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

Bohan Zhang*, Mohammad A. Salahuddin*, Peng Hu*†, Yunli Wang‡,
Noura Limam*, Bo Sun*, Diogo Barradas* and Raouf Boutaba*
*University of Waterloo, Canada, †University of Manitoba, Canada, ‡National Research Council, Canada
{b327zhan, m2salahu, peng.hu, b24sun, n2limam, dbarrada, rboutaba}@uwaterloo.ca, {yunli.wang}@nrc-cnrc.gc.ca

***Abstract*—Low Earth orbit (LEO) satellites-based non-terrestrial networks (NTN) are envisioned to complement the fifth-generation (5G) terrestrial networks (TN), enabling global cellular services. However, the high mobility and large coverage of these satellites result in frequent and numerous inter-satellite handovers, leading to signalling storms that degrade the satellite gNodeB services. To address this, we mathematically formulate the handover problem and propose a novel signalling load-aware handover protocol based on conditional handover. We evaluate the effectiveness of the protocol using a customized discrete-event simulator and compare it against a set of baseline conditional handover schemes. Our findings show that the proposed protocol significantly reduces signalling peaks and balances the load more effectively, enhancing the robustness and efficiency of handover in 5G NTN. The simulator is made publicly available.**

***Index Terms*—5G NTN, conditional handover, signalling storm**

I. Introduction

Terrestrial networks (TN) cover roughly 15% of the Earth's surface, leaving vast regions, including seas, mountains, and rural areas without coverage. These areas incur high capital and operational expenses with expected unequal revenue. Low Earth orbit (LEO) satellite constellations, a key part of non-terrestrial networks (NTN), are expected to complement fifth-generation (5G) TN, offering low-latency and ubiquitous connectivity that can benefit numerous verticals, such as ocean freight, fishery, oil exploration, farming, aviation, and emergency services. For instance, in the event of TN outages due to disasters, NTN can temporarily provide services and aid in disaster recovery. According to Global Market Insights [1], the market value of 5G NTN is projected to reach 79.8 billion USD by 2032, which is 19 times more compared to 2023.

To support the aforementioned vertical applications, mobile and satellite operators need to work closely to achieve an integrated solution for 5G NTN. In 5G NTN, satellites should carry regenerative payload to serve user equipments (UEs) as radio access network (RAN). The regenerative payload enables encoding, modulation, switching, and routing, which supports gNodeB (gNB) onboard the satellite to meet 5G requirements [2]. The payload also supports inter-satellite communications through inter-satellite links. Recently, T-Mobile and SpaceX have launched LEO satellites with onboard 4G eNodeB to provide Direct to Cell services. Although only SMS is currently supported, calling and data services for phones and IoT devices with common 4G standards are expected by 2025 [2], which is a significant milestone towards 5G NTN.

To achieve 5G NTN, numerous challenges exist, with one drawing significant attention from the industry, i.e., the signalling storm that occurs during the handover of many UEs [3], [4]. In 5G NTN, handovers can be classified into intra-satellite and inter-satellite handovers, while a satellite cell can operate in an earth-moving or earth-fixed fashion. For intra-satellite, the need for handover depends on the cell movement configuration and beam management of a satellite gNB [5]. However, the service link switch between satellites is unavoidable because satellites will move out of a UE's line of sight [4]. Therefore, this work focuses on inter-satellite handover.

The inter-satellite handover happens between two satellite gNBs and is also known as the 5G Xn handover. There are two types of Xn handover protocols, the baseline handover protocol (BHO) and the conditional handover protocol (CHO). In BHO, a source satellite communicates with only one satellite, and a UE detaches from the source and accesses a target satellite 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 BHO, CHO significantly increases the robustness of the handover procedure by preparing more target satellites and decoupling the preparation phase and the execution phase (c.f., Section II).

For the Xn handover, some remarkable differences exist between 5G TN and 5G NTN. First, a LEO satellite gNB covers a much larger area than a 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 interval. 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 consequentially 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 location information. Lastly, these satellites have limited computing, radio, and energy resources. With the future deployment of 5G UEs to reach densities of

978-3-903176-66-9

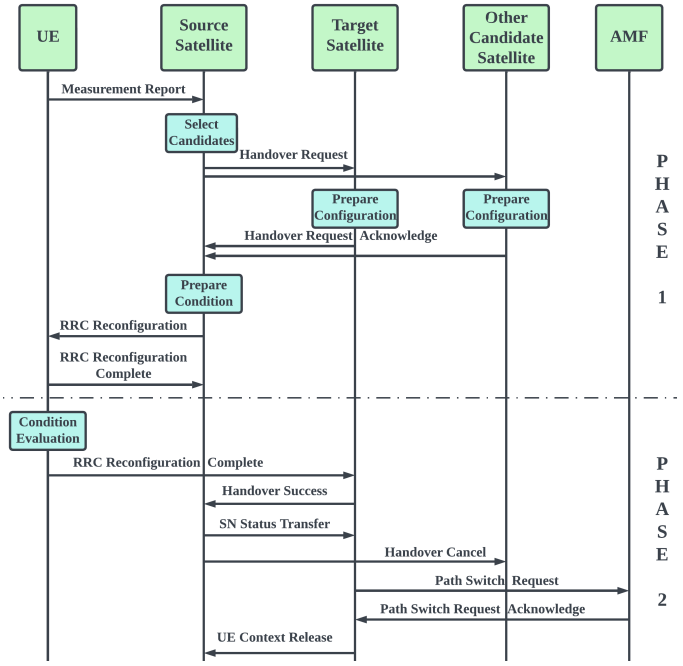

Fig. 1: Conditional handover protocol

10^6 UEs/km^2 (e.g., IoT), as suggested in 3GPP TS 22.261-6.4.2, optimizing inter-satellite handovers should be a priority.

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:

- 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.
- 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.
- We evaluate the performance on a customized discrete-event simulator for 5G NTN, which has been made publicly available at https://github.com/zbh888/congestionHO to facilitate future research.

II. 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.

A. 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. 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.
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 the access and mobility management function (AMF) for a path switch. After receiving the response, the target satellite asks the source satellite to release the UE context, which completes `Phase2`.

B. 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. 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.

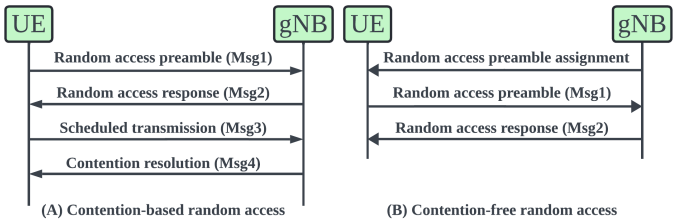

Fig. 2: Random access procedure

In CFRA, as shown in Fig. 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.

C. Related works

Previous works investigated CHO in TN networks [6]–[8]. In [6], authors focused on reducing unnecessary CHO measurement report transmission in TN by applying a convolutional neural network-based classifier on the UE. This classifier assists the UE in making handover decisions based on signal strength. In [7], 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 [8], 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 [9]–[11]. In [9], the authors modeled the mega-satellite constellation and deduced that the rate of handovers significantly increased with the number of satellites. In [10], 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 [11], 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 study, BHO or CHO was considered as an optimization problem with different objectives and solved using heuristics or reinforcement learning (RL) methods in [12]–[14]. In [12], 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, [13] 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 [14], 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-existing complete CHO formulation and the non-existing study of CHO signalling intensity. Previously, we proposed a group handover protocol for BHO to lower signalling intensity [15]. 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 [4], which we plan to explore in the future.

III. 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.

A. Problem

Given a set of UEs $\mathcal{U} = \{u_1, u_2, ...\}$ and a set of satellites $\mathcal{S} = \{s_1, s_2, ...\}$. We divide the time to be a list of consecutive discrete time slots $\mathcal{T} = \{t_1, t_2, ...\}$. 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:

- 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 [3], 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.

TABLE I: Inputs and variables

Notation	Meaning
u	UE device
s	satellite
t	time slot
	(INPUTS)
D	distance threshold for location-based handover triggering
$l^t_{u,s}$	$l^t_{u,s}$=1: u will trigger handover if connected to s at t
$c^t_{u,s}$	$c^t_{u,s}$=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)
$m^t_{u,s}$	$m^t_{u,s}$=1: u is (implicitly) served by s at t
$x^t_{u,s}$	$x^t_{u,s}$=1: s is the source satellite of u in `Phase1` at t
$h^t_{u,s}$	$h^t_{u,s}$=1: s is the target satellite of u in `Phase1` at t
$y^t_{u,s}$	$y^t_{u,s}$=1: s is the other candidate satellite of u in `Phase1` at t
$\hat{x}^t_{u,s}$	$\hat{x}^t_{u,s}$=1: s is the source satellite of u in `Phase2` at t
$\hat{h}^t_{u,s}$	$\hat{h}^t_{u,s}$=1: s is the target satellite of u in `Phase2` at t
$\hat{y}^t_{u,s}$	$\hat{y}^t_{u,s}$=1: s is the other candidate satellite of u in `Phase2` at t
$\delta^t_{u,s}$	$\delta^t_{u,s}$=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)

- The source satellite in `Phase1` will select a fixed number of candidates, $N_{candidate}$.
- We use CFRA to avoid RA failure. Thus, we follow 3GPP [4] to consider a timer-based execution condition [4]. Because dedicated RA preambles and RA occasions are limited resources [3], 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.
- 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 [8] to perform a CFRA to other candidates.
- 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 [16]. Therefore, we assume no additional reconfiguration will occur within $T_{decouple}$.

B. Mathematical formulation

We summarize the inputs and variables of our problem formulation in Table I. For inputs, let $c^t_{u,s} \in \{0,1\}$ denote the location relation between the UE u and the coverage center of the satellite s at time t, where $c^t_{u,s} = 1$ if the distance is less than or equal to D, and $c^t_{u,s} = 0$ otherwise. Thus, $c^t_{u,s} = 1$ implicitly shows u can be served by s at time t, and $c^t_{u,s} = 0$ otherwise. To simplify, we let $l^t_{u,s} \in \{0,1\}$ denote if the handover triggering condition of u from source s is satisfied, i.e., $l^t_{u,s} = 1$ if and only if $c^t_{u,s} = 1$ and $c^{t+1}_{u,s} = 0$. When $l^t_{u,s} = 1$, u and s should perform `Phase1` if u is served by s.

For decision variables, we first define $m^t_{u,s} \in \{0,1\}$ to denote the actual or implicit serving relationship between u and s. For instance, $m^t_{u,s} = 1$ 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 $m^t_{u,s} = 0$ otherwise. $m^t_{u,s}$ follows the property in (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 (2) ensures u must always, actually or implicitly, be served by a satellite with distance to the coverage center smaller or equal to D.

$$\sum_{s' \in \mathcal{S}} m^t_{u,s'} = 1, \forall u \in \mathcal{U}, t \in \mathcal{T} \tag{1}$$

$$m^t_{u,s} \leq c^t_{u,s}, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \tag{2}$$

Phase1: Let $h^t_{u,s} \in \{0,1\}$ denote the implicit handover action of u to s, where $h^t_{u,s} = 1$ if s, as target satellite, will be the next serving satellite of u, and $h^t_{u,s} = 0$ otherwise. We define (3) to ensure that u can implicitly handover to only one target satellite or not perform implicit handover at t. Let $x^t_{u,s} \in \{0,1\}$ denote handover action of u from s, where $x^t_{u,s} = 1$ if s is the source satellite, and $x^t_{u,s} = 0$ otherwise. Constraint (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 $h^t_{u,s} = 1$ only if $c^t_{u,s} = 1$, as denoted in (5).

$$\sum_{s' \in \mathcal{S}} h^t_{u,s'} \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T} \tag{3}$$

$$\sum_{s' \in \mathcal{S}} h^t_{u,s'} = \sum_{s' \in \mathcal{S}} x^t_{u,s'}, \forall u \in \mathcal{U}, t \in \mathcal{T} \tag{4}$$

$$h^t_{u,s} \leq c^t_{u,s}, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \tag{5}$$

Also, the handover will change the actual/implicit serving satellite. Constraint (6) captures that, at t, if no handover happened at $t-1$ (i.e., $(\sum_{s' \in \mathcal{S}} h^{t-1}_{u,s'}) = 0$), the serving relation will not change. Otherwise, the serving satellite should switch.

$$\begin{aligned} m^t_{u,s} = &(\sum_{s' \in \mathcal{S}} h^{t-1}_{u,s'}) \times h^{t-1}_{u,s} \\ &+ (1 - \sum_{s' \in \mathcal{S}} h^{t-1}_{u,s'}) \times m^{t-1}_{u,s}, \forall u \in \mathcal{U}, t \in \mathcal{T}, \forall s \in \mathcal{S} \end{aligned} \tag{6}$$

We let $y^t_{u,s} \in \{0,1\}$ denote the candidates selection other than the target satellite, where $y^t_{u,s} = 1$ when the source satellite selects s as candidate during the handover of u, $y^t_{u,s} = 0$ otherwise. Constraints (7) and (8) ensure that the selected candidate cannot be source satellite and target satellite.

$$y^t_{u,s} + h^t_{u,s} \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \tag{7}$$

$$y^t_{u,s} + x^t_{u,s} \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \tag{8}$$

In turn, (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 (10), $y^t_{u,s} = 1$ only if $c^t_{u,s} = 1$.

$$\sum_{s' \in \mathcal{S}} y^t_{u,s'} = (N_{candidates} - 1) \sum_{s' \in \mathcal{S}} h^t_{u,s'}, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (9)$$

$$y^t_{u,s} \leq c^t_{u,s}, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (10)$$

Phase2: One challenge of modelling the conditional handover is to maintain the satellite roles between `Phase1` and `Phase2` across time. We define $\hat{x}^t_{u,s} \in \{0,1\}, \hat{y}^t_{u,s} \in \{0,1\}, \hat{h}^t_{u,s} \in \{0,1\}$ to denote the satellite roles in `Phase2`, which correspond to their roles in `Phase1`. We first discuss the case for target satellite. We let $\hat{h}^t_{u,s} = 1$ when s is the target satellite in `Phase2` to which u performs random access, and $\hat{h}^t_{u,s} = 0$ otherwise. In our model, the target satellite will eventually be the next serving satellite and continue to serve UE until $l^t_{u,s} = 1$. 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 (11). Obviously, $\hat{h}^t_{u,s} = 1$ only if $c^t_{u,s} = 1$, as denoted in (12). To ensure that u cannot access s while performing handover from s, we define (13).

$$\sum_{t' \in \mathcal{T}} \hat{h}^{t'}_{u,s} = \sum_{t' \in \mathcal{T}} h^{t'}_{u,s}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (11)$$

$$\hat{h}^t_{u,s} \leq c^t_{u,s}, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (12)$$

$$\hat{h}^t_{u,s} + x^t_{u,s} \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (13)$$

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 (14), the access time cannot be the same as the selection time, while (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.

$$\hat{h}^t_{u,s} + h^t_{u,s} \leq 1, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (14)$$

$$0 \leq \sum_{t' \in \mathcal{T}} (t' \times \hat{h}^{t'}_{u,s}) - \sum_{t' \in \mathcal{T}} (t' \times h^{t'}_{u,s}) \leq T_{decouple}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (15)$$

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 (16) enforces this property.

$$\sum_{u' \in \mathcal{U}} \hat{h}^t_{u',s} \leq MAX_{access}, \forall t \in \mathcal{T}, s \in \mathcal{S} \quad (16)$$

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 (11) and (15), we define (17) and (18).

$$\sum_{t' \in \mathcal{T}} \hat{x}^{t'}_{u,s} = \sum_{t' \in \mathcal{T}} x^{t'}_{u,s}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (17)$$

$$0 \leq \sum_{t' \in \mathcal{T}} (t' \times \hat{x}^{t'}_{u,s}) - \sum_{t' \in \mathcal{T}} (t' \times x^{t'}_{u,s}) \leq T_{decouple}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (18)$$

It is also important to ensure time consistency in `Phase2`. We define constraint (19) to ensure this property.

$$\sum_{s' \in \mathcal{S}} \hat{x}^t_{u,s'} = \sum_{s' \in \mathcal{S}} \hat{h}^t_{u,s'}, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (19)$$

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 (20).

$$\sum_{s' \in \mathcal{S}} \hat{y}^t_{u,s'} = (N_{candidates} - 1) \sum_{s' \in \mathcal{S}} \hat{h}^t_{u,s'}, \forall u \in \mathcal{U}, t \in \mathcal{T} \quad (20)$$

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 (21).

$$\sum_{t' \in \mathcal{T}} \hat{y}^{t'}_{u,s} = \sum_{t' \in \mathcal{T}} y^{t'}_{u,s}, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (21)$$

To ensure the ordering correctness, we utilize an ancillary binary variable $\delta^t_{u,s} = \{0,1\}$, which acts as a lock. When $\delta^t_{u,s} = 1$, s cannot be a candidate of u in `Phase1`, but can be a candidate of u in `Phase2`, and $\delta^t_{u,s} = 0$ otherwise. This functionality is enforced using constraints (22) and (23).

$$y^t_{u,s} \leq (1 - \delta^t_{u,s}), \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (22)$$

$$\hat{y}^t_{u,s} \leq \delta^t_{u,s}, \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (23)$$

To make the lock behave as expected we introduce (24), where $\delta^t_{u,s}$ switches to 0 at t when s is a candidate of u in `Phase2` at $t-1$, and $\delta^t_{u,s}$ switches to and continue to be 1 when s is a candidate of u in `Phase1` at $t-1$. Moreover, (25) is necessary to ensure `Phase2` will always happen after `Phase1`.

$$\delta^t_{u,s} = (1 - \hat{y}^{t-1}_{u,s})(\delta^{t-1}_{u,s} + y^{t-1}_{u,s}), \forall u \in \mathcal{U}, t \in \mathcal{T}, s \in \mathcal{S} \quad (24)$$

$$\delta^{t=0}_{u,s} = 0, \forall u \in \mathcal{U}, s \in \mathcal{S} \quad (25)$$

Combining constraints (1)-(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 (26). The coefficient before each decision variable is based on the received signalling count in Fig. 1.

$$\min_{m,y,h,\hat{x},\hat{y},\hat{h},\delta} \max \sum_{u' \in \mathcal{U}} ((2 + N_{candidates}) x^t_{u',s} + y^t_{u',s} + h^t_{u',s} + 2\hat{x}^t_{u',s} + \hat{y}^t_{u',s} + 3\hat{h}^t_{u',s}), \forall t \in \mathcal{T}, s \in \mathcal{S} \quad (26)$$

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 $l^t_{u,s} = 1$, and there is only one time slot for any (u, s) pair across all time slots that has $l^t_{u,s} = 1$. Thus, we can remove those t where $\sum_s l^t_{u,s} = 0$ 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.

IV. Load-aware Conditional Handover

A. 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 (26).

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 in the far future could be more inaccurate. Also, these loads will be frequently updated and will be used to help real-time decision. 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. 1, `Select Candidates`, `Prepare Configuration`, and `Prepare Condition`. Load management is also included for illustration.

Algorithm 1 Select Candidates (SOURCE; `Phase1`)

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:** $candidates$, $candidatesLoads$

B. 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 $reportedSatellites$ 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 $O(reportedSatellite\texttt{.length})$, and details are in Alg. 1.

C. 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 is aware of the loads of other candidates, it can compute the maximum signalling count of each time slot within

Algorithm 2 Prepare Configuration (CANDIDATE; `Phase1`)

Input: $storedLoads$, α, β, γ $myLoad$, t^l, $candidates$, $candidateLoads$
$MaxLoads$ = []
for t in $T_{decouple}$ **do**
  $tLoad$ = []
  $myActualLoad$: Find actual load in $myLoad$ at t.
  $myPotentialLoad$: Find potential load in $myLoad$ at t.
  $tLoad.append(myActualLoad + \alpha \times myPotentialLoad)$
  for s in $candidates$ **do**
    $sActualLoad$: Find actual load in $storedLoads[s]$ at t.
    $sPotentialLoad$: Find potential load in $storedLoads[s]$ at t.
    $tLoad.append(sActualLoad + \alpha \times sPotentialLoad)$
  end for
  $tMaxLoad$: Find max value in $tLoad$.
  $MaxLoads.append(tMaxLoad)$
end for
$availableSlots$: Find time slots with available RA preambles.
$N = \gamma \times len(availableSlots)$
$bestSlots$: Based on $availableSlots$, $MaxLoads$, select N best slots.
$weights = [(bestSlots.length)^\beta, (bestSlots.length - 1)^\beta, ...]$
$bestSlot$: Based on $weights$ and $bestSlots$, perform random selection.
$myLoad.potentialLoad[t^l] += 2 + N_{candidate}$
$myLoad.potentialLoad[bestSlot] += 3$
Increment other candidates' $storedLoads.potentialLoad[bestSlot]$ by 1.
$fActualLoad = myLoad.actualLoad[t^l]$
$fPotentialLoad = myLoad.potentialLoad[t^l]$
$futureLoad = (fActualLoad, fPotentialLoad)$
SEND SOURCE $bestSlot$, $futureLoad$

Algorithm 3 Prepare Condition (SOURCE; `Phase1`)

Input: $futureLoad$, $bestSlot$, α, $candidates$, $myLoad$
Compute aggregated future load for each candidate following $futureLoad.fActualLoad + \alpha \times futureLoad.fPotentialLoad$.
$target$: Pick the candidate with minimum aggregated future load.
$bestSlot$: Find the corresponding timer.
$myLoad.actualLoad[bestSlot] += 2$
SEND UE $target$, $bestSlot$

$T_{decouple}$ 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 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. 2) is bounded by $O(N_{candidate}T_{window})$.

D. 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 $O(N_{candidate})$, and the details are shown in Alg. 3. Overall, LCH has a linear complexity, which shows the potential to be applied in practice.

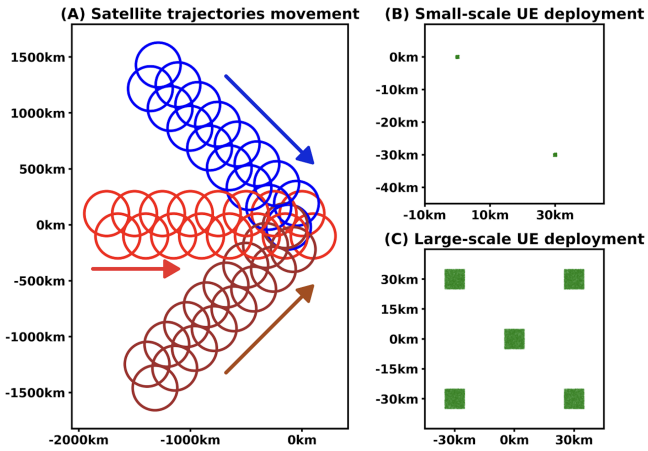

Fig. 3: Experiment scenarios

V. Performance Evaluation

A. 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. 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., $D = 200km$). 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 $1{,}100km$, $7.5km/s$ at $700km$, and $7.7km/s$ at $300km$, respectively.

We compare LCH with combinations of different options of `Select Candidates`, `Prepare Configuration`, `Prepare Condition` as CHO benchmarks, shown in Table II. OPT indicates the optimal solution from the solver. LCH corresponds to our heuristic algorithm, where $\alpha = 1$, $\beta = 0$, and $\gamma = 0.2$. 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 LCH in small- and large-scale scenarios, and analyze the trade-offs among reservation time, serving time, and load balancing.

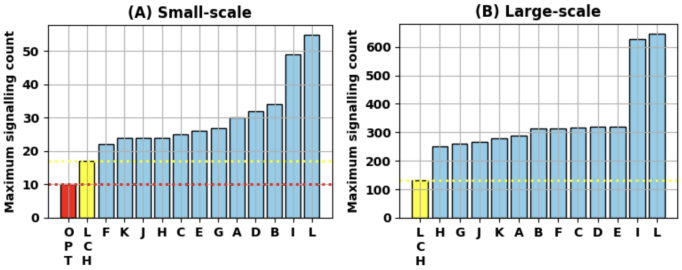

Fig. 4: Signalling peak evaluation

TABLE II: 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

B. 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. 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 solver to derive the optimal solution, and our simulator leverages theinitial 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. 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. 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. 4(B) shows 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, LCH reduces maximum signalling count by roughly 50% in all cases.

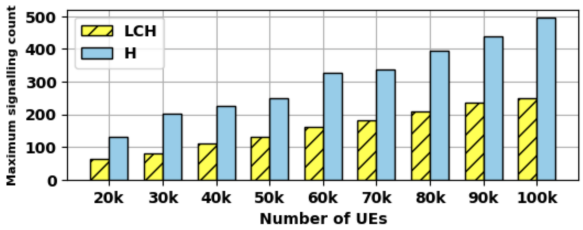

Fig. 5: Signalling peak comparison with different UE density

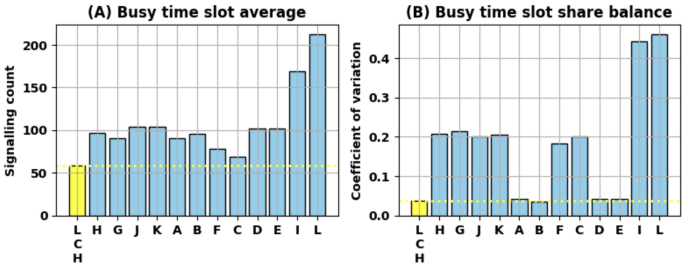

Fig. 6: Busy time slot evaluation

C. 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.

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. 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. 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. 7(A) shows that LCH leads to additional handovers. In practice, this could increase the handover failure rate. Also, more 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. 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,

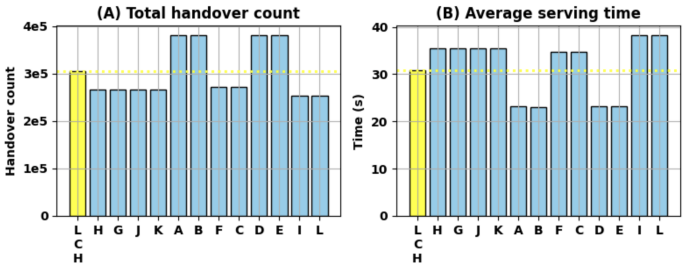

Fig. 7: Handover count and serving time evaluation

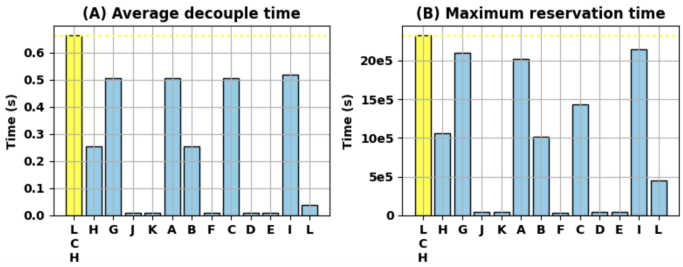

Fig. 8: Decouple time and reservation evaluation

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. 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. 8(B) shows satellites using LCH generally reserves higher 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.

VI. 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 and use cases 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.

VII. 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.

Acknowledgement

We acknowledge the support of the NRC-Waterloo Collaboration Center (project reference number 090755).

References

[1] P. Wadhwani, "5G NTN Market," Global Market Insights Inc., Tech. Rep., 2 2024, Accessed: 07-11-2024. [Online]. Available: https://www.gminsights.com/industry-analysis/5g-ntn-market

[2] SpaceX, "Starlink Direct to Cell," https://www.starlink.com/business/direct-to-cell, Accessed: 07-11-2024.

[3] 3rd Generation Partnership Project, "Solutions for NR to Support Non-Terrestrial Networks (NTN)," Tech. Rep. 33.821 v16.2.0, 04 2023, Accessed: 07-11-2024. [Online]. Available: https://tinyurl.com/2fufe4tk

[4] Ericsson - 3GPP, "3GPP TSG-RAN WG2 Meeting #112: Impacts of Earth Fixed and Moving Beams," Tech. Rep., 11 2020, Accessed: 07-11-2024. [Online]. Available: https://tinyurl.com/9wfwuhrz

[5] A. Sattarzadeh, Y. Liu, A. Mohamed *et al.*, "Satellite-based Non-Terrestrial Networks in 5G: Insights and Challenges," *IEEE Access*, vol. 10, pp. 11 274–11 283, 2021.

[6] A. Gharouni, U. Karabulut, A. Enqvist *et al.*, "Signal Overhead Reduction for AI-assisted Conditional Handover Preparation," in *Mobile Communication-Tech. and Appl.; 25th ITG-Symposium*, 2021, pp. 1–6.

[7] S. B. Iqbal, A. Awada, U. Karabulut *et al.*, "On the Modeling and Analysis of Fast Conditional Handover for 5G-Advanced," in *IEEE Symp. on Personal, Indoor and Mobile Radio Comm.*, 2022, pp. 595–601.

[8] J. Stańczak, U. Karabulut, and A. Awada, "Conditional Handover Modelling for Increased Contention Free Resource Use in 5G-Advanced," in *IEEE Symp. on Pers., Indoor and Mobile Radio Comm.*, 2023, pp. 1–6.

[9] S. Ji, D. Zhou, M. Sheng *et al.*, "Mega Satellite Constellations Analysis Regarding Handover: Can Constellation Scale Continue Growing?" in *IEEE Global Communications Conference*, 2022, pp. 4697–4702.

[10] L. Liu, Y. Li, H. Li *et al.*, "Democratizing Direct-to-Cell Low Earth Orbit Satellite Networks," in *21st USENIX Symposium on Networked Systems Design and Implementation*, 2024, pp. 791–808.

[11] J.-H. Lee, C. Park, S. Park *et al.*, "Handover Protocol Learning for LEO Satellite Networks: Access Delay and Collision Minimization," *IEEE Transactions on Wireless Communications*, 2023.

[12] S. He, T. Wang, and S. Wang, "Load-Aware Satellite Handover Strategy Based on Multi-Agent Reinforcement Learning," in *IEEE Global Communications Conference*, 2020, pp. 1–6.

[13] N. Badini, M. Jaber, M. Marchese *et al.*, "Reinforcement Learning-Based Load Balancing Satellite Handover Using NS-3," in *IEEE International Conference on Communications*, 2023, pp. 2595–2600.

[14] F. Wang, D. Jiang, Z. Wang *et al.*, "Seamless Handover in LEO-Based Non-Terrestrial Networks: Service Continuity and Optimization," *IEEE Transactions on Communications*, vol. 71, no. 2, pp. 1008–1023, 2022.

[15] B. Zhang, P. Hu, A. Azirani *et al.*, "Secure and Efficient Group Handover Protocol in 5G Non-Terrestrial Networks," in *IEEE International Conference on Communications*, 2024, pp. 5063–5068.

[16] H. Martikainen, I. Viering, A. Lobinger *et al.*, "On the Basics of Conditional Handover for 5G Mobility," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2018, pp. 1–7.