

Mitigating Signalling Storms in 5G

by

Bohan Zhang

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

© Bohan Zhang 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following list serves as a declaration of the Versions of Record for works included in this thesis:

1. *Portions of Chapter 3*: **Zhang, Bohan**, Zeinaty, P., Limam, N., Boutaba, R. “Mitigating Signaling Storms in 5G with Blockchain-assisted 5GAKA”, IEEE International Conference on Network and Service Management (IEEE CNSM), 2023 [[1](#)]
2. *Portions of Chapter 4*: **Zhang, Bohan**, Hu, P., Azirani, A., Salahuddin, M. A., Barradas, D., Limam, N., Boutaba, R. “Secure and Efficient Group Handover Protocol in 5G Non-Terrestrial Networks” IEEE International Conference on Communication (IEEE ICC), 2024 [[2](#)]
3. *Portions of Chapter 5*: **Zhang, Bohan**, Salahuddin, M. A., Hu, P., Wang, Y., Limam, N., Sun, B., Barradas, D., Boutaba, R. “Signalling Load-aware Conditional Handover in 5G Non-Terrestrial Networks”, IEEE International Conference on Network and Service Management (IEEE CNSM), 2024 (under review)

Abstract

Over the past 45 years, cellular networks have evolved from a luxury service for the few to an indispensable necessity for everyone, giving rise to countless new applications. This remarkable technological advancement has fundamentally transformed everyone's life. The latest fifth generation (5G) promises unprecedented performance in terms of latency, throughput, and subscriber density, as well as groundbreaking new services like autonomous driving, remote surgery, smart cities, and ubiquitous coverage. However, numerous challenges exist in achieving these promises, with one of the most significant being the vulnerability to signalling storm.

A signalling storm occurs when the volume of control signals exceeds the network's processing capacity, leading to service disruptions. During these disruptions, retries from the users and network entities amplify the problem, creating a snowball effect. This persistent threat, present since the era of the third generation (3G), remains a major concern in 5G networks. Over the past decade, signalling storms have caused severe outages worldwide, affecting millions of lives, resulting in financial losses amounting to millions of dollars, and millions of subscriber hours lost.

With the transition from the fourth generation (4G) to 5G, network elements have been virtualized and assigned more granular tasks, significantly increasing their numbers and the number of signals required to complete system procedures. From the subscriber's perspective, the enriched diversity of application scenarios and the increase in subscriber numbers have exponentially raised the complexity of network resource management and orchestration. Notably, the massive number of globally roaming cellular IoT devices with limited security features are susceptible to being compromised to execute low-cost Distributed Denial-of-Service (DDoS) attacks on operators worldwide. Due to the significant losses in the past, governments, operators, and academia consider signalling storms a major threat to 5G.

This thesis investigates past cases and numerous industry white papers, research articles, and 3GPP standards, summarizing the causes, scenarios, and mitigation solutions for signalling storms in 5G. Following these investigations, we propose a blockchain-assisted 5G Authentication and Key Agreement protocol to mitigate registration signalling storms. We also introduce a secure and efficient group handover protocol to address handover signalling storms in 5G Low Earth Orbit satellite non-terrestrial networks (NTN). Furthermore, we design a signalling load-aware conditional handover to reduce signalling peak in 5G NTN.

Network capacity has always been a driving force behind the evolution of cellular networks. We hope this work will shed light on the future design of 5G and beyond cellular networks.

Acknowledgements

I would like to express my gratitude to Raouf Boutaba for the continuous funding¹, support and guidance. Especially, I am grateful that he offered me this opportunity to learn 5G technology. I would like to express my gratitude to Noura Limam and Mohammad Ali Salahuddin for their continuous, tireless assistance in my research over the past two years.

They have provided great advice on my research, life, and future. They never put limitations on my research direction, allowing me to explore subjects that interest me. When I asked for a discussion, they were always available.

I would also like to express my gratitude to my other co-authors: Peng Hu, Paul Zeinaty, Ahmad Akbari Azirani, Diogo Barradas and Yunli Wang.

These people helped me establish my understanding of academia.

¹We acknowledge the support from IDEaS program from DND and the support from NRC-Waterloo Collaboration Centre.

Dedication

Dedication to myself and my family.

Table of Contents

Author’s Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	v
Dedication	vi
List of Figures	xi
List of Tables	xiii
1 Introduction	1
2 Background	3
2.1 Evolution of Cellular Network	3
2.2 5G Architecture	8
2.2.1 Overview	8
2.2.2 5G RAN	9
2.2.3 5G core network	12
2.2.4 Network Slicing	17

2.3	5G Non-Terrestrial Networks	18
2.3.1	Overview	18
2.3.2	Deployment Options	20
2.4	Signalling Storm in 5G	21
2.4.1	Past cases and causes	21
2.4.2	5G signalling storm scenarios	26
2.4.3	Mitigation of signalling storm	30
3	Mitigating Signalling Storm in 5G with Blockchain-assisted 5GAKA	35
3.1	Introduction	35
3.2	Background	37
3.2.1	SUPI and USIM	38
3.2.2	5G Authentication Protocol	38
3.3	Related Work	41
3.4	Problem statement	42
3.4.1	Attack Model	43
3.4.2	Attack Outcome	44
3.5	Blockchain-assisted 5G Authentication	44
3.5.1	Cryptographic Primitives	45
3.5.2	Key Ideas	46
3.5.3	The Setup	47
3.5.4	Blockchain-assisted 5G Authentication Workflow	48
3.5.5	Comparison with Baseline Mitigation Procedure	48
3.5.6	Security Analysis	50
3.6	Implementation and Evaluation	52
3.7	Discussion	54
3.8	Conclusions	56

4	Secure and Efficient Group Handover in 5G Non-Terrestrial Networks	57
4.1	Introduction	57
4.2	Related Work	59
4.3	Problem Statement	60
4.4	Secure and Efficient Group Handover	61
4.4.1	High-level Overview	61
4.4.2	Protocol Description	62
4.4.3	Protocol Analysis	65
4.5	Performance Evaluation	66
4.5.1	Satellite Handover Simulator	67
4.5.2	Experimental Setup	67
4.5.3	Experimental Results	69
4.6	Conclusions	71
5	Signalling Load-aware Conditional Handover in 5G Non-Terrestrial Networks	72
5.1	Introduction	72
5.2	Background and Related Works	74
5.2.1	Conditional handover protocol	74
5.2.2	Random access procedure	76
5.2.3	Related works	76
5.3	Mitigating Signalling Storm	78
5.3.1	Problem	78
5.3.2	Mathematical formulation	79
5.4	Load-aware Conditional Handover	84
5.4.1	Overview	84
5.4.2	Select candidates	85
5.4.3	Prepare configuration	86

5.4.4	Prepare condition	88
5.5	Performance Evaluation	88
5.5.1	Experiment setup	88
5.5.2	Experiment results	90
5.5.3	Analysis	92
5.6	Discussion and Future Work	94
5.7	Conclusions	94
6	Conclusions	96
	References	98

List of Figures

2.1	5G RAN protocol stack [3]	10
2.2	5G service based architecture [4]	12
2.3	5G core data plane protocol stack [4]	13
2.4	5G core control plane protocol stack [4]	13
2.5	Communication models for NF/NF services interaction [4]	15
3.1	5G service-based architecture [5]	37
3.2	5G-AKA Initialization	39
3.3	5G-AKA Authentication	40
3.4	System Overview	47
3.5	Blockchain-assisted 5G Authentication: Initialization phase	50
3.6	Deployment Model	53
3.7	Registration Processing Time Comparison for 4 Scenarios	53
3.8	CDF of registration time for the 4 Registration Storm Scenarios	54
4.1	Handover scenario with Earth moving cells	61
4.2	Sequence diagram of the proposed NTN handover protocol (the symbol $\diamond$ indicates a communication channel is secure; symmetric encryption with a key K is denoted $(*)_K$)	63
4.3	Message processing model	67
4.4	Scenario of our experiment	68
4.5	Message count every 200 ms	70

5.1	Conditional handover protocol	75
5.2	Random access procedure	77
5.3	Experiment scenarios	89
5.4	Signalling peak evaluation	91
5.5	Signalling peak comparison with different UE density	91
5.6	Busy time slot evaluation	92
5.7	Handover count and serving time evaluation	93
5.8	Decouple time and reservation evaluation	94

List of Tables

3.1	Smart Contract Standard	49
4.1	Simulation parameters and settings	68
4.2	Simulation results (baseline HO vs. GHO)	69
5.1	Inputs and variables	80
5.2	Algorithms notations	90

Chapter 1

Introduction

By the end of 2023, around 280 mobile operators had launched commercial 5G services, with Ericsson, Huawei, and Nokia holding the majority market share in terms of network equipments. Global 5G mobile subscriptions reached approximately 1.6 billion, accounting for 18 percent of all mobile subscriptions. Ericsson forecasts that by 2029, global 5G mobile subscriptions will exceed 5.3 billion, making up 58% of all mobile subscriptions [6]. Research and Markets, the world’s largest market research store, estimates that the global 5G services market will grow from about 122 billion USD in 2023 to approximately 10 trillion USD by 2028, at a compound annual growth rate of 52.4% [7].

This substantial market growth is driven by 5G’s promise of unprecedented network performance, including low latency, high throughput, and massive user density. These metrics enable applications previously unimaginable, such as autonomous driving, smart cities, virtual reality/augmented reality (VR/AR), and autonomous industries. Additionally, 5G will leverage satellites to achieve global network coverage. Consequently, there are high expectations for 5G to significantly boost economic development [8]. Furthermore, 5G is expected to enhance military activities and national defense capabilities [9] by enabling smart warehousing, drones, distributed command and control, satellite-enabled monitoring, and improved rural area connections. Recognizing its strategic importance, countries are investing heavily in this technology. As the US Department of Defense states, “5G is a critical strategic technology: those nations that master advanced communications technologies and ubiquitous connectivity will have a long-term economic and military advantage [10].”

However, these ambitious visions rely on a fundamental requirement: a reliable system without disruption. According to ENISA’s report [11], the number of global subscriber

hours lost has been increasing since 2020. Moreover, in recent years, several large-scale and prolonged outages have occurred [12, 13, 14, 15, 16]. During these outages, people were unable to communicate, use web services, or access emergency services. This level of disruption is undoubtedly unacceptable, especially in the context of 5G services.

This thesis investigates one of the primary causes of cellular network outages: signalling storms. A signalling storm happens when the volume of control signals exceeds the network's processing capacity, leading to service disruptions. During these disruptions, retries from users and network entities amplify the problem, creating a snowball effect. This persistent threat, present since the era of 3G, has caused several severe outages worldwide [17, 14, 18, 16, 19, 20], affecting millions of lives, resulting in financial losses amounting to millions of dollars, and millions of subscriber hours lost. In the 5G service-based architecture, the number of control signals is expected to be as much as 100 times greater, or even more, compared to 4G Diameter signalling [21]. From the subscriber perspective, the enriched diversity of application scenarios and the increase in subscriber numbers have exponentially raised the complexity of network resource management and orchestration. Notably, the massive number of globally roaming of cellular IoT devices makes them susceptible to being compromised to execute low-cost DDoS attacks on operators worldwide. Therefore, operators and government agencies consider signalling storms, either from legitimate traffic or malicious attacks, a major threat to 5G [22, 23, 24, 21, 25, 26, 27, 28].

The contributions of this thesis are as follows:

- We performed a comprehensive review of various academic research and industry reports of signalling storm threat in 5G networks. We summarized the causes, mitigation strategies, and various scenarios that may lead to a signalling storm in 5G networks.
- We investigated two scenarios and proposed our solutions. The first scenario is that a massive number of devices perform initial registration. And the second scenario is that a massive number of handovers caused by the mobility of LEO satellites in 5G non-terrestrial networks (NTN).
- We experimentally demonstrated the performance of our proposed solutions. All the code is made publicly available for the reproducibility of the results.

Chapter 2

Background

2.1 Evolution of Cellular Network

In 1865, James Clerk Maxwell invented Maxwell's equations, which describe how electric and magnetic fields interact and propagate, laying the foundation of electromagnetism. Based on this theory, he predicted the existence of electromagnetic waves, and Hertz's Experiment later proved it by generating radio waves in 1888. Tesla, Marconi, and Popov later independently created radiotelegraphy, which can transmit text using radio waves, indicating the beginning of wireless communication technology. Because radiotelegraphy is not restricted by the deployment of electric cables, it enables flexible point-to-point long-distance communication on extreme terrains such as sea and mountains. Therefore, it played a critical role in commercial, diplomatic, and military affairs throughout the first half of the 20th century. Compared to text transmission, sound transmission could convey more information. In 1900, Fessenden first discovered amplitude modulation radio, which enables sound transmission. This breakthrough later led to the development of radio broadcasting and walkie-talkies, which were extensively utilized during World War II and have continued to play important roles in today's society.

After World War II, the Bell system developed the Mobile Telephone Service (MTS) in 1946, and AT&T commercialized MTS to the civilian world, now considered pre-cellular or 0G. Back then, the user equipment (UE) was the car phone, which was heavy and expensive and usually mounted in cars or trucks, demonstrating this service's luxury. MTS operates in the VHF band, and one central base station would cover a very large area (e.g., a city) because VNF waves can propagate over a long distance. Thus, handover features such as changing the frequency channel were not supported. However, such a design can cause

network congestion and signal interference because the number of frequency channels is limited. As a result, only three customers in a city could make calls simultaneously, [29], which can be treated as the early signalling storm caused by the growth of UEs. To address the channel limitations in 0G and to support more subscribers, Ring first introduced the cellular concept in 1947. He proposed dividing the country into small hexagonal geographic locations called cells; each cell is covered and controlled by one base station. In his design, no two adjacent cells would use the same set of frequency channels to achieve frequency re-use, which could support more subscribers in 0G [30]. Unfortunately, this design must require handover between cells, which cannot be implemented with the technology then. So this vision was not realized in 0G [31]. At the near end of the 0G era, Cooper invented the first handheld phone, DynaTAC 8000X, which aimed to replace the heavy car phone, and made the first mobile phone call in 1973.

In 1977, Bell Lab built a trial cellular network around Chicago. After the trial's success, NTT started to build the first commercial cellular network in Tokyo and later expanded it to cover Japan's whole population, known as the 1G cellular network. Later, in 1983, DynaTAC 8000X was manufactured and commercialized, which forged the foundation of the modern mobile cellular network as we know it today. Due to the lack of a 1G standard, each country had its own 1G system in Europe, which was incompatible with each other. This caused difficulty in providing roaming services. Also, 1G transmits voice calls using analog modulation with frequency-division multiple access (FDMA), which is prone to eavesdropping and requires a lot of wireless spectrum. As mobile phone prices and 1G service charges decreased, the subscriber base grew heavily and soon reached the capacity of 1G, leading to more dropped ongoing calls or busy signals when trying to initiate a new call [31]. In 1991, the standard Global System for Mobiles (GSM) was published by the European Telecommunications Standards Institute (ETSI), known as 2G. Instead of analog modulation, 2G GSM used digital modulation with time-division multiple access to achieve encryption, more efficient spectrum utilization, and increased capacity. Because of a common standard, 2G allows international roaming. Moreover, 2G introduced the SIM card with an authentication scheme and a new popular variant of communication called short message service (SMS), although it could only transmit 160 characters at most.

Like from 1G to 2G, new mobile generations have appeared approximately every ten years. Since the Internet became a commercial medium, data service has become a persistent driving force in the evolution of cellular networks. The Universal Mobile Telecommunications System (3G UMTS), which was developed by the 3rd Generation Partnership Project (3GPP) and finalized in 2001, fully enables fast multimedia messaging service (MMS) and mobile internet service. Compared to 64 kbps in 2G, 3G achieved a network bandwidth of about 2 Mbps using Wideband Code Division Multiple Access (WCDMA),

increasing the channel frequency bandwidth from 200 kHz to 1.25 MHz. Also, 3G added a new licensed frequency band (2100 MHz) to mitigate the network congestion caused by increased subscribers and service usage. Importantly, 3G specifies a complete cellular network architecture, including RAN and the core network. The 3G core contains a packet-switched network functions domain that connects to the external IP network to build data sessions and a circuit-switched network functions domain that connects to an external public-switched telephone network to build voice sessions.

By the end of 2007, 3G provided services to 295 million people. However, the smartphone revolution brought by the iPhone and apps and the rise of streaming services such as YouTube introduced new demands for higher data rates. Industry soon found that the 3G network capacity was insufficient to support the bandwidth-intensive services planned by mobile operators, even with only a small amount of subscribers utilizing the service within a service area [32]. Also, severe outages happened worldwide because of the increasing number of smartphones [33, 34] has made the operators realize 3G is not enough. As a continuous progress to improve 3G, 3GPP finalized the standard, Long Term Evolution (LTE), in 2008 and later considered as 4G in 2010. 4G entirely abandoned circuit-switched network and transmits voice in packets through Voice over Internet Protocol (VoIP), which makes the 4G core network a complete packet-switched network. Compared to the hierarchical architecture of the 3G core, the fully IP-based and flattened architecture in the 4G core reduced the complexity and overhead associated with routing and management, resulting in a lower network latency from about 100~500 ms to about 30~50 ms. The low latency feature enabled latency-sensitive applications, such as real-time video calls, livestreaming, and online gaming. To further improve the spectrum utilization, 4G LTE installed MIMO antenna, increased the spectrum bandwidth, and replaced WCDMA with Orthogonal frequency-division multiplexing (OFDM) with up to 256-QAM (Quadrature Amplitude Modulation) to theoretically achieve network bandwidth up to 100 Mbps. However, although the data capacity issue was relieved, another cellular network threat was rising - the signalling storm. In the past, operators could reasonably manage the number of control signals because the types of services were limited, and each person was expected to have about only one cell phone. The advent of the iPad and apps began to make a change [35]. One person was no longer considered to carry only one cellular device, and multiple running apps on their devices will constantly chat with the Internet through the cellular network. The concurrent network sessions caused many more control signals, including authentication, handover, and billing, and added burdens to both RAN and core network, which can cause large-scale outages [18]. Since then, the operators began to lose their full control of signal management.

5G New Radio (NR) began to be deployed worldwide in 2019. The two main deploy-

ment options are the standalone option, which completely abandons the 4G component, and the non-stand-alone option, in which 5G components and 4G components co-exist to reduce deployment cost. Compared to the 4G's low-band radio waves (below 2.6 GHz), 5G additionally uses microwave in mid-band and high-band (FR1: 410 MHz to 7125 MHz [36], FR2: 24.25 GHz to 71.0 GHz [37]). Also, with carrier aggregation, 5G increased the maximum channel bandwidth from 20 MHz to 100 MHz in FR1 and 400 MHz in FR2, significantly increasing the network bandwidth. However, high-frequency waves, such as millimeter waves, exhibit shorter propagation distances due to higher path loss. Thus, 5G also introduced small base stations to provide more focused coverage in dense urban environments. Compared to the 4×4 MIMO in 4G, 5G installed massive MIMO antennas in the base station, which allows more simultaneous communications and enhances beamforming performance.

Notably, 5G made a big change in the RAN and core network design with the rise of cloud service and virtualization technologies. Previous RAN architecture was a centralized monolithic design with limitations in scalability and flexibility. 5G supports splitting the gNB into distributed units (DU) and centralized units (CU) along with RAN function split options to increase efficiency flexibly based on the different environment needs. The 4G core functions are also divided into more different network functions in 5G, resulting in higher granularity in capability and management. The 5G network elements, including DU, CU, and network functions in the core, are designed to be virtualized as VMs or containers managed using the NFV MANO framework so they can be deployed in the cloud and enable function migration [38]. Benefiting from virtualization and clearly defined interfaces, 5G functions can achieve interoperability between vendors because functions run in software, which is no longer coupled with hardware. As RAN contributes the most to the QoS of the communication, such interoperability boosted organizations like O-RAN ALLIANCE to work on the open-RAN architecture. The open RAN mainly aims to bring RAN intelligence by introducing non-RT RIC in the orchestration layer and near-RT RIC in the DU with xApp to perform monitoring, smart control, and management. 5G is also designed to separate the control and data planes in RAN and the core network to increase scalability and reliability. With the highly decoupled functions and flexible deployment options brought by virtualization and design, 5G can lower the latency to under 10ms by processing the data at the network's edge (multi-access edge computing). Also, more suitable services can be provided flexibly according to different application scenarios, which changes the previous "one size fits all" cellular network to a "one size per service" network. Thus, a completely new business model is appearing - network slicing. The main network slice services that 5G aims to provide are as follows [39]:

- **Enhanced Mobile Broadband (eMBB):** This service aims to provide high net-

work bandwidth for data transmission. The theoretical peak data rate required is 20 Gbps and 10 Gbps for downlink and uplink, respectively, with mobility support up to 500 km/hour. Data-intensive applications such as augmented reality and virtual reality can be achieved in this service.

- **Ultra Reliable Low Latency Communications (URLLC):** This service aims to achieve extremely low latency, about less than 1 ms for the data plane and 10 ms for the control plane. Latency-sensitive and mission-critical applications such as autonomous driving and remote surgery can be achieved in this service.
- **Massive Machine-Type Communications (mMTC):** This service aims to connect large numbers of UEs. Compared to 4G, which can support under 0.1 million UEs/km², 5G aims to support 1 million UEs/km². With such high-density connectivity, smart cities and remote sensors that significantly require massive IoT devices can be achieved.

Towards these goals, 3GPP has finished Rel 15-18 and is actively working on Rel 19, which aims to end before 2026. Rel 19, as a baseline for future 6G, will focus on IoT, NTN, extended reality applications, and network energy efficiency and will apply intelligent control and automation to enhance mobility performance and load balancing [40]. By 2023, about 280 providers around the world have launched commercial 5G services, and more than 40 have deployed or launched 5G standalone [6].

The advance of the cellular network is built upon countless breakthroughs in various fields, including materials science, physics, and computer science. However, as the cellular network becomes bigger and more complex as an engineering product, the management inevitably shows some difficulties, such as a longer time lag in applying, analyzing, and testing new technologies and higher communication costs in developing the standard, resulting in higher possibility of ambiguity and inconsistency among thousands of documents [41, 42]. As cellular networks advance and cell phones become necessities from an optional luxury service in only 45 years, network reliability and availability have become a huge promise, and the role of mobile operators in society has been significantly enhanced. However, maintaining persistent reliability every moment is difficult for an increasingly large system, and outages can happen from various sources, which are difficult to foresee. Extreme weather conditions or human accidents can damage the physical infrastructure, the misconfiguration of network servers and devices can disrupt the expected services, and the sudden load increase or malicious DDoS attacks can cause network congestion. When such faults happen, the unpredictable behavior from the unhealthy cellular system can cause large-scale and long-time outages, which can result in significant financial loss to both operators and millions of subscribers in today's world [43, 12, 13, 11]. Moreover, public safety is also threatened; it was reported that people have difficulty accessing the emergency service

during hours of outages [44] because they do not have a landline. Unlike previous generations, 5G highly relies on third-party data centers to deploy essential network functions. These data centers are reported to have increased and more frequent outages and have already caused millions of dollars in loss to tech companies like Amazon and Apple. Thus, the trend of the possibility of 5G outages in the future may not be optimistic. Also, 5G aims to enable more applications such as autonomous driving, intelligent transportation, and smart cities, which may change people’s lives once again. Like landlines are replaced by cellular networks [45], it is unclear if 5G may replace some existing products, although they may be capable of providing a more reliable service. If so, the expense of a service disruption such as outages may significantly increase. Although the improvement of cellular networks still has challenges in capacity and reliability, judging from its evolution and outcome, countless life-changing products and services have been born; numerous different vertical industries have benefited, making the cellular network one of the most successful and vital products in human history.

2.2 5G Architecture

2.2.1 Overview

As the successor to 4G, 5G also adopts the logical cellular network architecture containing RAN, backhaul network, and the core network. The RAN directly provides the wireless access point to the UEs over radio waves, which stays at the edge of the cellular network. It manages the spectrum usage and tries to guarantee the radio bearer quality (Channel quality indicator) in the physical layer and network performance (5G quality of service identifier) in the network layer of many UEs. The backhaul network connects the RAN and the core network, and the core network connects to the Internet to realize the data service between UE and the Internet. Unlike WiFi, a best-effort, contention-based service focused on the local area, cellular networks focus on service continuity in the wide area and performance guarantee for their subscribers. Thus, access, mobility, service quality, billing, and subscriber privacy are some of the main basic requirements of the cellular network. These requirements are mainly realized by the core network, making the core network the heart and brain of the cellular network. And the message transmitted in the cellular network to realize these requirements and control the system is called “signalling” or, more precisely, “control signal.” Because backhaul is mainly for routing packets and is not involved in control management, it does not exist in the design of signalling protocol and 5G architecture. Thus, the logical communication interfaces are defined between RAN

and core without backhaul. In 5G, the communication interfaces between RAN and core are called NG Interface. The interface between UE and the RAN is called the Uu interface (or air interface). From the perspective of cellular network evolution, each generation of architectural changes has led to more efficient system control to increase network scalability, reliability, and capacity. For 5G, the architecture has mainly benefited from virtualization. Next, we will introduce the architecture of the 5G RAN and 5G core network, along with the difference between 4G, to reveal how the 5G architecture can enable more efficient control of the cellular system.

From a global view, 5G is designed to separate the control plane and data plane (or user plane), an idea inherited from software-defined networking (SDN). This ensures the system flexibility because two planes can adjust their resources independently of each other. Also, this design helps to reduce latency by allowing the data plane node in the core network to move closer to the network edge without affecting the control plane nodes. Thus, the protocol stack for both RAN and core in 5G is separated into the control plane and data plane stack. Also, functions in RAN and core are virtualized, so 5G also adopts the NFV MANO framework to orchestrate and manage NFs and network slices.

2.2.2 5G RAN

5G RAN (or NG RAN) comprises many logical nodes, called gNB, previously called eNB in 4G. gNB is designed to communicate with UE using an air interface protocol stack from the bottom up, as shown in Fig. 2.1. The main functions of these layers are as follows [3]:

- **Low PHY, High PHY (L1):** Translate between analog or digital signals from RF and transport blocks on the transport channel. Responsible for coding and modulation, error detection and correction, channel access control, and power control.
- **Low MAC, High MAC (L2):** Translate between transport blocks and MAC service data unit (SDU) on the logical channel (DTCH, BCCH, PCCH, CCCH, DCCH). Responsible for HARQ, mapping between logical channels and transport channels, multiplexing, scheduling information reporting, and logical channel prioritization.
- **Low RLC, High RLC (L2):** Translate between MAC SDU and RLC PDU on the RLC channel. Responsible for ARQ, concatenation, segmentation, and reassembly. It operates in 3 modes of operation: Transparent Mode (TM), Unacknowledged Mode (UM), and Acknowledged Mode (AM).
- **PDCP (L2):** Translate between RLC PDU and PDCP PDU on radio bearers (DRB for data plane, SRB for control plane). Responsible for compressing and decompressing IP headers, ciphering and deciphering, integrity protection and verification, main-

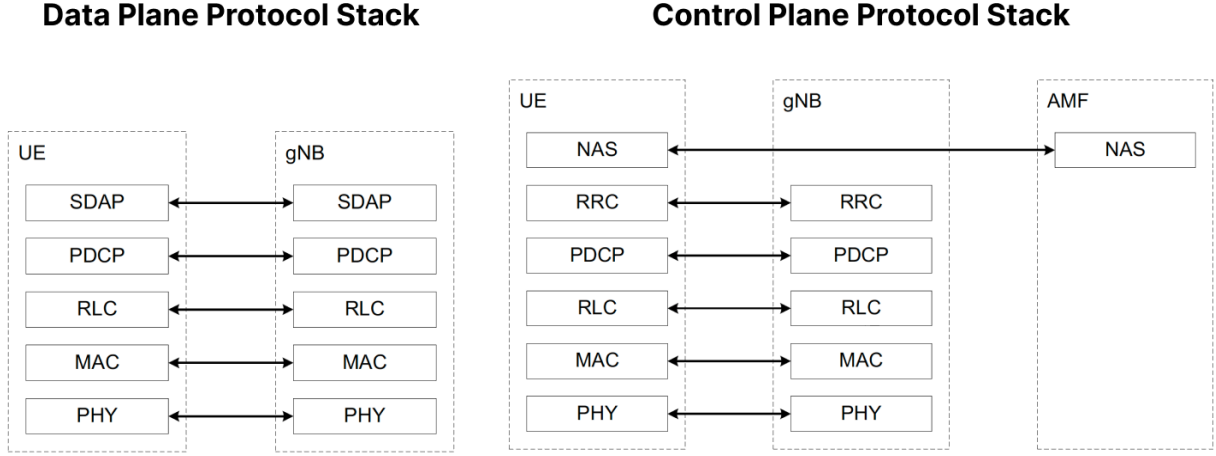


Figure 2.1: 5G RAN protocol stack [3]

tenance of PDCP SNs, duplication, reordering and in-order delivery, and out-of-order delivery.

- **SDAP (L2):** Responsible for mapping between a QoS flow and a data radio bearer and marking QFI. Marking QoS flow ID (QFI) in both DL and UL packets. SDAP only exists in the data plane and does not exist in 4G.
- **RRC:** Responsible for management related services. Main services include broadcasting system information, paging, connection management, mobility management, QoS management, and NAS message translation. RRC only exists in the control plane.

In 4G, eNB is comprised of the antenna, RRU, and BBU. The antenna and RRU handle the RF function, which processes radio waves into digital signals. They are mounted in a cell base tower with a feeder between the antenna and RRU. BBU handles the whole air interface protocol stack from PHY to SDAP/RRC. Fronthaul is the interface between RRU and BBU, and backhaul is the interface between BBU and the core network. One BBU could handle multiple RRU, but no central entity manages multiple BBUs. In 5G, due to the massive MIMO, more feeder ports are required on RRU, which is difficult to achieve. Thus, the antenna and RRU are integrated into AAU.

The BBU is also split into DU and CU, which can be virtualized and decoupled from hardware. The fronthaul is then the interface between AAU and DU, and the backhaul becomes the interface between CU and the core network. Due to the separation, F1 (midhaul) is introduced as the interface between DU and CU. One DU can manage multiple AAUs, and one CU can manage multiple DUs in a centralized manner. CU can also be

further split into one CU-CP for the control plane and multiple CU-UP for the data plane. The interface between CU-CP and CU-UP is called E1 interface [46]. The interface between two gNBs is called the Xn interface, which is also the interface between two gNB-CUs (Xn-C interface). DU-CU separation brings the following main benefits:

- **Increased capacity:** DU-CU separation was originally proposed as a solution to avoid the impact of traffic surge. Combining the baseband resource of multiple BBUs or deploying CU in a data center could effectively increase the capacity of a gNB. A gNB could handle more cells and reduce the interaction between gNBs (e.g., more intra-gNB handover and less Xn handover). More resources can also be allocated to high-traffic cells to avoid congestion.
- **Higher flexibility:** DU-CU separation also brings at least eight possible functional split options [47]. A split option refers to cutting the packet processing pipeline in the RAN protocol stack into two and assign to DU and CU. For example, in option 7 (Intra PHY split), AAU and DU would handle RF and Low PHY, and CU would handle the other upper layers. Different split options introduce trade-offs among energy consumption, deployment cost, latency, and throughput, which offers the RAN flexibility to dynamically satisfy the need for different use case scenarios [48].

Although the DU-CU separation could bring the above benefits, it also introduces higher management complexity, optimization difficulty, deployment cost, and risk from a single point of failure [49]. Thus, a “monolithic” gNB containing one DU and one CU is still a feasible deployment.

Open RAN architecture [50, 51] is a shift towards virtualization, cloudification, multi-vendor interoperability, and data-driven management and orchestration, built on top of the 3GPP architecture. O-RAN ALLIANCE mainly drives this project. In Open RAN, software and hardware are decoupled, and software connects through open interfaces, which helps avoid operators being locked from a single infrastructure provider. DU and CU run in a cloud-native manner, which increases capacity, reduces operating expenses, and enables big-data applications. Importantly, near-RT (10ms to 1s) RIC is introduced in RAN with xApp onboarded, and non-RT (more than 1s) RIC is introduced with rApp onboarded. The controllers mainly monitor the network and provide data-driven guidance to gNBs. These controllers are functionally similar to NWDAF in the core network but are closer to the network edge and only responsible for RAN. With AI capabilities and history data analysis, these controllers can be useful for various use cases [52], such as improving handover performance for vehicles on the highway, interference detection and optimization, and detection and restricting signalling storms.

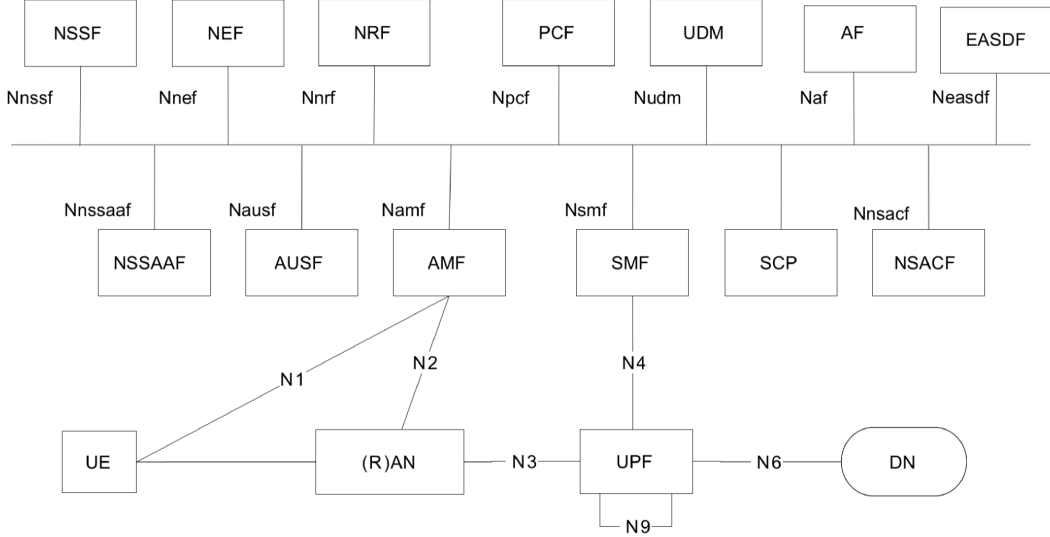


Figure 2.2: 5G service based architecture [4]

2.2.3 5G core network

The main changes in the 5G core network from 4G are adopting service-based architecture and NFV. From 4G to 5G, monolithic network functions, such as HSS, MME, and PCRF, are split into more discrete network functions; each provides micro-services with a higher granularity, which opens the door to network slicing [4]. In 4G, network functions were coupled with proprietary hardware. This poses a challenge for the network to scale flexibly because the deployment comes with the hardware. 5G network functions are virtualized to containers or VMs for a more flexible deployment and capacity scaling. Due to the virtualization, these NFs are managed by the orchestration and management layer. The 5G SBA is shown in Fig. 2.2, which contains some main 5G NFs. A complete NF list can be found in [53]. For the communication service (VoNR), 5G reuses the IMS architecture with the original session initiation protocol (SIP) and session description protocol (SDP). The IMS provides interfaces to 5G core NFs, 4G core NFs, and 3G core NFs to achieve interworking [54].

In the data plane, UPF is the only 5G core network function type that routes traffic between the RAN and the data network (Internet). It is responsible for traffic measurement reports and enforcing QoS policies made from the core network [4]. Benefiting from virtualization, UPF can be deployed closer to the RAN and connect MEC servers in the data network to lower the latency. Different UPFs can connect to each other through the

N9 interface to achieve traffic classification. N6 terminates at the PDU Session Anchor UPF. Figure 2.3 shows the data plane protocol stack.

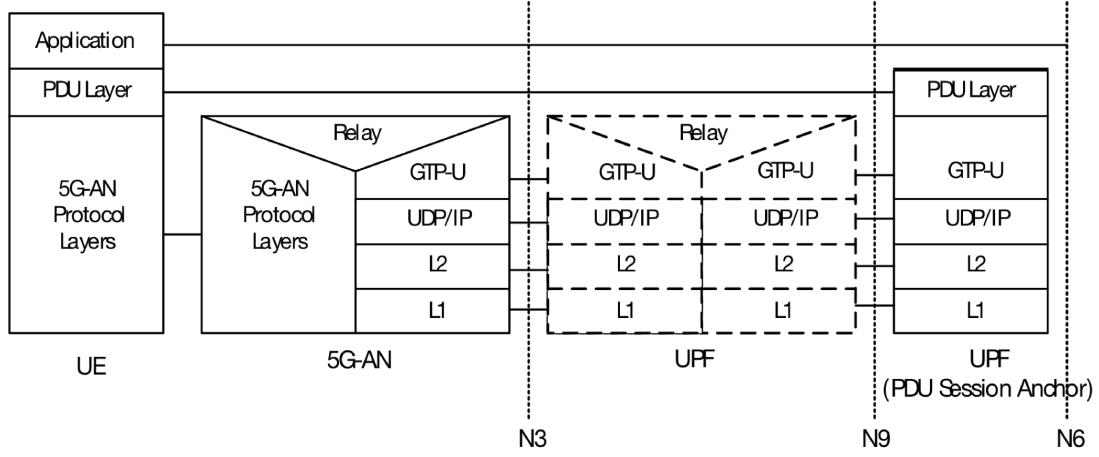


Figure 2.3: 5G core data plane protocol stack [4]

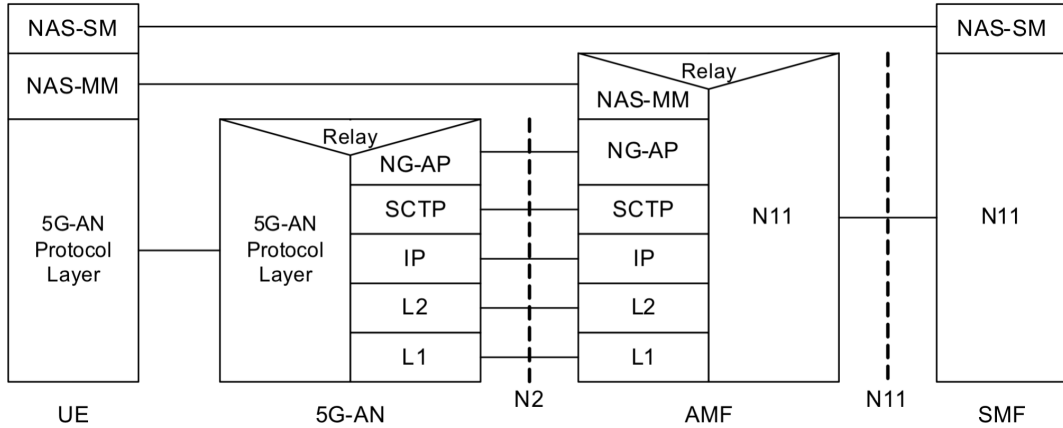


Figure 2.4: 5G core control plane protocol stack [4]

In the control plane, network functions communicate and collaborate to make control decisions. These decisions are sent to the RAN control plane by AMF through the N2 interface, to the core data plane by SMF through the N4 interface, and to the UE through the N1 interface by AMF with RAN as a relay. The UE communicates with AMF over an application layer protocol named NAS protocol [55] as shown in Figure 2.4. In several

cases, such as location service, SMS, and session management signalling, UE needs to communicate with other 5G NFs, excluding AMFs. Thus, the NAS protocol comprises a NAS-MM and a NAS-SM components [4]. The functions of some main NFs are as follows (more details can be found in [4]):

- **AMF:** Access and registration management. Mobility management. Location services management. Connection and reachability management.
- **SMF:** Session establishment and management. UE IP address allocation and management. Control aspects of QoS.
- **UDM:** Subscription management. SUCI de-concealment. Generation of 3GPP AKA Authentication Credentials. User identification handling.
- **NEF:** Exposure of capabilities and events to external third-party:
- **NRF:** Maintains the NF profile of NF instances. NF service discovery. Generate access tokens for NF services.
- **SCP:** Indirect NF communication. Message forwarding and routing between NFs. Load-balancing [21].
- **NWDAF:** Data collection from NFs, AFs, and orchestration management. Support AI modeling training and provisioning. Provisioning of data analysis to NFs.
- **SEPP:** Message filtering and policing on inter-PLMN control plane interfaces. Topology hiding.

5G control plane NF communication

5G NFs in the control plane communicate with each other using service-based interfaces. The service-based interfaces use HTTP/2 protocol with JSON as the application layer serialization protocol. For security protection at the transport layer, all 3GPP NFs shall support TLS [56]. Thus, digital certificates are required to achieve mutual authentication between NFs [57].

The end-to-end interaction between two NFs follows a consumer-producer paradigm; that is, the consumer NF consumes the service produced by the producer NF. Within this paradigm, two communication mechanisms adopted are as follows [4]:

- **Request-response:** Consumer NF calls the RESTful API provided by producer NF, and a one-time response from the producer to this request is expected within a certain period.
- **Subscribe-notify:** Consumer NF can subscribe to the event exposure service provided by the producer NF by calling the RESTful API. The producer NF will notify the consumer NF when the event happens or the state changes.

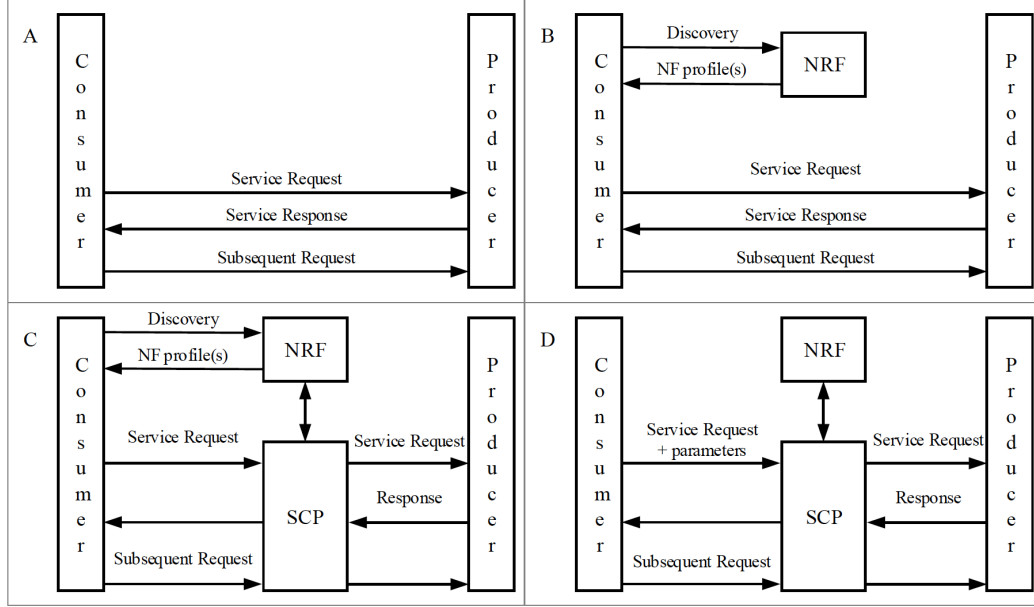


Figure 2.5: Communication models for NF/NF services interaction [4]

Before NFs can communicate, NF needs to register its NF profile, which contains NF information such as NF Instance ID, IP address, NF Type, and heartbeat timer, to the NRF (A complete list of the NF profile is in [53]). So, NRF maintains information of available NF instances and their supported services. When the consumer intends to ask for some producer service, it must find which producer NFs are available. This process is called NF discovery. The consumer NF will obtain a set of producers, and it can select one NF to communicate, which is called NF selection. After the discovery and selection, the producer needs to acknowledge that the consumer's requests are authorized by NRF. This is known as the authorization of NF service access, and 3GPP recommends using OAuth 2.0 framework [58]. Under this framework, NRF shall be the OAuth 2.0 authorization server, the consumer shall be the OAuth 2.0 client, and the producer shall be the OAuth 2.0 resource server. An access token will be given to the consumer by the NRF if the request is legit. The consumer will then use the token to request service from the producer.

The call flows of these processes vary slightly in different interaction modes. However, OAuth roles never change, and the NRF will always be responsible for generating the access token [58]. NRF is also always responsible for NF registration. Overall, the four interaction modes, shown in Fig. 2.5, for NF communications are defined by 3GPP as follows [4]:

- (A) **Direct communication without NRF interaction:** Consumers are configured with producers' NF profiles and directly communicate with a producer of their choice. No authorization is required because NRF is not present. This option can be used in a test environment but not in production.
- (B) **Direct communication with NRF interaction:** Consumers perform NF discovery by querying the NRF. Based on the result, the consumer performs the selection. The consumer sends the request to the selected producer. This option is a common model in specifications and articles.
- (C) **Indirect communication without delegated discovery:** Consumers perform discovery by querying the NRF and then perform NF selection. The consumer sends the request to the SCP containing the producer's address. The SCP routes the request and response between the consumer and the producer. In this option, SCP can perform load balancing and signalling policy enforcement.
- (D) **Indirect communication with delegated discovery:** SCP is responsible for NF discovery. Consumers add any necessary discovery and selection parameters, which are used to find a suitable producer, to the service request. The SCP uses the request address and the included parameters in the request message to route the request to a suitable producer instance. The SCP can perform discovery with an NRF and obtain a discovery result. This option first reduced some signalling. Also, SCP can optimize signalling further because it handles discovery and selection.

5G system procedures

5G system management is achieved by completing system procedures. Each system procedure is defined to accomplish a particular mission, such as registration, handover, system interworking, and session establishment. It involves control signal transactions between NFs, the RAN, and UEs. So, system procedures are also known as signalling protocol. Any system behavior, including failure handling, should correspond to a system procedure. Those procedures can be initiated both by UEs and the network side. A complete list of system procedures can be found in [59]. When describing the system procedure or signalling protocol, people generally abstract the communication model and assume end-to-end communication between network entities.

For better API management, each producer shall provide multiple services, and each service provides multiple service operations. Each service operation follows the 'Request-response' or 'Subscribe-notify' paradigm and has clear customer types [59]. Thus, within a system procedure, the control signal exchanged is actually the service operation calls. For example, SMF as a producer can provide three services named `Nsmf_PDUSession`,

`Nsmf_EventExposure`, and `Nsmf_NIDD`. Within the service `Nsmf_PDUSession`, SMF provides 13 service operations. One operation is `CreateSMContext`, which follows 'Request-response,' and the consumer is AMF. It can be used in many system procedures, including PDU session establishment, handover, and SMF context transfer. A complete NF services and operations list is available in [59]. A collection of the 3GPP-defined RESTful API for these operations of main NFs is available in [60].

2.2.4 Network Slicing

Network slicing is a paradigm where logical networks/partitions are created with appropriate isolation, resources, and optimized topology to serve a purpose, service category, or customer. Slices are end-to-end overlay networks with different isolation levels that share the same physical infrastructures [4]. This is a new business model, which is expected to be integral to unlocking the enterprise opportunity for the 5G because MNOs can provide network slice as a service (NSaaS) to the industry verticals (often referred to as 'tenants') to satisfy their specific requirements rather than a one-size-fits-all approach in 4G. To prepare a network slice, a service level agreement between the MNO and tenants must be signed following the network slice template. The template [61] defines slice attributes and slice requirements such as maximum PDU sessions, UE density, delay, throughput, energy efficiency, spectrum, scalability, and isolation level. Importantly, it is allowed in NSaaS to expose network management capabilities to the tenants [62]. Also, tenants may provide their customized network functions [61]. Each network slice is identified by S-NSSAI (Single-network slice selection assistance information), which is comprised of an 8-bit slice type (SST) and an optional 24-bit slice differentiator (SD). SST defines an expected slice behavior in terms of features and services, including eMBB, URLLC, mMTC, V2X, HMTTC, and HDLLC. SD complements the SST to differentiate amongst multiple slices of the same slice type [4]. During registration, the UE may send an NSSAI, which contains up to 8 S-NSSAIs. The network side will return S-NSSAIs that the UE is allowed to access. The PDU session will then be established within a certain slice instance.

Each slice instance of an end-to-end network slice is called a network slice instance (NSI). It is simply a deployment of a network slice. The network slice subnet instance (NSSI) represents a group of network functions (including their corresponding resources) that form part or complete constituents of an NSI. In general, NSSI could be created as RAN NSSI, core NSSI, and transport NSSI. Different NSIs may share the same NSSI. Also, the grouping of the network functions allows the management of each NSSI to be conducted independently of the NSI [63]. An NSI Identifier (NSI ID) uniquely identifies an NSI [4] from the network side. Note that one NSI obviously can be associated with multiple

S-NSSAIs. One S-NSSAI can also be associated with multiple NSIs because one NSI may satisfy the requirements of multiple S-NSSAI. The network may simultaneously serve one UE with one or more NSIs of different S-NSSAI. The AMF serving the UE logically belongs to each of the NSIs serving the UE [4]. However, one PDU session uniquely belongs to one NSI. In general, one UE can be served by up to 8 NSIs and can have up to 15 PDU sessions at the same time.

The lifecycle of an NSI includes commissioning, operation, and decommissioning phases [63]. The commissioning phase includes NSI configuration creation or modification. The operation phase includes activation to support communication services, run-time resource and topology planning, monitoring and modification, and deactivation to stop supporting services. The decommissioning phase is to remove the NSI. The lifecycle of an NSI is managed by the OAM layer. In addition to MANO entities (NFVO, VNFM, VIM), 3GPP defined customized management entities in the OAM layer to manage network slices. For example, NSMF manages NSIs, and NSSMF manages NSSIs. CSMF converts the communication service-related requirements to the network slice-related requirements and communicates with NSMF [64, 65].

In the control plane, 3GPP introduced the following NFs to support network slicing-related system procedures [4]:

- **NSSF:** Selecting the set of NSIs serving the UE. Provisioning of slice information to AMF. Selecting the AMF that can serve the UE.
- **NSSAAF:** Support for network slice-specific authentication and authorization [59] with a AAA server, which is hosted by HPLMN or DN third-party.
- **NSACF:** Monitoring and controlling the number of registered UEs and the number of PDU sessions. Provisioning/Notifying network slice status to AFs.

2.3 5G Non-Terrestrial Networks

2.3.1 Overview

Although cellular networks already covers 95% of the population, 90% of which is covered by 4G/5G networks [66], there is still around 85% of the earth that remains without coverage [67], including rural areas, mountains, islands, and sea. Providing coverage to those areas requires high capital expenditure and operating expenses with expected low revenue. Thus, the industry is considering building 5G non-terrestrial networks with satellites and integrating NTN with TN to provide 5G global coverage. Integrating 5G NTN brings the following main use cases and benefits [68, 69, 70]:

- **Ubiquitous and continuous connectivity:** By providing global coverage from space, NTN can directly provide high throughput service to remote areas without deploying base stations and fibers. Without a doubt, this could improve the lives of the people who live or work in remote areas. This will also benefit businesses such as environmental monitoring, ocean freight, fishery, oil exploration, farming, and airplanes. It can also enable outdoor adventure and emergency services globally.
- **Enhanced broadband and temporary use:** The UE is allowed to connect to and switch between TN and NTN; this will increase the total bandwidth and connection capacity of an area, especially when the capacity in rural areas is not enough to support the service need or traffic surge (e.g., traveling). When the TN experiences outages from disasters, such as floods or earthquakes, the NTN can be used temporarily for services and disaster recovery.
- **Optimal routing:** The NTN can also be used as a backhaul network to optimize the routing to lower the latency.

Although NTN is not limited to LEO, MEO, GEO satellites, and UAV, the industry has spent more effort considering LEO satellites mainly because LEO satellites can provide low latency services and high stability. Starting in 1997, Motorola launched the first satellite communication constellation containing 66 LEO satellites - Iridium. However, due to the technology limitations and limited use cases, the Iridium project turned out to be a complete failure in 2000. Today, with the advancement in technology and new application scenarios, LEO satellite communication has drawn great market attention, and the success of Starlink has proved its feasibility. However, more satellites in LEO orbit means a higher possibility of collision and interference and more difficulty in managing, especially satellites from different operators from different countries. Thus, to satisfy a certain network performance constraint, the number of LEO satellites has a limit [71, 72]. Also, the LEO satellites sweep the Earth's surface, including all the countries, which can be a powerful tool for military activities such as tracking, monitoring, and accurate remote control of military weapons. Such strategic advantage is also displayed in the Russo-Ukrainian war [73]. Therefore, countries and operators attach great importance to building or benefiting from LEO constellations. As LEO satellites are envisioned to complement 5G, mobile operators and satellite operators are working closely to build the integration [74]. From Global Market Insights, the market value of 5G NTN was 4.2 billion USD and is predicted to be 79.8 billion USD by 2032. Although the LEO satellites provide communication services similar to the cellular network, their system design and technology are generally proprietary [70] and cannot directly apply to the 5G system. Therefore, the 5G NTN standard established by 3GPP [75, 76] is extremely important and valuable for the research in consistently integrating satellites into the 5G ecosystem.

2.3.2 Deployment Options

Satellites are deployed on three kinds of orbits based on the altitudes. GEO orbit is at 35,786 km, and the satellite's position is fixed regarding elevation/azimuth with respect to a given earth point. The MEO orbit altitude is from 7000 – 25000 km. The LEO orbit altitude in 5G NTN is designed to be 300 - 1500 km [75], and the satellite moves in a circular fashion around the Earth with a high speed of around 7.1 - 7.9 km/s. Different satellite shells can be designed based on inclination and altitude. Each satellite shell could contain many circular orbital planes, which contain the satellites [77, 78]. For 5G NTN, the allocated frequency band is L/S-band, targeting the handheld devices (Direct-to-cell), and Ka-band, targeting the devices with higher-gain antennas [79]. LEO satellites are also designed with earth-moving cells and earth-fixed cells through steering beams. Both options are feasible but will cause different impacts on the cell/tracking area design and information broadcasting management (SSB beam) [75, 80].

The satellites can carry two kinds of payload, which are transparent payload and regenerative payload. The transparent payload enables RF filtering, frequency conversion, and amplifications, which acts as an analog RF repeater. The transparent satellite and NTN gateway can act as a remote radio unit in the NR-Uu interface between UE and gNB on the ground, with no modification to the NG-RAN architecture [75]. However, the high delay between RF and PHY would result in such deployment not meeting the 5G requirement [81]. The regenerative payload, in addition, enables encoding, decoding, modulation, demodulation, switch, and/or routing, which is effectively equivalent to having all or part of gNB functions on board the satellite [69]. This could also enable communication between satellites through the inter-satellite link (ISL). The regenerative satellite can support the following architectures in the RAN [75]:

- **Full gNB onboard:** The satellite acts as a full gNB node to provide NR-Uu interface directly to the UE. The feeder link will be used in the NG interface connecting the 5G core network. Xn-interface between satellites can be built through ISL or possibly through the feeder link if ISL is not present [75]. Xn-interface between satellite and terrestrial gNB should go through the feeder link.
- **gNB-DU onboard:** The satellite acts as a gNB-DU component and provides the NR-Uu interface directly to the UE. The CU is assumed to be on the ground, and the feeder link will be used in the F1 interface. Due to the DU-CU separation, Open RAN features such as xApp and RIC can be implemented. Also, RAN functional split options can be considered to optimize the performance [82].

Importantly, both options would cause an impact on mobility (TR 38.821-8.3), network identity (TR 38.821-8.5), and feeder link switch (TR 38.821-8.7) management regarding

network performance, especially latency. The protocol design, intelligent control, and optimization across all layers in the protocol stack should be carefully adjusted considering the use case scenario, high-speed satellite movement, complex satellite positions, and satellite deployment options. The UE can also be dual-connected to two satellites (either transparent or regenerative) or one satellite and one ground base station.

Satellites can also carry certain core network functions, such as UPF, to enable edge computing. Traffic from UE to UE can also be locally routed without going to the ground [83]. With either payload, the satellites can be part of the backhaul network [83, 68]. The ISL (for regenerative satellites) can also be utilized to realize routing between satellites. However, the NTN component and a terrestrial component may have inconsistencies, such as QoS information [84]. Also, the satellite backhaul category needs to be reported to SMF because satellite networks are dynamic, and the delay requires real-time monitoring and control.

2.4 Signalling Storm in 5G

2.4.1 Past cases and causes

Signalling storm refers to a situation when control signal intensity exceeds the cellular network processing capacity, which lowers the network performance and stability, resulting in service disruption. During the disruption, the retries will further flood the network and cause a huge snowball effect. It has been a persistent cellular network threat since 3G and has caused several severe, large-scale outages, costing operators millions of dollars and user hours in losses. Although operators usually stay silent about outage causes, the following cases have demonstrated the power of signalling storms:

2010, Telecom New Zealand, 72 hours

In 2010, Telecom New Zealand experienced four major outages within three months. The second outage on Jan 27th, which the CEO said was caused by thousands of users suddenly re-registering after a separate network routing fault took down some cell sites. Alcatel Lucent, the company that helped build the network, described this outage as a “strange signalling storm, an unprecedented situation.” It was to take three days before service was fully restored and cost Telecom New Zealand 5 million dollars in compensation to subscribers [20].

2011, Telenor, 18 hours

On Jun 10th, 2011, Telenor, Norway’s largest telecom company, experienced a significant outage. The problem started with a boot-up and software update regarding Telenor’s internet services. Although the reason was unclear, the boot-up triggered unusually high internal activity between servers in its telephone network, described by Telenor as a signal storm. This outage left 3 million users without coverage for up to 18 hours and resulted in 18 million dollars loss to Telenor. The minister for transport and communications described the situation as “totally unacceptable,” and Telenor’s CEO described this as “a telephone operator’s worst nightmare [18].”

2012, NTT Docomo, 4.5 hours

From Jun 2011 to Jan 2012, NTT Docomo, Japan’s largest mobile operator, experienced a series of outages. On Jan 25th, 2012, the outage left 2.5 million subscribers unable to make voice or data calls for about four and a half hours. The severity of the frequent outages led the Ministry of Internal Affairs and Communications to order the network operators to conduct a comprehensive investigation. In the operator’s reports [19, 85], they revealed that their core network failed to complete the processing of control signals that increased due to the rapid proliferation of applications (VoIP, etc.), which caused congestion in the core network elements including authentication server, user management server, mail information server, and packet switch system, resulting in network malfunction. After the outages, they planned a 160 billion JPY budget to increase their network capacity in three years. Also, NTT Docomo requested Google and Android app developers to transmit control signals, such as pinging servers from Android devices, less frequently [86].

2020, T-Mobile, 12 hours

On Jun 15th, 2020, T-Mobile in the US experienced a 12-hour outage, disrupting calling and texting services nationwide, including 911 service. The Federal Communications Commission (FCC) report [14] revealed that the outage started with a link failure that caused massive call signals to congest a misconfigured router. Because the signalling traffic could not pass, all LTE users in the Atlanta market lost connectivity to LTE data, VoLTE, and the 3G and 2G circuit switch network, which disrupted voice, text, and data services in the Southeast.

To initiate a VoLTE phone call in the T-Mobile network, either through a base station or

WiFi, the signal will be transmitted to the EPC and then to the IP Multimedia Subsystem (IMS) where the device is registered and eventually be routed towards the destination.

During the isolation, the devices tried to re-register with the IMS using WiFi, which was designed to be feasible as an alternative method. The IMS was also well designed to route the registration to the IMS where the device was last registered to enhance efficiency. However, a software error that did not affect the system before the outage prevented the registration from being completed without routing to the correct IMS. The IMS repeatedly rerouted after failing to route the registration signals, and the devices repeatedly attempted to register, which caused a registration signalling storm, congesting the IMS. After the link failure was resolved in only 12 minutes, more devices also attempted to register through base stations. Unfortunately, T-Mobile thought the previously failed link was continuing to cause disruptions and manually shut down the link, which they had already repaired. This mistake again isolated the Atlanta market, and the devices continued contributing to congesting the system through WiFi registration. After one hour, T-Mobile resolved the same link problem they had resolved before. However, because the IMS in other regions cannot reach the IMS in Atlanta, external incoming signals were routed to all regions in the US, and the software error prevented the registration from being complete.

The congestion soon spread to other IMS and created a nationwide registration storm. The vast majority of devices soon fell back to 3G and 2G circuit-switched networks to make and receive calls while the device continued its registration attempts to the VoLTE network. The large number of devices attempting to fall back caused congestion and overwhelmed the computing resources in the 2G/3G network. When 2G/3G failed, the 911 service that relied on the same resources in 2G/3G also failed. FCC estimated that the outage caused at least 23,621 failed calls to 911. Eventually, T-Mobile agreed to pay 19.5 million dollars in a settlement with the FCC.

2021, NTT Docomo, 29 hours

On Oct 14th, 2021, starting at about 5 pm, NTT Docomo experienced a large-scale outage. Their report revealed [16] that the incident started from updating their core network location management server. However, the remote roaming IoT devices had software incompatible issues with the new server that caused connection failure. After Docomo rolled back to the old server, the IoT devices triggered massive registration requests to update the location, which caused congestion in the location management server and spread to the whole core network. The 4G/5G network was restored after 12 hours, and the 3G network was restored after 29 hours. This outage affected around 2 million users' voice and data services.

2022, KDDI, 86 hours

On Jul 2nd, 2022, starting at about 1:35 am, KDDI, Japan’s second-largest mobile operator, experienced one of the most significant outages in history. From KDDI’s report [15], the incident started when routers in the core network were replaced during maintenance. However, the new router was incorrectly configured, causing only a one-way signal to be transmitted from devices to the network. Because the devices could not receive a response, the device soon started to re-transmit the location registration requests. The large volume of these control signals caused congestion at VoLTE nodes and subscriber DB. Although KDDI immediately rolled back the old router setting after only 15 minutes, things were already out of control. The VoLTE nodes at other sites soon became congested because distributed processing was performed via the nationwide transport network. Due to the congestion, data inconsistencies appeared at the subscriber DB, triggering a cascading effect that continued the service interruption. Although KDDI previously implemented a mechanism to reduce the load on subscriber DB and correct subscriber DB data inconsistencies in such a situation, the congestion continued where the load was not sufficiently reduced.

It took 86 hours before KDDI announced that the network was fully restored. This outage affected around 22.78 million voice services and more than 7.65 million users’ data services. Over three days, the Meteorological Agency was unable to send weather data at hundreds of stations, dozens of cash machines were disabled in central Japan, and Tokyo and other municipalities had trouble reaching COVID-19 patients subject to health monitoring at home. KDDI eventually estimated a total compensation of 7.3 billion JPY, with only 200 JPY for each user. “It is extremely regrettable that mobile phone service, which is a crucial infrastructure for social and economic activity, has been unusable for such a long time,” said Deputy Chief Cabinet Secretary Seiji Kihara [17].

Causes

From the network side, 5G introduces more network functions and adopts a much more decentralized control plane architecture. Such a design will obviously introduce more signalling than 4G. In 2021, Huawei reported that some Chinese operators are experiencing as much as 100 times or even more control signalling in 5G SBA compared to 4G Diameter signalling [21]. Also, from some recent surveys [27, 28], many operators believe signalling storms will be more common in 5G and it will be a significant challenge.

From the UE side, this concern is exacerbated as the number of connected devices keeps increasing with a goal of supporting 1 million UEs/km² density. Besides the normal

subscribers, the cellular IoT device is a big concern. It was reported that 3 billion cellular IoT devices have already been connected by the end of 2023, and Ericsson predicted there will be around 6 billion IoT devices connected by 2029, and many of them are capable of global roaming [6]. These IoT devices are manufactured from various sources of the supply chain, which are out of the operator’s control. The software and hardware security has hence become a big concern [22]. On Oct 21st, 2016, Mirai (“future” in Japanese) launched a record-breaking DDoS attack on DNS provider Dyn and caused major Internet platforms to be unavailable to large swathes of users in Europe and North America. First, Mirai compromised around 0.6 million IoT devices by exploiting their zero-day vulnerability, such as default usernames and passwords. Then, the attack to overload the DNS server was carried out through DNS flooding, sending massive DNS lookup requests from tens of millions of IP addresses [87]. Although the attack does not require high hacking skills, it could easily cause network congestion and service interruption. In 2014, Telenor experienced network degradations caused by a limited number of connected devices. After the investigation, they found that as few as 500 aggressive devices could create a network signalling storm, which can cause congestion [88]. They proposed that operators separate control signals of roaming IoT devices from customers to mitigate social impact because it could take up to four hours to restore service once the signalling storm happened. This 2014 recommendation was adopted by NTT Docomo in 2021 as one of the signalling storm prevention solutions after the outage caused exactly by roaming IoT devices [16].

Due to huge losses in the past, industry and government agencies consider signalling storm, either due to legitimate traffic or malicious attacks, as a major 5G threat [22, 23, 24, 21, 25, 26]. Clearly, signalling storm can happen from various causes, but the causes can be roughly categorized into the following four classes:

- **The human error and system failure:** Human error is involved in many incidents in Cybersecurity and is difficult to avoid. In past cases of signalling storms, we noticed some outages that started from the misconfiguration of network elements. As the 5G system becomes more complex, virtualized, and distributed, the burden of the OAM to correctly configure the network becomes heavier, and more prone to human error.
- **The network design and implementation:** Clearly, failing to implement the network elements may cause unexpected congestion. When the signalling storm appears, correcting the implementation mistake in time is difficult. Although 5G network elements are decoupled from the hardware, the compatibility among products from different vendors poses a risk of system failure. Even when the implementations are correct, network architecture and system procedures design may fail to handle the congestion under certain scenarios.
- **The dynamic user activities:** As the number of new services and use case scenar-

ios such as V2X, AR/VR, and NTN increases, users' behavior becomes more complex. Such complexity will undoubtedly introduce challenges to network resource management for avoiding signal surges based on different deployments. Although network slice isolation is designed to ease this complexity, one UE can simultaneously be served by multiple slices, which means the user activity can affect multiple slices. Resource management considering user activity will also be more difficult as the number of slices scales.

- **The malicious activities:** By exploiting device vulnerabilities, malicious parties could control massive devices to send control signals to perform a DDoS attack in the control plane. Through the years, DDoS, as an almost skill-free attack, has been more frequent [89]. Notably, launching a large-scale DDoS has become a business (DDoS-for-hire), and the cost is getting cheaper, which lowers the barrier of performing DDoS [90]. For an always-on service like 5G, a short period of congestion could have a huge knock-on effect, as demonstrated by KDDI's outage in 2022.

2.4.2 5G signalling storm scenarios

We reviewed industry standards and academic research and listed the following scenarios that could lead to a signalling storm in 5G (The list is not exhaustive, and we assume that 5G network elements are trusted).

UE access related

- A.1 In the primary authentication with SUCI, the network can only find the identity of the UE after executing multiple round trips of signals and performing ECIES decryption. If the attacker launches/replays massive registrations (potentially with fake credentials), it will cause the processing power of the UDM to decrease, and the response to the request of the legitimate UE to be slow [91, 92]. Obviously, the attack could also be performed simply through registration/de-registration with or without SUCI [93].
- A.2 Malicious UEs could repeatedly perform initial attach procedure to a gNB [93]. Identifiers such as C-RNTI are assigned after this procedure, so no identifier is needed to perform this procedure. Initial attachment requires contention-based random access for uplink synchronization. This would result in normal UEs failing to attach because of preamble collision. We expect such an attack will be more powerful in the NTN setting when gNB moves and initial attach is more common. Also, the rogue base

station may connect legitimate UEs. If the rogue station simultaneously releases the UEs, it could cause UEs to simultaneously attach the real gNB [94].

- A.3** In some scenarios, such as sports games, concerts, or several-day-long events like industry fairs/conferences. Local hosting networks may be provided as temporary local services to PLMN users. When the event is over, the local hosting network may be switched off. This may lead to a signalling peak in the HPLMN and result in longer time scales for users to re-register to/re-select their home PLMN [95].

UE mobility related

- B.1** In some scenarios, such as high-speed rail, car platoon, and transport systems in densely populated areas, many UEs may perform handover simultaneously, which could downgrade the performance of gNB. Such handover includes both intra-gNB handover and inter-gNB handover. Also, the handover could trigger mobility registration update and re-authentication if the UEs left the tracking area, creating a suddenly increased signal load to the core network [96]. The problem is more severe in 5G NTN because gNB is moving, and the speed is much higher than a high-speed rail. The gNB is potentially required to frequently handle the handover procedures for a large number of devices even when they are not moving [75, 81]. This could pose a challenge to both intra-satellite handover and inter-satellite handover.
- B.2** The broadcasting information SIB from gNB is not integrity protected. In this case, the attacker could modify the tracking area code in the broadcasting information to trigger massive UEs to perform mobility registration updates towards the HPLMN [97].
- B.3** When a UE requires a guaranteed bit rate (GBR) service and the QoS level is not maintained, it has to be restored as soon as possible. When the UE needs handover, the source gNB must release the UE even if the target gNB cannot handle the requirement. In this case, the target gNB cannot establish the QoS flow because of the admission control. Also, the gNB will not notify the core when the QoS flow can be established. The core needs to repeatedly attempt to re-establish the GBR service. These re-attempts involve a considerable number of signalling messages and are sent without any awareness of gNB congestion or potential link quality, which could result in performance degradation in gNB [98]. This scenario is more severe in certain scenarios, such as handing over many unmanned vehicles in a platoon.

UE state change related

- C.1** After AS security activation, the malicious UEs could send RRC signals such as re-establishment, release, and resume to switch the RRC states between CONNECTED, IDLE, and INACTIVE. This will consume the gNB resource and energy [99] and lead to a DoS [93, 100, 101]. Also, the procedures involve signalling among AMF, SMF, and UPF, which will threaten the core network.
- C.2** After PDU session establishment (up to 15 sessions for one UE), the UEs are allowed to modify the PDU session parameters, such as QoS requirement. This procedure involves a considerable amount of signalling among gNB, AMF, UDM, SMF, UPF, and PCF. Many simultaneous PDU session modifications would lead to a signalling storm. An example scenario for non-malicious behavior is a large number of people watching a live sports event at a stadium and trying to catch the action replay videos, videos of player profiles, and use of Augmented Reality (AR) glasses or headsets to get augmented information at the real-time related to the event and players [96]. Also, from the network side, modification of default session rules, such as session AMBR and default QoS, requires communicating with UEs. A simultaneous change of the session AMBR and default QoS for many UE(s) may introduce a signalling storm in the core [102]. In addition, UEs could also switch between non-3GPP access (e.g., Wi-Fi) and 3GPP access. In such a case, the network must handover the PDU session. This procedure [59] involves PDU establishment and PDU session release with many signalling exchanges among N3IWF, AMF, SMF, UPF, and RAN. If UEs frequently perform this procedure either from malicious behavior or mobility requirements, it may lead to a signalling storm [103].

UE application-level communication related

- D.1** UE to UE communication also may cause a signalling spike. In some scenarios like Christmas or disaster, there are expected to be massive voice calls and SMS in a short period. These voice calls and SMS setup involve many signalling exchanges among AMF, UDM, SMSF, PCF, UPF, SMF [59], IMS, and RAN paging, which can lead to system overload [103]. Also, in 5G LAN, identifier update or identity discovery may cause massive paging signalling and link establishment on the gNB/UE-UE relay [104, 105]. Moreover, the botnet could be used to send simultaneous SMS messages targeting a specific cell to potentially congest the paging channel [106].
- D.2** It is worth noting that application-level heartbeat messages from a large number of UEs in a region could cause frequent IDLE-CONNECTED state transitions in the RAN [96]. Also, it triggers PDU session establishment and release. These messages

are small in size but generate many control signals. In 2013, China Mobile claimed that the heartbeat messages from WeChat occupied 60% of their signalling resources but only with 10% data in the 3G network, and they intended to charge extra fees for the signalling. Although the charging plan failed, the controversy has sparked a large debate and analysis of signalling storm [107]. Furthermore, Cloud-based applications worsen the problem since they must keep their local data constantly in sync with the server, resulting in frequent additional signalling in the network [108, 109].

Network architecture and design related

- E.1** In 5G NTN, new challenges are brought due to the dynamic neighbor relationship among satellite gNBs and terrestrial gNBs. In terrestrial gNBs, setting up the Xn interface is not frequent because the number of neighbors is small, and they are fixed on the ground. So, gNBs have a relatively fixed neighbor relationship (ANR). In the LEO satellite network, the neighbor satellites change, so the ANR will frequently change and require a more frequent Xn-interface setup. Also, dual connectivity and handover between satellite gNB and terrestrial gNB are required. In this case, setting up the Xn interface becomes more challenging because the satellite's coverage is huge. When the satellite appears at the horizon, Xn-interface setups may be triggered for many terrestrial gNBs. The setup involves rounds of communications, including certificate verification and key exchange. The opposite is when a satellite moves below the horizon, all its Xn interfaces to terrestrial gNBs will become unavailable, and this may trigger subsequent actions at application protocol and/or SCTP level in the relevant terrestrial gNBs. Considering there would be more satellites to provide global coverage (Mega constellation), the signalling surge due to ANR change will be more severe [75].
- E.2** NRF is a centralized NF that would be frequently involved in many NF communications. During registration, discovery, and notification processes, full NF profiles of huge sizes have to be submitted or retrieved. In the large-scale network, this information, which contains millions of UE ranges, leads to huge-size (typically several million bytes) frame overhead and long processing time. Also, even a tiny change in a subscribed NF can trigger large-scale notifications. To distribute the full NF profile to all related subscribers simultaneously, the NRF can easily trigger a push storm and become a performance bottleneck [110].
- E.3** Some NFs support EventExposure service. This allows any authorized NF to subscribe for event notifications. Depending on the type of event, some events may occur very frequently for a large number of UEs (e.g., location change, UE presence in an area of interest report). This can lead to a lot of event notifications on the SBI,

leading to signalling overload [96].

- E.4** Certificate update is essential for the always-on cellular network. Failing to update them can cause serious outages, which has been demonstrated in O2 and Softbank's outage in 2018 [111]. Due to the broader attack surface brought by virtualization and third-party cloud, the certificate's validity cannot be long. In certain scenarios, such as approaching expiration date, a common CA certificate revocation, or compromise of a used crypto algorithm, the system may trigger certificate updates for many related entities at the same time [112]. This can cause congestion and temporary service disruption.

System failure related

System restarts, or failure can also cause a signalling storm. HPLMN RAN may fail because of disasters. In this case, massive users may access the backup PLMN (with roaming agreement) and potentially cause a signalling storm to the backup PLMN [113]. When HPLMN is restored, the backup PLMN needs to release the connections. In this case, simultaneously releasing all UEs will cause a signalling surge, and this may again cause outages in HPLMN [113] (This case is similar to A.3. However, the scale is expected to be larger). One example of an NF restart is when UDR restarts, UDR might lose UE's authentication status in the serving network. In such a case, the UDM cannot retrieve the authentication status and may let AMF re-authenticate UEs. This can potentially trigger many paging signals and UEs re-register signals, leading to a signalling storm [110]. More scenarios can be found in [114]. In general, a restart of an NF can result in a large number of signalling on the SBI within a short interval of time [96].

2.4.3 Mitigation of signalling storm

Methods for mitigating signalling storms can be classified into prevention methods before the storm and recovery methods after the appearance of the storm. In this section, we will discuss the main approaches to achieve them.

Monitoring, prediction and detection

With various application scenarios and complex network architecture, automation, zero-touch, and data-driven network resource management is essential for 5G and beyond cellular networks, although they are not always reliable. Such data-driven approaches enables

prediction of the future traffic and detection of anomaly UE/network behaviour, which can mitigate the signalling storm.

With traffic prediction, operators could prepare resources in advance. For example, based on the trend on social media or news, Huawei predicted the number of people who would gather in one certain area to chase popular events. In this way, they can prepare the resources in advance [115]. This also applies to some special festivals such as Christmas or New Year. Also, from historical data, operators could train forecasting models or digital twins and use them to adjust resources and expenses in the short term or assist in real-time decision-making. This will especially help to prevent the signalling storm related to UE mobility behaviours [116]. Some applicable scenarios include improving handover quality and reducing unnecessarily repeated signalling in dense urban traffic with small base stations [117], potentially removing the uplink measurement report signalling, and decreasing the random access collision rate in LEO satellite handover [118]. Also, in [119], authors used time series to forecast the signalling load and adjust the signalling storm alarm threshold to avoid false alarms.

The detection model can be trained to learn the normal behavior of the UE, so during persistent DDoS attacks, the network is capable of denying the request from malicious UEs. Such an approach is highly recommended in C1, C2, D1, and D2 because the UE is generally assumed to be identified by the network. For example, in [120], several classic ML models are trained using success count and timestamp to detect malicious UEs performing re-registration and session establishment. In [100], the authors construct a Markov model to represent the RRC state transition of the UE and a queuing model for the RAN to represent the delay. Using the two models, they showed the impact of the RRC signalling storm. The models were also re-used in [121] and [122]; the authors adjust a counter value to detect malicious UE who exploit session state transition and heartbeat message. The threshold needs to be carefully selected because if the threshold is small, then the false positive rate will be high, and if the threshold is large, the snowball effect may emerge before denying malicious UEs. In addition, the authors considered bandwidth an important factor because these signalling attacks usually do not consume much bandwidth resources, which can be used as a threshold. We expect their modeling could also provide insights to detect malicious UEs in other attack scenarios, but the parameter adjustment is a challenge when applied in the real world.

Unfortunately, the data-driven approach generally requires that the UE activity must be linked to a persistent UE identifier. Otherwise, the model would find difficult to classify UEs. In the RAN, identifiers like C-RNTI frequently change, and some UE identifiers like SUPI and IMEI can only be visible in the core [52]. So, scenario A.2 requires additional protection mechanisms (Note that A.1 also requires additional mechanisms because iden-

tifiers can only be revealed after the attack). For example, in [123], a detection mechanism has been proposed to identify malicious IoT devices that perform repeated initial attachments to gNB. Due to the lack of C-RNTI, the authors proposed using a timing advance to identify the static IoT (A.2). However, it is expected that the timing advance may not be sufficient because the IoT can manipulate the timing advance. In [52], O-RAN ALLIANCE proposed to use signal quality and timing advance to represent a fingerprint of the UE. Still, such an approach under many scenarios can be highly inaccurate considering the mobility, environment, and malicious modification of UE parameters.

For the model deployment, O-RAN ALLIANCE proposed to deploy the model in near-RT RIC xApps for faster response. As mentioned, some identifiers can only be visible in the core. Thus, they also proposed to deploy the model in the non-RT RIC rApps because this allows the model to obtain information from the core [52]. In the core network, 3GPP allows NWDAF to detect abnormal network function load, signalling storm, and malicious UEs [124, 125].

Protocol, congestion control and network design

Communication protocols can be redesigned to achieve higher resource efficiency under certain scenarios and assumptions. Importantly, when redesigning the protocol, the designer must consider the transition of security-related information, such as identifiers and session keys. Without careful consideration, the new protocol may open more severe vulnerabilities, such as impersonation. Also, the designed protocol should support large-scale use cases because of the nature of cellular networks. Without these considerations, the new protocol may be difficult to leverage in the real world.

Over the years, researchers have focused on UE access-related and mobility-related scenarios and proposed new communication protocols. For example, using blockchain to achieve decentralized registration to avoid DDoS [126, 127]. However, the scalability issue is unavoidable if millions of devices must rely on blockchain to register the network. In [128], a decentralized blockchain-based certificate management frame is proposed to ease the PKI automation complexity in 5G. The proposed framework could reduce signalling during certificate updates. For the initial attachment, 3GPP proposed merging the 4-step RACH procedure into a 2-step one to reduce latency and signalling overhead [129]. In [130], authors proposed to use a relay in high-speed rail to avoid handover signalling congestion. In [131], group mobility registration update after handover has been proposed for the platoon. In [132], group pre-authentication has been proposed to avoid massive mobility registration updates in high-speed rail. In [133], an enhancement of NF resource authorization has been proposed to reduce the load of NRF by utilizing the side-car proxies

to remove using an access token. 3GPP also proposed using SCP (NF communication mode C, D) to assist signalling load balancing among NFs [21].

Different congestion control mechanisms in the protocol can also be implemented according to the use cases and scenarios. For example, a back-off timer can be optionally implemented on the UE side or sent to UE as a network instruction. This will help to avoid congestion in registration, random access, and session establishment [59]. Access class barring or extended access barring can also be optionally implemented to limit the simultaneous random access. PCF is also useful in avoiding congestion by delivering policy to the NFs. For example, using PCF can avoid simultaneous AMBR/default QoS change [102]. Also, it can decide the UE releasing rate when UEs are served by VPLMN during a disaster. It is also possible to control the load in the network element that is closer to the edge, such as gNB and AMF, to avoid signalling congestion in the center of the core network [134]. For these control mechanisms, parameter tuning in different scenarios and service requirements is a challenge because there is usually a trade-off between congestion rate, resource utilization, and network performance. For example, in [101], the authors decide to randomize the inactive timer to limit the attacker to performing RRC state transition pulse wave DDoS. In [135], a UE decided backoff timer in random access is implemented to achieve a better balance between network throughput and collision rate.

Resource scaling and redundancy backup

In general, most outages are caused by capacity issues. After the outage, the operator normally will increase the network resources with a consideration of their operating expense and profit. Scaling resources in advance is the basic and mandatory way to make the network capacity catch up with the rapidly increasing number of subscribers to prevent congestion. However, during the signalling storm, the number of signals will increase exponentially, which is expected to exceed the original capacity. At that point, increasing the capacity makes it difficult to mitigate the storm in a short time. For some cloud-native functions, the elastic scaling feature of cloud computing will help mitigate the storm by providing more resources. However, an attempt/re-attempt from the UE or NFs could be triggered within a period in an order of milliseconds. The snowball effect can be faster than the cloud scaling [108]. This issue is exacerbated when the scaling requires costly data replication or redistribution. Besides, if the signalling storm is caused by a malicious party (e.g., DDoS), then simply relying on the cloud will make it difficult to mitigate the storm because the attack can be persistent. This is the reason why people implement DDoS protection even when their servers are deployed in the cloud. Also, operators must be aware that the cloud is not free and the cloud resource is not infinite [136].

Preparing redundancy backup is important to prevent system failure-related signalling storms. In practice, any component in a network has a possibility of failing. For a complex product like a cellular network, system failure is deemed to happen. When it happens, the redundancy backup will be useful to avoid service disruption. For example, operators could prepare redundant UDM with duplicated subscriber data when the serving UDM is experiencing system failure. After one UDM fails, the operator can also feed the subscriber data to the 4G HSS and switch the users from 5G to 4G to try to avoid a signalling storm [137].

Chapter 3

Mitigating Signalling Storm in 5G with Blockchain-assisted 5GAKA

3.1 Introduction

Control signals are essential to manage network elements, allocate resources, and enable operations like handover, authentication, and billing. They precede any adaptation to changing conditions, performance optimization, and improvement to users' experience. Control traffic volume is typically measured in the number (and size) of transactions or messages exchanged between network elements. In 5G cellular networks, control signals are transmitted in the control plane (CP), while the data plane (DP) transmits actual user data. When the control plane signals' intensity exceeds capacity, it becomes a signaling storm [109].

The signalling storms has been a common problem in 3GPP mobile broadband network technologies, dating back to the 3G mobile networks [138, 139]. It has attracted a lot of attention from the research community. Because control procedures between mobile devices and base stations typically involve massive volumes of radio signals, such as Radio Resource Control (RRC), Random Access Channel (RACH), and Paging, research on this particular challenge has mainly focused on the Radio Access Network (RAN) [100, 140]. However, experience shows that signaling storms are also a threat to the Core Network (CN). Telenor, the largest mobile network operator in Norway, reported an outage for calls and SMS due to the centralized mobile broadband server being overloaded by a signaling storm triggered by a simple reconfiguration of one voice server [18] in 2011. The outage impacted 3 million users for up to 18 hours, costing Telenor an 18 million USD loss. A straightforward way to

prevent such outages is to increase the network capacity. For example, DoCoMo, Japan’s largest telecom operator, spent 160 billion JPY in 2012 to increase the network capacity after an outage affecting 2.5 million users for 4 hours due to a signaling overhead [19]. However, in 2021, DoCoMo experienced another outage affecting 2 million users for 12 hours [16, 141]. The outage occurred after several connections with IoT devices failed, and the devices were forced to re-register with the network, which flooded the CN. This caused the registration procedure to time out and triggered subsequent re-attempts, gradually amplifying the signaling traffic and eventually overloading the network. It took DoCoMo hours to process the backlog. This outage exemplifies the danger of signaling storms and demonstrates that simply increasing network capacity to meet the rapid growth of edge devices is not an effective solution to signaling storm threats. This event also proved that re-registration had been a weak link in the cellular network system and could be exploited by malicious parties to threaten the security of today’s 5G network. It was reported that 3 billion cellular IoT devices have already been connected by the end of 2023, and Ericsson predicted there will be around 6 billion IoT devices connected by 2029, and many of them are capable of global roaming. The global cellular IoT market size was valued at USD 3.9 billion in 2021 and is projected to reach USD 15.4 billion by 2027, growing at a compound annual growth rate of 25.7% [142]. Cellular IoT devices can be weaponized to conduct signaling storm attacks in the context of 5G networks and incur significant damage, the same way their counterparts (e.g. IoT Mirai Botnet [87]) were used to launch various network attacks in the past. Indeed, this is a legitimate concern that must be addressed.

The contributions of this work are as follows:

- We formalize the registration signaling storm attack model. We show the importance of this problem through experimental evidence.
- We propose a blockchain-assisted registration protocol to mitigate this threat effectively. Our protocol can be integrated with both 5GAKA and EAP-AKA procedures and does not affect authentication performance when there is no attack.
- We validate through rigorous performance evaluation the efficiency of our solution.

The remainder of this chapter is organized as follows. We introduce the 5G architecture and 5GAKA procedure in Section 3.2. Section 3.3 surveys and discusses related works. The registration signaling storm attack model is formalized and its potential damage is discussed in Section 3.4. Our proposed protocol and architecture are described and analyzed in Section 3.5. Section 3.6 provides the implementation details and evaluation results. Section 3.7 discusses the solution and outlines future challenges. Finally, Section 3.8 concludes this chapter.

3.2 Background

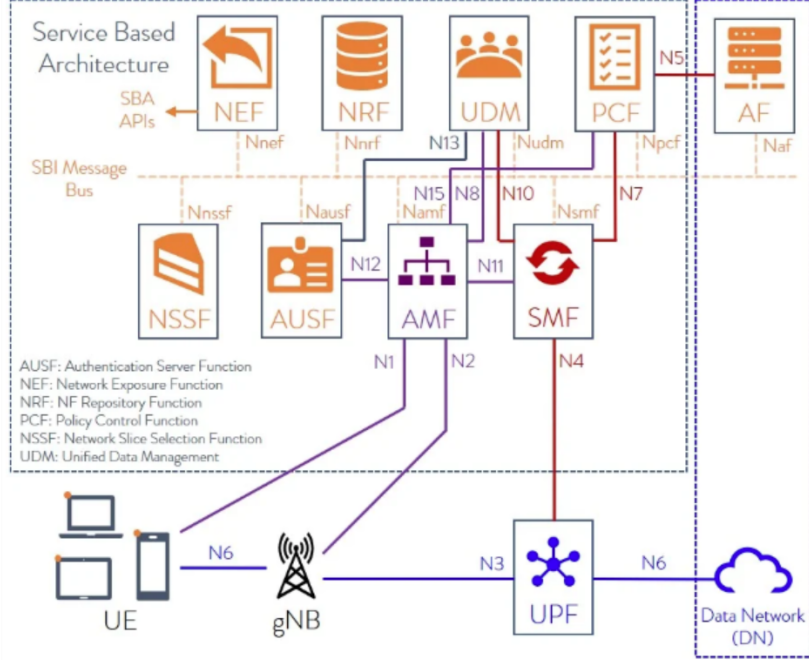


Figure 3.1: 5G service-based architecture [5]

One particular aspect of 5G is the 3GPP-defined Service-Based Architecture (SBA) in which the CP functionalities of the 5G network are delivered through a set of interconnected Network Functions (NFs), deployed using software from various sources and suppliers, with each NF authorized to access services of other NFs. Compared with prior 3GPP technologies, the SBA brings to 5G CN greater modularity, flexibility, and scalability.

Figure 3.1 shows the CP and DP separation in the 5G architecture, where DP is responsible for handling the delivery of user data, while CP manages network signaling, authentication, and resource allocation. As such, the resources dedicated to the CP are expected to be more constrained than those allocated to DP [109].

In the following, we will focus on the CP NFs and other components involved in the authentication procedure and describe the protocol.

3.2.1 SUPI and USIM

UE is composed of two essential components: the Mobile Equipment (ME) and the Universal Integrated Circuit Card (UICC), which contains the Universal Subscriber Identity Module (USIM). The USIM is intended to securely store the subscriber's security-related context, including a globally unique Subscription Permanent Identifier (SUPI) and a long-term key K , and handle the computation incurred by the authentication protocol [143]. Once in possession of one's SUPI and the corresponding long-term key K , any party can authenticate itself to the 5G network on behalf of the legitimate UE. According to the 3GPP TS-33.501, the length and format of the SUPI are arbitrary. However, for legacy reasons, it is commonly assumed that the SUPI is equivalent, in length and format, to 4G International Mobile Subscriber Identity (IMSI). The IMSI is usually implemented as a 15-digit number. The first three digits are the Mobile Country Code (MCC), and the next three are the mobile network code (MNC). They, together, indicate a specific mobile operator in the world. The remaining digits are the Mobile Identification Number (MSIN), which uniquely identifies a subscriber to the operator's network. In 4G, when the UE registers to the network, the IMSI is sent in plaintext over the air. As such, the IMSI can be easily captured with the infamous IMSI catcher [144]. After one's IMSI (or SUPI) is leaked, the user's privacy is jeopardized because the attacker can exploit the Signaling System 7 (SS7) protocol to intercept calls and track the user's location [145]. To avoid this, the SUPI in 5G is first encrypted into the Subscriber Concealed Identifier (SUCI) and then sent over the air. To be more specific, a UE's MCC and MNC are sent in plaintext, but its MSIN is encrypted.

3.2.2 5G Authentication Protocol

The 3GPP defined two standard protocols for UE registration, i.e., the 5GAKA and the Extensible Authentication Protocol Authentication and Key Agreement (EAP-AKA), with minor differences. In this section, we describe the workflow of the 5GAKA protocol following the 3GPP TS-33.501.

The NFs involved in the 5GAKA protocol are the Security Anchor Function(SEAF), the Authentication Server Function(AUSF), and the Unified Data Management Function(UDM). Because the SEAF is hosted in the Access and Mobility Management Function (AMF), AMF and SEAF are considered interchangeable in this work.

5GAKA Initialization Phase

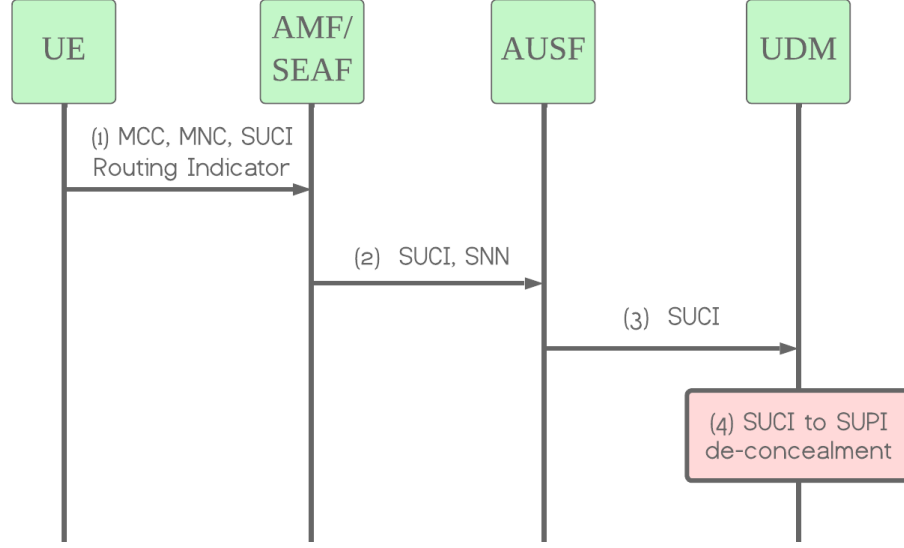


Figure 3.2: 5G-AKA Initialization

Algorithm 1 Elliptic Curve Integrated Encryption Scheme

Input: SUPI, PK_{HN} , KDF

- 1: Randomly generate key pair SK_{UE}, PK_{UE} .
 - 2: $TK = PK_{HN} \cdot SK_{UE}$
 - 3: $EK || ICB || MK = KDF(TK)$
 - 4: $Cipher = Encrypt_{AES}(SUPI, EK)$
 - 5: $MAC_{UE} = HMAC(Cipher, MK)$
 - 6: $SUCI = PK_{UE} || Cipher || MAC_{UE}$
 - 7: **Return** SUCI
-

The initialization phase workflow is shown in Figure 3.2. The UE performs a SUPI to SUCI concealment following the Elliptic Curve Integrated Encryption Scheme (ECIES) showed in Algorithm 1 using the home network (HN)’s public key PK_{HN} and a local key generator function KDF . The UE sends to the SEAF the SUCI, MCC, MNC, and a routing indicator indicating which UDM should serve the UE. The SEAF then forwards the SUCI and the serving network name (SNN) to the AUSF. The AUSF temporarily stores the SNN and forwards the request to the UDM. Because the UDM holds the HN’s secret key SK_{HN} , it performs a SUCI to SUPI de-concealment following ECIES decryption. After that, the UDM checks if the SUPI is a valid subscriber number of the HN.

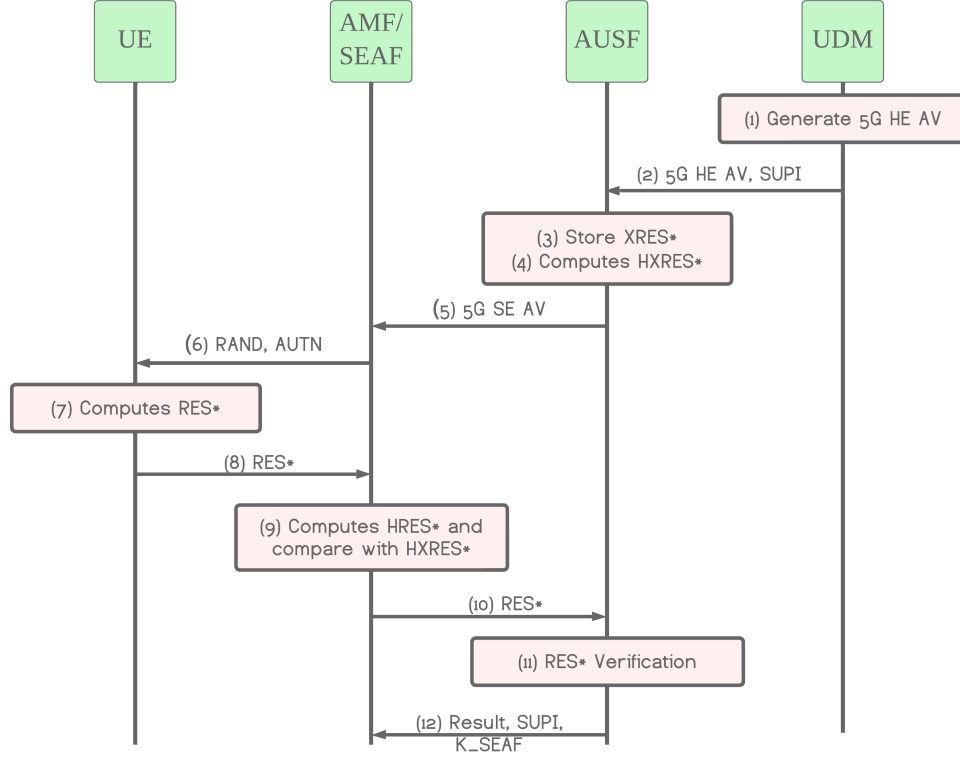


Figure 3.3: 5G-AKA Authentication

5GAKA Authentication Phase

The authentication phase follows a challenge-response authentication scheme shown in Figure 3.3. The UDM first prepares a random challenge $RAND$. Next, the UDM takes the input value $RAND$, K , and a synchronized sequence number; and feeds them to the key derivation function, resulting in $AUTN$ and session keys. The UDM sends $SUPI$ along with a Home Environment Vector containing $RAND$, $AUTN$, and the expected response $XRES^*$ to the AUSF. The AUSF then sends the SEAF a Serving Environment Vector (SEV) containing $RAND$, $AUTN$, and $HXRES^*$, which is the hash of $XRES^*$. Upon receiving $RAND$, $AUTN$ from the SEAF, the UE computes the response RES^* and sends it back to the CN. The AUSF sends SEAF the UE's $SUPI$ and a session key K_{SEAF} after AUSF and SEAF verify the correctness of RES^* , indicating successful authentication.

3.3 Related Work

Several vulnerabilities and exploits against the 5G-AKA protocol were reported in the literature [146, 147]. These include the SUCI replay attack, linkability attack, and the lack of perfect forward secrecy vulnerability.

Although the aforementioned works have addressed different vulnerabilities of 5GAKA, the persistent registration signaling storm attack has yet to be thoroughly studied. Indeed, the more *centralized* nature of the UDM [148] makes 5GAKA vulnerable to such an exploit. Due to the distributed nature of blockchains, blockchain-based authentication protocols were suggested in the literature. [149] proposes to simply replace the UDM with a blockchain serving as a distributed database. In practice, such a method is hard to deploy because the UDM needs to manage many subscriber-related data, including billing, user location, and security-related UE context. Hence, a blockchain-based UDM brings synchronization and storage challenges to the operator. [150] also addresses the registration storm issue by solving the SUCI replay attack in a non-roaming scenario using a private blockchain. The main idea is that the UE will send the SUCI and a hash commitment of the next SUCI in the initialization phase. The commitment will be saved on the blockchain so that SUCI cannot be replayed. Although their results show the method is robust to signaling storm attacks, their approach is unrealistic to be considered. They completely removed the authentication phase in 5GAKA along with the long-term key K , and solely relied on gNB querying blockchain to perform the mutual authentication, resulting in the UE can hardly trust the HN. Unlike our solution, their change in 5GAKA will cause changes in other protocols, such as handover and PDU session establishment, because AMF and UDM no longer save the UE security context. Moreover, any compromised gNB will leak the HN secret key and can permanently deny the UE from the network by polluting the blockchain. Most blockchain-based existing solutions to 5GAKA vulnerabilities require writing at least one record on the blockchain per registration. This will undoubtedly introduce a significant storage overhead, and blocked synchronization may delay the authentication procedure.

In [123], signaling storm attacks are addressed at the RAN, without using blockchains, but rather using the O-RAN xApp framework. The proposed solution uses the timing advance (TA) parameter during the UE random access phase to identify a unique IoT device assuming the device is not moving. The xApp is responsible for making statistical analyses for every incoming random access request based on the history data. Such an idea can be compelling, considering the signal is stopped at the very beginning of the registration. However, the usage is very limited as it requires both gNB and UE to stay static, and UE must connect to the same gNB. Even under such an assumption, gNB does not have a promising role in identifying the malicious UE as the attack to the 5G core is

distributed, and gNB can only have a partial view of the attack. Moreover, identifying a device using TA is inaccurate because the malicious device can easily change this value by tampering with the round trip time measurement.

Instead of detecting malicious users at the gNB, where the gNB can hardly have a global view, [120] proposes a machine learning detection system in the 5G core against 5G signaling storm attacks, including registration attacks. The 5G core is responsible for classifying malicious SUPIs using the history packets data feature, such as the time interval between two packets and the total count of packets during a given time. Although the result shows that the detection is ideal, such a method requires storing the history feature data of all UEs, which will introduce a huge storage burden. Moreover, they did not discuss how to mitigate the attack after classifying SUPI.

3.4 Problem statement

In the following, we consider the 3GPP 5GAKA authentication procedure, even though our attack model and methodology apply both to 5GAKA and EAP-AKA. We aim to propose a 3GPP 5GAKA-compliant, efficient, and realistically deployable solution by 5G operators. We also assume the following:

1. SN and HN can securely communicate.
2. SN and HN securely store their cryptographic key pairs.
3. UE securely stores SUPI and long-term key K in USIM, unless compromised.
4. The air interface between UE and gNodeB is not secure. An attacker can monitor and tamper with the information transmitted on the N1 interface, the same as an adversary in the formal Dolev-Yao (DY) model [151].
5. As all cryptographic primitives, in the UE, are computed in the hardware USIM, using more cryptographic primitives would make the USIM more resourceful and more expensive. Thus, we assume the UE cannot do asymmetric decryption [152].
6. The UE may or may not be in a roaming scenario. The UE will first communicate with SN's SEAF function in a roaming scenario. The UE will first communicate with HN's AMF function in a non-roaming scenario.
7. Most mobile operators implement UDM as a relatively centralized cloud-native function as it requires saving a huge amount of data. AMFs, on the other hand, are normally assumed to be relatively distributed as they can be easily deployed at the edge of the 5G CN.

3.4.1 Attack Model

With the increasing number of 5G cellular devices, it is realistically feasible for an adversary to compromise an important number of UEs (*e.g.*, cellular IoT devices) and use them to target an operator's UDM with a simple and powerful attack leveraging 5GAKA, and reproduce the damage induced by the Telenor [18] or DoCoMo incidents [19, 16], only to cite a few.

We may think of different attack scenarios leveraging 5GAKA to generate bogus registrations storm using compromised UEs. It is important to note that the compromised UEs can be geographically distributed. A registration signaling storm attack against an operator's network can be launched from another operator's network (a.k.a. roaming network). The attacker can simply specify the MCC, MNC, and routing indicator of the target network, and the roaming network will direct the registration message to the right place. Also, the attacker may use valid or invalid MSINs. For convenience, an MSIN is said to be valid (respectively invalid) if it exists (respectively does not exist) in the operator's UDM database. We distinguish three possible bogus registration scenarios:

1. The UE uses a valid MSIN, and then the UE can legally perform a full 5GAKA procedure.
2. The UE is replaying a previous SUCI, then the UE can make the UDM perform ECIES decryption and authentication vector generation and let the UDM hold the communication session.
3. The UE uses an invalid MSIN and carefully chooses the MCC and MNC, then the UE can make the UDM perform ECIES decryption.

Bogus registrations with valid MSIN (i.e., scenario 1) are relatively more harmful as they result in a higher overhead on the UDM and the CN. In fact, not only do they make the UDM perform a full 5GAKA procedure, but also they trigger several message exchanges between different functions in the 5G CN. However, such a method is relatively slower than the other two bogus registration scenarios where the UE does not need to interact with the HN. For the other two scenarios, the attacker can send massive requests within a short time frame and force the UDM to compute the ECIES, which is required to verify the subscriber's identity, hence consuming its processing power [92]. Such carefully crafted attacks can be triggered from multiple SNs, and the bogus registration signals will be ultimately forwarded to the targeted UDM. The attacker can also select a random subset of the compromised devices to register periodically or at random times, making them harder to detect.

3.4.2 Attack Outcome

When the 5G CN is under a registration storm attack, it is expected to experience a spike in traffic due to the increase in the number of registration requests. Besides, the AMF, SMF, and AUSF may experience a temporary memory spike as they need to maintain temporary session management data. As these NFs are virtualized, their computation power and storage may not be designed to handle unexpected traffic, leading to the risk of crashing. More importantly, the SUPI can only be revealed at the more "centralized" UDM; the massive number of incoming requests will require significant cryptographic computation, draining the UDM's computation power away from the legitimate users' requests. As the 5GAKA protocol happens entirely in the CP, which has more constrained resources than the DP, the massive amount of incoming traffic also causes traffic congestion, delaying legitimate users' requests and leading to a denial of service for legitimate users.

Overall, a registration storm attack is a distributed denial of service (DDoS) attack whose volume and impact can be amplified in two ways:

- Legitimately: by legitimate UEs whose registrations are delayed and repeatedly re-attempt registration, after a timeout.
- Adversely: once the 5GAKA procedure is completed (registration scenario 1), the adversary can trigger PDU session establishment and then trigger the de-registration procedure, causing more signals inside the CP. Each PDU session establishment involves the RAN and 5 NFs to exchange 17 messages in the CP. Moreover, one UE is allowed to trigger up to 15 PDU sessions. Finally, the de-registration procedure involves the RAN and 5 NFs exchanging 15 messages to release the UE context.

The registration delay is critical in 5G because authentication precedes any service. When the UDM is under attack, a legitimate UE may be denied a handover; when a UDM is under attack, the users' QoE is impacted. 5G aims to offer ultra-reliable low latency communications (URLLC) and support mission-critical services. However, as long as the CN is vulnerable to registration storm attacks, none of these services can be guaranteed. As such, the network must be able to detect and block bogus registrations at their earliest phase.

3.5 Blockchain-assisted 5G Authentication

In this section, we present our blockchain-assisted 5G authentication protocol. The solution applies to both 5GAKA and EAP-AKA. We first provide some cryptographic primitives before delving into the details of the solution.

3.5.1 Cryptographic Primitives

Blockchain and Smart Contract:

Blockchain is a decentralized digital ledger technology that allows multiple parties to record and maintain a shared database securely, transparently, and immutably. The blockchain can be categorized into public, private, and consortium-based. In the public blockchain, anyone can perform read and write operations. In contrast, the private blockchain is a permissioned ledger that only one organization can write and may allow other organizations to read the blockchain. A consortium blockchain combines public and private blockchains, where multiple organizations form a collaborative alliance that can both write and read the blockchain. Performance-wise, Ethereum, a representative of the public blockchain, can only handle 15 transactions per second, while a consortium blockchain can theoretically achieve 20,000 transactions per second [153]. A smart contract is a self-executing computer program that automatically executes, controls, or documents events when specific conditions are met. They are stored on a blockchain, providing a secure and transparent way of execution, which reduces the need for trusted intermediates as well as losses due to fraud.

ECDSA and ecRecover

The Elliptic Curve Digital Signature Algorithm (ECDSA) uses a private-public key pair $(sk, pk = sk * G)$ derived from an elliptic curve with order n and generator G . The signer first takes a message msg and computes its hash $z = SHA1(msg)$. The signer will then generate a random nonce k , calculate the curve point $(x_1, y_1) = k \cdot G$, derives $R = x_1 \bmod n$, and then computes $S = k^{-1}(z + R \cdot sk) \bmod n$. The signature (R, S) can be verified using pk and msg .

After receiving the signature, the verifier computes

$$\begin{aligned}(z \cdot S^{-1} \cdot G + R \cdot S^{-1} \cdot pk) &= (z + R \cdot sk) \cdot S^{-1} \cdot G \\ &= (z + R \cdot sk)(z + R \cdot sk)^{-1}(k^{-1})^{-1} \cdot G \\ &= k \cdot G\end{aligned}$$

So, checking the result with R verifies the signature. From signature (R, S) and msg , one can recover possible public key by trying point P to compute $-zR^{-1} \cdot G + S \cdot R^{-1} \cdot P$, where the x-axis of P is one of $R + n, R + 2n, \dots$. If $P == k \cdot G$, then the result will be

$$\begin{aligned}
& -zR^{-1} \cdot G + S \cdot R^{-1} \cdot k \cdot G \\
& = (-z \cdot R^{-1} + (z \cdot R^{-1} + sk)) \cdot G \\
& = sk \cdot G = pk
\end{aligned}$$

In Ethereum, an additional signature identifier V is also included in the signature to allow one to recover pk . This is so-called the *ecRecover*. The result signature (R, S, V) will be 65 bytes long using a 256-bit long key providing a 128-bit security level.

3.5.2 Key Ideas

The ultimate goal of the mitigation solution is to block the malicious UEs' registrations from overloading the UDM. Because the UDM manages the subscribers' data and has a global view of registrations no matter where they come from, it is an ideal choice to detect and identify malicious SUPIs. Because the AMFs/SEAFs are in the CN, which has a high-security requirement, and they are relatively distributed near the edge compared to UDM, they are best fitted to block the malicious UE's requests. The UDM should inform each AMF/SEAF about the malicious SUPI. However, it is inefficient to let a single UDM inform every AMF/SEAF from different operators. Besides, AMF/SEAF are not designed to receive the SUPI before the HN authenticates the UE. Thus, we propose letting blockchain act as a trusted platform so that UDM can efficiently and securely share the information with every AMF/SEAF once it detects a malicious SUPI. Because AMF/SEAF cannot reveal the SUPI, we add a second identity to the UE, which is the blockchain address. Unlike the SUPI, which is used in different protocols such as SS7, resulting in linkability issues and jeopardizing user privacy, the blockchain address should only be used in the authentication protocol to mitigate such concerns. As SUPI and long-term key K are the only necessary confidential information needed for authentication, the long-term key K is an ideal choice to be the blockchain private key that can generate the corresponding blockchain address and be updated following the Long Term Key Update Process (LTKUP). Since each UE is assigned an on-chain address, the UDM can flag the malicious UEs' addresses on the blockchain. Figure 3.4 shows our system model. In this example, two malicious UEs subscribed to operator B are trying to register through AMFs from operators A (visited network) and B. The UDM of operator B, UDM-B, sends the transaction to the smart contract and delegates the reject decision to AMF. The AMF checks the contract-UDM-B and sees that UDM-B has banned those UEs. As such, the AMF rejects any registration from the two UEs.

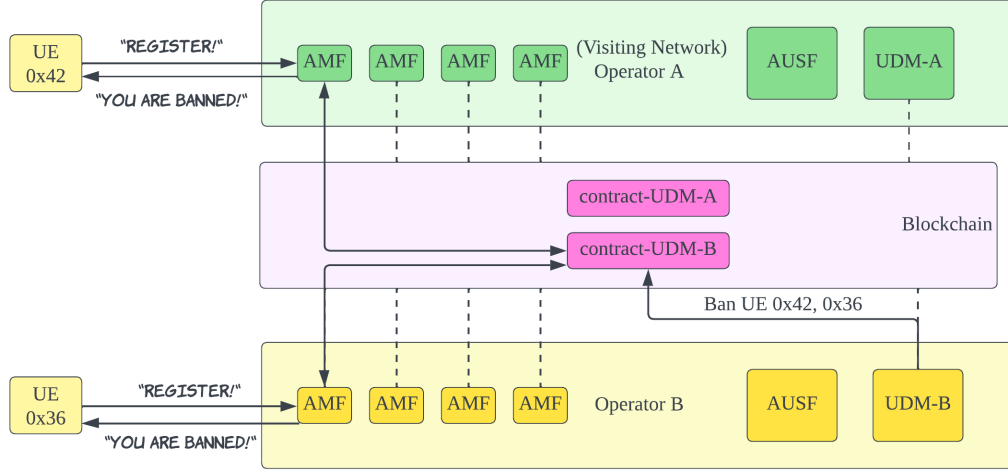


Figure 3.4: System Overview

In practice, adding any extra information to the protocol can affect the performance of the 5G authentication procedure. Thus, we add an optional mode called *delegation mode* in the authentication protocol that the CN can activate when the UDM is under attack. When the delegation mode is enabled, the SEAF/AMF shall ask for UE's blockchain identity after it receives the SUCI, a slight registration delay to avoid potential high latencies. We name this extra information from the UE *Blockchain-Vector (BV)*. Below are the security requirements and functional desiderata of *BV* to prevent a UE from bypassing the AMF if it is flagged as malicious by the UDM:

1. *BV* reveals the UE's blockchain address to the SEAF/AMF but not to the public.
2. *BV* allows SEAF/AMF to verify that the sender holds a valid blockchain account issued by its home UDM.
3. *BV* has a one-to-one mapping to SUCI, so one valid Blockchain-Vector cannot be replayed on any SUCI.
4. *BV* together with *SUCI* cannot be replayed.
5. *BV* should not introduce any linkability issue.

3.5.3 The Setup

Operators with roaming agreements could set up a consortium blockchain $blockchain_{consortium}$; each operator should maintain at least one blockchain node. They should also agree on a list of public and secret key pairs (PK_{SEAF}, SK_{SEAF}) and store them in each AMF/SEAF.

Each operator O_i should deploy a smart contract on $blockchain_{consortium}$ for certain UDM U_j with address $contract_j$. All these smart contracts should provide functionalities following the standard as specified in Table-3.1

During USIM production, operator O_i should generate the long-term key K as a random 256-bit number with a valid corresponding point on Curve25519 following the Ethereum standard. The UE holding K should be able to generate the corresponding blockchain address $address_{UE}$. In addition, operator O_i should generate another 256-bit random number $salt_{UE}$ and save it in the USIM. Then O_i updates this information in $blockchain_{consortium}$ by sending a transaction to call `updateSalt(addressUE, saltUE)` in the smart contract $contract_j$. O_i should also let the UE keep a list of PK_{SEAF} .

3.5.4 Blockchain-assisted 5G Authentication Workflow

We now show the proposed Blockchain-assisted 5G Authentication protocol workflow using 5GAKA as an example. We modified the initialization phase of 5GAKA and kept the authentication phase unchanged. This way our methodology can be applied to EAP-AKA as it has the same initialization phase as 5GAKA.

Figure 3.5 shows our initialization phase workflow. The UE shall first send the SUCI, MCC, MNC, and a routing indicator to the SEAF. The SEAF follows Algorithm 2. First, it checks if UDM is in delegation mode. If not, the workflow follows 5GAKA. If the UDM has enabled the delegation mode, SEAF should send a random number to the UE and await the response. The UE will create the response following Algorithm 3. After receiving UE's response, the SEAF will either (i) reject the UE's registration request, or (ii) send AUSF/UDM the SUCI along with the UE's on-chain address. If UDM finds that the concealed SUPI does not match the address, then UDM bans the address and sends the transaction to the blockchain. In practice, the attacker is not motivated to sign other UE's SUCI as their account will be banned during the storm attack. In the protocol, the SEAF/AMF can periodically monitor the UDM status to avoid frequent unnecessary queries if the delegation mode is disabled. As EAP-AKA and 5GAKA share the same initialization phase, our methodology also applies to EAP-AKA.

3.5.5 Comparison with Baseline Mitigation Procedure

The baseline methodology is to keep the authentication protocol untouched and have the UDM process all registration requests and send a registration reject message to malicious UEs after decrypting SUCI following ECIES. Indeed both the baseline mitigation procedure

Table 3.1: Smart Contract Standard

Name	Variable/Function	Input	Caller	Description
owner	variable (addr)	NA	NA	An address indicating the contract owner
delegate	variable (bool)	NA	NA	A boolean indicating if the attack is happening
blackList	variable (addr → bool)	NA	NA	A mapping from address to boolean. The boolean is true if the address is banned
saltList	variable (addr → uint256)	NA	NA	A mapping from address to integer. The number is not 0 if the address is valid
changeUDMStatus()	function	NA	owner	A write operation changing the boolean delegate
changeOwner()	function	addr	owner	A write operation changing the owner address
updateSalt()	function	addr [], uint256 []	owner	A write operation updating the values of addresses in saltList
banUser()	function	addr []	owner	A write operation changing the values of addresses in blackList to True
recoverUser()	function	addr []	owner	A write operation changing the values of addresses in blackList to False
getSaltStatus()	function	addr []	Any	A read operation returns the values of addresses in saltList and blackList
getUDMStatus()	function	NA	Any	A read operation returns the value of delegate

and the proposed Blockchain-assisted 5G authentication solution assume that the UDM is capable of detecting if a UE is compromised and being enrolled in a signalling storm attack. Compared to this baseline solution, ours has the following advantages:

1. **Eliminating the possible traffic congestion:** As the UDM is a more centralized NF, the massive amount of HTTP requests generated by the signalling storm will likely cause traffic congestion resulting in packet loss which may cause authentication failure for legitimate UEs. The massive malicious requests shall not go to the UDM in our solution and will be dealt with in a distributed fashion, so traffic congestion should not occur.
2. **Reducing the computation load of the UDM:** 3GPP indicates that the network side cannot determine the sender's identity before computing ECIES. The massive requests would exhaust the processing power of the target UDM and slow down the response to legitimate UEs [92]. In our solution, as the SUCI shall not go to UDM, this concern is no longer relevant.
3. **Reducing the concurrent session management load of the AUSF and UDM:** Managing concurrency is always challenging for network servers. The authentication protocol requires the AUSF and UDM to save temporary identifiers to keep the com-

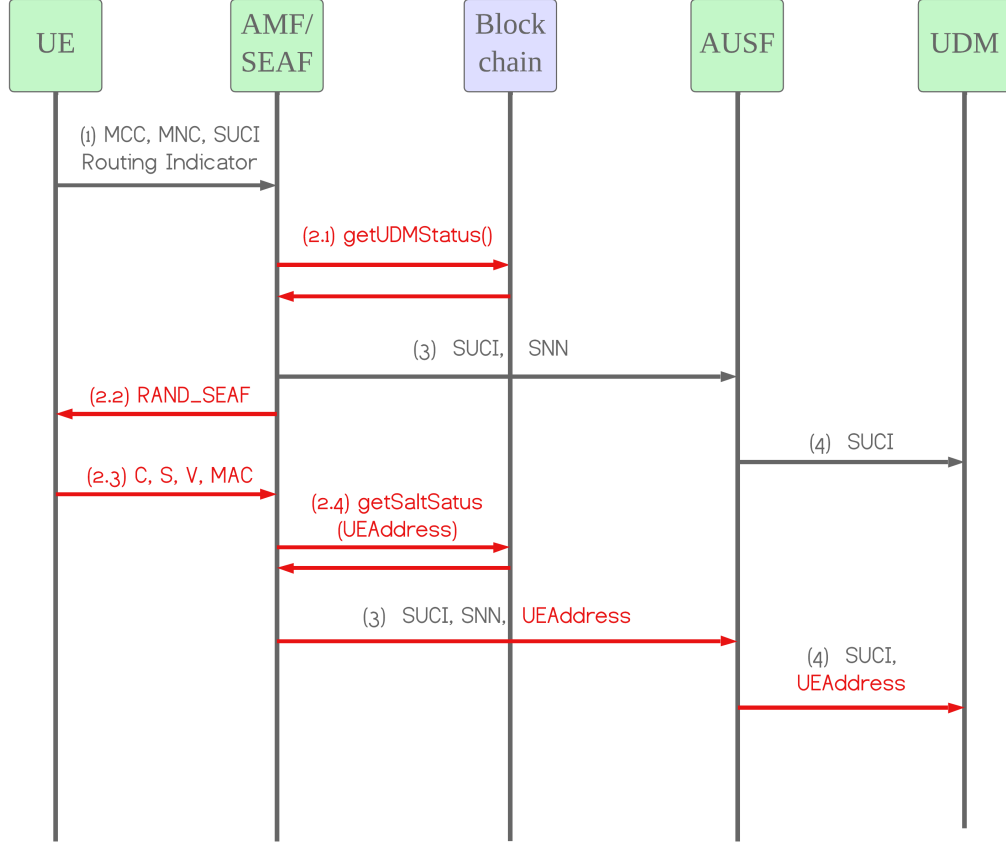


Figure 3.5: Blockchain-assisted 5G Authentication: Initialization phase

munication session between UE. The spike in concurrent sessions may overwhelm the servers. In our solution, this concern is mitigated.

3.5.6 Security Analysis

In this section, we evaluate whether our solution fulfills the requirements and desiderata stated earlier. It is essential to understand that the purpose of *BV* is not to prevent an attacker from attempting to fail someone's authentication but to prevent an attacker from legally bypassing the AMF/SEAF check.

1. **BV reveals the UE's blockchain address to the SEAF/AMF but not to the**

Algorithm 2 AMF/SEAF Processing Logic

Input: PK_{UE} , SK_{SEAF} , MCC, MNC, Routing Indicator, SUCI

- 1: Locally find the UDM *contractAddress* based on MCC, MNC, and Routing Indicator.
 - 2: $underAttack = contractAddress.getUDMStatus()$
 - 3: **if not** *underAttack* **then**
 - 4: Send SUCI to AUSF; **End Process**
 - 5: **end if**
 - 6: Generate random number $RAND_{SEAF}$; send it to UE.
 - 7: $mask = SHA256(PK_{UE} \cdot SK_{SEAF} || RAND_{SEAF})$
 - 8: Waiting to receive $BV = (C, S, V, MAC)$ from UE.
 - 9: $R = mask \oplus C$
 - 10: $address = ecRecover(R || S || V, SUCI || RAND_{SEAF})$
 - 11: **if** $address == 0x0$ **then**
 - 12: Send UE "Recover Address Failure"; **End Process**
 - 13: **end if**
 - 14: $salt, ban = contractAddress.getSaltStatus(address)$
 - 15: **if** $MD5(SUCI || R || salt) \neq MAC$ **then**
 - 16: Send UE "Blockchain MAC Failure"; **End Process**
 - 17: **end if**
 - 18: **if** *ban* **then**
 - 19: Send UE "Registration Reject"; **End Process**
 - 20: **end if**
 - 21: Send SUCI, *address* to AUSF; **End Process**
-

public: As the SEAF holds the SK_{SEAF} , only SEAF can recover R . The attackers holding (C, S, V) cannot recover the blockchain address.

2. **BV allows SEAF/AMF to verify that the sender holds a valid blockchain account issued by its home UDM:** The salt is not 0 indicating the HN acknowledges this address matches one of its subscribers.
3. **BV has a one-to-one mapping to SUCI, so one valid Blockchain-Vector cannot be replayed on any SUCI:** BV is generated from $SUCI$, a previously sent BV cannot be used on another $SUCI$.
4. **BV together with $SUCI$ cannot be replayed:** BV is also generated from $RAND_{SEAF}$. An attacker holding a valid $SUCI, BV$ pair cannot bypass the check because each registration requires a different $RAND_{SEAF}$.
5. **BV should not introduce any linkability issue:** The ECDSA signature algo-

Algorithm 3 UE Signing Process

Input: $SUCI, K, RAND_{SEAF}, PK_{SEAF}$

- 1: $msg = SUCI || RAND_{SEAF}$
 - 2: Sign msg based on Ethereum ECDSA using K resulting in (R, S, V) .
 - 3: $mask = SHA256(SK_{UE} \cdot PK_{SEAF} || RAND_{SEAF})$
 - 4: $C = mask \oplus R$
 - 5: $MAC = MD5(SUCI || R || salt)$
 - 6: **Return** $BV = (C, S, V, MAC)$
-

rithm ensures that the signature will differ even with the same message. An attacker who sends UE the same $RAND_{SEAF}$ will receive a different response.

6. **The Purpose of Salt:** In our solution, we add a random number as salt saved in `saltList` other than a boolean indicating whether the UE is a valid subscriber. Although it is theoretically infeasible to compute an existing valid blockchain account, [154] has launched a blockchain private key guessing attack that discovered 732 active private keys in Ethereum. Applying the salt prevents the attacker from bypassing the blockchain checking by guessing a valid address account.

3.6 Implementation and Evaluation

We implemented a proof-of-concept blockchain-assisted 5G authentication procedure in an emulated environment using Kubernetes. Free5GC v3.3.0 and UERANSIM v3.2.6 were used to emulate the 5G CN and the 5G RAN respectively. The emulation of the blockchain is based on Ganache which supports the Ethereum Virtual Machine. To integrate the blockchain into Free5GC and enable smart contract functionality, we used the go-ethereum module v1.10.26 and solidity v0.8.4. Our deployment is shown in Figure 3.6; two sets of UEs, each containing malicious and benign UEs, connect to two AMFs through two gNBs, and the two AMFs connect to one AUSF and blockchain. Our code is publically available at <https://github.com/pzeina/5g-bv-storm.git>.

In our experiment, each set of UEs contains 220 benign and 30 malicious UEs. Each of the 440 benign UEs will register to the network exactly once at a randomly selected time t ranging between 1 and 250 seconds. The registration storm consists of the remaining 60 malicious UEs who will register to the network in bursts occurring once every 60 seconds. In practice, such attacks can be more continuous in time and more frequent. The smart contract was deployed in Ganache, and the AMF can query the contract storage

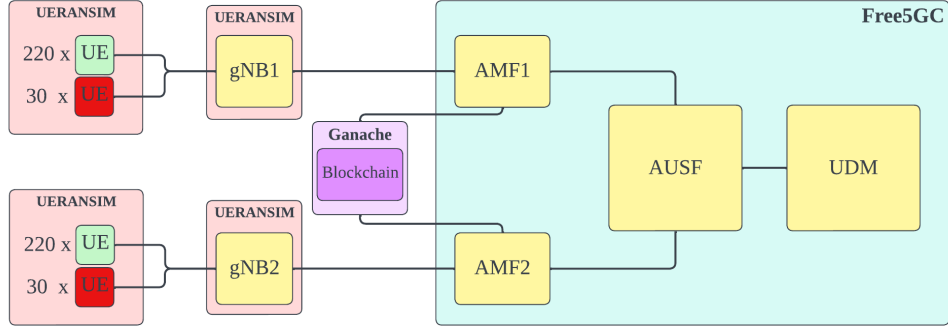


Figure 3.6: Deployment Model

information. We assume the UDM has previously tagged the malicious UEs and sent the transaction to the blockchain to ban these and delegate the rejection to the AMFs. Four scenarios were defined to evaluate our solution:

1. **No storm:** There is no registration storm attack.
2. **Storm with no defense:** Malicious UEs successfully register to the network and are not blocked.
3. **Storm with baseline defense:** Malicious UEs attempt to register to the network but their registration is rejected by the UDM after decryption of the SUCI.
4. **Storm with blockchain-assisted 5GAKA:** Malicious UEs are blocked by the AMFs.

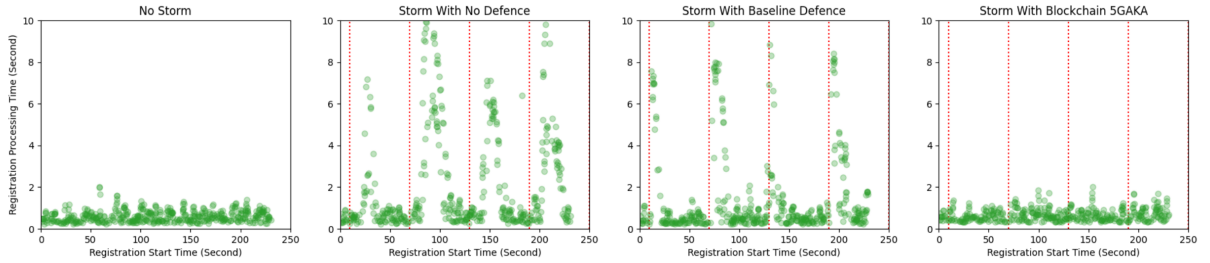


Figure 3.7: Registration Processing Time Comparison for 4 Scenarios

Figure 3.7 shows the impact of the registration storm on the registration processing time under the 4 different scenarios. We measured the 440 benign UEs' registration processing time which is the difference between the authentication success timestamp and the registration start timestamp. The x-axis represents the registration start timestamps. The

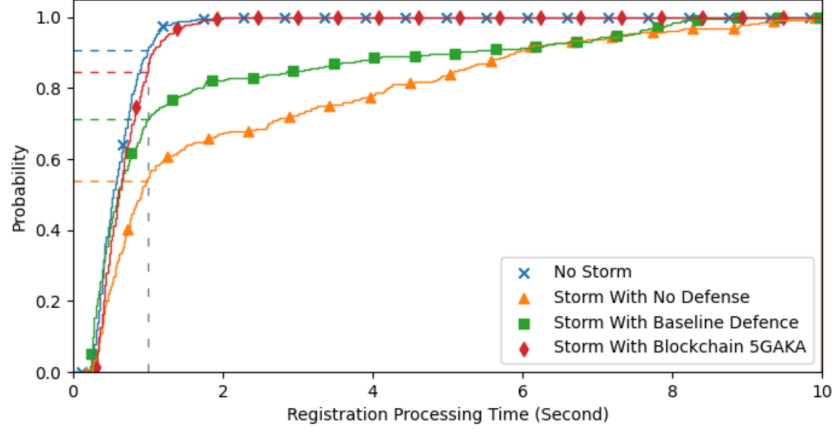


Figure 3.8: CDF of registration time for the 4 Registration Storm Scenarios

red vertical lines represent the start of each registration storm. The 4 sets of experiments follow the same randomly selected UEs' registration start times for fairness. The results show that the storms significantly increase the UE registration time in the absence of any defense mechanism. This is also noticeable in the presence of the baseline defense approach, as the bogus registrations still overload the UDM increasing the overall registration processing time. In contrast, our approach is more robust for it is not sensible to the storms. In fact, with our solution, we see that the registration times are very close to those in the absence of any storm. These findings are emphasized in Figure 3.8, which showcases the CDF of the registration time under all 4 scenarios. We see that the baseline solution mitigates the attack at a certain level but does not ideally address the problem, whereas our methodology successfully mitigates the attack at the cost of a very small delay.

3.7 Discussion

Billing difficulty: The transparency of billing between different operators is a challenge in mobile systems that remains unsolved in our solution. Because no public record can show SEAF/AMF spent their computation power to ban the malicious UEs, billing transparency is hard to achieve. However, this may be mitigated by building a consortium alliance with a proper financial agreement. Detection and classification of malicious UEs are also essential parts that our research does not address. As the attack may follow any pattern with different subsets of malicious UEs and as the UDM shall prevent a false alarm introduced

by group gathering activities, we expect identifying attacks and classifying malicious UEs to be challenging. Some ML-enabled methods may be helpful but would bring unnecessary storage and computation burdens to the UDM. Future work would be to design an efficient and accurate way to identify the attack and classify the malicious UEs.

Computation cost analysis: When the attack is not happening, AMF can periodically monitor the status of UDMs, so the extra cost to each registration can be optimized to be ignored. Assuming the attack is happening, the extra cost for each registration involves computation from UE ($Compute_{UE}$), AMF ($Compute_{AMF}$) and blockchain node ($Compute_{blockchain}$) plus the propagation delay of UE-AMF $Delay_{UE-AMF}$ communication and AMF-blockchain communication $Delay_{AMF-chain}$. As the hashing and XOR operation is $O(1)$ complexity and the input size is short, we ignore these types of operations. $Compute_{blockchain}$ contains a query time of the smart contract storage slot and a query time of variables in the contract. As the smart contract storage slot management is usually implemented as a Merkle Patricia Trie, the time complexity of querying given the smart contract address will be $O(\log n)$, where n is the number of contracts in total. The smart contract storage itself is a 256-bit array and the query time complexity will be $O(1)$ using hash. $Compute_{UE}$ involves an ECDSA signing procedure which mainly contains one multiplication operation for computing R and one multiplication operation for computing S . The key agreement with AMF involves one multiplication operation. From the UE side, The key agreement and R computation can be prepared before sending $SUCI$. $Compute_{AMF}$ involves an ECDSA recover public key procedure which mainly contains two multiplication operations. As the key agreement can be done by waiting for UE's signature, this computation can be ignored. $Delay_{blockchain}$ mainly depends on the distance between AMF and the blockchain node. This can be optimized by introducing more blockchain nodes. Overall, the total extra cost can approximately be represented by $2 \times Delay_{UE-AMF} + 2 \times Delay_{AMF-chain} + 3$ multiplication operations + $O(\log n)$, where n is the number of contracts in total.

Blockchain storage The performance of blockchain is highly dependent on the framework in use and is out of the scope of this work. From a design perspective, although each UE is assigned an on-chain address, they are not externally owned accounts (EOA). In practice, different UDM will have a different number of subscribers to manage, they will have different robustness against potential attacks. Our method offers the flexibility to let operators decide which UDM needs such protection. Thus the blockchain storage size will only be affected by the number of subscribers in the protected UDMs. Suppose an operator only needs to protect one UDM, then the other UEs will not be saved on the chain.

3.8 Conclusions

In this work, we reviewed the 5G authentication protocol. We concluded that the de-concealment and centralized nature of the UDM can be exploited by malicious parties to initiate a registration signaling storm, overwhelming the control signal resources of the CP and leading to a denial of service for the 5G CN. To mitigate this attack, we propose a consortium blockchain-assisted solution that leverages the decentralized nature of AMF deployment to distribute the registration procedure and alleviate the load on the UDM in a trusted manner. Importantly, our approach applies to both 5GAKA and EAP-AKA protocols without affecting the performance of the authentication protocol, making it a realistic and flexible option for operators to consider. To validate the effectiveness of our solution, we conducted empirical evaluations using emulations. The experiments provide evidence that our proposed solution outperforms the baseline mitigation approach and enhances the security and resilience of 5G networks.

Chapter 4

Secure and Efficient Group Handover in 5G Non-Terrestrial Networks

4.1 Introduction

The wide-spreading LEO satellite constellations have become indispensable to the 5G and beyond cellular networks for closing the connectivity gap. LEO-based non-terrestrial networks (NTNs) can enable the International Telecommunication Union (ITU-R)'s envisioned framework for the new mobile communication generation, named IMT-2030, by supporting wide coverage, ubiquitous connectivity, and massive communication use case scenarios. Integrating NTNs with 5G can extend the service coverage to a truly global scale and facilitate a wide range of applications, including disaster response and recovery, public safety, vehicular connectivity, and the Internet of Things (IoT). In 5G and beyond NTNs, handover is pivotal in ensuring service continuity when UE signals must be transferred between orbiting non-terrestrial radio base stations, also known as gNBs.

The 3GPP standard, as specified in release 18 of TS 23.502, defines an Xn-based handover mechanism, where two gNBs can communicate, and an N2-based handover mechanism, where the 5G Access and Mobility Management Function (AMF) communicates with two gNBs. However, satellite gNBs introduce various challenges, making these handover schemes, designed for terrestrial networks (TNs), difficult to apply in the integrated TN and NTN scenario [75]. First of all, due to the high dynamics and movements of the LEO satellites, numerous users will trigger handover requests around the same time, resulting in a large number of messages that need to be handled between UEs and satellites and between multiple satellites in a short period of time. This can easily cause a

signaling storm [100] in an NTN radio access network (RAN). This can be exacerbated by the expected support for massive machine-type communication (mMTC). Second, the signal strength used as an indicator for TN handover cannot be applied to the NTNs as the signal strength will experience a very small change within the satellite coverage [75]. Third, different from TNs, the ground-space and space-space links in NTNs are subject to much higher latency, which poses a significant limitation on the use of re-transmissions in an NTN handover scheme. Fourth, a satellite gNB has limited computing resources, posing real-world limitations on the processing capability of handover requests. Last but not least, an efficient NTN handover protocol must tackle all of the above challenges while operating securely, resisting attacks even in the case where UE devices may be compromised and weaponized. Therefore, it is critical to address these challenges when designing NTN handover protocols.

Most existing works are based on TNs and address scenarios where handover poses a challenge, e.g., for UEs with high mobility that quickly change serving cells on platoons [131] and high-speed rails [155]. However, in a 5G NTN setting, the UE's mobility is overshadowed by the wide coverage of the satellite gNB [75]. Moreover, due to unique challenges in 5G NTNs, such as high mobility of NTN gNBs relative to ground UEs, large number of UE connections, and security risks, traditional handover schemes for TNs cannot be applied to NTNs, and new handover protocols are needed. The existing NTN handover solutions are focused on the physical-layer handover [156], handover procedure scheduling, and UE clustering [157, 158] without discussing the essential protocol design based on the latest 3GPP development. This work bridges the gap in the NTN handover protocol design above the physical layer, while achieving efficiency and security. The main contributions are:

- We propose a novel secure, efficient, and 3GPP-compliant Xn-based group handover protocol for NTNs that mitigates signalling storms and malware security risks on UEs.
- Our proposed protocol is implemented and validated through discrete-event simulations, where the key metrics such as success rate, message overheads, drop rate, and latency are significantly improved compared to the baseline 5G handover protocol [159, 59].
- We develop an open-source simulator where the NTN handover protocols can be implemented and evaluated. The code can be found here <https://github.com/zbh888/LEOhandover>.

4.2 Related Work

The state-of-the-art handover solutions are mainly focused on TN scenarios, while NTN handover has recently emerged as a new research topic. In the following, we discuss the relevant works in TNs with UE mobility and recent advancements in NTN handover. Most of the following works are focused on the security challenges of group handover, while performance-related issues are less studied.

In [159], a comprehensive security analysis demonstrates that 5G handover would not introduce confidential information leakage as long as the gNBs are not compromised. However, in NTN, this assumption is weakened and the mutual trust between satellites needs to be achieved in Xn-based handover. To accomplish this, a lightweight group key-based handover protocol is introduced in [160], where satellites in a constellation negotiate a shared group key. The protocol allows a source satellite to send a handover ticket to a UE, and the UE can directly contact the target satellite using this ticket to achieve efficiency. However, the authors do not discuss how the target satellite can obtain necessary configuration messages from the source satellite, such as service session ID and C-RNTI. An identity-based encryption method is also proposed in [161] that allows a UE to generate a symmetric key directly with a target gNB and ensure the source gNB does not know the future session key.

In 5G networks, each AMF handles the tracking area where a UE is registered. Once UEs enter a new tracking area after handover, they need to be re-authenticated by AMF through the mobility registration update, as depicted in 3GPP TS 23.502 -4.2.2. In this case, the efficiency of processing many such re-authentication requests must be considered. Group authentication was proposed to be used before or after Xn-based handover [162, 163] and N2-based handover [131, 155] to reduce message overheads. In this case, UEs form groups and are assigned a group head (GH) to aggregate the information and perform group authentication. In NTN, re-authentication is rarely expected after the handover as UEs do not move to another tracking area within a very short time. Also, the existing works assume that a GH is completely trusted and that group members do not leave a group during the handover procedure. These assumptions are unrealistic in the NTN handover scenarios due to the high mobility of LEO satellites and various malware attacks faced by UE devices especially IoT devices.

UE clustering algorithms for LEO satellite networks are discussed in [157, 158], where the handover time scheduling problem is formulated to optimize performance metrics such as handover latency and quality of experience. However, the group handover protocols are not discussed in [157, 158] and a compromised GH can interrupt services of the group

members. In our work, we consider the key characteristics of NTN and propose an Xn-based handover protocol that enables NTN gNB to flexibly configure UEs to perform group/individual handover based on real-time situations without trusting a single GH. Our work will also address the signaling storm in handover procedures that can overload NTN gNB processors.

4.3 Problem Statement

An important consideration based on the aforementioned challenges is the mobility of the LEO satellite. Compared to a 5G TN scenario where only a fraction of UEs requires handover within a short time, the rapid movement of LEO satellites (i.e., ~ 7.56 km/s [75]) requires handover for a large number of UEs within a short period of time. Because LEO satellites cover much larger areas than ground stations, the excessive handover control signals from a large number of UEs can lead to a signaling storm, which can downgrade the network performance and delay the response to handover requests. To avoid losing connectivity, UEs will usually retry their handover when they do not receive a response, which amplifies the signaling storm. To highlight this challenge, we calculate the expected number of UEs requiring handover within a cell (i.e., cell-level handover depicted in 3GPP TS 38.300-9.2.3) in a given time in the following paragraphs.

Handover scenario description: Assume that the speed of a satellite is V_{sat} , cell radius is R_c , and all satellites have the same speed and cell radius. Variable d represents the moving distance. Assuming the density of UEs is uniform, we calculate the number of UEs that must be handed off in a given period ΔT . We denote the intersection area of two circles in Fig. 4.1 as $A_{intersect}$ and the cell area as $A_{circle} = \pi R_c^2$. The red area requiring the satellite to hand off UEs, $A_{hand-off} = A_{circle} - A_{intersect}$, can be calculated as follows:

$$\begin{aligned} A_{intersect} &= 2R_c^2 \arccos(d/2R_c) - d\sqrt{R_c^2 - d^2/4} \\ A_{hand-off} &= \pi R_c^2 - 2R_c^2 \arccos(d/2R_c) + d\sqrt{R_c^2 - d^2/4}. \end{aligned}$$

Assuming a satellite connects to N UEs uniformly in a cell, the UE density is N/A_{circle} . Within ΔT , the number of UEs requiring to be handed off, $N_{hand-off}$, will be:

$$\begin{aligned} N_{hand-off} &= N \cdot A_{hand-off} / A_{circle} \\ &= N - 2N \arccos(d/2R_c) / \pi + dN(\sqrt{4R_c^2 - d^2}) / 2\pi R_c^2. \end{aligned}$$

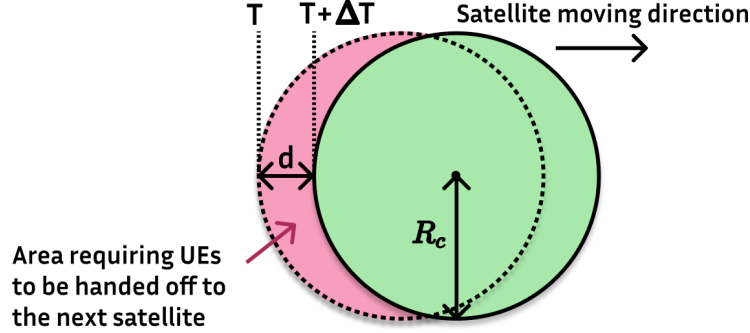


Figure 4.1: Handover scenario with Earth moving cells

In the Starlink setting [164] where $R_c \approx 12.07 \text{ km}$, $V_{sat} \approx 7.56 \text{ km/s}$, and $N = 65519$ (the maximum C-RNTI value [75]), the UE density is about 36 UEs/km^2 , requiring the satellite to handover $2.6 \times 10^4 \text{ UEs/second}$. This challenge is exacerbated in the context of mMTC, where each LEO satellite serves a massive number of UEs (i.e., 10^6 UEs/km^2 , as depicted in 3GPP TS 22.261-6.4.2), requiring the satellite to handover $1.8 \times 10^8 \text{ UEs/second}$. Moreover, Sony indicated that handover signaling storms should be addressed in the first release of 3GPP NTN solution TS 38.821 [165], while 3GPP indicates that this challenge will be addressed in the future [75].

4.4 Secure and Efficient Group Handover

4.4.1 High-level Overview

The objective of mitigating the handover signaling storm in an NTN is to reduce the workload of the source satellite (S-Satellite) by reducing its communication overhead. To achieve this goal, UEs near each other can form a group and a GH is selected to perform handover from the source satellite on behalf of the group. However, a compromised GH can maliciously start an early, late, or PingPong [166] handover to disrupt the group member's service and increase the satellite's workload. To avoid this issue, we use additive threshold secret sharing to allow the source satellite to know if the majority of the group members approve the GH's handover request. This scheme also ensures that the group handover succeeds even when some group members switch to inactive mode.

In our solution, AMF is responsible for forming UE groups in its tracking area and delivering the group information to the UEs and satellites. Different secret shares are

generated by the source satellite and delivered to each UE in the previous handover. The source satellite needs to select more than one GH from each UE group, called *Group Aggregator (GA)*, to decrease the risk of late handover and then deliver the hash commitment of the group member's share to each *GA*. Furthermore, as the UE behavior is unpredictable, the source satellite must be capable of notifying or canceling the group-based handover based on real-time situations. For example, some UEs in a group switch to idle mode or de-register from the network, making the group unsuitable for group handover. During a group handover, a non-*GA* member does not send the request to the satellite but publicly broadcasts its secret share to inform the *GA* of its intention to handover. The *GA* verifies the correctness of each received share by checking the hash commitment. After receiving enough shares above a given threshold, the *GA* can generate the group handover ticket and initialize the group handover request to the source satellite. In such a way, the *GA* cannot initialize an early or PingPong handover request. After receiving the group handover ticket, the source satellite first verifies the correctness of the group handover ticket. Then, it communicates with the target satellite (T-Satellite) to perform the handover for the group.

4.4.2 Protocol Description

In this protocol, we assume satellites and AMF can securely communicate with each other. Considering a tracking area with a dense UE deployment, the corresponding AMF places UEs in different groups based on geometric location and UE attributes, such that group members will require a handover at a similar time [158]. AMF can decide to leave some UEs out of the group if their attributes are not suitable for group handover with neighboring UEs. If a UE is assigned to a group, it receives a groupID, *GID*, from the serving AMF during the registration or periodic update. AMF also informs the satellite of the group's geographic information and the expected number of UEs when the satellite performs the tracking area identity update. Each satellite with *RAN-ID* generates a digital signature key pair (SK_{RAN-ID}, PK_{RAN-ID}).

Fig. 4.2 shows our group handover sequence diagram. The source satellite monitors the group's suitability for group handover. If suitable, the source satellite follows Alg. 4-Thread1 to notify the group members, selects the *GAs*, and sends each *GA* the required threshold and the *CommitmentMap*, which contains the hash commitments of all group members. The group member should use PK_{RAN-ID} to verify the notification message and then wait to perform the group handover. When a group member requests handover, it broadcasts the *GID* and its secret share. The *GA* follows Alg. 5 to verify the correctness

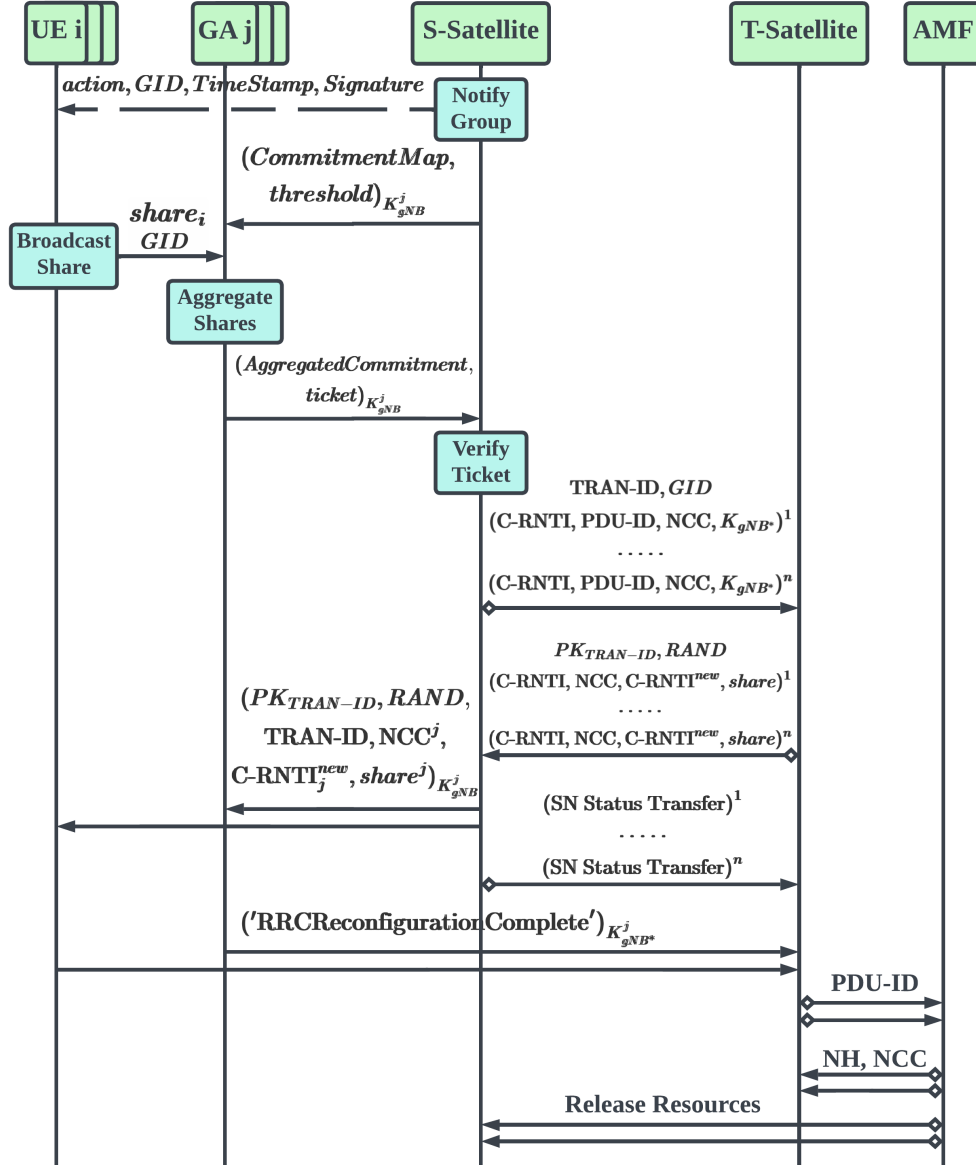


Figure 4.2: Sequence diagram of the proposed NTN handover protocol (the symbol $\diamond$ indicates a communication channel is secure; symmetric encryption with a key K is denoted $(*)_K$)

of the share, aggregate the received shares, and send the handover request with an aggregated handover ticket to the source satellite. The source satellite follows Alg. 4-Thread2 to

Algorithm 4 Source Satellite Processing Logic

Input: $RAN-ID$, $RAND$, GID , $CommitmentShareMap$, $CommitmentMap$, SK_{RAN-ID}

- 1: **Parallel Thread1:**
- 2: **while** Monitoring group GID connected number **do**
- 3: $nonce = RAN-ID || RAND || GID || TimeStamp$
- 4: **if** Group GID is suitable for group handover **then**
- 5: $action = \text{'SwitchToGroupHandover'}$
- 6: $m = action || nonce$
- 7: $Sig = \text{Sign}(m)$ using SK_{RAN-ID}
- 8: Broadcast ($RAN-ID || GID || action || TimeStamp || Sig$)
- 9: Decide $threshold$ based on the connected number
- 10: Select Group Aggregators $AG_1, \dots, AG_n$
- 11: Send ($CommitmentMap, threshold$) to each AG_i
- 12: **end if**
- 13: **if** Group GID does not need group handover **then**
- 14: $action = \text{'CancelGroupHandover'}$
- 15: $m = action || nonce$
- 16: $Sig = \text{Sign}(m)$ using SK_{RAN-ID}
- 17: Broadcast ($RAN-ID || GID || action || TimeStamp || Sig$)
- 18: **end if**
- 19: **end while**
- 20: **Parallel Thread2:**
- 21: **while** Listening to group handover request **do**
- 22: Receiving ($ticket, AggregatedCommitment$)
- 23: Get share $share_1, \dots, share_t$ in $CommitmentShareMap$ based on $AggregatedCommitment$ and $CommitmentMap$
- 24: **if** $ticket == share_1 \oplus \dots \oplus share_t$ **then**
- 25: Send group handover request to the target satellite
- 26: **end if**
- 27: **end while**

verify the correctness of the ticket and requests the target satellite to handover the group. The target satellite then returns the configuration information along with a random nonce $RAND_{TRAN-ID}$, its public key $PK_{TRAN-ID}$, and the new shares for group members. The target satellite saves the new shares and the corresponding commitments in a map called $CommitmentShareMap$. The source satellite then securely delivers the configuration,

share, $RAND_{TRAN-ID}$ and $PK_{TRAN-ID}$ to group members individually. Each group member then attaches to the target satellite through random access and uses the new K_{gNB^*} to achieve mutual trust.

4.4.3 Protocol Analysis

In this section, we analyze the proposed protocol’s security and efficiency. We also discuss potential security and efficiency trade-offs of alternative solutions.

Security: Below, we analyze three important security aspects of our protocol.

- *Privacy preserving*—The UE will only publicly broadcast the group ID and the secret. As there is no associated 5G identifier and the groupID cannot be used to distinguish group members, the UE will not leak private information that could be used, e.g., for a linkability attack.
- *Avoid early/late/PingPong group handover*—The GA cannot initialize group handover without most members agreeing to handover, so an early or malicious Ping-Pong handover can be avoided. We assume at least one aggregator is benign, which means it will send the request as soon as the ticket is ready. Therefore, a malicious GA cannot postpone the handover.
- *Replay resistant*—A compromised party may try to forge/replay broadcasting information to maliciously inform UEs to perform group handover or cancel the group handover. The signature ensures that the broadcasting information cannot be forged in our protocol. The secret random nonce and timestamp ensure that the message cannot be replayed even if satellites periodically cover the UE.

Message overhead: Consider a group GID has N_G UEs and K_G GAs. To perform a group handover, the source satellite receives K_G messages from GAs, one from the target satellite, and one from AMF, resulting in $2 + K_G$ messages. Compared to the 5G baseline handover, which requires the satellite to receive three messages per handover, our protocol significantly reduces the number of received messages from $3N_G$ to $2 + K_G$.

Security-efficiency trade-off: Below, we list some security concerns and alternative solutions that sacrifice efficiency.

- *Share distribution*—In our protocol, the shares are generated by the target satellite and sent to UE by the source satellite. If the source satellite is compromised, the attacker who obtained shares can initialize an early handover to disrupt the service. To solve this, an alternative solution is that the source satellite distributes and sends the share and group handover notification to each UE during connection by replacing

Algorithm 5 Group Aggregator Processing Logic

Input: *threshold*, *CommitmentMap*, *GID*

```
1: ticket = 0
2: AggregatedCommitment = []
3: while Monitoring broadcasting information do
4:   Receive broadcasting message (GIDi, Sharei)
5:   if GIDi == GID then
6:     Commitment = Hash(GID||RAND||Sharei)
7:     if Commitment exists in CommitmentMap then
8:       ticket = ticket  $\oplus$  Sharei
9:       i = CommitmentMap.index(Commitment)
10:      Add i to AggregatedCommitment
11:    end if
12:  end if
13:  Count = Length of AggregatedCommitment
14:  if Count > threshold then
15:    Send handover request to the source satellite attaching (ticket,
      AggregatedCommitment)
16:  end if
17: end while
```

the broadcast notification message. However, this will add a computation overhead on the source satellite and decrease the overall performance, which was verified on our simulator.

- *Share broadcasting*—In our protocol, the broadcast message from a UE is not protected and stays anonymous. A potential DoS attack on the aggregators may be carried out if the attacker knows the *GID*. The attacker can force the *GA* to keep computing the hash. To mitigate this, one alternative solution is to enable a secure machine-to-machine communication channel between group members and *GA*. However, this will introduce identity leakage and lower efficiency due to decryption and channel establishment.

4.5 Performance Evaluation

This section showcases the efficacy of our group handover protocol in mitigating signaling storms, and compares its performance to a baseline 5G handover protocol.

4.5.1 Satellite Handover Simulator

We implement our protocol on a custom simulator that relies on the popular discrete-event simulation package **SimPy**, and provide a proof-of-concept evaluation. Our simulator is designed for simulating handover protocols in NTN, allowing us to perform experiments with an arbitrary number of satellites and UEs with mobility support. UEs are implemented as state machines where each state represents a different stage in the handover protocol. The simulator measures handover performance by tracking UE states at different timesteps and can generate animations.

Message processing model: Fig. 4.3 shows our message processing model. Our simulator computes the end-to-end communication delay by computing the following delays:

- *Propagation delay:* This delay is computed based on the physical distance between the sender and receiver.
- *Transmission delay:* The size of a control signal is less than 3,000 bytes, while 5G supports large bandwidth up to 20 Gbps. In these conditions, the delay is about 1 μs .
- *Queuing delay:* Messages wait in a queue until a processor is available to handle handover requests, with the possibility of requests being dropped if the queue is full.
- *Processing delay:* We divide message processing time into physical layer, logic execution, and cryptography tasks (encryption, decryption, signatures, and hashing), favouring batch hashing for efficiency purposes.

4.5.2 Experimental Setup

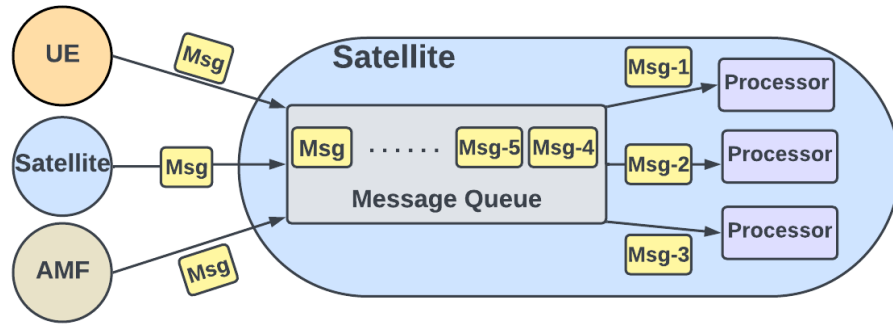


Figure 4.3: Message processing model

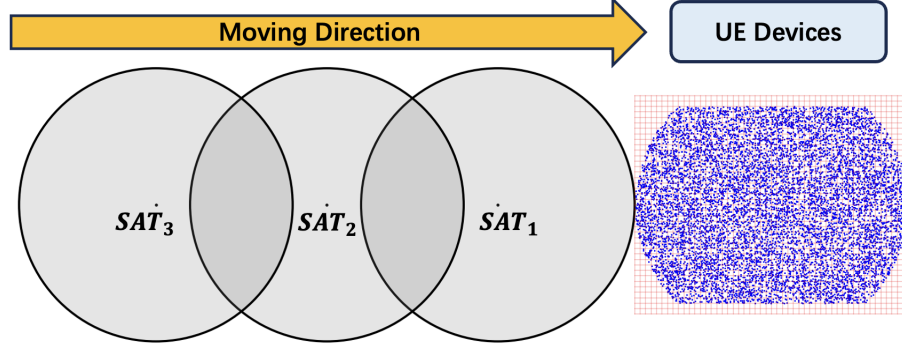


Figure 4.4: Scenario of our experiment

Table 4.1: Simulation parameters and settings

Parameter	Value	Parameter	Value
Satellite speed	7.56 km/s	HO timeout	30 ms
Footprint radius	25 km	GHO timeout	35 ms
Inter-satellite distance	30 km	Square width	1 km
Inter-satellite delay	1 ms	Physical-layer processing time	0.05 ms
Ground-satellite delay	3 ms	Logic execution time	0.05 ms
5G core-satellite delay	10 ms	Encryption/Decryption	0.1 ms
Transmission delay	1 μ s	Signature sign/verify	0.3 ms
Message queue size	500	Single hash	0.05 ms
Max re-transmission	15	Batch hash	0.1 ms
Packet size	3,000 bytes	Number of processors	4

We now outline our scenario and introduce the configurations for the baseline handover and group handover protocols.

Handover scenario: Without loss of generality, we deploy three satellites with one logical cell each to simulate a generalized and continuous inter-satellite handover scenario, as shown in Fig. 4.4. To better study the relationship between the density and handover performance, we randomly deploy 10,000 to 70,000 UEs and assign them to fixed square groups. We assume all UEs are connected to SAT_1 at the start of the simulation, and we report the performance of UE handover from SAT_1 to SAT_2 . Our simulation also incorporates optimizations expected in real deployments, such as prioritizing messages, including messages from other satellites, responses from the core, and attachment requests from UE. Furthermore, we assume that a UE will require a handover when it is closer to another satellite, which is a location-based handover trigger [75]. We ran the simulation five times with different seeds (10, 20, 30, 40, 50) to generate multiple UE deployment

configurations.

Baseline HO vs. GHO protocol: In the baseline handover protocol (denoted HO) [159, 59], a UE individually sends handover requests to the source satellite. Upon receiving the request, this satellite asks the target satellite for configuration data and then delivers it back to the UE. In our group handover protocol (denoted GHO), we allow the source satellite to send UE requests in batches to reduce message overhead.

We ran simulations with both handover protocols, and Table 4.1 shows the parameters we used for each protocol. For GHO, the ground space is partitioned into square areas, each with a fixed width. Within each square, UEs are organized into groups. Both HO and GHO are configured with a re-transmission timeout so that a UE will re-transmit the handover request after waiting a pre-defined time.

4.5.3 Experimental Results

Table 4.2: Simulation results (baseline HO vs. GHO)

UE Number	Success Rate (%)		Total Messages		UE Messages		Drop rate (%)		Successful UE WT (ms)*		Failed UE WT (ms)**	
	HO	GHO	HO	GHO	HO	GHO	HO	GHO	HO	GHO	HO	GHO
10,000	100.00	100.00	30,000	16,558	10,000	4,787	0.00	0.00	8.99±0.00	16.1±0.18	—	—
20,000	100.00	100.00	60,000	26,820	20,000	5,007	0.00	0.00	9.24±0.00	19.45±0.15	—	—
30,000	100.00	100.00	102,060	36,938	42,060	5,089	0.98	0.00	27.94±0.08	23.08±0.13	—	—
40,000	81.25	100.00	414,555	47,379	349,456	5,355	74.53	0.00	214.23±0.70	25.55±0.12	858.48±3.35	—
50,000	66.56	100.00	602,981	67,275	536,317	12,497	82.28	0.00	246.20±0.71	71.25±0.21	887.84±5.22	—
60,000	56.26	92.86	777,770	121,982	710,143	47,664	86.17	1.09	263.77±0.72	338.67±0.93	901.55±1.66	437.15±2.03
70,000	48.70	82.70	943,105	137,979	874,797	58,135	88.57	2.60	274.47±0.72	482.52±1.19	909.38±1.40	568.81±1.68

* The time difference between when sending a handover request and when the receiving the configuration response.

** The time difference between when sending a handover request and when losing connectivity without receiving the configuration response.

Table 4.2 shows the results of our experiment for SAT_1 in responding to handover requests. The success rate indicates the fraction of UEs that successfully complete the handover procedure. We also include the total messages the satellite received and the total UE handover/retransmit request messages. The drop rate indicates the number of dropped messages with respect to the total messages. UE waiting time is the difference between when UE sends a handover request or broadcasts the share and when the UE receives the configuration from the source satellite. Next, we discuss our main findings.

GHO achieves a higher successful handover rate: In HO, the satellite is capable of processing the request when the number of UEs are 10,000 and 20,000. However, at 30,000

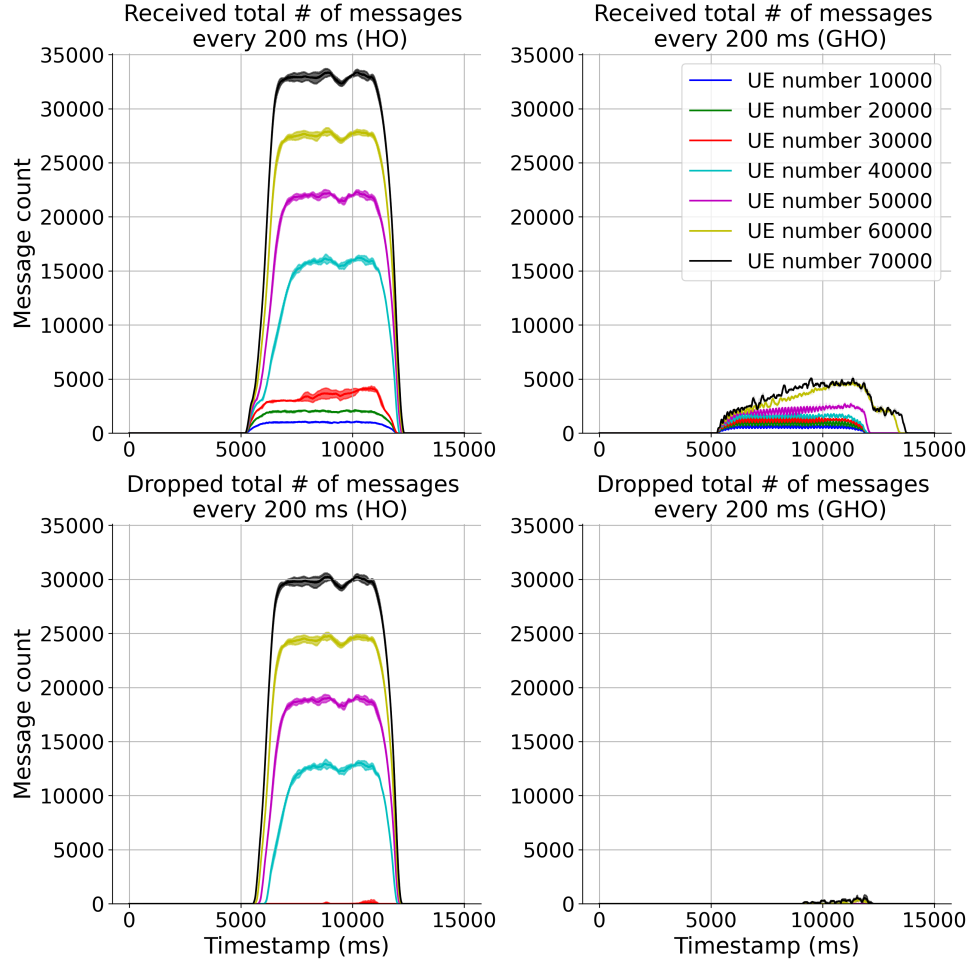


Figure 4.5: Message count every 200 ms

UEs, the UE waiting time increases, indicating that the UE started to re-transmit handover requests. We see the total received messages are increasing non-linearly, and the satellite drops messages, indicating a signaling storm overwhelms the satellite. Consequently, from 30,000 UEs to 70,000 UEs, the handover success rate decreases, and the UE waiting time increases. For GHO, we see that the total messages are smaller than in HO, which reduces the burden on the satellite, and the success rate is much better than HO. However, at 60,000 UEs, GHO also experiences delay because SAT_2 needs to take over a massive number of UEs and delays the response to the messages from SAT_1 , which is inevitable in a handover protocol. This is because a UE must individually access the satellite and achieve mutual

trust using the new session key. We also observe an overall higher waiting time in GHO when the satellite is not overwhelmed.

GHO reduces satellites’ message processing burden: To analyze this dynamic process, we also collected the received and dropped messages every 200 ms through time. Fig. 4.5 shows the result from SAT_1 . Overall, due to the circular shape of a satellite’s coverage, the total number of received messages and drop rate gradually increase and decrease. SAT_1 starts to handover UEs at 5,300 ms, and SAT_2 starts to handover UEs at 9,500 ms. We can see that when SAT_1 is overloaded and drops messages, the total number of messages dramatically increases because UE re-transmission messages amplify the signaling storm. Compared to HO, our GHO protocol significantly reduces the number of received messages. Even when the satellite is overloaded, as the number of GAs is much smaller than the total UEs, the re-transmission messages do not cause a significant amplification of messages in GHO.

4.6 Conclusions

We examined the signaling storm problem in 5G NTN handover and proposed a secure and efficient Xn-based GHO protocol to mitigate this issue. We implemented a discrete-event simulator to demonstrate that our GHO protocol helps reduce message overhead and handover additional UEs. However, we find that the overall handover process is still afflicted by a large number of access requests. Formal security verification and evaluation with multi-cells will be performed in the future. We will also study GHO for NTN with an Earth-fixed cell, where massive UEs in a cell will require handover simultaneously. Moreover, we will investigate signaling storm due to conditional handover protocol in 5G NTNs.

Chapter 5

Signalling Load-aware Conditional Handover in 5G Non-Terrestrial Networks

5.1 Introduction

Terrestrial cellular networks cover only about 15% of the Earth’s surface, leaving vast areas, including seas, mountains, and rural regions without coverage. These areas incur high capital and operational expenses with expected unequal revenue. LEO satellite constellations, as a key part of a 5G NTN, are expected to complement a 5G terrestrial network (TN), offering low-latency services everywhere. This ubiquitous coverage can greatly benefit verticals, such as environmental monitoring, ocean freight, fishery, oil exploration, farming, and aviation. It also supports global outdoor adventures and emergency services. Additionally, the 5G NTN can enhance TN by providing additional bandwidth to local areas. In the event of TN outages due to disasters, the NTN can temporarily provide services and aid in disaster recovery. According to Global Market Insights Inc., the market value of 5G NTN was 4.2 billion USD in 2023 and is projected to reach 79.8 billion USD by 2032 [167].

As the US Department of Defense states [10], “5G is a critical strategic technology: those nations that master advanced communications technologies and ubiquitous connectivity will have a long-term economic and military advantage.” Recognizing the importance of this mission and to achieve the aforementioned unprecedented applications, mobile operators and satellite operators are working closely to achieve 5G NTN. In 5G NTN, satellites should carry regenerative payload to serve UEs as RAN. The regenerative payload

enables encoding, modulation, switching, and routing, which could effectively offer gNB onboard to meet the 5G requirements [168]. It will also support inter-satellite communications through inter-satellite link (ISL). Recently, T-Mobile and SpaceX have launched LEO satellites with onboard 4G LTE eNodeB to provide Direct to Cell services. Although they currently only support SMS service, they expect to support calling and data services for mobile phones and IoT devices with common 4G LTE standards by 2025 [168]. This is a significant milestone towards achieving global cellular services and an important step towards 5G NTN.

To achieve 5G NTN, numerous challenges exist, with one drawing significant attention from the industry: the signalling storm that occurs during the handover of many UEs [75, 169]. In 5G NTN, handover can be roughly classified into intra-satellite handover and inter-satellite handover, and the satellite cell could be operating in an earth-moving cell and earth-fixed cell. For intra-satellite, the handover necessarily depends on the cell movement configuration and beam management of a satellite gNB [80]. However, the service link switch between satellites is unavoidable in any scenario [169]. Thus, this work focuses on inter-satellite handover, also known as 5G Xn handover, between two satellite gNBs. There are two types of Xn handover protocol: BHO and the conditional handover protocol (CHO). In BHO, the source satellite communicates with only one satellite, and the UE detaches from the source and accesses the target immediately after receiving the configuration. In CHO, the source satellite communicates with more than one satellite and sends the UE a condition to access the target satellite. The UE maintains the connection with the source satellite until the condition is met. Compared to the baseline handover, CHO significantly increases the robustness of the handover procedure by preparing more target satellites and decoupling the preparation and the execution phases.

Considering the Xn handover, there are some remarkable differences between 5G TN and 5G NTN. First, the LEO satellite gNB covers a much larger area than the TN gNB. Due to the high mobility of satellites, moving at speeds around 7.56 km/s, the satellite gNB may need to handover the service link to other satellite gNBs for many UEs in a short time. This could cause a sudden increase in signalling intensity, leading to performance degradation and affecting all ongoing sessions. Also, the random access (RA) requests from many UEs will significantly increase the possibility of preamble collision, leading to access failure. Second, due to the larger cell size, RA occasion becomes longer to tolerate the delay difference between near cell center UE and near cell edge UE. This equivalently decreases the RA opportunities within a unit of time. Third, because of the high altitude, the signal strength difference is small between the cell center and cell edge. This makes the signal strength less useful for UE handover measurement compared to using location information. Last but not least, these satellites have limited computing,

radio and energy resources. With the future deployment of more UEs, targeted by 5G to reach densities of 10^6 UEs/km², as suggested in 3GPP TS 22.261-6.4.2, optimizing the inter-satellite handover should be a priority in 5G NTN.

Previous 5G NTN studies focus on reducing the handover count or transmitted signals, but overlook that signalling intensity causes performance degradation. In this work, we aim to decrease the signalling peak in CHO to avoid a signalling storm. Although CHO enhances the robustness of the handover, it increases signalling overhead and leads to a higher probability of signalling peaks. We argue that the increased inter-satellite communications could allow individual satellites to obtain more accurate signalling load information of the constellation if they can predict and share the load in real-time. In this way, each satellite could perform local optimization to decrease the signalling peak globally. Our main contributions are as follows:

1. We mathematically formulate CHO to alleviate signalling storm in 5G NTN. To the best of our knowledge, this is the first comprehensive formulation of CHO.
2. We propose an efficient, signalling load-aware heuristic for large-scale scenarios, and evaluate the effectiveness against a set of baseline CHO schemes. Our protocol significantly reduces signalling peaks and balances the load, enhancing the robustness and efficiency of handover in 5G NTN.
3. We evaluate the performance on a customized discrete-event simulator for 5G NTN, made publicly available at <https://github.com/zbh888/congestionH0>.

5.2 Background and Related Works

We first present the workflow of CHO and RA. Then, we describe related work and identify the research gaps we address.

5.2.1 Conditional handover protocol

This section presents CHO in 5G NTN following 3GPP TS 38.300-9.2.3. The signalling flow is depicted in Fig. 5.1, where **Phase1** represents CHO preparation phase and **Phase2** represents CHO execution phase. CHO works as follows:

1. The UE measures multiple cells until the handover triggering condition is satisfied. The UE sends the source satellite a **Measurement Report** containing multiple best cells.

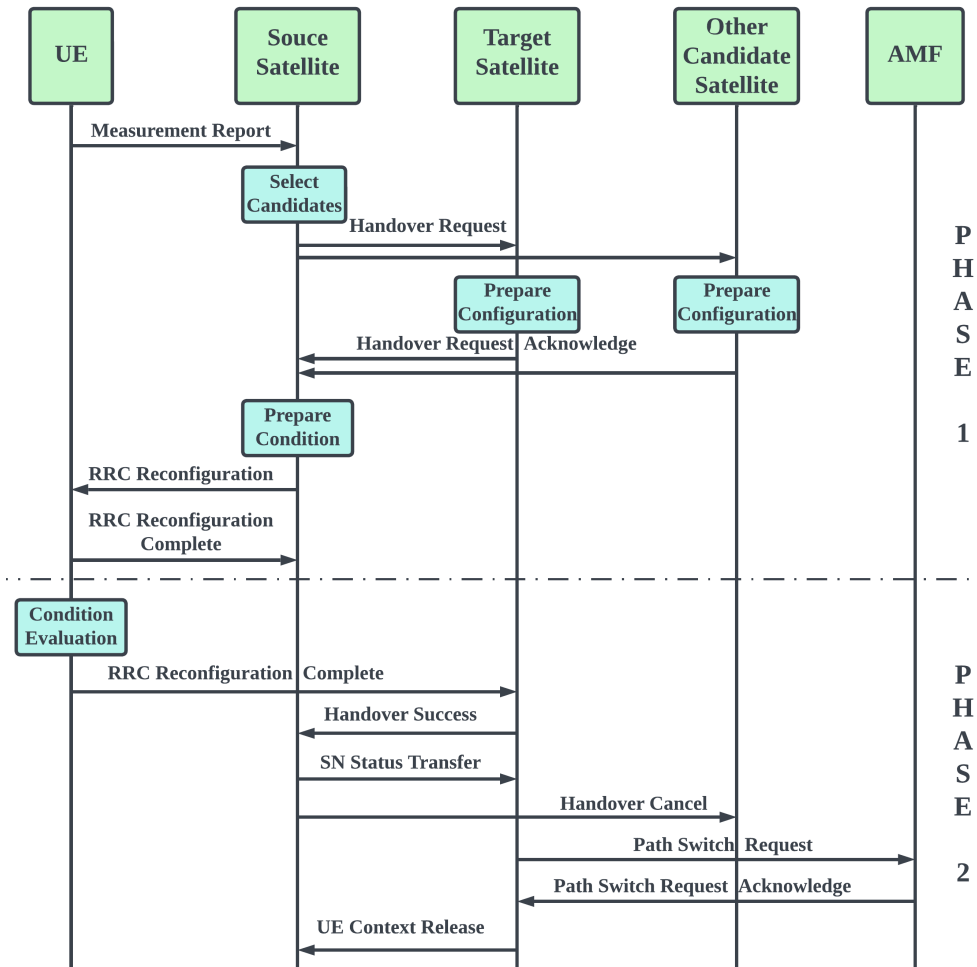


Figure 5.1: Conditional handover protocol

2. The source satellite selects multiple candidate satellites from the **Measurement Report**. A **CHO Handover Request** is sent to each candidate.
3. The candidate satellites reserve resources for UEs and reply to the source satellite with **Handover Request ACK**, containing the new UE configuration.
4. The source satellite sends UE an **RRC Reconfiguration** message containing the candidates configurations and CHO execution conditions to access candidates.
5. The UE sends an **RRC Reconfiguration Complete** message to the source satellite, acknowledging the reception. This completes **Phase1**.
6. The UE maintains the connection with the source satellite and starts evaluating the

CHO execution conditions for the candidates. If the condition is satisfied, the UE detaches from the source satellite, applies the stored corresponding configuration for that selected target cell, synchronizes to that candidate cell, and completes CHO by sending **RRC Reconfiguration Complete** message to the target satellite. The UE releases stored CHO configurations after completing the handover procedure.

7. The target satellite sends the source satellite **Handover Success** message after the UE has successfully accessed the target satellite. The source satellite replies with the **SN Status Transfer** message. The source satellite also sends the other candidates **Handover Cancel** message to release their resource reservation.
8. The target satellite requests AMF for a path switch. After receiving the response, the target satellite asks the source satellite to release the UE context, which completes **Phase2**.

5.2.2 Random access procedure

In CHO, the RA procedure is used when the UE needs to access the target gNB, involving uplink synchronization with gNB. There are two types of RA procedures, contention-based RA (CBRA) and contention-free RA (CFRA). In CBRA, the UE selects the RA preamble and sends it to the gNB in Msg1, as shown in Fig. 5.2(A). Then, gNB sends an RA response in Msg2, containing temporary C-RNTI and UL-Grant, that are used for the UE to transmit the next message while maintaining the same identity in Msg1. Next, the UE sends the gNB Msg3, which contains the purpose of initializing RA procedures such as handover, initial connection, or link re-establishment. Because temporary C-RNTI is calculated using the time and RA preamble, if two UEs send the same preamble simultaneously (i.e., same RA occasion), then two UEs will receive the same identity in Msg2. In this case, transmitting Msg4 indicates the success of the procedure. Without Msg4, the UE considers a RA failure due to preamble collision and restarts CBRA.

In CFRA, as shown in Fig. 5.2(B), the dedicated preamble is assigned to the UE in advance, so the random access will not fail due to preamble collision. In CHO, a dedicated preamble assignment could be included in the configuration in **Handover Request ACK**. The Msg1 is included in the **RRC Reconfiguration Complete** message.

5.2.3 Related works

Previous works investigated CHO in TN networks [170, 171, 172]. In [170], authors focused on reducing unnecessary CHO measurement report transmission in TN by applying a

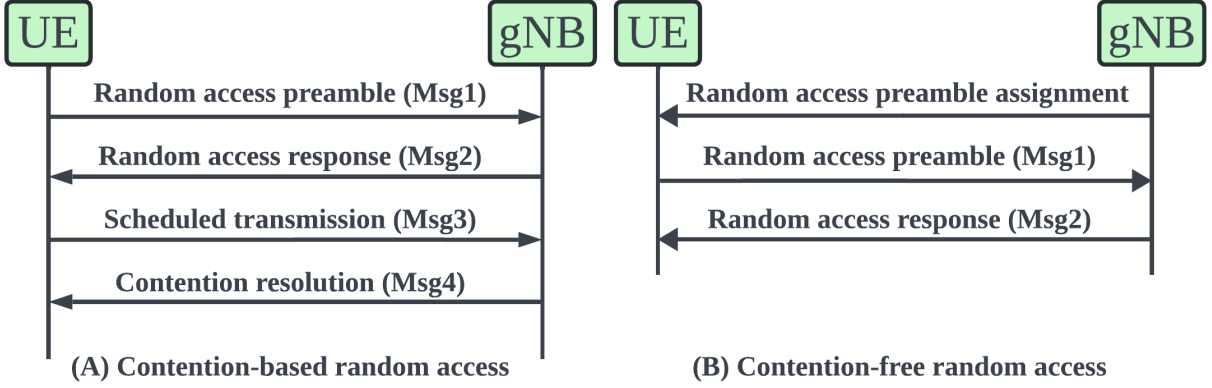


Figure 5.2: Random access procedure

convolutional neural network-based classifier on the UE. This classifier assists the UE in making handover decisions based on signal strength. In [171], authors evaluated fast CHO in TN with small cell gNBs, allowing for candidate and UE to reserve configuration after CHO success. This enables the UE to access a pre-prepared candidate if needed, thereby reducing signalling overhead. However, fast CHO may be unnecessary in 5G NTN due to large satellite coverage and corresponding long reservation time. In [172], authors proposed allowing the UE to request candidate gNBs to update RA resources, including preambles, to adapt the CHO decouple time and achieve more robust CFRA in CHO. However, this introduces extra signalling messages.

Other works focused on either BHO or CHO in NTN [72, 81, 118]. In [72], the authors modeled the mega-satellite constellation and deduced that the rate of handovers significantly increased with the number of satellites. In [81], the authors focused on a multi-satellite provider scenario and proposed replacing the inter-satellite handover procedure with an access token. Although this reduces inter-satellite communication, the massive RA requests are expected to cause a signalling storm. Additionally, removing handovers significantly impacts session continuity and network resource management, which was not addressed. In [118], the authors addressed handover delay and CBRA collisions in BHO, proposing a deep reinforcement learning model for the source satellite to predict resources on the target satellite. In this design, the UE does not send a measurement report. Instead, the trained model in the satellite decides whether to perform handover and distribute requests based on predicted resources on the target satellite. However, the scalability of this method is not assessed in a large-scale deployment and the consequences of removing measurement reports are not discussed.

Most similar to our work, BHO or CHO was considered as an optimization problem with different objectives and solved using heuristics or reinforcement learning (RL) methods in [173, 174, 175]. In [173], the authors proposed minimizing the number of handovers while maintaining available communication channels in BHO using multi-agent RL. In this approach, the UE, as an agent, needs to obtain global information from other UE agents to decide handover actions. Their experiment, using only six UEs, showed a reduction in the average number of handovers. As an incremental work, [174] built upon their approach, focusing on channel balancing. However, a complete optimization formulation was not provided, and dynamically obtaining other UEs' information along with the RL algorithm is impractical for scalability. Also, the authors did not consider that it is the signalling intensity, not the total number of handovers, that causes the signalling storm. In [175], the authors considered path loss and aimed to maximize total throughput while maintaining bandwidth capacity in CHO by proposing a two-stage target selection algorithm with long-term considerations. Although their heuristic increases throughput, it only considers individual UE optimization without accounting for all UEs, making it unclear how to maintain capacity constraints. Due to their time slot of 70s-150s, they cannot consider signalling intensity and real-time control.

Our work fills the research gap of the non-existent comprehensive CHO formulation and the non-existent study of CHO signalling intensity. In Chapter 4, we proposed a group handover protocol for BHO to lower signalling intensity [2]. However, we observed that the massive number of individual RA procedures could cause gNB performance degradation. In this work, we use the decoupling feature of CHO and CFRA to avoid this issue. Although this work focuses on optimizing individual CHO, we anticipate that group mobility could further reduce signalling intensity in CHO [169], which we plan to explore in the future.

5.3 Mitigating Signalling Storm

In this section, we start by defining our problem and assumptions on CHO. Then, we present a comprehensive mathematical model of CHO in 5G NTN.

5.3.1 Problem

Given a set of UEs $\mathcal{U} = \{u_1, u_2, \dots\}$ and a set of satellites $\mathcal{S} = \{s_1, s_2, \dots\}$. We divide the time to be a list of consecutive discrete time slots $\mathcal{T} = \{t_1, t_2, \dots\}$. As time goes, the relative locations among UEs and satellites will change due to mobility, and the serving

relation will change following CHO. Thus, among all $s \in \mathcal{S}$ in each $t \in \mathcal{T}$, there will be a maximum signalling count, representing the total number of signalling messages that some $s \in \mathcal{S}$ received at some $t \in \mathcal{T}$. We are aiming to minimize this number. In this way, we minimize the signalling peak of the satellite constellation. In the context of CHO, we make the assumptions and definitions considering real-world constraints and model simplicity as follows:

1. The UE measures and reports all the possible satellites that could provide cellular service in **Phase1**. Because the change of signal strength in 5G NTN is not as significant as in the TN scenario [75], we consider location-based handover triggering such that the UE carrying constellation ephemeris, will trigger handover when the distance between UE and the center of the satellite coverage exceeds a threshold D .
2. The source satellite in **Phase1** will select a fixed number of candidates, $N_{candidate}$.
3. We use CFRA to avoid RA failure. Thus, we follow 3GPP[169] to consider a timer-based execution condition [169]. Because dedicated RA preambles and RA occasions are limited resources [75], we let MAX_{access} be the maximum access opportunities of $s \in \mathcal{S}$ in each $t \in \mathcal{T}$. In practice, the value is dependent on the configuration of satellites.
4. The condition prepared by the source satellite is defined as a combination of access priorities and timer. The source will suggest the best target satellite and the UE will access the satellite using the target satellite-defined timer. In practice, upon access failure, the UE could use CBRA or follow [172] to perform a CFRA to other candidates.
5. Signalling count in each $t \in \mathcal{T}$ will not introduce congestion delay. This helps us to compare different algorithms fairly. So, **Phase1** or **Phase2** should finish within each time slot $t \in \mathcal{T}$. And **Phase1** and **Phase2** for each $u \in \mathcal{U}$ cannot happen within the same time slot. The decoupling time between **Phase1** and **Phase2** for each $u \in \mathcal{U}$ is bounded by a predefined threshold $T_{decouple}$. This helps to avoid situations where the source satellite fails to provide services to the UE during the decoupling time due to satellite mobility. Additionally, with longer decoupling times, the UE may request to replace or remove the previous measurement report, leading to unnecessary signalling between the source and candidate satellites [176]. Therefore, we assume no additional reconfiguration will occur within $T_{decouple}$.

5.3.2 Mathematical formulation

We summarize the inputs and variables of our problem formulation in Table 5.1. For inputs, let $c_{u,s}^t \in \{0, 1\}$ denote the location relation between the UE u and the coverage

Table 5.1: Inputs and variables

Notation	Meaning
u	UE device
s	satellite
t	time slot
(INPUTS)	
D	distance threshold for location-based handover triggering
$l_{u,s}^t$	$l_{u,s}^t=1$: u will trigger handover if connected to s at t
$c_{u,s}^t$	$c_{u,s}^t=1$: u can be served by s at t without handover
MAX_{access}	maximum RA threshold
$T_{decouple}$	maximum decoupling time threshold
$N_{candidate}$	the number of candidates, including the target satellite
(DECISION VARIABLES)	
$m_{u,s}^t$	$m_{u,s}^t=1$: u is (implicitly) served by s at t
$x_{u,s}^t$	$x_{u,s}^t=1$: s is the source satellite of u in Phase1 at t
$h_{u,s}^t$	$h_{u,s}^t=1$: s is the target satellite of u in Phase1 at t
$y_{u,s}^t$	$y_{u,s}^t=1$: s is the other candidate satellite of u in Phase1 at t
$\hat{x}_{u,s}^t$	$\hat{x}_{u,s}^t=1$: s is the source satellite of u in Phase2 at t
$\hat{h}_{u,s}^t$	$\hat{h}_{u,s}^t=1$: s is the target satellite of u in Phase2 at t
$\hat{y}_{u,s}^t$	$\hat{y}_{u,s}^t=1$: s is the other candidate satellite of u in Phase2 at t
$\delta_{u,s}^t$	$\delta_{u,s}^t=1$: s cannot be the other candidates of u in Phase1 , and s can be the other candidates of u in Phase2 (serves as a lock)

center of the satellite s at time t , where $c_{u,s}^t = 1$ if the distance is less than or equal to D , and $c_{u,s}^t = 0$ otherwise. Thus, $c_{u,s}^t = 1$ implicitly shows u can be served by s at time t , and $c_{u,s}^t = 0$ otherwise. To simplify, we let $l_{u,s}^t \in \{0, 1\}$ denote if the handover triggering condition of u from source s is satisfied, i.e., $l_{u,s}^t = 1$ if and only if $c_{u,s}^t = 1$ and $c_{u,s}^{t+1} = 0$. When $l_{u,s}^t = 1$, u and s should perform **Phase1** if u is served by s .

For decision variables, we first define $m_{u,s}^t \in \{0, 1\}$ to denote the actual or implicit serving relationship between u and s . For instance, $m_{u,s}^t = 1$ could represent an actual serving relationship when u is being served by s at t before and during **Phase1**, or an implicit serving relationship after s is selected to be the target satellite of u in **Phase1** but t is within $T_{decouple}$, and $m_{u,s}^t = 0$ otherwise. $m_{u,s}^t$ follows the property in (5.1). For simplicity, we do not consider dual connectivity and assume that across all s at any given t , there is only one actual/implicit serving satellite. Constraint (5.2) ensures u must always, actually or implicitly, be served by a satellite with distance to the coverage center smaller or equal to D .

$$\sum_{s' \in \mathcal{S}} m_{u,s'}^t = 1, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (5.1)$$

$$m_{u,s}^t \leq c_{u,s}^t, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.2)$$

Phase1: Let $h_{u,s}^t \in \{0, 1\}$ denote the implicit handover action of u to s , where $h_{u,s}^t = 1$ if s ,

as target satellite, will be the next serving satellite of u , and $h_{u,s}^t = 0$ otherwise. We define (5.3) to ensure that u can implicitly handover to only one target satellite or not perform implicit handover at t . Let $x_{u,s}^t \in \{0, 1\}$ denote handover action of u from s , where $x_{u,s}^t = 1$ if s is the source satellite, and $x_{u,s}^t = 0$ otherwise. Constraint (5.4) ensures that implicit handover to the target satellite happens at the same time when u sends a measurement report to the source satellite. We can see that $h_{u,s}^t = 1$ only if $c_{u,s}^t = 1$, as denoted in (5.5).

$$\sum_{s' \in \mathcal{S}} h_{u,s'}^t \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (5.3)$$

$$\sum_{s' \in \mathcal{S}} h_{u,s'}^t = \sum_{s' \in \mathcal{S}} x_{u,s'}^t, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (5.4)$$

$$h_{u,s}^t \leq c_{u,s}^t, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.5)$$

Also, the handover will change the actual/implicit serving satellite. Constraint (5.6) captures that, at t , if no handover happened at $t - 1$ (i.e., $(\sum_{s' \in \mathcal{S}} h_{u,s'}^{t-1}) = 0$), the serving relation will not change. Otherwise, the serving satellite should switch.

$$\begin{aligned} m_{u,s}^t &= \left(\sum_{s' \in \mathcal{S}} h_{u,s'}^{t-1} \right) \times h_{u,s}^{t-1} \\ &+ \left(1 - \sum_{s' \in \mathcal{S}} h_{u,s'}^{t-1} \right) \times m_{u,s}^{t-1}, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \end{aligned} \quad (5.6)$$

We let $y_{u,s}^t \in \{0, 1\}$ denote the candidates selection other than the target satellite, where $y_{u,s}^t = 1$ when the source satellite selects s as candidate during the handover of u , $y_{u,s}^t = 0$ otherwise. Constraints (5.7) and (5.8) ensure that the selected candidate cannot be source satellite and target satellite.

$$y_{u,s}^t + h_{u,s}^t \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.7)$$

$$y_{u,s}^t + x_{u,s}^t \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.8)$$

In turn, (5.9) ensures that the number of selected candidates matches the pre-defined candidate number. Moreover, it restricts candidate selection to only take place during handover. As denoted in (5.10), $y_{u,s}^t = 1$ only if $c_{u,s}^t = 1$.

$$\sum_{s' \in \mathcal{S}} y_{u,s'}^t = (N_{\text{candidates}} - 1) \sum_{s' \in \mathcal{S}} h_{u,s'}^t, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (5.9)$$

$$y_{u,s}^t \leq c_{u,s}^t, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.10)$$

Phase2: One challenge of modelling the conditional handover is to maintain the satellite roles between **Phase1** and **Phase2** across time. We define $\hat{x}_{u,s}^t \in \{0, 1\}, \hat{y}_{u,s}^t \in \{0, 1\}, \hat{h}_{u,s}^t \in \{0, 1\}$ to denote the satellite roles in **Phase2**, which correspond to their roles in **Phase1**. We first discuss the case for target satellite. We let $\hat{h}_{u,s}^t = 1$ when s is the target satellite in **Phase2** to which u performs random access, and $\hat{h}_{u,s}^t = 0$ otherwise. In our model, the target satellite will eventually be the next serving satellite and continue to serve UE until $l_{u,s}^t = 1$. Therefore, any UE u can select s as target satellite only once in **Phase1** throughout the simulation time. We also ensure that u random access the same target satellite only once during the simulation time, which is enforced in (5.11). Obviously, $\hat{h}_{u,s}^t = 1$ only if $c_{u,s}^t = 1$, as denoted in (5.12). To ensure that u cannot access s while performing handover from s , we define (5.13).

$$\sum_{t' \in \mathcal{T}} \hat{h}_{u,s}^{t'} = \sum_{t' \in \mathcal{T}} h_{u,s}^{t'}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (5.11)$$

$$\hat{h}_{u,s}^t \leq c_{u,s}^t, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.12)$$

$$\hat{h}_{u,s}^t + x_{u,s}^t \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.13)$$

We also need to ensure that the difference between the selection time in **Phase1** and the accessing time in **Phase2** is within $T_{decouple}$. We take advantage of the fact that s can be selected and accessed by u only once to simplify the model. In (5.14), the access time cannot be the same as the selection time, while (5.15) ensures that the access follows the selection when s was selected by u as target satellite at some time. If s was not ever selected by u as target satellite, then the difference is 0.

$$\hat{h}_{u,s}^t + h_{u,s}^t \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.14)$$

$$0 \leq \sum_{t' \in \mathcal{T}} (t' \times \hat{h}_{u,s}^{t'}) - \sum_{t' \in \mathcal{T}} (t' \times h_{u,s}^{t'}) \leq T_{decouple}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (5.15)$$

Because the random access preambles are limited, we define MAX_{access} as the maximum random access that a satellite can allow at any time. Constraint (5.16) enforces this property.

$$\sum_{u' \in \mathcal{U}} \hat{h}_{u',s}^t \leq MAX_{access}, \forall t \in \mathcal{T}, s \in \mathcal{S} \quad (5.16)$$

Next, we discuss the source satellite in **Phase2**. Because s can be source satellite to u only once in **Phase1** and **Phase2**, the modeling is similar. Following (5.11) and (5.15), we

define (5.17) and (5.18).

$$\sum_{t' \in \mathcal{T}} \hat{x}_{u,s}^{t'} = \sum_{t' \in \mathcal{T}} x_{u,s}^{t'}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (5.17)$$

$$0 \leq \sum_{t' \in \mathcal{T}} (t' \times \hat{x}_{u,s}^{t'}) - \sum_{t' \in \mathcal{T}} (t' \times x_{u,s}^{t'}) \leq T_{decouple}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (5.18)$$

It is also important to ensure time consistency in **Phase2**. We define constraint (5.19) to ensure this property.

$$\sum_{s' \in \mathcal{S}} \hat{x}_{u,s'}^t = \sum_{s' \in \mathcal{S}} \hat{h}_{u,s'}^t, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (5.19)$$

Lastly, we must maintain the role for candidates. This is especially challenging because u may be related to the same candidate multiple times throughout the simulation time. So, the modeling methodology for the source and target satellites does not fully apply. As a result, we had to increase the model complexity. First, we ensure the time consistency and correct number of candidates using (5.20).

$$\sum_{s' \in \mathcal{S}} \hat{y}_{u,s'}^t = (N_{candidates} - 1) \sum_{s' \in \mathcal{S}} \hat{h}_{u,s'}^t, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (5.20)$$

Also, although s can be a candidate satellite of u multiple times, the number of times s is a candidate of u in **Phase1** should match in **Phase2**, which is enforced in (5.21).

$$\sum_{t' \in \mathcal{T}} \hat{y}_{u,s}^{t'} = \sum_{t' \in \mathcal{T}} y_{u,s}^{t'}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (5.21)$$

To ensure the ordering correctness, we utilize an ancillary binary variable $\delta_{u,s}^t = \{0, 1\}$, which acts as a lock. When $\delta_{u,s}^t = 1$, s cannot be a candidate of u in **Phase1**, but can be a candidate of u in **Phase2**, and $\delta_{u,s}^t = 0$ otherwise. This functionality is enforced using constraints (5.22) and (5.23).

$$y_{u,s}^t \leq (1 - \delta_{u,s}^t), \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.22)$$

$$\hat{y}_{u,s}^t \leq \delta_{u,s}^t, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.23)$$

To make the lock behave as expected we introduce (5.24), where $\delta_{u,s}^t$ switches to 0 at t when s is a candidate of u in **Phase2** at $t - 1$, and $\delta_{u,s}^t$ switches to and continue to be 1 when s is a candidate of u in **Phase1** at $t - 1$. Moreover, (5.25) is necessary to ensure **Phase2** will always happen after **Phase1**.

$$\delta_{u,s}^t = (1 - \hat{y}_{u,s}^{t-1})(\delta_{u,s}^{t-1} + y_{u,s}^{t-1}), \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (5.24)$$

$$\delta_{u,s}^{t=0} = 0, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (5.25)$$

Combining constraints (5.1)-(5.25), a feasible solution whose logic follows CHO and our assumptions is guaranteed. As our objective is to avoid signalling peak, we minimize the total signalling received by each satellite at each time slot, as shown in (5.26). The coefficient before each decision variable is based on the received signalling count in Fig. 5.1.

$$\begin{aligned} \min_{m,y,h,\hat{x},\hat{y},\hat{h},\delta} \max_{u' \in \mathcal{U}} \sum & ((2 + N_{candidates})x_{u',s}^t + y_{u',s}^t \\ & + h_{u',s}^t + 2\hat{x}_{u',s}^t + \hat{y}_{u',s}^t + 3\hat{h}_{u',s}^t), \forall t \in \mathcal{T}, s \in \mathcal{S} \end{aligned} \quad (5.26)$$

The formulated mixed integer programming problem has a large number of variables, corresponding to the number of UEs, the number of satellites, and the number of time slots, i.e., in the order of $(|\mathcal{U}| \times |\mathcal{S}| \times |\mathcal{T}|)$. Indeed, the problem grows as the order increases, which can lead to intractable solving times, even for small-scale scenarios. We observe that in most of the time slots, however, the handover will not happen because UE is being continuously served by one satellite and no decisions are required. Therefore, we can remove those time slots to reduce the number of variables.

We also took advantage of the fact that one UE can be served by one satellite only once. This means the handover decisions in **Phase1** can only be made when $l_{u,s}^t = 1$, and there is only one time slot for any (u, s) pair across all time slots that has $l_{u,s}^t = 1$. Thus, we can remove those t where $\sum_s l_{u,s}^t = 0$ for any u . This further reduces the number of variables to the order of $(|\mathcal{U}| \times |\mathcal{S}|^2)$. Considering the decoupling time for **Phase1** and **Phase2**, we could extend each of those filtered time slots to $T_{decouple}$ time slots to allow **Phase2** decisions. Overall, this reduces the number of variable to the order of $(|\mathcal{U}| \times |\mathcal{S}|^2 \times T_{decouple})$, which is independent of the number of simulation time slots.

5.4 Load-aware Conditional Handover

5.4.1 Overview

Due to the formulation complexity, requirement of a centralized entity and the dynamic nature of 5G NTN, it is difficult to leverage an optimization solver at scale. Thus, we propose a load-aware conditional handover protocol (LCH), which utilizes the frequent communications among satellites to allow each satellite to swiftly gain awareness of the

constellation load in a decentralized manner. Without extra signalling, satellites can make intelligent decisions to avoid signalling peaks.

In LCH, we first allow each satellite to locally keep a record of future signalling count of each time slot for itself. This record should be long enough to predict UE's handover time. More specifically, each satellite maintains two records namely `myLoad.actualLoad` and `myLoad.potentialLoad`. The actual load represents the signalling load that satellites are sure to incur. For example, when UE is being served by the source satellite, then the source satellite knows that the handover must happen at some point. The potential load represents the signalling load that satellites anticipate to happen. For example, the candidate prepared the timer condition but is unsure if the UE will access it, as the decision is made by the source satellite.

Next, we define a time window T_{window} . For any signalling call among satellites, the sender should send the actual load and potential load within T_{window} . Thus, each satellite maintains a load storage mapping, `storedLoads`, and the stored load should be frequently updated when receiving the latest load from other satellites. Note that this window could not be long because predicted load far in the future could be more inaccurate. Also, these loads will be frequently updated and will be used to help real-time decisions. With shorter T_{window} , the inter-satellite link bandwidth and computing resources should be saved. However, we expect T_{window} to be greater than or equal to $T_{decouple}$. When the source satellite sends an **Handover Request** to the candidate, we allow the source satellite to share all candidates' identities and their loads in `storedLoads` to help decide the timer, although the stored loads may not be the latest. Thus, one must identify the latest load information. Consider that s_1 and s_2 are selected by both s_3 and s_4 as candidate satellites. Here, s_1 will receive the load of s_2 from both s_3 and s_4 . Then, s_1 must identify and keep track of the latest load of s_2 . This could be done through a timestamp. In our discrete time slot setting, we define a `priority` variable, which is set to 0 at the beginning of the time slot. When satellites update their `myLoad.actualLoad` and `myLoad.potentialLoad`, the `priority` increments. Using `priority`, satellites can keep track of the latest loads of other satellites. Next, we will describe our heuristic algorithms for three decisions in Fig. 5.1, **Select Candidates**, **Prepare configuration**, and **Prepare condition**. Load management is also included for illustration.

5.4.2 Select candidates

During candidate selection, the current loads of all satellites are important and we want the source satellite to avoid sending **Handover Request** to candidates that are already experiencing high signalling load. Thus, the source should choose $N_{candidates}$ from *reportedSatellites*

Algorithm 6 Select Candidates (SOURCE)

Input: *reportedSatellites*, *storedLoads*, α

```
1: allSatCurrentLoads = []
2: for s in reportedSatellites do
3:   sActualLoad : Find current actual load in storedLoads[s].
4:   sPotentialLoad : Find current potential load in storedLoads[s].
5:   sLoad = sActualLoad +  $\alpha \times$  sPotentialLoad
6:   allSatCurrentLoads.append(sLoad)
7: end for
8: candidates : Find  $N_{candidate}$  satellites with smallest current signalling based on
   allSatCurrentLoads. For those with the same load, perform random selections.
9: candidateLoads = []
10: for s in candidate do
11:   Increment cActualLoad[s] by 1 at current time.
12:   cActualLoad : Find future actual load in storedLoads[s].
13:   cPotentialLoad : Find future potential load in storedLoads[s].
14:   candidateLoads.append(cActualLoad, cPotentialLoad)
15: end for
16: SEND CANDIDATE: candidates, candidateLoads
```

in the UE's **Measurement Report**, who currently have the lowest signalling load. However, the source may use the same load information of other satellites to perform multiple rounds of candidate selection. This is also an important factor to consider in practice when source satellite has not yet received **Handover Request ACK**. Thus, satellites should locally update other candidates' loads. The source satellite should update the loads when candidates send **Handover Request** or **Handover Cancel**, as these two messages cannot be predicted. Also, we assign a weight parameter α to *potentialLoad* because those predicted signalling may not eventually happen. The algorithm complexity is bounded by $O(\text{reportedSatellite.length})$, and details are in Alg. 6.

5.4.3 Prepare configuration

Using timer-based condition, the candidate satellite should prepare available RA preambles and ensure the UE will access at the prepared time slot. When the UE accesses the target satellite, the other candidates will be informed with **Handover Cancel**, thus, the access time slot should be carefully chosen to avoid a signalling peak. Because each candidate

Algorithm 7 Prepare Configuration (CANDIDATE)

Input: *storedLoads*, α , β , γ *myLoad*, t^l , *candidates*, *candidateLoads*

```
1: MaxLoads = []
2: for  $t$  in  $T_{decouple}$  do
3:    $tLoad$  = []
4:   myActualLoad : Find actual load in myLoad at  $t$ .
5:   myPotentialLoad : Find potential load in myLoad at  $t$ .
6:    $tLoad.append(myActualLoad + \alpha \times myPotentialLoad)$ 
7:   for  $s$  in candidates do
8:      $sActualLoad$  : Find actual load in storedLoads[ $s$ ] at  $t$ .
9:      $sPotentialLoad$  : Find potential load in storedLoads[ $s$ ] at  $t$ .
10:     $tLoad.append(sActualLoad + \alpha \times sPotentialLoad)$ 
11:   end for
12:    $tMaxLoad$  : Find max value in  $tLoad$ .
13:   MaxLoads.append(tMaxLoad)
14: end for
15: availableSlots : Find time slots with available RA preambles.
16:  $N = \gamma \times len(availableSlots)$ 
17: bestSlots : Based on availableSlots, MaxLoads, select  $N$  best slots.
18:  $weights = [(bestSlots.length)^\beta, (bestSlots.length - 1)^\beta, \dots]$ 
19: bestSlot : Based on weights and bestSlots, perform random selection.
20:  $myLoad.potentialLoad[t^l] += 2 + N_{candidate}$ 
21:  $myLoad.potentialLoad[bestSlot] += 3$ 
22: Increment other candidates' storedLoads.potentialLoad[bestSlot] by 1.
23:  $fActualLoad = myLoad.actualLoad[t^l]$ 
24:  $fPotentialLoad = myLoad.potentialLoad[t^l]$ 
25:  $futureLoad = (fActualLoad, fPotentialLoad)$ 
26: SEND SOURCE bestSlot, futureLoad
```

is aware of the loads of other candidates, it can compute the maximum signalling count of each time slot within $T_{decouple}$ across all candidates. Then, the candidate finds all the available slots that have available dedicated preambles. Finally, the candidate randomly selects a slot from N available slots that have the lowest maximum signalling count. Let $\gamma = N/availableSlots.length$, where γ is a percentage.

Because smaller decouple time will result in shorter resource reservation, we could adopt weighted random sampling to allow the satellite to pick a smaller timer. The corresponding

Algorithm 8 Prepare Condition (SOURCE)

Input: *futureLoad*, *bestSlot*, α , *candidates*, *myLoad*

- 1: Compute aggregated future load for each candidate following *futureLoad.fActualLoad*
+ $\alpha \times \text{futureLoad.fPotentialLoad}$.
 - 2: *target* : Pick the candidate with minimum aggregated future load.
 - 3: *bestSlot* : Find the corresponding timer.
 - 4: *myLoad.actualLoad*[*bestSlot*]+ = 2
 - 5: **SEND UE** *target*, *bestSlot*
-

decay factor is defined as β . The local load management for the candidate should follow a hypothesis that the UE will eventually access it. Finally, the candidate satellite should be able to predict the time slot t^l when the UE will require handover based on the mobility of the satellite and UE. The UE can also report t^l in the **Measurement Report**. Therefore, the candidate should also attach the predicted load at t^l in **Handover Request ACK** signalling. The complexity of this algorithm (c.f., Alg. 7) is bounded by $O(N_{\text{candidate}}T_{\text{window}})$.

5.4.4 Prepare condition

During the condition preparation, the source satellite should suggest the UE to access the candidate to which the UE can connect in order to avoid the occurrence of a signalling storm. The source satellite could decide a more adequate target satellite by choosing the candidate with the lowest predicted signalling load (i.e., *futureLoad*). Through this procedure, the source satellite can prevent any future target satellite from having the need to handle massive amounts of UEs attempting to simultaneously perform handover. The algorithm complexity is bounded by $O(N_{\text{candidate}})$, and the details are shown in Alg. 8. Overall, LCH has a linear complexity, which shows the potential to be applied in practice.

5.5 Performance Evaluation

5.5.1 Experiment setup

To evaluate the performance of LCH, we implemented a discrete-event simulator based on Simpy, which is capable of performing large-scale simulations. We also implemented our formulation using Gurobi solver to derive the optimal solution. The simulation time step is 10ms. As shown in Fig. 5.3(A), we deployed six trajectories, each in groups of two, with

the same moving direction and altitude. Within one trajectory, the inter-satellite distance is 250km , and the circle, which has a diameter of 400km , represents the coverage when the UE does not require handover (i.e., $D = 200\text{km}$). Within one trajectory group, the inter-trajectory distance is 200km . The brown, blue, and red trajectory groups, moving at angles of 45° up right, 45° bottom right, and 0° respectively, have speeds of approximately 7.3km/s at altitudes of $1,100\text{km}$, 7.5km/s at 700km , and 7.7km/s at 300km , respectively. For LCH, we set $\alpha = 1$, $\beta = 0$ and $\gamma = 0.2$.

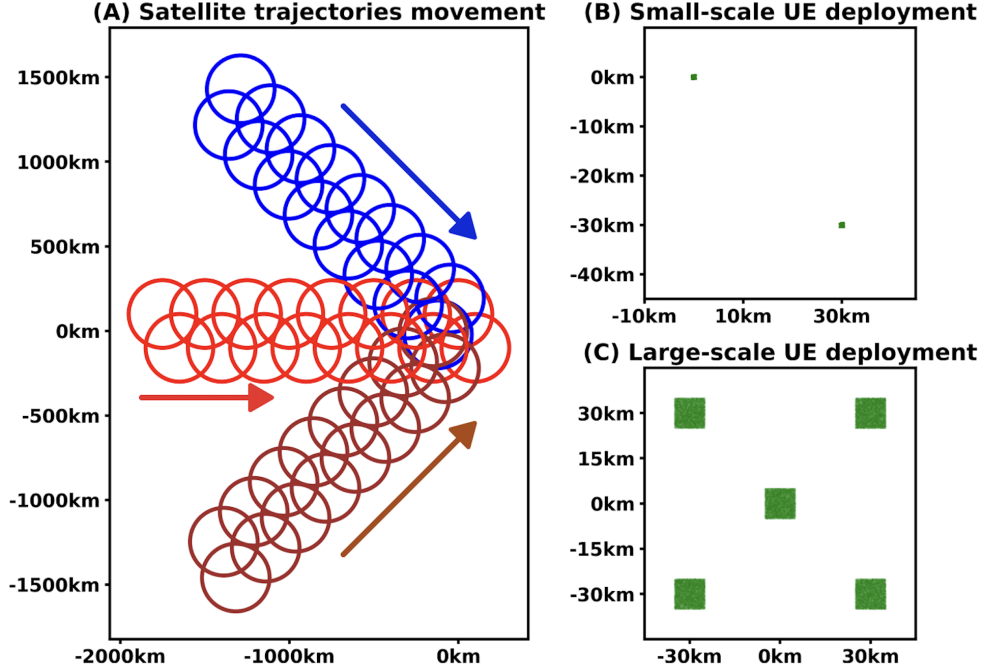


Figure 5.3: Experiment scenarios

We compare LCH with combinations of different options of **Select Candidates**, **Prepare Configuration**, **Prepare Condition** as CHO benchmarks, shown in Table-5.2. OPT indicates the optimal solution from the solver. LCH corresponds to our heuristic algorithm. A through K are simple heuristic algorithms for selecting candidate, preparing configuration and preparing condition. **Random** represents random selection. **Longest in Select Candidate** represents that the source selects the candidates with longest serving time. **Longest in Prepare Condition** represents that the source suggests the target satellite with the longest serving time. **Earliest in Prepare Configuration** represents the candidate will prepare the earliest access slot. **Earliest in Prepare Condition** represents that the source selects the target satellite with the earliest access slot. Next, we evaluate

Table 5.2: Algorithms notations

Label	Select Candidates	Prepare Configuration	Prepare Condition
OPT	solver	solver	solver
LCH	LCH	LCH	LCH
A	Random	Random	Random
B	Random	Random	Earliest
C	Random	Random	Longest
D	Random	Earliest	Random
E	Random	Earliest	Earliest
F	Random	Earliest	Longest
G	Longest	Random	Random
H	Longest	Random	Earliest
I	Longest	Random	Longest
J	Longest	Earliest	Random
K	Longest	Earliest	Earliest
L	Longest	Earliest	Longest

LCH in small- and large-scale scenarios, and analyze the trade-offs among reservation time, serving time, and load balancing.

5.5.2 Experiment results

For the small-scale scenario, we deployed two rectangular grids, each with a length of $1km$, located at coordinates $(0km, 0km)$ and $(30km, -30km)$, as shown in Fig. 5.3(B). Each grid contains approximately 100 randomly distributed static UEs. We set $T_{decouple} = T_{window} = 20$, $MAX_{access} = 2$, and $N_{candidate} = 3$. The simulation time is 100 seconds, equivalent to 10,000 time slots. In this experiment, we use the Gurobi solver to derive the optimal solution, and our simulator leverages the initial UE assignment from the optimal solution, and run LCH and benchmarks. We measure the maximum signalling count across all satellites and all time slots. Fig. 5.4(A) shows that our heuristic outperforms all the benchmarks, but there is room for further optimization.

For the large-scale scenario, we deployed five rectangular grids, each with a length of $10km$, located at coordinates $(0km, 0km)$, $(30km, 30km)$, $(30km, -30km)$, $(-30km, 30km)$, and $(-30km, -30km)$, as shown in Fig. 5.3(C). Each grid contains approximately 10,000 randomly distributed static UEs. This scenario is similar to a remote factories setting. We set $T_{decouple} = T_{window} = 100$, $MAX_{access} = 56$, and $N_{candidate} = 3$. The simulation time is 200 seconds, equivalent to 20,000 time slots. Due to the complexity of the formulation, we are unable to use the solver to derive the optimal solution for this scenario. In this experiment, we will use a random initial assignment and ignore the result of the first 50 seconds to remove the impact of a bad initial assignment. Fig. 5.4(B) shows

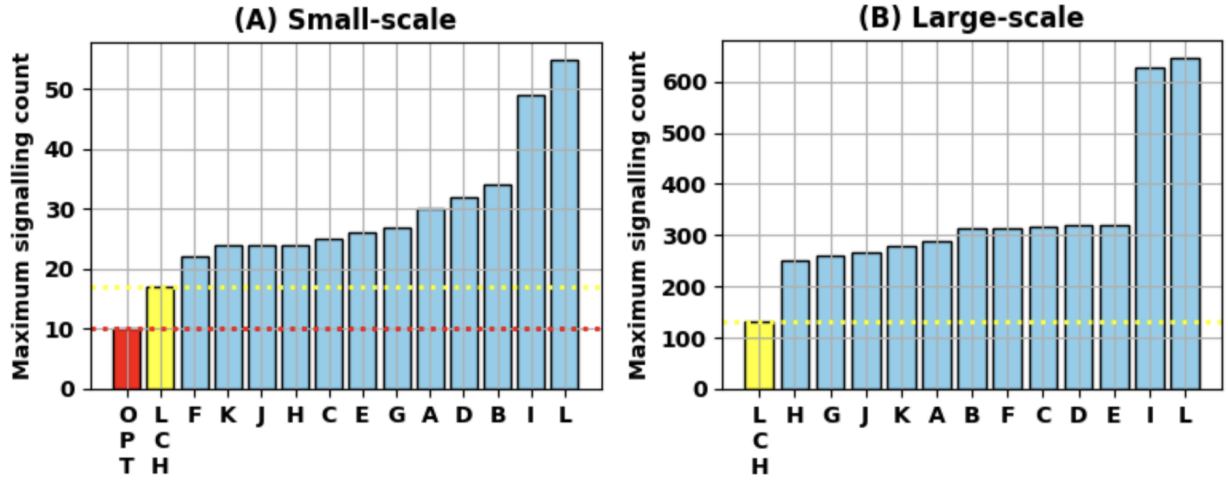


Figure 5.4: Signalling peak evaluation

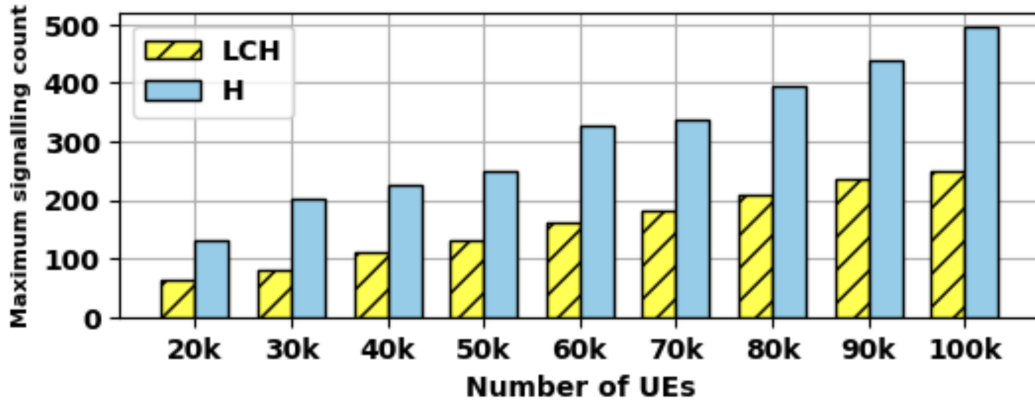


Figure 5.5: Signalling peak comparison with different UE density

that LCH significantly lowers the signalling peak, as it is designed to lower the signalling count in each decision. Then, we compare LCH with the best benchmark H with varying UE densities. As shown in Fig. 5.5, LCH reduces maximum signalling count by roughly 50% in all cases.

5.5.3 Analysis

In addition to minimizing the signalling peak, we study how LCH impacts the other performance KPIs. An algorithm with higher load balancing, lower handover and less resource reservation time is desirable. We now present our main findings.

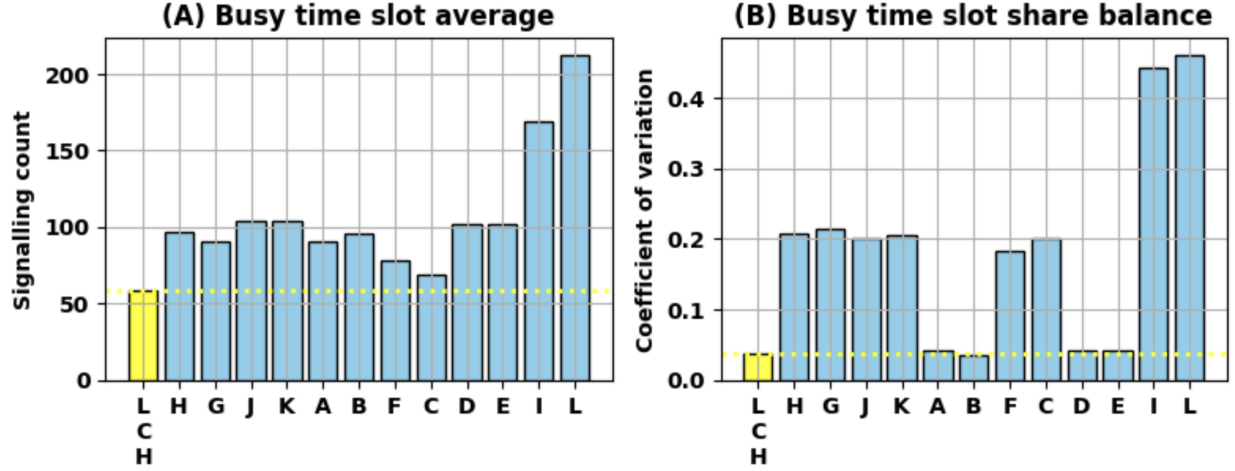


Figure 5.6: Busy time slot evaluation

LCH balances and lowers the signalling peaks: We define ‘busy time’ as the top 20% time slots with the highest signalling count. We calculate the average signalling count of these busy times as shown in Fig. 5.6(A). We then calculate how many busy slots each satellite has and compute the coefficient of variation (CV). CV, defined as the ratio of the standard deviation to the mean, measures the dispersion of data points around the mean, which can best capture the busy time balance among satellites. From the result in Fig. 5.6(B), we observe that LCH achieves a low average signalling count and CV, i.e., LCH not only prevents signalling peak but also distributes the signalling across satellites to balance the load. Method C also achieves a low average signalling count because picking a satellite providing a long service could reduce the handover and, hence, reduce the possibility of creating a high signalling slot. However, such a method causes certain satellites to handle more busy time slots. Unlike C, we notice that A, B, D, and E have low CV scores. This is because UEs near each other do not compete for the longest serving time using these methods.

LCH causes more handovers and shorter serving time: We measure the total handover count through simulation. The result in Fig. 5.7(A) shows that LCH leads to additional handovers. In practice, this could increase the handover failure rate. Also, more

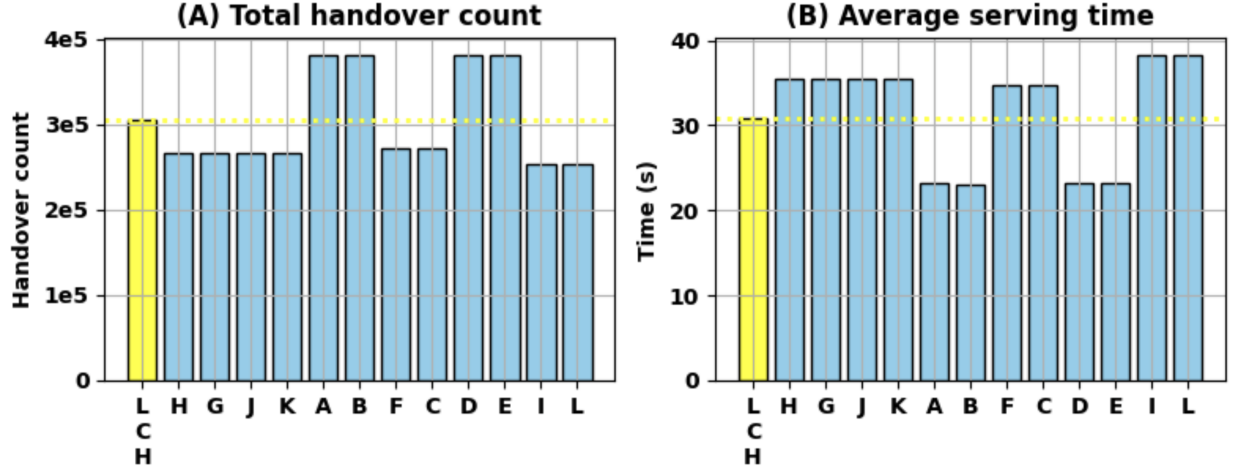


Figure 5.7: Handover count and serving time evaluation

handovers equivalently mean that the constellation will experience more signalling and result in more energy consumption. Moreover, we measure the average serving time that a satellite serves the UE. The result in Fig. 5.7(B) shows that LCH is around 7 seconds shorter compared to the best benchmark. In contrast, those methods, considering the UE serving length, create a smaller handover count and longer serving time. Thus, to decrease the handover count and increase the serving time, the future design of LCH could consider utilizing the predicted serving length as a decision factor.

LCH reserves more resources: We measure the average decouple time that the UE receives. The result in Fig. 5.8(A) shows LCH generates higher decouple time. This is caused by the fact that the near future is easier to predict than the far future in T_{window} . Also, when the satellite almost flies over the UE group, it will cause a large vacancy at the end of T_{window} . In practice, with high decouple time, the UE may require replacing the measurement and causing more signalling. We also measure the time that the candidate reserves the resources until receiving handover cancellation or random access. The result in Fig. 5.8(B) shows satellites using LCH reserves more resources. This issue is mainly caused by the high decoupling time. Thus, to reduce reservation, the future design of LCH should reduce the decoupling time, where the decay factor β could help. Alternatively, a candidate could utilize the predicted load to calculate the possibility of being selected as the target satellite and assign the RA slot more intelligently.

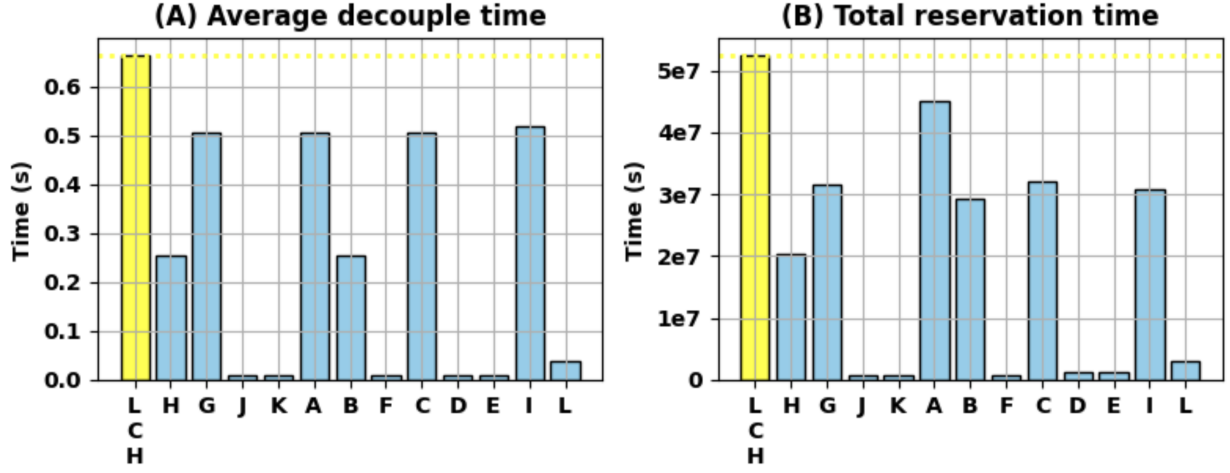


Figure 5.8: Decouple time and reservation evaluation

5.6 Discussion and Future Work

In this work, we assumed a 2-D environment due to deploying UEs in a small area. However, future research should evaluate more complex scenarios with larger geographic regions and dynamic densities using a 3-D environment. Also, LCH should be adjusted and tested considering the movement of UEs. Additionally, handover is not the only task that gNB performs. Future work should consider other procedures, such as paging and radio link establishment, to simulate a dynamic signalling environment to test the robustness of LCH. Moreover, the current formulation is limited in scale, and future work should focus on developing more efficient formulations with relaxed constraints to solve larger-scale problems. Finally, We assumed a short discrete time slot setting in this work, and **Phase1** or **Phase2** to happen in the same slot to reduce the problem complexity. Future research could decouple the signalling calls in **Phase1** or **Phase2** to further refine LCH.

5.7 Conclusions

This work investigated the signalling storm problem during inter-satellite CHO. We derived a comprehensive mathematical formulation and proposed an efficient heuristic, LCH, designed for efficiently solving large-scale problems. The effectiveness of LCH is evaluated against CHO baselines on small-scale and large-scale scenarios. Our results show that LCH significantly reduces signalling peaks and achieves effective load balancing across satellites.

However, it introduces additional handovers and total signalling operations, leading to shorter serving time, and long resource reservation time. Future work should focus on refining LCH to mitigate these drawbacks while maintaining a low and balanced signalling peak.

Chapter 6

Conclusions

In this thesis, we examined the threat of signalling storms in 5G networks. We found that the causes of signalling storms are primarily due to system failures, protocol design issues, an increasing number of UEs, dynamic UE behavior, and malicious attacks. Operators can mitigate or avoid signalling storms by utilizing intelligent monitoring and detection, optimizing protocol design, and leveraging redundancy. We hypothesized that signalling storms caused by UE access-related behaviors and mobility behaviors should be addressed as priorities in 5G.

To mitigate the registration signalling storm, we proposed a consortium blockchain-assisted 5G AKA protocol to offload the centralized UDM processing to decentralized AMF/SEAF within all PLMNs in a trusted manner. For the 5G NTN handover scenario, we proposed an Xn-based group baseline handover protocol that does not rely on a trusted group head. Also, we derived a comprehensive mathematical formulation for conditional handover, which can also be used to model the baseline handover protocol. Moreover, we proposed a signalling load aware conditional protocol to make individual satellites aware of the global constellation signalling load and make intelligent decisions to decrease the signalling peak in real time. All three protocols can be optionally implemented without affecting existing 5G protocols, providing the flexibility to address scenario-specific issues.

Looking ahead, formal security verification can be leveraged to identify potential security vulnerabilities. While our work primarily focuses on the network layer, we believe that cross-layer analysis and consideration of physical layer characteristics would be beneficial in mitigating signalling storms. This approach could help detect malicious sources, make more intelligent decisions, and assist in protocol design. Additionally, large-scale experiments, particularly in 5G NTN, should be conducted to further examine the impact and

key solutions for signalling storms.

Network capacity has always been a driving force behind the evolution of cellular networks. We hope this work will shed light on the future design of 5G and beyond cellular networks.

References

- [1] B. Zhang, P. Zeinaty, N. Limam, and R. Boutaba, “Mitigating signaling storms in 5g with blockchain-assisted 5gaka,” in *2023 19th International Conference on Network and Service Management (CNSM)*. IEEE, 2023, pp. 1–9.
- [2] B. Zhang, P. Hu, A. A. Azirani, M. A. Salahuddin, D. Barradas, N. Limam, and R. Boutaba, “Secure and Efficient Group Handover Protocol in 5G Non-Terrestrial Networks,” *arXiv preprint arXiv:2403.13936*, 2024.
- [3] 3GPP, “NR; NR and NG-RAN Overall Description; Stage-2,” 3rd Generation Partnership Project, Technical Specification (TS) 38.300 v18.1.0, 4 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/27daen9z>
- [4] —, “System Architecture for the 5G System (5GS),” 3rd Generation Partnership Project, Technical Specification (TS) 23.501 v18.5.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/3t6mb5vn>
- [5] N. Singh, “What is the 5G Service-Based Architecture (SBA),” <https://tinyurl.com/2cpnh8m>, Accessed: 14-4-24.
- [6] Ericsson, “Ericsson Mobility Report, November 2023,” Ericsson, Technical Report, 11 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5a9kfvrz>
- [7] Research and Markets, “5G Services Market 2023,” Research and Markets, Market Report, 12 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/3uxvzk36>
- [8] CAICT, “The Development of 5G in China and Its Economic and Social Impact,” China Academy of Information and Communications Technology, Technical Report, 12 2023, Accessed: 15-4-24. [Online]. Available: <http://www.caict.ac.cn/kxyj/qwfb/bps/202401/P020240326603524901232.pdf>
- [9] Cyber Security and Information Systems Information Analysis Center, “The State of 5G Technology and Applications to the DoD and Military,” U.S. Department of Defense, Technical Report, 6 2022, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4rpseesn>

- [10] U.S. Department of Defense, “Department of Defense 5G Strategy Implementation Plan Advancing 5G Technology & Applications Securing 5G Capabilities,” Washington, D.C., Technical Report, 12 2020, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/2m4jfk7h>
- [11] ENISA, “Telecom Security Incidents 2022,” The European Union Agency for Cybersecurity, Annual Report, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/35dv7ch2>
- [12] R. Clark, “China Telecom Network Crash Takes Down Millions,” <https://tinyurl.com/ynezdsp9>, 2023, Accessed: 15-4-24.
- [13] M. Zadikian and I. Poshnjari, “Rogers Pledges Five-day Credits as Bay Street Weighs Outage Impact,” <https://tinyurl.com/4uwvxhpj>, 2022, Accessed: 15-4-24.
- [14] Public Safety and Homeland Security Bureau, “June 15, 2020 T-Mobile Network Outage Report, PS Docket No. 20-183,” Federal Communications Commission, Technical Report, 10 2020, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/ysxm3tdj>
- [15] KDDI, “The July 2 Communication Failure and Our Response,” KDDI, Technical Report, 7 2022, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/bdd73eca>
- [16] NTT Docomo, “Briefing Session Regarding the Status of Response to Communication Failures,” NTT Docomo, Technical Report, 11 2021, Accessed: 15-4-24. [Online]. Available: <http://ngt.idc.nttdocomo.co.jp/20211110.10.pdf>
- [17] M. Yamaguchi, “Outages Disrupt Services at Japan’s No. 2 Telecoms Carrier,” <https://tinyurl.com/y7rw3wch>, 2022, Accessed: 15-4-24.
- [18] A.-D. Fisher, “Signal Storm Caused Telenor Outages,” <https://tinyurl.com/4wcufra7>, 2011, Accessed: 15-4-24.
- [19] NTT Docomo, “Response to Series of Network Malfunctions,” NTT Docomo, Technical Report, 01 2012, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4ftb9jrk>
- [20] IT Brief New Zealand, “XT Outages: What We Know,” <https://tinyurl.com/5eh6jznd>, 2010, Accessed: 15-4-24.
- [21] Huawei, “5G Signalling and Control Plane Traffic Depends on Service Communications Proxy,” Huawei, Technical Report, 12 2021, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4bjtazs6>
- [22] 5G Americas, “The Evolution of Security in 5G: A Slice of Mobile Threats,” 5G Americas, Technical Report, 7 2019, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/phua43w7>

- [23] ENISA, “ENISA Threat Landscape for 5G Networks,” The European Union Agency for Cybersecurity, Technical Report, 12 2020, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/22syuj8j>
- [24] NSA and CISA, “5G Network Slicing: Security Considerations for Design, Deployment, and Maintenance,” National Security Agency and Cybersecurity & Infrastructure Security Agency, Technical Report, 7 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/kdedrr4u>
- [25] GSMA, “Mobile Telecommunications Security Landscape,” Global System for Mobile Communications, Technical Report, 2 2022, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5n8wc585>
- [26] IMT-2020 (5G) Promotion Group, “5G Security Knowledge Base,” China Academy of Information and Communications Technology, Technical Report, 12 2021, Accessed: 15-4-24. [Online]. Available: <http://www.caict.ac.cn/kxyj/qwfb/ztbg/202112/P020211210558556929664.pdf>
- [27] J. Hodges, “Securing 5G Networks: Addressing Control Plane Challenges,” <https://tinyurl.com/27recspb>, 2019, Accessed: 14-4-24.
- [28] Kaleido Intelligence, “Signalling Security - Sizing the Operator Threat,” Cellsys, Security Survey Report, 1 2021, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/ykwnpn9r>
- [29] G. Gow and R. Smith, *Mobile And Wireless Communications: An Introduction*. McGraw-Hill Education (UK), 2006.
- [30] N. Linge and A. Sutton, “The Road to 4G,” *J. Inst. Telecommun. Prof.*, vol. 8, no. 1, pp. 9–16, 2014.
- [31] P. Launiainen, *A Brief History of Everything Wireless: How Invisible Waves Have Changed the World*. Springer, 2018.
- [32] F. A. Saeed, “Capacity Limit Problem in 3G Networks,” <https://tinyurl.com/47cz9pa6>, 2010, Accessed: 14-4-24.
- [33] Guelph Mercury, “UK Mobile Phone Operator O2 Apologizes for Service Problems Caused by Smartphones,” <https://tinyurl.com/3db4jrj5>, 2010, Accessed: 15-4-24.
- [34] Jenna Wortham, “Customers Angered as iPhones Overload AT&T,” <https://tinyurl.com/4dupfsze>, 2009, Accessed: 15-4-24.
- [35] E. Savitz, “How The New iPad Creates Signalling Storm For Carriers,” <https://tinyurl.com/yws2thjv>, 2012, Accessed: 15-4-24.
- [36] 3GPP, “NR; User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone,” 3rd Generation Partnership Project, Technical

- Specification (TS) 38.101-1 v18.5.0, 04 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/478jbv3z>
- [37] —, “NR; User Equipment (UE) Radio Transmission and Reception; Part 2: Range 2 Standalone,” 3rd Generation Partnership Project, Technical Specification (TS) 38.101-2 v18.5.0, 04 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4v2ca4yw>
- [38] —, “Management and Orchestration; Provisioning,” 3rd Generation Partnership Project, Technical Specification (TS) 28.531 v18.5.0, 04 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/25asbb4f>
- [39] —, “Study on Scenarios and Requirements for Next Generation Access Technologies,” 3rd Generation Partnership Project, Technical Report (TR) 38.913 v18.0.0, 04 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/2suddxad>
- [40] Ericsson, “The Next Wave of 5G – 3GPP Release 19,” <https://tinyurl.com/4s7bwz2n>, 2023, Accessed: 15-4-24.
- [41] M. Akon, T. Yang, Y. Dong, and S. R. Hussain, “Formal Analysis of Access Control Mechanism of 5G Core Network,” in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 666–680.
- [42] A. Al Ishtiaq, S. S. S. Das, S. M. M. Rashid, A. Ranjbar, K. Tu, T. Wu, Z. Song, W. Wang, M. Akon, R. Zhang *et al.*, “Hermes: Unlocking Security Analysis of Cellular Network Protocols by Synthesizing Finite State Machines from Natural Language Specifications,” *arXiv preprint arXiv:2310.04381*, 2023.
- [43] A. Gregg, “FCC Opens Formal Investigation into Massive AT&T Outage,” <https://tinyurl.com/bfxm64xt>, 2024, Accessed: 15-4-24.
- [44] S. Neuman, “AT&T Says Cell Service is Back after a Widespread Outage and Some Disrupted 911 Calls,” <https://tinyurl.com/2nn56dzs>, 2024, Accessed: 15-4-24.
- [45] H. Kelly, “Landlines Are Dying Out, But to Some, They’re a Lifeline,” <https://tinyurl.com/5dfwrtz6>, 2024, Accessed: 15-4-24.
- [46] 3GPP, “NG-RAN; Architecture Description,” 3rd Generation Partnership Project, Technical Specification (TS) 38.401 v18.1.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4tkcsrak>
- [47] —, “Study on New Radio Access Technology: Radio Access Architecture and Interfaces,” 3rd Generation Partnership Project, Technical Report (TR) 38.801 v14.0.0, 4 2017, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mpw8wcfp>

- [48] S. Matoussi, I. Fajjari, S. Costanzo, N. Aitsaadi, and R. Langar, “5G RAN: Functional Split Orchestration Optimization,” *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, pp. 1448–1463, 2020.
- [49] Quad Critical and Emerging Technology Working Group, “Open RAN Security Report,” National Telecommunications and Information Administration, Technical Report, 5 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mtkkhdtx>
- [50] M. Polese, L. Bonati, S. D’oro, S. Basagni, and T. Melodia, “Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376–1411, 2023.
- [51] O-RAN Work Group 1, “O-RAN Architecture Description,” O-RAN ALLIANCE, Technical Report v11.0, 2 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5d3u3mam>
- [52] —, “O-RAN Use Cases Analysis Report,” O-RAN ALLIANCE, Technical Report v13.0, 2 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5d3u3mam>
- [53] 3GPP, “5G System; Network Function Repository Services; Stage 3,” 3rd Generation Partnership Project, Technical Specification (TS) 29.510 v18.6.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/8jcbmpkx>
- [54] 3GPP, “Communication Services (VoLTE/VoNR),” <https://tinyurl.com/yvej96t>, 2024, Accessed: 14-4-24.
- [55] 3GPP, “Non-Access-Stratum (NAS) Protocol for 5G System (5GS); Stage 3,” 3rd Generation Partnership Project, Technical Specification (TS) 24.501 v18.6.0, 4 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/2s4kmc2a>
- [56] —, “5G System; Technical Realization of Service Based Architecture; Stage 3,” 3rd Generation Partnership Project, Technical Specification (TS) 29.500 v18.5.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/26256kwr>
- [57] —, “Network Domain Security (NDS); Authentication Framework (AF),” 3rd Generation Partnership Project, Technical Specification (TS) 33.310 v19.0.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5n8j3ndu>
- [58] —, “Security Architecture and Procedures for 5G System,” 3rd Generation Partnership Project, Technical Specification (TS) 33.501 v18.5.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/3pfbra6z>
- [59] —, “Procedures for the 5G System (5GS),” 3rd Generation Partnership Project, Technical Specification (TS) 23.502 v18.5.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mrd8jwz9>

- [60] J. de Gregorio, “5GC APIs,” https://github.com/jdegre/5GC_APIs.git, 2024, Accessed: 14-4-24.
- [61] GSMA, “Generic Network Slice Template,” Global System for Mobile Communications, Technical Report v8.0, 1 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4b4xm826>
- [62] 3GPP, “Study on Network Slice Management Capability Exposure,” 3rd Generation Partnership Project, Technical Report (TR) 28.824 v18.0.1, 7 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mr2pauh5>
- [63] —, “Management and Orchestration; Concepts, Use Cases and Requirements,” 3rd Generation Partnership Project, Technical Specification (TS) 28.530 v18.0.0, 1 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/ydcn5hcu>
- [64] —, “Study on Management and Orchestration of Network Slicing for Next Generation Network,” 3rd Generation Partnership Project, Technical Report (TR) 28.801 v15.1.0, 1 2018, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/46j3e778>
- [65] —, “Management and Orchestration; Architecture Framework,” 3rd Generation Partnership Project, Technical Report (TR) 28.533 v18.1.0, 4 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/45xmpdpt>
- [66] International Telecommunication Union, “Mobile Network Coverage,” <https://tinyurl.com/mpa7bjae>, 2021, Accessed: 14-4-24.
- [67] Iridium, “How To Turn Your Cell Phone into a Satellite Phone,” <https://tinyurl.com/bdxfhkmv>, 2021, Accessed: 14-4-24.
- [68] 3GPP, “Study on Using Satellite Access in 5G,” 3rd Generation Partnership Project, Technical Report (TR) 22.822 v16.0.0, 07 2018, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4mesk6rj>
- [69] 5G Americas, “5G & Non-Terrestrial Networks,” 5G Americas, Technical Report, 2 2022, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5f5zpypt>
- [70] —, “Update on 5G Non-Terrestrial Networks,” 5G Americas, Technical Report, 7 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/47vda3pv>
- [71] M. Lifson, A. D’Ambrosio, D. Arnas, and R. Linares, “How Many Satellites Can We Fit in Low Earth Orbit?: Capacity Integrating Risk-Based and Intrinsic Methods,” in *Astrodynamics Specialist Conference*. AMOS conference, 2022, pp. 1–19.
- [72] S. Ji, D. Zhou, M. Sheng, D. Chen, L. Liu, and Z. Han, “Mega Satellite Constellations Analysis Regarding Handover: Can Constellation Scale Continue Growing?” in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 4697–4702.

- [73] N. P. Walsh, A. Marquardt, F. Davey-Attlee, and K. Gak, “Ukraine Relies on Starlink for Its Drone War. Russia Appears to Be Bypassing Sanctions to Use the Devices Too,” <https://tinyurl.com/3c3utuer>, 2024, Accessed: 14-4-24.
- [74] M. Sheetz, “SpaceX and T-Mobile Send First Texts via Starlink Satellites,” <https://tinyurl.com/4fum8zws>, 2024, Accessed: 14-4-24.
- [75] 3GPP, “Solutions for NR to Support Non-Terrestrial Networks (NTN),” 3rd Generation Partnership Project, Technical Specification (TS) 33.821 v16.2.0, 04 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/2fufe4tk>
- [76] —, “Study on New Radio (NR) to Support Non-terrestrial Networks,” 3rd Generation Partnership Project, Technical Report (TR) 38.811 v15.4.0, 10 2020, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/7ujfhcek>
- [77] S. Cakaj, “The Parameters Comparison of the “Starlink” LEO Satellites Constellation for Different Orbital Shells,” *Frontiers in Communications and Networks*, vol. 2, p. 643095, 2021.
- [78] T. Sesnic, “Starlink Group 6-53 — Falcon 9 Block 5,” <https://tinyurl.com/5dbcztj5>, 2024, Accessed: 14-4-24.
- [79] S. Euler, X. Fu, S. Hellsten, C. Kefeder, O. Liberg, E. Medeiros, D. S. Erik Nordell, P. Synnergren, E. Trojer, and I. Xirouchakis, “Using 3GPP Technology for Satellite Communication,” Ericsson, Technical Report, 06 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/5n7sbnt8>
- [80] A. Sattarzadeh, Y. Liu, A. Mohamed, R. Song, P. Xiao, Z. Song, H. Zhang, R. Tafazolli, and C. Niu, “Satellite-based Non-terrestrial Networks in 5G: Insights and Challenges,” *IEEE Access*, vol. 10, pp. 11 274–11 283, 2021.
- [81] L. Liu, Y. Li, H. Li, J. Yang, W. Liu, J. Lan, Y. Wang, J. Li, J. Wu, Q. Wu *et al.*, “Democratizing Direct-to-Cell Low Earth Orbit Satellite Networks,” in *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, 2024, pp. 791–808.
- [82] R. Campana, C. Amatetti, and A. Vanelli-Coralli, “RAN Functional Splits in NTN: Architectures and Challenges,” *arXiv preprint arXiv:2309.14810*, 2023.
- [83] 3GPP, “Study on 5G System with Satellite Backhaul,” 3rd Generation Partnership Project, Technical Report (TR) 23.700-27 v18.0.0, 12 2022, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/bdehhvtw>
- [84] Y. Drif, E. Lavinal, E. Chaput, P. Berthou, B. T. Jou, O. Grémillet, and F. Arnal, “Slice Aware Non-terrestrial Networks,” in *2021 IEEE 46th Conference on Local Computer Networks (LCN)*. IEEE, 2021, pp. 24–31.

- [85] NTT Docomo, “Summary of Measures Taken to Prevent Mobile Communications Service Interruptions, Protect Secrecy of Communication and Appropriately Manage Personal Information,” NTT Docomo, Technical Report, 03 2012, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/42a6vnam>
- [86] M. Brian, “Japan’s NTT DoCoMo to Request that Google Reduces Android’s Data Usage,” <https://tinyurl.com/ydr4wr7v>, 2012, Accessed: 15-4-24.
- [87] Cloudflare, “Inside the Infamous Mirai IoT Botnet: A Retrospective Analysis,” <https://tinyurl.com/4cd7a4xf>, 2017, Accessed: 14-4-24.
- [88] Ericsson, “Analyzing IoT Device Performance,” Ericsson, Technical Report, 11 2016, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mtk9j4fa>
- [89] Canadian Centre for Cyber Security, “Defending against Distributed Denial of Service,” Communications Security Establishment Canada, Technical Report, 2 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/yc2czwyn>
- [90] Imperva, “Booters, Stressers and DDoSers,” <https://tinyurl.com/bdw3wc98>, 2024, Accessed: 14-4-24.
- [91] 3GPP, “Study on Security Aspects of Enhanced Support of Non-Public Networks (NPN) Phase 2,” 3rd Generation Partnership Project, Technical Report (TR) 33.858 v18.1.0, 9 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/52j22mca>
- [92] —, “Study on Authentication Enhancements in the 5G System (5GS),” 3rd Generation Partnership Project, Technical Report (TR) 33.846 v17.0.0, 12 2021, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/3j8mwvrs>
- [93] —, “Study on Evolution of Cellular Internet of Things (CIoT) Security for the 5G System,” 3rd Generation Partnership Project, Technical Report (TR) 33.861 v16.1.0, 9 2020, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/bp7shckm>
- [94] A. Tabiban, H. A. Alameddine, M. A. Salahuddin, and R. Boutaba, “Signalling Storm in O-RAN: Challenges and Research Opportunities,” *IEEE Communications Magazine*, 2023.
- [95] 3GPP, “Study on 5G Networks Providing Access to Localized Services,” 3rd Generation Partnership Project, Technical Report (TR) 22.844 v18.2.0, 12 2021, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/47tufy5e>
- [96] —, “Study on Load and Overload Control of 5GC Service Based Interface,” 3rd Generation Partnership Project, Technical Report (TR) 29.843 v16.0.0, 9 2019, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mrtde23m>

- [97] H. Yang, S. Bae, M. Son, H. Kim, S. M. Kim, and Y. Kim, “Hiding in Plain Signal: Physical Signal Overshadowing Attack on {LTE},” in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 55–72.
- [98] 3GPP, “Study on Enhancement of Ultra-Reliable Low-Latency Communication (URLLC) Support in the 5G Core Network (5GC),” 3rd Generation Partnership Project, Technical Report (TR) 23.725 v16.2.0, 6 2019, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/mdhy4dp3>
- [99] F. Francois, O. H. Abdelrahman, and E. Gelenbe, “Impact of Signalling Storms on Energy Consumption and Latency of LTE User Equipment,” in *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*. IEEE, 2015, pp. 1248–1255.
- [100] G. Gorbil, O. H. Abdelrahman, M. Pavloski, and E. Gelenbe, “Modeling and Analysis of RRC-based Signalling Storms in 3G networks,” *IEEE Transactions on Emerging Topics in Computing*, vol. 4, no. 1, pp. 113–127, 2015.
- [101] R. Ettiane and R. E. Kouch, “Mitigating Denial of Service Signalling Threats in 5G Mobile Networks,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.
- [102] 3GPP, “5G System; Session Management Policy Control Service; Stage 3,” 3rd Generation Partnership Project, Technical Specification (TS) 29.512 v18.5.1, 4 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/bdev6cc5>
- [103] —, “3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control Plane (GTPv2-C); Stage 3,” 3rd Generation Partnership Project, Technical Specification (TS) 29.274 v18.6.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/wv4a2e2p>
- [104] —, “Feasibility Study on LAN Support in 5G,” 3rd Generation Partnership Project, Technical Report (TR) 22.821 v16.1.0, 6 2018, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/ymcww3wy>
- [105] —, “Study on Security Aspects of Proximity Based Services (ProSe) in 5G System (5GS) Phase 2,” 3rd Generation Partnership Project, Technical Report (TR) 33.740 v18.1.0, 9 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/2a5n59tv>
- [106] K. Fang and G. Yan, “Paging Storm Attacks against 4G/LTE Networks from Regional Android Botnets: Rationale, Practicality, and Implications,” in *Proceedings of*

the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks, 2020, pp. 295–305.

- [107] Y. Luo, W. Huang, J. Wang, and Q. Xu, “Technical Experiment and Analysis of WeChat’s Impact on the 3G Cellular Network,” *Information and Communications Technologies*, vol. 4, pp. 16–20, 2013. [Online]. Available: <https://www.mobibrw.com/2014/1658>
- [108] Ericsson, “Handling of Signalling Storms in Mobile Networks,” Ericsson, Technical Report, 8 2015, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/bdzayykz>
- [109] M. Q. Khan, “Signalling Storm Problems in 3gpp Mobile Broadband Networks, Causes and Possible Solutions: A Review,” in *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*. IEEE, 2018, pp. 183–188.
- [110] 3GPP, “Study on Restoration of Profiles Related to UDR,” 3rd Generation Partnership Project, Technical Report (TR) 29.821 v17.0.0, 12 2021, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/2rd9bme6>
- [111] C. Reichert, “Ericsson: Expired Certificate Caused O2 and SoftBank Outages,” <https://tinyurl.com/5eyh9xkk>, 2018, Accessed: 14-4-24.
- [112] 3GPP, “Study on Automated Certificate Management in Service-Based Architecture (SBA),” 3rd Generation Partnership Project, Technical Report (TR) 33.876 v18.0.1, 7 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/nvn85pfa>
- [113] —, “Study on Support for Minimization of Service Interruption,” 3rd Generation Partnership Project, Technical Report (TR) 22.831 v17.1.0, 10 2019, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/22mwmz6h>
- [114] —, “5G System; Restoration Procedures,” 3rd Generation Partnership Project, Technical Specification (TS) 23.527 v18.3.0, 3 2024, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/bdddc6jc>
- [115] Huawei, “Striding Towards the Intelligent World,” Huawei, Technical Report, 2023, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/4twmezp6>
- [116] C. Hu, W. Fan, E. Zeng, Z. Hang, F. Wang, L. Qi, and M. Z. A. Bhuiyan, “Digital twin-assisted real-time traffic data prediction method for 5g-enabled internet of vehicles,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 4, pp. 2811–2819, 2021.
- [117] J. Sun, Y. Zhang, and M. Trik, “PBPHS: a Profile-based Predictive Handover Strategy for 5G networks,” *Cybernetics and Systems*, pp. 1–22, 2022.

- [118] J.-H. Lee, C. Park, S. Park, and A. F. Molisch, “Handover Protocol Learning for LEO Satellite Networks: Access Delay and Collision Minimization,” *IEEE Transactions on Wireless Communications*, 2023.
- [119] D. Feng, S. Li, Z. Yang, Y. Xiang, J. Zheng, and X. He, “Research of Deep Learning and Adaptive Threshold based Signalling Storm Prediction and Top Cause Tracking,” *IEEE Access*, 2023.
- [120] S. Park, B. Cho, D. Kim, and I. You, “Machine Learning based Signaling DDoS Detection System for 5G Standalone Core Network,” *Applied Sciences*, vol. 12, no. 23, p. 12456, 2022.
- [121] E. Gelenbe, O. H. Abdelrahman, and G. Gorbil, “Detection and Mitigation of Signalling Storms in Mobile Networks,” in *2016 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2016, pp. 1–5.
- [122] M. Pavloski, “Detecting and Mitigating Storm Attacks in Mobile Access to the Cloud,” in *2019 IEEE International Conference on Fog Computing (ICFC)*. IEEE, 2019, pp. 53–58.
- [123] M. Hoffmann and P. Kryszkiewicz, “Signalling storm detection in iiot network based on the open ran architecture,” in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2023, pp. 1–2.
- [124] D. M. Manias, A. Chouman, and A. Shami, “An NWDAF Approach to 5G Core Network Signalling Traffic: Analysis and Characterization,” in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 6001–6006.
- [125] 3GPP, “Study on Core Network Enhanced Support for Artificial Intelligence (AI) / Machine Learning (ML),” 3rd Generation Partnership Project, Technical Report (TR) 23.700-84 v0.3.0, 4 2024, Accessed: 15-5-24. [Online]. Available: <https://tinyurl.com/5n8n877h>
- [126] Z. Haddad, M. M. Fouda, M. Mahmoud, and M. Abdallah, “Blockchain-based Authentication for 5G Networks,” in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIOT)*. IEEE, 2020, pp. 189–194.
- [127] M. Hojjati, A. Shafieinejad, and H. Yanikomeroglu, “A Blockchain-based Authentication and Key Agreement (AKA) Protocol for 5G Networks,” *IEEE Access*, vol. 8, pp. 216 461–216 476, 2020.
- [128] J. Yan, B. Yang, L. Su, S. He, and N. Dong, “Decentralized Certificate Management for Network Function Virtualization (NFV) Implementation in 5G Networks,” in *Mobile Multimedia Communications: 14th EAI International Conference, Mobimedia 2021, Virtual Event, July 23-25, 2021, Proceedings 14*. Springer, 2021, pp. 81–93.

- [129] C.-C. Tseng, H.-C. Wang, J.-R. Chang, L.-H. Wang, and F.-C. Kuo, "Design of Two-Step Random Access Procedure for URLLC Applications," *Wireless Personal Communications*, vol. 121, pp. 1187–1219, 2021.
- [130] M.-S. Pan, T.-M. Lin, and W.-T. Chen, "An Enhanced Handover Scheme for Mobile Relays in LTE-A High-speed Rail Networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 2, pp. 743–756, 2014.
- [131] X. Yan, M. Ma, and R. Su, "Efficient group handover authentication for secure 5g-based communications in platoons," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3104–3116, 2022.
- [132] R. Ma, J. Cao, D. Feng, H. Li, and S. He, "FTGPHA: Fixed-trajectory Group Pre-handover Authentication Mechanism for Mobile Relays in 5G High-speed Rail Networks," *IEEE transactions on vehicular technology*, vol. 69, no. 2, pp. 2126–2140, 2019.
- [133] P. Sharma, T. Atalay, H.-A. Gibbs, D. Stojadinovic, A. Stavrou, and H. Wang, "5g-wave: A core network framework with decentralized authorization for network slices," *arXiv preprint arXiv:2404.13242*, 2024.
- [134] Y. Yin, W. Guan, X. Zhang, and X. He, "Research on Effective Protection against Signalling Storms through End-to-end Flow Control," *The Journal of Telecommunications and Information Technology*, 2023.
- [135] C. Zhang, X. Sun, J. Zhang, X. Wang, S. Jin, and H. Zhu, "Distributive throughput optimization for massive random access of m2m communications in lte networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 828–11 840, 2020.
- [136] G. Somani, M. S. Gaur, D. Sanghi, M. Conti, and R. Buyya, "DDoS Attacks in Cloud Computing: Issues, Taxonomy, and Future Directions," *Computer communications*, vol. 107, pp. 30–48, 2017.
- [137] H. Zhu, J. Li, J. Hu, and W. Li, "Failure-aware and Automated Disaster Backup in the 5G Core Network," in *2022 International Communication Engineering and Cloud Computing Conference (CECCC)*. IEEE, 2022, pp. 48–53.
- [138] P. P. Lee, T. Bu, and T. Woo, "On the detection of signalling dos attacks on 3g wireless networks," in *IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications*. IEEE, 2007, pp. 1289–1297.
- [139] M. Pavloski, "Signalling attacks in mobile telephony," in *Security in Computer and Information Sciences: First International ISCIS Security Workshop 2018, Euro-CYBERSEC 2018, London, UK, February 26-27, 2018, Revised Selected Papers 1*. Springer, 2018, pp. 130–141.

- [140] T.-H. Chen, J.-W. Chang, and H.-Y. Wei, “Dynamic Inter-channel Resource Allocation for Massive M2M Control Signalling Storm Mitigation,” in *2016 IEEE 84th vehicular technology conference (VTC-Fall)*. IEEE, 2016, pp. 1–5.
- [141] GlobalData Technology, “Docomo Outage Demonstrates the Danger of Signalling Storms,” <https://tinyurl.com/yu5m83ah>, 2021, Accessed: 14-4-24.
- [142] MarketsandMarkets, “Cellular IoT Market,” <https://tinyurl.com/mprcrzum>, 2023, Accessed: 14-4-24.
- [143] M. Pauliac, “USIM in 5G Era,” *Journal of ICT Standardization*, pp. 29–40, 2020.
- [144] D. Strobel, “IMSI Catcher,” *Chair for Communication Security, Ruhr-Universität Bochum*, vol. 14, 2007.
- [145] K. Ullah, I. Rashid, H. Afzal, M. M. W. Iqbal, Y. A. Bangash, and H. Abbas, “SS7 Vulnerabilities—A Survey and Implementation of Machine Learning vs Rule Based Filtering for Detection of SS7 Network Attacks,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1337–1371, 2020.
- [146] D. Basin, J. Dreier, L. Hirschi, S. Radomirovic, R. Sasse, and V. Stettler, “A Formal Analysis of 5G Authentication,” in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 1383–1396.
- [147] J. Cao, M. Ma, H. Li, R. Ma, Y. Sun, P. Yu, and L. Xiong, “A survey on security aspects for 3gpp 5g networks,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 170–195, 2019.
- [148] C. Craven, “What is unified data management,” <https://tinyurl.com/54kve73s>, 2019, Accessed: 14-4-24.
- [149] Z. Gao, D. Zhang, J. Zhang, Z. Liu, H. Liu, and M. Zhao, “BC-AKA: Blockchain based Asymmetric Authentication and Key Agreement Protocol for Distributed 5G Core Network,” *China Communications*, vol. 19, no. 6, pp. 66–76, 2022.
- [150] M. C. Chow and M. Ma, “A Secure Blockchain-based Authentication and Key Agreement Scheme for 3GPP 5G Networks,” *Sensors*, vol. 22, no. 12, p. 4525, 2022.
- [151] D. Dolev and A. Yao, “On the security of public key protocols,” *IEEE Transactions on information theory*, vol. 29, no. 2, pp. 198–208, 1983.
- [152] A. Koutsos, “The 5g-aka authentication protocol privacy,” in *2019 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2019, pp. 464–479.
- [153] C. Gorenflo, S. Lee, L. Golab, and S. Keshav, “FastFabric: Scaling Hyperledger Fabric to 20000 Transactions per Second,” *International Journal of Network Management*, vol. 30, no. 5, p. e2099, 2020.

- [154] I. S. Evaluators, “Ethercombing: Finding Secrets in Popular Places,” <https://tinyurl.com/59bn2eus>, accessed: 2019-04-23.
- [155] Y. Yang *et al.*, “FHAP: Fast Handover Authentication Protocol for High-Speed Mobile Terminals in 5G Satellite-Terrestrial Integrated Networks,” *IEEE Internet of Things Journal*, 2023.
- [156] E. Juan, M. Lauridsen, J. Wigard, and P. Mogensen, “Handover Solutions for 5G Low-Earth Orbit Satellite Networks,” *IEEE Access*, vol. 10, pp. 93 309–93 325, 2022.
- [157] K. Zhu, C. Hua, P. Gu, and W. Xu, “User Clustering and Proactive Group Handover Scheduling in LEO Satellite Networks,” in *IEEE Computing, Communications and IoT Applications*, 2020, pp. 1–6.
- [158] L. Yang, X. Yang, and Z. Bu, “A Group Handover Strategy for Massive User Terminals in LEO Satellite Networks,” in *IEEE Vehicular Technology Conference*, 2022, pp. 1–6.
- [159] A. Peltonen, R. Sasse, and D. Basin, “A Comprehensive Formal Analysis of 5G Handover,” in *ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, pp. 1–12.
- [160] K. Xue, W. Meng, H. Zhou, D. S. Wei, and M. Guizani, “A Lightweight and Secure Group Key based Handover Authentication Protocol for the Software-Defined Space Information Network,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3673–3684, 2020.
- [161] Q. Kong, R. Lu, and F. Yin, “Achieving Efficient and Secure Handover in LEO Constellation-assisted Beyond 5G Networks,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 641–653, 2021.
- [162] C. Lai and Y. Ma, “A Novel Group-Oriented Handover Authentication Scheme in MEC-Enabled 5G Networks,” in *IEEE/CIC International Conference on Communications in China*, 2021, pp. 29–34.
- [163] C. Lai, Y. Ma, R. Lu, Y. Zhang, and D. Zheng, “A Novel Authentication Scheme Supporting Multiple User Access for 5G and Beyond,” *IEEE Transactions on Dependable and Secure Computing*, 2022.
- [164] S. Pekhterev, “The Bandwidth of the StarLink Constellation and the Assessment of its Potential Subscriber Base in the USA,” <https://tinyurl.com/mr496t53>, Accessed: 14-4-24.
- [165] 3GPP, “Signalling Storm During HOs and Timer Based Trigger Details,” 3rd Generation Partnership Project, Technical Report R2-2108065, 08 2021, Accessed: 11-10-23. [Online]. Available: <https://tinyurl.com/mwvfjmhd>

- [166] K. Ghanem, H. Alradwan, A. Motermawy, and A. Ahmad, “Reducing Ping-Pong Handover Effects in Intra EUTRA Networks,” in *International Symposium on Communication Systems, Networks & Digital Signal Processing*, 2012, pp. 1–5.
- [167] P. Wadhvani, “5G NTN Market,” Global Market Insights Inc., Market Report, 2 2024, Accessed: 18-6-24. [Online]. Available: <https://tinyurl.com/y6a4687p>
- [168] SpaceX, “Starlink Direct to Cell,” <https://www.starlink.com/business/direct-to-cell>, Accessed: 18-6-24.
- [169] Ericsson, “3GPP TSG-RAN WG2 Meeting #112: Impacts of Earth Fixed and Moving Beams,” 3rd Generation Partnership Project, Meeting Document, 11 2020, Accessed: 15-4-24. [Online]. Available: <https://tinyurl.com/9wfwuhrz>
- [170] A. Gharouni, U. Karabulut, A. Enqvist, P. Rost, A. Maeder, and H. Schotten, “Signal Overhead Reduction for AI-assisted Conditional Handover Preparation,” in *Mobile Communication-Technologies and Applications; 25th ITG-Symposium*. VDE, 2021, pp. 1–6.
- [171] S. B. Iqbal, A. Awada, U. Karabulut, I. Viering, P. Schulz, and G. P. Fettweis, “On the Modeling and Analysis of Fast Conditional Handover for 5G-Advanced,” in *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2022, pp. 595–601.
- [172] J. Stańczak, U. Karabulut, and A. Awada, “Conditional Handover Modelling for Increased Contention Free Resource Use in 5G-Advanced,” in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2023, pp. 1–6.
- [173] S. He, T. Wang, and S. Wang, “Load-Aware Satellite Handover Strategy Based on Multi-Agent Reinforcement Learning,” in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.
- [174] N. Badini, M. Jaber, M. Marchese, and F. Patrone, “Reinforcement Learning-Based Load Balancing Satellite Handover Using NS-3,” in *ICC 2023-IEEE International Conference on Communications*. IEEE, 2023, pp. 2595–2600.
- [175] F. Wang, D. Jiang, Z. Wang, J. Chen, and T. Q. Quek, “Seamless Handover in LEO-Based Non-Terrestrial Networks: Service Continuity and Optimization,” *IEEE Transactions on Communications*, vol. 71, no. 2, pp. 1008–1023, 2022.
- [176] H. Martikainen, I. Viering, A. Lobinger, and T. Jokela, “On the basics of conditional handover for 5G mobility,” in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2018, pp. 1–7.