 5.10.3 Simulation Setup

To empirically evaluate the required time to discovery and relocating ($T\text{dis}_i$ and $T\text{rel}_i$), we use (5.9) and (5.10) for the simulation and evaluation. Based on the above assumptions, we set the following parameters: $K = 20$ (i.e., 20 *ref-gNBs* in a RoI—$K$ is also the number of the formed clusters, as detailed in Section 5.6.1.1), $v_r = 1\ m/s$, and $\mathbf{r}_i^* = 70\ m$. The value of $\mathbf{r}_i^* = 70\ m$ is typical of small picocells or large femtocells. The speed of 1 $m/s$ for the UAV we set above represents a rather slow-flying vehicle; we use this value to make our experimental scenario very conservative, generating experimental results conservatively (worse than can be expected in practice). More practically realistic values for the speed would result in even better results than we show here.

According to the LTE system specification, we set $TTI = 1\ msec.$ and $SFN = 1024$—this is the maximum value that should be assigned to the *SFN* and refers to the total number of the system

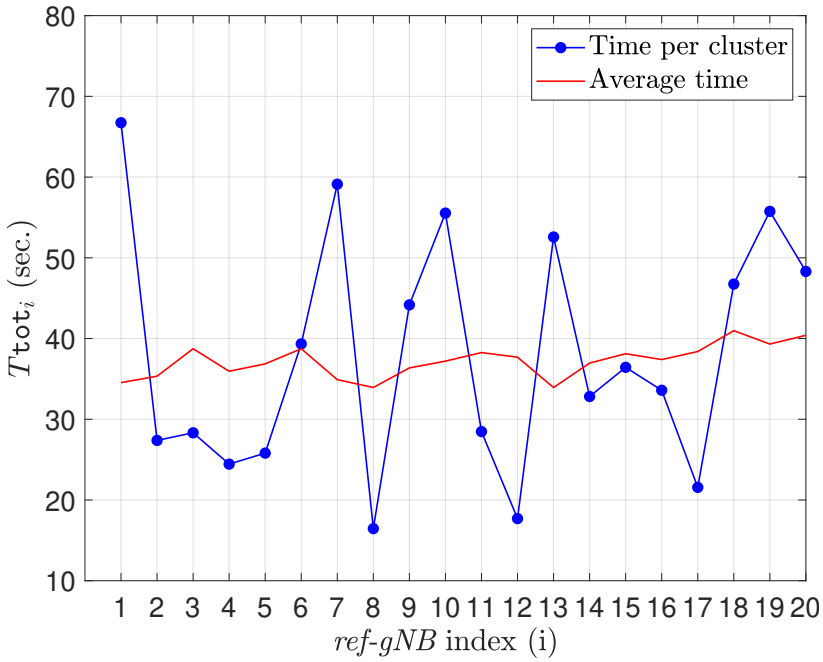



**Figure 5.8:** Discovery and relocation time

frames, which are necessary to acquire all the system information, resulting in $T_{\text{csi}} = 10.24$ *sec*. For $\eta_i$ and $\mathbf{r}_i$, we set their values according to uniform distributions in the range of 0.1–0.9 and 10–70 *m*, respectively.

Figure 5.8 illustrates the variation of $T\text{tot}_i$. We can see that the values of $T\text{tot}_i$ lie in the range of 16.45–66.73 *sec*. This is a reasonable time to act in emergency situations. The average time to complete the two processes is $T_{\text{ave}} = 38.06$ *sec*. Of course, if the deployed *UA-gNBs* were to have higher speed ($v_r$) or higher processing capability (lower $\eta_i$), the time cost for discovery and relocation would be lower (as quantified in (5.9)).

## 5.11 Generating Crisis Maps, *UEBCMs*

After aggregating all the necessary data (as in Table 5.2) from the associated *UA-gNB*$_i$, each *ref-gNB*$_i$ will generate its own *UEBCMs* for the area around its cell edge and beyond (including the in-cell area), having information about the surviving UE distributions. That is, the generated *UEBCMs* should give sufficient awareness to the PSAs such that they have enough knowledge about where the survivors are collected, prioritizing the SAROs in a quick way. For that purpose,

we generate two types of *UEBCMs*, which we describe in Sections 5.11.2 and 5.11.3, after we introduce an illustrative scenario in the next subsection.

### 5.11.1 Illustrative Scenario Setup

To illustrate how to generate the *UEBCMs*, we consider a simple scenario with one *ref-gNB*$_i$ and its associated *UA-gNB*$_i$ (recall the searching procedure of Section 5.6.1). The same process applies to each *ref-gNB* in the RoI. To use the searching procedure of Section 5.6.2, we set the necessary parameters as follows. We set $\mathbf{c}_i = 40\ m$ and $\mathbf{r}_i^* = 70\ m$ (cell radius of *UA-gNB*$_i$ and *ref-gNB*$_i$, respectively), resulting in $\mathbf{b}_i = 32.74°$ and $J_i = 11$ (i.e., $\mathcal{L}_i = \{L_{i,j} : j = 0, 1, \ldots, 10\}$). The values we set for $\mathbf{c}_i$ and $\mathbf{r}_i^*$ are typical of small picocells or large femtocell [132].

After completing the searching process, we have all the information needed to produce the picture illustrated in Figure 5.9, which shows the attached UEs corresponding to each *SZ*, including the in-cell UEs. In this example, the solid blue circles refer to centers of the *SZs*, denoted by $R_{i,j}$—that is, for each $L_{i,j}$, there is a corresponding $R_{i,j} = (\mathbf{r}_i^* + \mathbf{c}_i)\underline{/\, j \cdot \mathbf{b}_i}$ (similar to the formula in (5.4)). The UEs around each $R_{i,j}$ are associated with the corresponding *SZ* $j$, $z_i^j$ (i.e., where they are screened and discovered). Each $z_i^j$ has its own UEs as shown in the multi-colored circles in Figure 5.9.

To further clarify, we now characterize this illustrative scenario using the detailed notation from Section 5.5.2. The set of *ref-UEs* is equal to $\mathcal{G}_i = \{ue_{i,1}^{\text{ref}}, ue_{i,2}^{\text{ref}}, \ldots, ue_{i,63}^{\text{ref}}\}$. The set of *UA-UEs* within the *SZ* $j = 0$ is equal to $\mathcal{S}_i^0 = \{ue_{i,1}^0, ue_{i,2}^0, \ldots, ue_{i,80}^0\}$. Similarly, the set of *UA-UEs* within zone $j = 1$ is equal to $\mathcal{S}_i^1 = \{ue_{i,1}^1, ue_{i,2}^1, \ldots, ue_{i,12}^1\}$, and so on for the others $z_i^j$. This represents the distribution of UEs in the *SZs*, which is defined by $\mathcal{Z}_i = \{z_i^0, z_i^1, \ldots, z_i^{10}\}$. With the information now gathered, we are ready to produce two types of *UEBCMs*, as described in the next two subsections: one to show UE densities, and a second one to show UE RSRP levels.

### 5.11.2 *UEBCM* for UE Densities

Based on the preceding information, the corresponding *ref-gNB* generates a *UEBCM* for the impacted area, giving visual information about the potential survivor distribution. Specifically, the

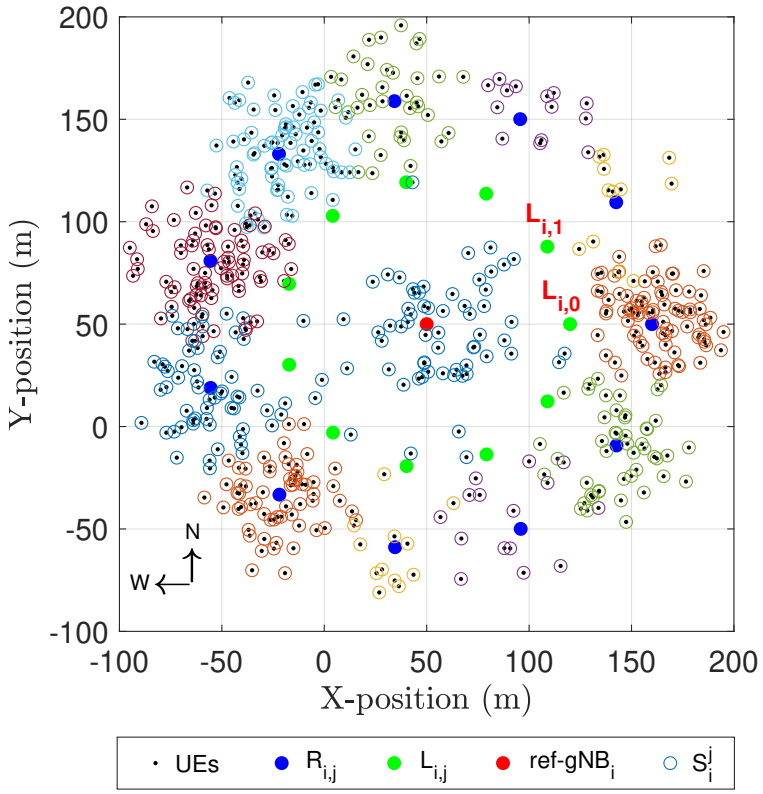



**Figure 5.9:** Attached UEs corresponding to each *SZ*

*ref-UEs* and *UA-UEs* in *UE-based SAROs* have become human sensors for the survivors in the RoI. Accordingly, the *ref-gNB* will generate a map for the survivor density distribution (shown in Figure 5.10). As we can see from this example map, based on the previous illustrative scenario, the majority of individuals are clustered toward the north-west, far from the center of the *ref-gNB*, which is located at $(50, 50)$. Furthermore. another area with a significant number of individuals is located to the east of the *ref-gNB*. These two areas contain about 75% of the individuals; these should have higher priority than other regions to be considered for SAROs. The areas with darker colors should be given lower priority, which are located to the far south and north-east (illustrated in dark blue in Figure 5.10). This will help the PSAs to act quickly, prioritizing their SAROs to find the majority of the disaster victims.

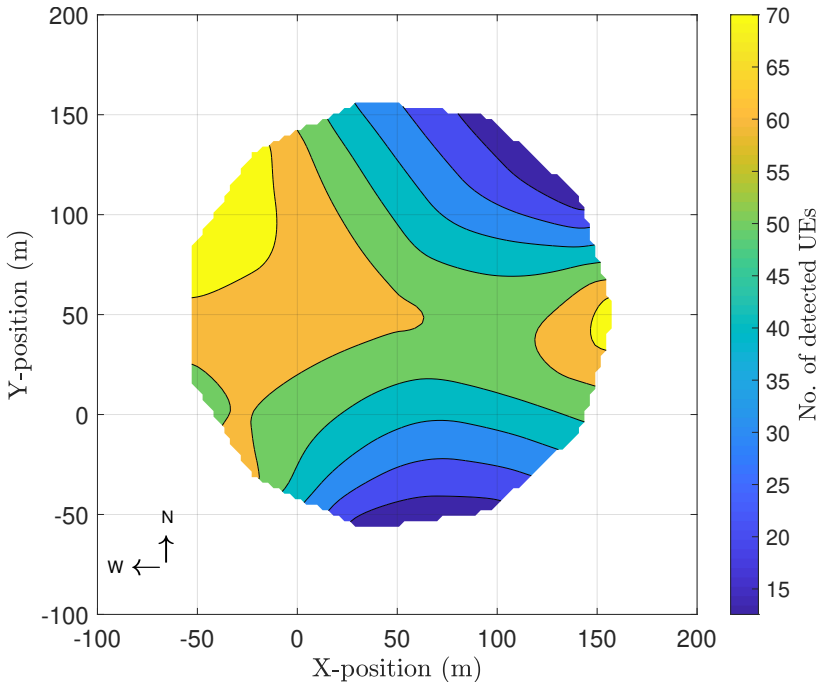



**Figure 5.10:** *UEBCM*: Density of surviving UEs

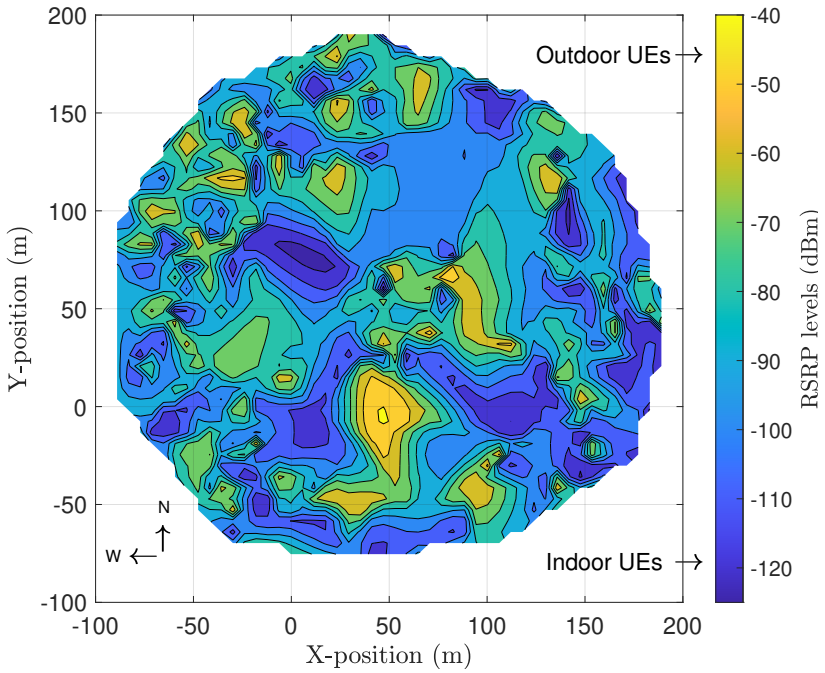



**Figure 5.11:** *UEBCM*: RSRP levels of the attached UEs

### 5.11.3 *UEBCM* for UE RSRP Levels

Based on the received RSRP levels during the screening procedure, the *ref-gNB* will generate another useful map, as Figure 5.11 illustrates. We can observe from this map that some UEs receive

**Table 5.4:** Information table for the *PDRFC*

| *SZ index* | *Priority index* | *UA-gNB locations (polar)* | *UE densities* |
|---|---|---|---|
| $z_i^0$ | 0 | $L_{i,0}$ | 80 |
| $z_i^5$ | 1 | $L_{i,5}$ | 77 |
| $z_i^4$ | 2 | $L_{i,4}$ | 75 |
| $z_i^7$ | 3 | $L_{i,7}$ | 62 |
| . | . | . | . |
| $z_i^1$ | 10 | $L_{i,1}$ | 12 |

high RSRP levels (approximately $-40$ to $-70$ *dBm*), especially in the yellowish (and yellowish-green) areas. These levels are (likely) associated with outdoor UEs (located in some particular areas, shown in Figure 5.11). Likewise, we can see areas in dark blue in Figure 5.11, representing low levels of RSRP. These levels are (likely) received from indoor UEs (experiencing high **PL**); they might be stranded inside buildings and need immediate help. This gives rise to a significant observation, as we illustrate in the following. As detailed in Figure 5.10, one area that has the majority of individuals is located toward the north-west. When comparing this specific area with the corresponding RSRP levels map (Figure 5.11), we notice that these UEs are likely to be indoors. Specifically, we can conclude that this particular area is high-density cluster of UEs that are experiencing high **PL**. This area would thus be given high priority for SAROs relative to other areas. Based on this, we can generate a *PDRFC*, as described next.

### 5.11.4 Building *Priority-Driven RFC (PDRFC)*

As defined in Section 5.5.1, the *PDRFC* is used to identify areas that need immediate RFC. Based on the extracted information from the preceding maps (Sections 5.11.2 and 5.11.3), Table 5.4 is built, providing priorities on where immediate RFC is needed (based on UE clustering). Clearly, $z_i^0$, $z_i^5$, $z_i^4$, and $z_i^7$ need immediate RFCs; the majority of survivors can be found there. This can be achieved by deploying dedicated *UA-gNBs* hovering over these zones, providing semi-permanent RFC.

In sum, the vital information provided by *UEBCMs* and *PDRFC* enables the PSAs to provide SAROs for the largest number of survivors in a prioritized way.

## 5.12 Summary

In this chapter, we have described a new framework for SAROs (called *UE-based SAROs*) to find and locate post-disaster survivors based on the idea that most individuals have their own UEs—potentially, they are still alive and need to be rescued. Our framework, *UE-based SAROs*, addresses the following concerns. 1) With the lack of RFC, how do we give the corresponding PSAs awareness of the individual locations in the RoI without their assistance (for life-saving purposes)? 2) What is the quickest way to recover the RFC right after disasters, exploiting the surviving gNBs (i.e., *ref-gNBs*), before it becomes too late, especially considering that most UEs are battery limited? 3) Based on the fact that finding and locating survivors is more important than providing RFC elsewhere, where are the majority of survivors located and how (indoors or outdoors)?

*UE-based SAROs* provide vital information to the PSAs to prioritize their operations and manage the available resources. By considering the surviving UEs as human-based sensors distributed in the RoI and are able to exchange signaling messages without active user participation, the *UE-based SARO* provides the following benefits: 1) right after disasters, it generates immediate visual crisis maps, *UECBMs*, showing the potential survivor distribution, 2) it provides quick vital information about which regions contain the majority of survivors, and 3) based on the preceding information, the PSAs can prioritize and manage their SAROs effectively, providing the necessary RFC accordingly.

Finally, *UE-based SAROs* provide PSAs with situational awareness about the disaster-impacted area quickly and even before they arrive at the scene, keeping the PSAs better informed about locations of the disaster victims. This enables the PSAs to serve the largest number of survivors in a timely manner even when the cellular communication infrastructure is partially dysfunctional.

# Chapter 6

# Conclusion

As we have seen throughout this dissertation, two essential MM procedures, *TAU* and *Paging*, are required to track and locate all UEs while moving within the network coverage area. These procedures are still used in LTE and 5G, although they are prone to multiple failures. This motivates considering how to provide a very fast way (close-to-zero latency) to track and locate all mobile UEs and minimize the power consumption in these devices (especially because most UEs are battery-limited). Furthermore, because of the tremendous increase in high-mobility UEs, the consequent swap can impact not only the network performance (cost more network resources and even lead to a congested network) but also the UE experience (increase power consumption in UEs). In Chapter 2, we have extensively examined and discussed a variety of solution schemes that have been proposed to mitigate the LTE MM overhead in terms of *TAU* and *Paging*.

In Chapter 3, we have specifically discussed MM solutions to achieve the critical requirements of 5G. In addition, we have investigated applying current LTE MM solution to 5G use cases. Based on our evaluation, the LTE MM solution schemes will not satisfy the 5G use cases because of their limitations owing to high implementation complexity, high latency, and high computation cost (e.g., these schemes do not maintain close-to-zero latency). Furthermore, we have highlighted the new 5G system architecture, which is designed based on legacy LTE systems. This new design is intended to reduce not only the *TAU* but also *Paging* signaling overhead and maintain the *Paging* latency to be extremely low (e.g., < 1 ms) relative to current LTE systems. Moreover, in Chapter 3, many new aspects in terms of 5G MM have been discussed, which include the NG-RAN, NG-RRC, RNA, RNAU, and the *Paging* DRX cycle configurations (5GC/NG-RAN-based *Paging*). According to these aspects, the envisioned 5G requirements would be achieved for not only the network performance but also UE experience. At the time of writing this dissertation, however, network operators and many research groups were still developing more MM solutions to satisfy 5G goals.

In Chapter 4, we have proposed a novel solution scheme to solve the MM problems. Our solution, called *gNB-based UeMT*, aims to support life-critical systems and real-time applications, which are crucial requirements for 5G. The *gNB-based UeMT* solution achieves the following essential features. 1) The mobile IoT/UEs will no longer trigger the *TAU/RNAU* to report their location changes, giving much higher power savings with no signaling overhead. 2) Instead, the network elements, gNBs, take over the responsibility of *Tracking* and *Locating* these IoT/UEs, giving always-known IoT/UE locations. 3) Our *Paging* procedure is markedly improved over the conventional one, providing very fast IoT/UE reachability with no *Paging* messages being sent simultaneously. 4) This solution guarantees lightweight signaling overhead with very low *Paging* delay; it achieves about 92% reduction in the corresponding signaling overhead. A potential future extension of our solution is to apply *gNB-based UeMT* to support what is called 5G vehicular (e.g., V2X) and D2D communications, exploiting the always-known IoT/UE locations feature.

Finally, in Chapter 5, we have proposed a new framework for SAROs, named *UE-based SAROs*, to find and locate post-hazard survivors based on the idea that most individuals have their own UEs. In our discussion of *UE-based SAROs*, we have highlighted and addressed the most critical situations. First, with the lack of RFC, how do we give the corresponding PSAs awareness of the *sur-UE* locations without explicit survivor involvement? Second, by taking advantage of the surviving gNBs, how do we provide a quick way to recover the RFC right after disasters before it becomes too late, especially considering that most UEs are battery limited? Third, how do we provide information on disaster victim locations and immediate environment (indoors or outdoors) without the ability to communicate? Our solution provides PSAs with situational awareness about the disaster-impacted area quickly and even before they arrive at the scene, keeping the PSAs better informed about locations of the disaster victims. A direction for future extension of our framework, *UE-based SAROs*, is to deal with the situation where the entire cellular-network infrastructure is completely dysfunctional, with no surviving gNBs. Moreover, our solution can be leveraged to track and monitor the mobility of first responders so that they can be relocated to where they are most needed.

# Bibliography


[1] E. Grigoreva, Jianghua Xu, and W. Kellerer, "Reducing mobility management signaling for automotive users in LTE advanced," in *2017 IEEE Int. Symp. Local Metrop. Area Networks*, Jun. 2017, pp. 1–6.

[2] 3GPP TS 23.401 v14.4.0, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access (release 14)," 2017.

[3] 3GPP TS 36.304 v14.3.0, "User Equipment (UE) procedures in idle mode (release 8)," 2017. http://www.3gpp.org

[4] Alcatel-Lucent, "The impact of small cells on MME signaling: Methods to reduce and optimize MME core signalling caused by small cells," Tech. Rep., 2013. http://www.tmcnet.com/tmc/whitepapers

[5] S. Parija, N. Nath, P. Sahu, et al, "Dynamic profile based paging in mobile communication," in *2015 IEEE Int. Conf. Microwave, Opt. Commun. Eng.*, Dec. 2015, pp. 342–345.

[6] Y. Xiao, H. Chen, and M. Guizani, "Performance evaluation of pipeline paging under paging delay constraint for wireless systems," *IEEE Trans. Mob. Comput.*, vol. 5, no. 1, pp. 64–76, Jan. 2006.

[7] M. Rumney, "Taking 5G from vision to reality," 2014. https://blog.3g4g.co.uk/2014/07/taking-5g-from-vision-to-reality.html

[8] 3GPP TS 23.501 v15.0.0, "System architecture for the 5G system; Stage 2 (release 15)," 2017. http://www.3gpp.org

[9] 3GPP TS 38.300 v15.0.0, "NR; NR and NG-RAN overall description; Stage 2 (release 15)," 2017. http://www.3gpp.org

[10] M. Säily, "Deliverable D6.1 draft asynchronous control functions and overall control plane design," *METIS-II/D6.1*, 2016. http://www.5g-ppp.eu

[11] I. Da Silva, G. Mildh, M. Säily, et al, "A novel state model for 5G radio access networks," in *2016 IEEE Int. Conf. Commun. Work.*, May 2016, pp. 632–637.

[12] Ericsson, "Ericsson mobility report," E-164 80 Stockholm, Sweden, Tech. Rep., 2017. https://www.ericsson.com/en/mobility-report

[13] Y. Heisler, "A huge 4G milestone: LTE is now available for 98% of Americans," Tech. Rep., 2015. http://bgr.com/2015/03/23/lte-coverage-map-united-states/

[14] 3GPP TR 21.905 v16.0.0, "Vocabulary for 3GPP specifications (release 16)," 2019. https://portal.3gpp.org

[15] O. B. Karimi, J. Liu, and C. Wang, "Seamless wireless connectivity for multimedia services in high speed trains," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 4, pp. 729–739, May 2012.

[16] A. A. R. Alsaeedy and E. K. P. Chong, "A review of mobility management entity in LTE networks: Power consumption and signaling overhead," *Int. J. Netw. Manag.*, vol. 30, no. 1, pp. 1–27, Jan./Feb. 2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/nem.2088

[17] R. Liou, Y. Lin, and S. Tsai, "An investigation on LTE mobility management," *IEEE Trans. Mob. Comput.*, vol. 12, no. 1, pp. 166–176, Jan. 2013.

[18] T. Deng, X. Wang, P. Fan, et al, "Modeling and performance analysis of a tracking-area-list-based location management scheme in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6417–6431, Aug. 2016.

[19] M. Olsson, S. Sultana, S. Rommer, et al, *SAE and the evolved packet core: Driving the mobile broadband revolution*, 1st ed. Elsevier/Academic Press, 2009.

[20] Nokia Siemens Networks, "Signaling is growing 50% faster than data traffic," Karaportti 3, ESPOO, Finland, Tech. Rep., 2012. http://docplayer.net/6278117-Signaling-is-growing-50-faster-than-data-traffic.html

[21] D. Nowoswiat, "Managing LTE core network signaling traffic," Nokia, Tech. Rep., 2013. https://insight.nokia.com/managing-lte-core-network-signaling-traffic

[22] 3GPP TS 32.410 v14.0.0, "Telecommunication management; Key Performance Indicators (KPI) for UMTS and GSM (release 14)," 2017. https://portal.3gpp.org

[23] 3GPP TS 32.426 v14.0.0, "Telecommunication management; Performance Management (PM); Performance measurements Evolved Packet Core (EPC) network (release 14)," 2017. https://portal.3gpp.org

[24] 3GPP TA 32.455 v14.0.0, "Key Performance Indicators (KPI) for the Evolved Packet Core (EPC); Definitions (release 14)," 2017. https://portal.3gpp.org

[25] C. Rose and R. Yates, "Minimizing the average cost of paging under delay constraints," *Wirel. Networks*, vol. 1, no. 2, pp. 211–219, Jun. 1995. http://link.springer.com/10.1007/BF01202543

[26] R. Kreher and K. Gaenger, *LTE signaling: Troubleshooting and performance measurement*, 2nd ed., John Wiley & Sons, 2016.

[27] A. A. R. Alsaeedy and E. K. P. Chong, "Tracking area update and paging in 5G networks: A survey of problems and solutions," *Mob. Networks Appl.*, vol. 24, no. 2, pp. 578–595, Apr. 2019. https://doi.org/10.1007/s11036-018-1160-6

[28] ——, "Tracking area update procedure unnecessary in 5G: Improving user experience and offloading signaling overhead," in *2018 9th IEEE Annu. Ubiquitous Comput. Electron. Mob. Commun. Conf. UEMCON 2018*. New York, NY, Nov. 8–10, 2018, pp. 854–860. https://doi.org/10.1109/UEMCON.2018.8796728

[29] ——, "Mobility management for 5G IoT devices: Improving power consumption with lightweight signaling overhead," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8237–8247, Oct. 2019. https://doi.org/10.1109/JIOT.2019.2920628

[30] ——, "Survivor-centric network recovery for search-and-rescue operations," in *2019 IEEE Resil. Week*, San Antonio, TX, Nov. 4–7, 2019, pp. 71–77. https://doi.org/10.1109/RWS47064.2019.8971818

[31] ——, "Post-hazard cellular network recovery by unmanned aerial vehicles and user equipment cooperation," *IEEE IT Prof.*, 2020. https://doi.org/10.1109/MITP.2020.2982648

[32] ——, "5G and UAVs for mission-critical communications: Swift network recovery for search-and-rescue operations," *Mob. Networks Appl.*, 2020. https://doi.org/10.1007/s11036-020-01542-2

[33] V. Nguyen, T. Do, and Y. Kim, "SDN and virtualization-based LTE mobile network architectures: A comprehensive survey," *Wirel. Pers. Commun.*, vol. 86, no. 3, pp. 1401–1438, Feb. 2016.

[34] D. Wang, L. Zhang, Y. Qi, et al, "Localized mobility management for SDN-integrated LTE backhaul networks," in *2015 IEEE 81st Veh. Technol. Conf.*, May 2015, pp. 1–6.

[35] L. Valtulina, M. Karimzadeh, G. Karagiannis, and et al, "Performance evaluation of a SDN/OpenFlow-based Distributed Mobility Management (DMM) approach in virtualized LTE systems," in *2014 IEEE Globecom Work. (GC Wkshps)*, Dec. 2014, pp. 18–23.

[36] E. Aqeeli, A. Moubayed, and A. Shami, "Towards intelligent LTE mobility management through MME pooling," in *2015 IEEE Glob. Commun. Conf.*, Dec. 2015, pp. 1–6.

[37] E. Dahlman, S. Parkvall, and J. Sköld, *4G: LTE/LTE-advanced for mobile broadband*. Elsevier/Academic Press, 2011.

[38] 3GPP TS 28.533 v15.1.0, “Management and orchestration; Architecture framework (release 15),” 2018.

[39] 3GPP TS 24.301 v14.4.0, “Non-Access-Stratum (NAS) protocol for Evolved Packet System (EPS); Stage 3 (release 14),” 2017. https://portal.3gpp.org

[40] 3GPP TS 36.331 v14.3.0, “Radio Resource Control (RRC); Protocol specification (release 14),” 2017.

[41] Federal Communications Commission, “Wireless Emergency Alerts (WEA),” 2017. https://www.fcc.gov/consumers/guides/wireless-emergency-alerts-wea

[42] 3GPP, “LTE compatible commercial mobile alert system,” 3GPP, Washington DC, Tech. Rep., 2010. https://www.3gpp.org/news-events/partners-news/1130-LTE-Compatible-Commercial-Mobile-Alert-System

[43] K. Kyamakya and K. Jobmann, “Location management in cellular networks: Classification of the most important paradigms, realistic simulation framework, and relative performance analysis,” *IEEE Trans. Veh. Technol.*, vol. 54, no. 2, pp. 687–708, Mar. 2005.

[44] N. Nath, S. Parija, P. Sahu, et al, “Brief comparison of sequential paging and concurrent paging in cellular technology,” in *2015 IEEE Int. Conf. Ind. Instrum. Control*, May 2015, pp. 1078–1082.

[45] S. M. Razavi, “Tracking area planning in cellular networks-optimization and performance evaluation,” Thesis, Linkoping University, 2011. http://liu.diva-portal.org/smash/get/diva2:402919/FULLTEXT01.pdf

[46] X. Zhang, *LTE optimization engineering handbook*, 1st ed., John Wiley & Sons, 2018.

[47] M. Hawley, “Reduce core network signaling with a field-proven MME,” Alcatel-Lucent 9471 WMM R&D, Tech. Rep., 2015. http://www.tmcnet.com/tmc/whitepapers

[48] S. Hailu and M. Säily, "Hybrid paging and location tracking scheme for inactive 5G UEs," in *2017 IEEE Eur. Conf. Networks Commun.*, Jun. 2017, pp. 1–6.

[49] L. Chen, H. Liu, Z. Fan, et al, "Modeling the tracking area planning problem using an evolutionary multi-objective algorithm," *IEEE Comput. Intell. Mag.*, vol. 12, no. 1, pp. 29–41, Feb. 2017.

[50] M. Nawaz, "Exploiting tracking area list concept in LTE networks," Ph.D. dissertation, Linköping University, 2013.

[51] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile users: To update or not to update?" *Wirel. Networks*, vol. 1, no. 2, pp. 175–185, Jun. 1995. https://doi.org/10.1007/BF01202540

[52] A. Roy, J. Shin, and N. Saxena, "Entropy-based location management in long-term evolution cellular systems," *IET Commun.*, vol. 6, no. 2, p. 138, 2012.

[53] H. Fu, P. Lin, H. Yue, et al, "Group mobility management for large-scale machine-to-machine mobile networking," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1296–1305, Mar. 2014.

[54] T. Hashimoto, T. Kubo, and Y. Kitatsuji, "Lightweight paging based on location update intervals in LTE network," in *2016 IEEE Int. Conf. Commun.*, May 2016, pp. 1–6.

[55] S. Parija, N. Nath, P. Sahu, et al, "Dynamic intelligent paging in mobile telecommunication network," *Sādhanā*, vol. 43, no. 2, pp. 1–16, Feb. 2018.

[56] H. Liu, F. Gu, and Q. Zhang, "Decomposition of a multiobjective optimization problem into a number of simple multiobjective subproblems," *IEEE Trans. Evol. Comput.*, vol. 18, no. 3, pp. 450–455, Jun. 2014.

[57] A. Roy, A. Misra, and S. K. Das, "Location update versus paging trade-off in cellular networks: An approach based on vector quantization," *IEEE Trans. Mob. Comput.*, vol. 6, no. 12, pp. 1426–1440, Dec. 2007.

[58] A. Bhattacharya and S. K. Das, "LeZi-update: An information-theoretic approach to track mobile users in PCS networks," in *Proc. 5th Annu. ACM/IEEE Int. Conf. Mob. Comput. Netw.* New York, NY, 1999, pp. 1–12.

[59] E. Clayirci and I. Akyildiz, "User mobility pattern scheme for location update and paging in wireless systems," *IEEE Trans. Mob. Comput.*, vol. 1, no. 3, pp. 236–247, Jul. 2002.

[60] N. Samaan and A. Karmouch, "A mobility prediction architecture based on contextual knowledge and spatial conceptual maps," *IEEE Trans. Mob. Comput.*, vol. 4, no. 6, pp. 537–551, Nov. 2005.

[61] B. Blanco, J. Fajardo, I. Giannoulakis, et al, "Technology pillars in the architecture of future 5G mobile networks: NFV, MEC and SDN," *Comput. Stand. Interfaces*, vol. 54, pp. 216–228, Nov. 2017.

[62] A. Tawbeh, H. Safa, and A. R. Dhaini, "A hybrid SDN/NFV architecture for future LTE networks," in *2017 IEEE Int. Conf. Commun.*, May 2017, pp. 1–6.

[63] SDxCentral, "What is OpenFlow? Definition and how it relates to SDN," 2018. https://www.sdxcentral.com/sdn/definitions/what-is-openflow/

[64] R. Guerzoni, R. Trivisonno, and D. Soldani, "SDN-based architecture and procedures for 5G networks," in *Proc. 1st Int. Conf. on 5G for Ubiquitous Connect.*, Akaslompolo, 2014, pp. 209–214. https://doi.org/10.4108/icst.5gu.2014.258052

[65] C. Reichert, "AT&T to launch 5G across 19 cities," Sep. 2018. https://www.zdnet.com/article/at-t-to-launch-5g-across-19-cities/

[66] G. Akpakwu, B. Silva, G. Hancke, et al, "A survey on 5G networks for the internet of things: Communication technologies and challenges," *IEEE Access*, vol. 6, pp. 3619–3647, 2018.

[67] L. D. Nguyen, "Resource allocation for energy efficiency in 5G wireless networks," *EAI Endorsed Trans. Ind. Networks Intell. Syst.*, vol. 5, no. 14, pp. 1–6, Jun. 2018.

[68] N. Vo, T. Duong, M. Guizani, et al, "5G optimized caching and downlink resource sharing for smart cities," *IEEE Access*, vol. 6, pp. 31457–31468, 2018.

[69] N. Vo, T. Duong, H. Tuan, et al, "Optimal video streaming in dense 5G networks with D2D communications," *IEEE Access*, vol. 6, pp. 209–223, 2018.

[70] C. Yuen, M. Elkashlan, Y. Qian, et al, "Energy harvesting communications: Part 1 [Guest Editorial]," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 68–69, Apr. 2015.

[71] N. Nguyen, T. Duong, H. Ngo, et al, "Secure 5G wireless communications: A joint relay selection and wireless power transfer approach," *IEEE Access*, vol. 4, pp. 3349–3359, 2016.

[72] N. Panwar, S. Sharma, and A. K. Singh, "A survey on 5G: The next generation of mobile communication," *Phys. Commun.*, vol. 18, pp. 64–84, Mar. 2016.

[73] R. Taranto, S. Muppirisetty, R. Raulefs, et al, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 102–112, Nov. 2014.

[74] A. Osseiran, F. Boccardi, V. Braun, et al, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[75] E. Hossain, M. Rasti, H. Tabassum, et al, "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective," *IEEE Wirel. Commun.*, vol. 21, no. 3, pp. 118–127, Jun. 2014.

[76] B. Bangerter, S. Talwar, R. Arefi, et al, "Networks and devices for the 5G era," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 90–96, Feb. 2014.

[77] 3GPP TS 22.261 v16.1.0, "Service requirements for the 5G system; Stage 1 (release 16)," 2017. http://www.3gpp.org

[78] P. Agyapong, M. Iwamura, D. Staehle, et al, "Design considerations for a 5G network architecture," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 65–75, Nov. 2014.

[79] N. Meng, H. Zhang, and B. Lin, "User-centric mobility management based on virtual cell in ultra-dense networks," in *2016 IEEE/CIC Int. Conf. Commun. China*, Jul. 2016, pp. 1–6.

[80] P. Fan, J. Zhao, and C. I, "5G high mobility wireless communications: Challenges and solutions," *China Commun.*, vol. 13, no. Supplement2, pp. 1–13, 2016.

[81] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile users: To update or not to update?" Ph.D. dissertation, Jun. 1995.

[82] X. Ge, J. Ye, Y. Yang, et al, "User mobility evaluation for 5G small cell networks based on individual mobility model," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 528–541, Mar. 2016.

[83] J. Chan, S. Zhou, and A. Seneviratne, "A QoS adaptive mobility prediction scheme for wireless networks," in *IEEE GLOBECOM 1998 (Cat. NO. 98CH36250)*, Sydney, New South Wales, Australia, 1998, pp. 1414–1419 vol. 3.

[84] C. Song, Z. Qu, N. Blumm, et al, "Limits of predictability in human mobility." *Sci. (New York, N. Y.)*, vol. 327, no. 5968, pp. 1018–1021, Feb. 2010.

[85] J. Scourias, "Dynamic location management and activity-based mobility modelling for cellular networks," Thesis, University of Waterloo (Canada), 1997.

[86] 3GPP TR 23.799 v14.0.0, "Study on architecture for next generation system (release 14)," 2016. http://www.3gpp.org

[87] J. Kim, D. Kim, and S. Choi, "3GPP SA2 architecture and functions for 5G mobile communication system," *ICT Express*, vol. 3, no. 1, pp. 1–8, Mar. 2017. https://www.sciencedirect.com

[88] M. Bagaa, T. Taleb, and A. Ksentini, "Efficient tracking area management framework for 5G networks," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 6, pp. 4117–4131, Jun. 2016.

[89] S. Jeon, S. Figueiredo, R. L. Aguiar, and H. Choo, "Distributed mobility management for the future mobile networks: A comprehensive analysis of key design options," *IEEE Access*, vol. 5, pp. 11423–11436, 2017.

[90] D. Kominami, T. Iwai, H. Shimonishi, et al, "A control method for autonomous mobility management systems toward 5G mobile networks," in *2017 IEEE Int. Conf. Commun. Work. (ICC Work.)*, Paris, May 2017, pp. 498–503.

[91] L. Nguyen, H. Tuan, and T. Duong, "Energy-efficient signalling in QoS constrained heterogeneous networks," *IEEE Access*, vol. 4, pp. 7958–7966, 2016.

[92] K. C. K. Cheng and R. H. C. Yap, "Search space reduction for constraint optimization problems," in *Princ. Pract. Constraint Program.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 635–639.

[93] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet Things J.*, 2017.

[94] S. Chen, J. Hu, Y. Shi, Y. Peng, J. Fang, R. Zhao, and L. Zhao, "Vehicle-to-Everything (V2X) services supported by LTE-based systems and 5G," *IEEE Commun. Stand. Mag.*, 2017.

[95] A. Meola, "What is the Internet of Things (IoT)? Meaning & definition," 2018. https://www.businessinsider.com/internet-of-things-definition

[96] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-advanced pro and the road to 5G*, 3rd ed. Academic Press, 2016.

[97] 3GPP TS 36.422 v15.0.0, "X2 signaling transport (release 15)," 2018. https://portal.3gpp.org

[98] 3GPP TS 36.423 v15.1.0, "X2 application protocol (X2AP) (release 15)," 2018. http://www.3gpp.org

[99] Cisco Systems, "LTE paging profile configuration mode commands." USA: Cisco, 2017, ch. 42. https://www.cisco.com

[100] O. Safecom, "Land Mobile Radio (LMR) 101 part 1: Educating decision-makers on LMR technologies," Tech. Rep., 2016. https://www.dhs.gov/sites/default/files/publications/LMR101_508FINAL.pdf

[101] W. Müller, H. Marques, J. Rodriguez, and B. Bouwers, "Next-generation infrastructure for public protection and disaster relief organizations," *SPIE Newsroom*, Jul. 2016. http://www.spie.org/x119789.xml

[102] M. Zarri, "Network 2020: Mission critical communications," GSMA, Tech. Rep., 2017. www.gsma.com/network2020

[103] Nokia, "LTE networks for public safety services," 2015. http://networks.nokia.com

[104] P. Ringqvist, "The promise of 5G for public safety," 2018. https://www.emsworld.com/commentary/1221807/promise-5g-public-safety

[105] AT&T Newsroom, "All 50 U.S. States, 2 Territories and the District of Columbia Opt-In to FirstNet," Dec. 2017. http://www.5gamericas.org/en/newsroom/member-news/all-50-us-states-2-territories-and-district-columbia-opt-firstnet/

[106] 3GPP TS 24.334 v15.2.0, "Proximity-Services (ProSe) User Equipment (UE) to ProSe function protocol aspects; Stage 3 (release 15)," 2018. http://www.3gpp.org

[107] 3GPP TS 22.179 v16.4.0, "Mission Critical Push To Talk (MCPTT); stage 1 (release 16)," 2018. http://www.3gpp.org

[108] 3GPP TS 23.468 v15.0.0, "Group Communication System Enablers for LTE (GCSE_LTE); stage 2 (release 15)," 2018. https://www.etsi.org

[109] 3GPP TR 22.862 v14.1.0, "Feasibility study on new services and markets technology enablers for critical communications; stage 1 (release 14)," 2016. https://portal.3gpp.org

[110] Public Safety and Homeland Security, "2017 Atlantic hurricane season impact on communications," Tech. Rep., 2018. https://docs.fcc.gov/public/attachments/DOC-353805A1.pdf

[111] 3GPP TS 36.213 v15.2.0, "Physical layer procedures (release 15)," 2018. https://portal.3gpp.org

[112] 3GPP TS 36.214 v14.4.0, "Physical layer; Measurements (release 14)," 2018. https://portal.3gpp.org

[113] S. Mcgrath, E. Grigg, S. Wendelken, et al, "ARTEMIS: A vision for remote triage and emergency management information integration," 2003. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.4959

[114] K. Lorincz, D. Malan, T. Fulford-Jones, et al, "Sensor networks for emergency response: Challenges and opportunities," *IEEE Pervasive Comput.*, vol. 3, no. 4, pp. 16–23, Oct. 2004.

[115] D. Williams, "Tactical medical coordination system (TacMedCS)," NAVAL HEALTH RESEARCH CENTER SAN DIEGO CA, Tech. Rep., 2007. https://apps.dtic.mil/docs/citations/ADA477535

[116] J. C. Tanner, "Nokia unveils LTE backpack for critical comms with bonus data aggregation," May 2017. https://disruptive.asia/nokia-lte-backpack-critical-communications/

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

[118] G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu, and N. Brahmi, "Device-to-Device communications for national security and public safety," *IEEE Access*, vol. 2, pp. 1510–1520, 2014.

[119] K. Ali, H. Nguyen, P. Shah, et al, "Architecture for public safety network using D2D communication," in *2016 IEEE Wirel. Commun. Netw. Conf. Work.*, Apr. 2016, pp. 206–211.

[120] K. Ali, H. Nguyen, Q. Vien, et al, "Disaster management using D2D communication with power transfer and clustering techniques," *IEEE Access*, vol. 6, pp. 14643–14654, 2018.

[121] Z. Kaleem, N. Qadri, T. Duong, et al, "Energy-efficient device discovery in D2D cellular networks for public safety scenario," *IEEE Syst. J.*, vol. 13, no. 3, pp. 2716–2719, Sep. 2019.

[122] A. Masaracchia, L. Nguyen, T. Duong, et al, "An energy-efficient clustering and routing framework for disaster relief network," *IEEE Access*, vol. 7, pp. 56520–56532, 2019.

[123] S. Sharma and H. M. Pandey, "Genetic algorithm, particle swarm optimization and harmony search: A quick comparison," in *2016 6th Int. Conf. - Cloud Syst. Big Data Eng.*, Noida, Jan. 2016, pp. 40–44.

[124] X. Lin, V. Yajnanarayana, S. Muruganathan, et al, "The sky is not the limit: LTE for unmanned aerial vehicles," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 204–210, Apr. 2018.

[125] S. Muruganathan, X. Lin, H. Maattanen, et al, "An overview of 3GPP release-15 study on enhanced LTE support for connected drones," May 2018. http://arxiv.org/abs/1805.00826

[126] S. Shakoor, Z. Kaleem, M. Baig, et al, "Role of UAVs in public safety communications: Energy efficiency perspective," *IEEE Access*, 2019.

[127] A. Merwaday and I. Guvenc, "UAV assisted heterogeneous networks for public safety communications," in *2015 IEEE Wirel. Commun. Netw. Conf. Work.*, Mar. 2015, pp. 329–334.

[128] M. Deruyck, J. Wyckmans, W. Joseph, et al, "Designing UAV-aided emergency networks for large-scale disaster scenarios," *EURASIP J. Wirel. Commun. Netw.*, vol. 2018, no. 1, p. 79, Dec. 2018. https://doi.org/10.1186/s13638-018-1091-8

[129] D. He, S. Chan, and M. Guizani, "Drone-assisted public safety networks: The security aspect," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 218–224, Apr. 2017.

[130] Google.org, "https://google.org/crisismap/weather_and_events."

[131] 3GPP TS 36.133 v14.5.0, "Requirements for support of radio resource management," 2017. http://www.etsi.org/standards-search

[132] Fujitsu Network Communications Inc., "High-capacity indoor wireless solutions: Picocell or Femtocell?" Tech. Rep., 2013. https://www.fujitsu.com/us/Images/High-Capacity-Indoor-Wireless.pdf

[133] D. Arthur, D. Arthur, and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discret. Algorithms*, 2007.

[134] 3GPP TS 36.942 v15.0.0, "Radio Frequency (RF) system scenarios (release 15)," 2018. http://www.3gpp.org

[135] M. Rowe, "Beam steering: One of 5G's components," 2018. https://www.edn.com/electronics-blogs/5g-waves/4460861/Beam-steering--One-of-5G-s-components

[136] M. M. Abusitta, Y. A. S. Dama, R. A. Abd-Alhameed, et al, "Beam steering of horizontally polarized circular antenna arrays," in *2011 IEEE Loughbrgh. Antennas Propag. Conf.*, Loughborough, Nov. 2011, pp. 1–4. https://doi.org/10.1109/LAPC.2011.6114126

[137] A. Singh, A. Kumar, A. Ranjan, et al, "Beam steering in antenna," in *2017 IEEE Int. Conf. Innov. Information, Embed. Commun. Syst.*, Coimbatore, Mar. 2017, pp. 1–4.

[138] P. Frenger, P. Moberg, J. Malmodin, et al, "Reducing energy consumption in LTE with cell DTX," in *2011 IEEE 73rd Veh. Technol. Conf.*, Yokohama, May 2011, pp. 1–5.