Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in ICT for Internet and Multimedia

Scheduling resources and managing HARQ retransmissions in Non-Terrestrial Networks

Master Candidate

Francesco Rossato

Student ID 2082121

Supervisor

Professor Michele Zorzi

University of Padova

Co-supervisor

Prof. Marco Giordani

University of Padova

Dr. Matteo Pagin

University of Padova

Academic Year
2023/2024

To my parents
and friends

Abstract

The advent of 5G technology has revolutionized the world of mobile communication, offering higher throughput, lower latency and high reliability. As the demand for connectivity continues to grow, non-terrestrial networks (NTNs) have been identified by the Third generation Partnership Project as crucial for future 5-6G deployments, complementing terrestrial networks with the aim of increasing resiliency and providing connectivity in remote areas. The characteristics of NTNs, such as the long distances involved, the high propagation delay and the larger cell footprint, introduce complexities not typically encountered in their terrestrial counterpart, and the current protocol stack was not designed with such challenges in mind. Through a comprehensive literature review and the use of ns-3 network simulator, this work identifies key characteristics and constraints of non-terrestrial networks, evaluating existing protocols and highlighting their strengths and limitations in this challenging scenario, before finally delving into the adaptation and optimization of 5G New Radio Layer-2 protocols for NTNs, proposing innovative solutions and paving the way for new applications and services that leverage the full potential of 5G technology beyond traditional terrestrial boundaries.

Contents

List of Figures

List of Acronyms

TN terrestrial networks

NTN non-terrestrial network

NTNs non-terrestrial networks

HAPs high-altitude platforms

HAP high-altitude platform

UAVs unmanned aerial vehicles

UAV unmanned aerial vehicle

MEO medium earth orbit

HARQ hybrid automatic repeat request

ARQ automatic repeat request

NR new radio

LEO low earth orbit

GEO geosynchronous equatorial orbit

UE user equipment

RACH random access channel

gNB gNodeB

3GPP third generation partnership project

MAC media access control

RTT round-trip time

ACK acknowledgement

TB transport block

TBS total block size

SINR signal plus interference to noise ratio

SNR signal-to-noise ratio

NACK negative acknowledgement

ISL inter-satellite links

FEC forward error correction

RV redundancy version

TDD time division duplexing

SR scheduling request

BSR buffer status report

1 Introduction

1.1 The non-terrestrial era

The third generation partnership project (3GPP)[1], the standardization body tasked with the development of protocols for mobile communication networks, recently put a great emphasis on the importance of the integration of different access technologies along with the existing terrestrial mobile telecommunication infrastructure [1].

In the specific, the envisioned future for mobile communications, starting with the already established 5G new radio (NR) and further expanding with the new sixth-generation cellular networks (6G), foresees the integration of a new non-terrestrial component. The latest 3GPP releases (Rel. 17 and Rel. 18) require 5G and 6G networks to be able to provide non-terrestrial satellite access complementing the already existing terrestrial access technologies [21] [16].

To give a proper definition, the term non-terrestrial networks (NTNs) refers to a category of networks where at least one link is routed via an aerial or space-borne vehicle such as high-altitude platforms (HAPs), unmanned aerial vehicles (UAVs) or telecommunication satellites.

[1] `3gpp.org`

1.2 Limitations of terrestrial networks

In order to explain the motivations behind the choice of expanding the current terrestrial infrastructure, the following paragraphs present a few scenarios where current terrestrial infrastructure has some limitations, failing to provide an adequate service, and NTNs can be used to provide a better coverage.

1.2.1 Remote places

While terrestrial networks (TN) make the well-established foundation of today's mobile communication infrastructure, their own nature poses some intrinsic limitations to their deployment in certain scenarios, especially in rural and remote areas. Conditions such as harsh terrain and geographical impediments act as natural barriers to the deployment of terrestrial infrastructure. Moreover, the need for ground equipments such as base stations and networking gear require the presence of an already established and reliable power grid, further driving up the costs that telecommunication companies would have to sustain.

The population density in remote and rural places is typically much lower than cities, and users are spaced in vast areas of land. Because of this reason, the investment that has to be made in order to provide coverage to a certain number of users would be much higher compared to a more urbanized scenario, where potential users are more densely distributed. The already high infrastructure cost will therefore hardly generate any profit, making this kind of market even more unattractive to private investors and further limiting the possibility for the people living there to access the internet, a resource which is becoming increasingly more important as time goes by.

As studied and documented in [27], the issue of an inadequate broadband coverage in rural regions is an enormous challenge. Providing connectivity to the half of the world population living in rural or underprivileged areas requires a colossal effort, but it would also be a unique opportunity to kick-start the economy of currently underdeveloped countries. Access to the Internet would provide the population a possibility to progress on the educational, environmental, business and health planes, promoting a more fair access to information and alleviating the problem of digital divide between different parts of the world.

1.2.2 Redundancy and additional capacity

Another limitation of the current terrestrial infrastructure is the lack of redundancy and robustness against natural disasters. Extreme events such as earthquakes, hurricanes, fires and floods, but also deliberate behaviors such as targeted attacks by terrorist organizations and sabotages can disrupt the connectivity, leading to outages that can last for a long period of time. This in turn can cause significant economical damage and hinder the already difficult rescue efforts, potentially leading to loss of lives.

The simple installation of a greater number of base stations is not a viable solution because an extreme event such as a tornado would easily be able to render all the additional equipment useless, and the cost of essentially doubling the existing access network to make it more resilient would be enormous.

In this scenario, NTNs can act as a redundant access methodology to decrease the downtimes of terrestrial infrastructure, providing emergency communication services and also additional capacity when required. Fig. 1.1, courtesy of [6], shows a scenario where terrestrial and non-terrestrial access technologies are used transparently to access the network.

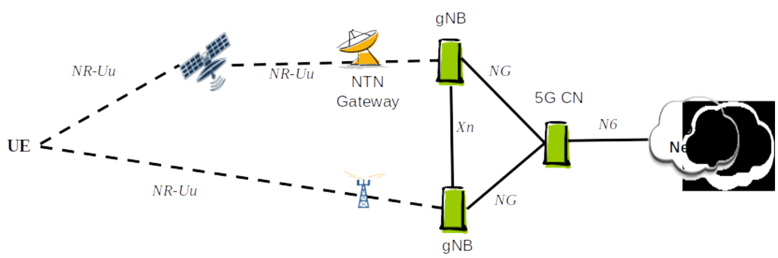

Figure 1.1: Transparent use of multiple radio access technology [6]

1.2.3 Long distances and sensors

Remote equipment, offshore plants and distribution grids will also benefit from the research carried out in this field, since providing terrestrial connectivity in those scenarios is a very challenging task. The installation of an underwater optical fiber link to serve a single endpoint, such as an offshore oil rig or power plant located in the open ocean, would bear a disproportional cost compared to the functions required, and providing maintenance would be another challenging and expensive task. The deployment of non-terrestrial networks would be the easiest way to provide connectivity on a global scale, therefore allowing internet access in isolated places without the need for an expensive dedicated connection.

Consider now the problem of connecting of a number of sensors placed in a large area. When distances are large, solutions may either be the densification of radio stations or the use of a lower frequency in order to have a less severe propagation loss. However, those approaches have their downsides and are not always feasible, requiring costly, ad-hoc solutions tailored for the specific scenario. In this case, the vast coverage area of NTNs will undoubtedly be useful to provide internet connection [25].

Other scenarios where NTNs can become useful in overcoming the limitation of their terrestrial counterpart are well described in [12] and [26].

1.3 Satellite types

Satellites are divided in three different categories depending on their orbiting altitude: geosynchronous equatorial orbit (GEO), medium earth orbit (MEO), and low earth orbit (LEO) satellites. Each one has its own characteristics, presenting some upsides and downsides as briefly described below. Fig. 1.2 illustrates the different orbiting altitudes as well as the approximate coverage area for each orbit.

1.3.1 GEO satellites

Orbiting at a height of 35.786Km, to an observer placed on the Earth surface GEO satellites appear stationary since their orbiting period is the same as the Earth rotational period.

Advantages Since GEO satellites are geostationary, continuous coverage to a designated area can be provided using as little as a single satellite, while the use of non-GEO satellites would require the deployment of a constellation, which is both more complex and more expensive.

This also vastly simplifies the tracking for the ground equipment. Since the satellite position is always known, once the position of the user equipment (UE) is established, the relative position of the satellite can be easily calculated.

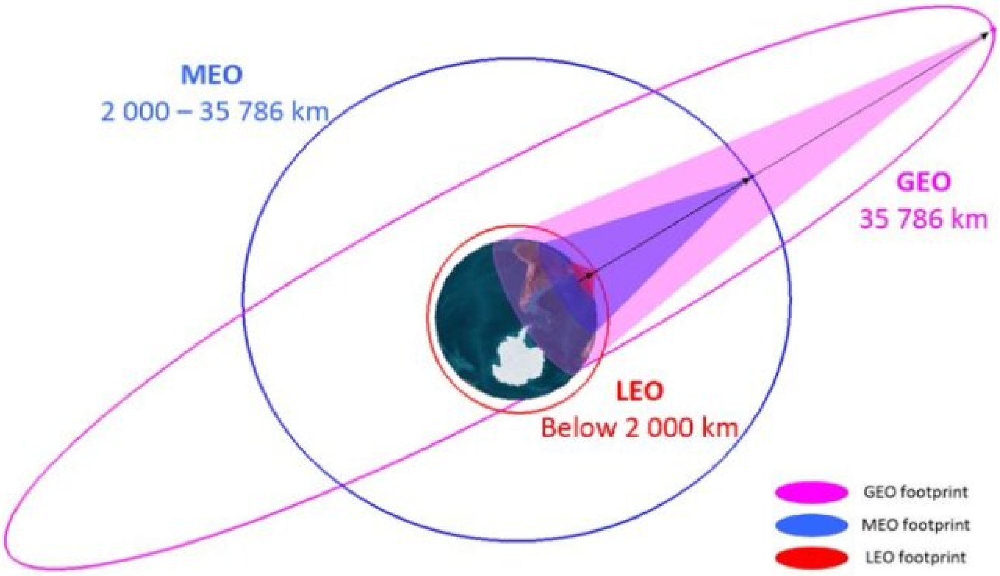

Figure 1.2: Height and approximate coverage areas for satellites at different orbits[15]

As shown in Fig. 1.2, the high altitude of GEO satellites creates a large cell footprint. While the deployment cost of a single GEO satellite is higher than both MEO and LEO ones, the cost per coverage area is overall lower, and an almost full coverage of the terrestrial globe can be achieved using only three equally spaced satellites [11].

Disadvantages The disadvantages of GEO satellites are mainly linked to the large distance with UEs places on the Earth surface: the transmission power and the antenna gain have to be high enough to overcome the greater propagation losses, and the propagation delay of the signal alone adds about 120ms to the overall latency. This means that if the UE sends a request to a server at $t = 0$ through a GEO link, the packet will be received by the destination node at least at $t = 240$ms. The response will then finally reach the UE after at least 480ms from the initial transmission, and these calculations do not factor in any delay related to medium access requests, packet transmission times and processing delays, which would further increased the overall latency.

In addition to the positive aspects previously discussed, the large cell footprint also brings some downsides with it. Due to the vast area, a single satellite will be required to serve a massive number of users, so the total available capacity will have to be shared between a bigger number of equipments, and the throughput experienced by each of them will be reduced.

Solutions to provide a greater capacity have been proposed and are currently in the early stage, such as the use of beamforming to divide the covered area in smaller cells and the employment of higher frequency bands towards Ku, K and Ka as depicted in Fig. 1.3 [10]. Moreover, the high number of users leads to a large rate of initial access requests, with the possibility of channel saturation as described in [6].

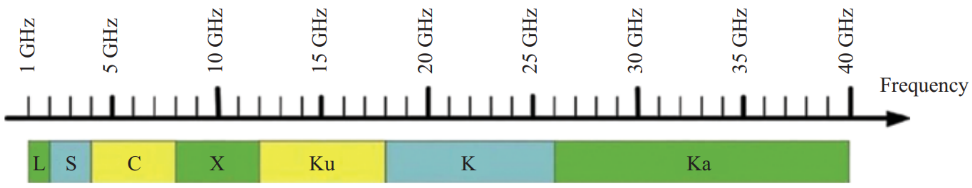

Figure 1.3: Satellite spectrum bands allocation [10]

1.3.2 MEO satellites

MEO orbit is comprised between LEO and GEO, therefore all satellites orbiting between 2000Km and 35.786Km are considered as MEO. This vast orbital space is mainly used by navigation systems such as GALILEO [11].

The propagation delay can vary a lot depending on the altitude, but it is larger than LEO and smaller than GEO. The same point can be made regarding the cell size and number of served users.

MEO satellites do require a constellation in order to provide continuous coverage over a designated area, since they are not geostationary. The numerology of the constellation, however, is smaller than the one required for LEO satellites, since each platform can serve a larger area.

Their peculiarity of presenting many of the downsides that characterize both GEO and LEO satellites, while being unable to offer any substantial benefit over their competitors besides the need for a smaller constellation, makes them less than ideal candidates for applications in non-terrestrial networks.

1.3.3 LEO satellites

Orbiting below the threshold altitude of 2.000Km, LEO satellites are the most promising solution in the realm of NTNs for a number of reasons hereby discussed.

Advantages The lower altitude entails a shorter propagation delay, on the order of about 6ms, and the smaller coverage area of each satellite depicted in Fig. 1.2 means that the total number of users that need to be served is smaller. This also allows the use of higher frequency bands and the constraints of high antenna gains are less stringent compared to GEO satellites, since the experienced path loss is much smaller. This in turn enables the achievement of higher overall throughputs, more suited to satisfy the requirements of modern days broadband connectivity, as detailed in [13].

Disadvantages The cost per deployed satellite is significantly smaller than GEO and MEO satellites, and multiple deployments within a single launch are possible, further driving the costs down. However, given that LEO satellites are not geostationary, a large constellation is needed to provide a continuous service, driving up the deployment costs significantly. As an example of how vast those constellations can become, Fig. 1.4 depicts the LEO satellites employed by Starlink[2], with 4.808 units in service at the time of writing[3].

Since low orbiting satellites remain in view of the user equipments only for a short period of time, with an average in-view duration of just 13 minutes as calculated in [18], all the connected users are expected to be handed over to the next available satellite within this time window. Such behavior would create a noticeable protocol overhead, consuming available channel capacity and potentially adding more latency. However, the predictable nature of this phenomenon might allow for a partial automation without requiring data to be exchanged.

The small coverage area means that more terrestrial gateways have to be deployed, since each satellite can only communicate with the ground via the terrestrial gateways that fall within its view. A different solution to the densification of gateways is the use of inter-satellite links: high-bandwidth links between different satellites of the constellation, capable of connecting satellites that do not have gateways in sight to ones that are connected to a gateway, allow-

[2]Starlink is a satellite internet constellation operated by Starlink Services, LLC, a wholly-owned subsidiary of American aerospace company SpaceX.

[3]Source: `satellitemap.space`

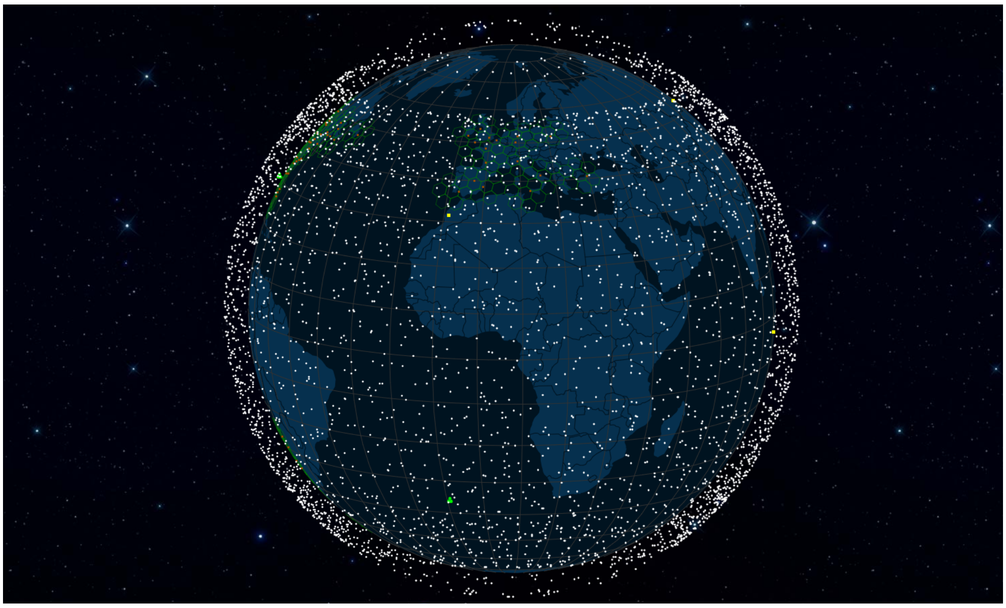

Figure 1.4: Starlink constellation as July 2024. Source: `satellitemap.space`

ing traffic to be routed to the ground with an additional hop. Inter-satellite links have to be implemented if coverage over the oceans is required. This ultimately adds up to the already high constellation deployment costs.

A problem affecting all the non-geostationary satellites involves their speed relative to the user equipment located on the ground. A LEO satellite moves with a speed of about 7.8Km/s [17], presenting therefore a noticeable Doppler shift. This has to be compensated for, and preliminary solutions in this sense require usage of UEs with GNSS capabilities, which is not always a reasonable assumption [13, 2].

1.3.4 Multilayered networks

The aforementioned solutions are not to be considered mutually exclusive. In fact, the multitude of possibilities that their combinations offer opens to the study of various different scenarios, in which the upsides of the space, air, and ground layers are orchestrated to improve quality of service. Fig. 1.5 showcases a highly sophisticated non-terrestrial multilayered network scenario where different access technologies are used.

In [26] it has been shown that the use of high-altitude platform (HAP) as relays between the ground segment of the network and the upper GEO satellite links can deliver up to six times the capacity, and better overall outage probabil-

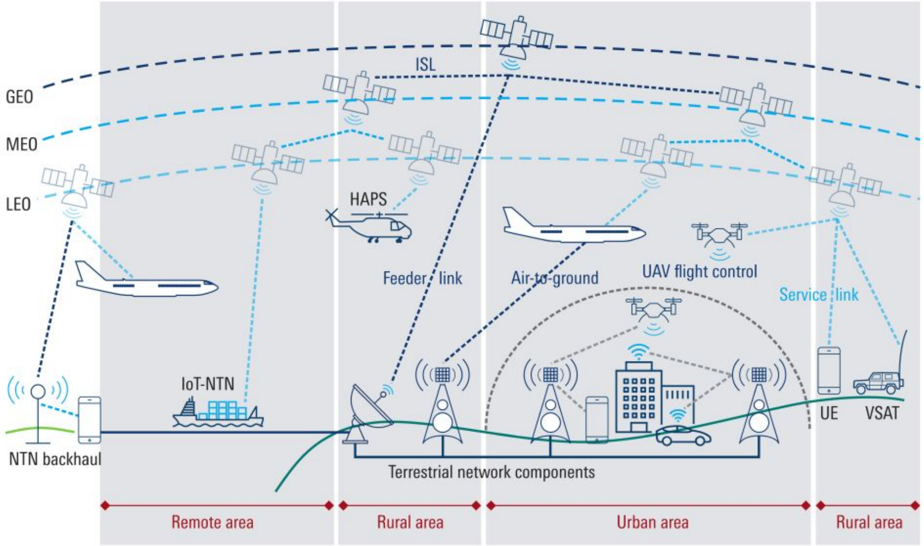

Figure 1.5: Complex multilayered NTN scenario [20]

ity, than point-to-point GEO transmissions.

1.4 Types of payloads

When implementing a non-terrestrial network, an important parameter to be decided is the type of payload to use.

1.4.1 Bent-pipe payload

This is the simplest approach, where the role of the satellite consist only of repeating the signal received from the UE on the ground towards the terrestrial gateway. Configurations such as the one depicted in Fig. 1.6 go by the name of bent pipe payloads and are characterized by the presence of a terrestrial g-NodeB, while the satellite has the sole purpose of providing a transparent link with the UE.

While such solutions are by far the simplest in terms of payload complexity, the main drawback is the even longer experienced latency, since communications between users served by the same satellite would also have to be routed through the terrestrial gateway, increasing the latency by at least two times the propagation delay. This scenario also poses strict bandwidth requirements on the feeder link, since all the traffic must necessarily pass through it. Other, smarter, solutions are able to route at least part of the inbound traffic autonomously, without routing everything back to earth.

Figure 1.6: NTN with access network based on satellites with bent pipe payload [6]

1.4.2 On-board g-NodeB

A slightly more sophisticated approach foresees the installation of g-NodeB capabilities directly onto the satellite payload. This has the benefit of reducing the experienced latency in some cases, and reduce the utilization of the feeder link. Certain protocols, designed to terminate at the gNodeB (gNB), can in this case reach their designated endpoint without necessitating to be routed back to the ground gateway.

Figure 1.7 from [6] shows the architecture of a NTN using access network based on satellites with gNB on board as payload.

Figure 1.7: NTN with access network based on satellites with gNB payload [6]

1.5 Available commercial solutions

Commercial solutions first concerning satellite-based phone calls and successively evolved to provide internet access have been available since a long time. However, the majority of this legacy infrastructure makes use of GEO satellites, therefore presenting all the limitations discussed in section 1.3, offering limited throughput and large delays.

While commercial solutions using LEO satellites have been successfully deployed, there are many proprietary protocols involved, and until now no internationally agreed standard has been defined.

2
Scheduling problems

Before starting to tackle the problems that the high propagation delay causes to the HARQ protocol, attention shall be put on the random access procedure and on the scheduler, both happening at a lower level, therefore necessary to be able to receive packets.

If those layers are not in working conditions, the communication will not be able to take place.

2.1 5G scheduler

As the name suggests, the main task of the scheduler is to allocate resources to the various connected users in the form of transmission and reception opportunities. In mobile communication networks, the scheduling is dictated by the network and the UE has to follow the provided indications.

5G channels

The NR standard comprises the usage of several channels depending on the type of data to be sent. The medium access control layer is a sort of intermediary, using channels provided by the underlying physical layer and providing the so-called logical channels to the upper layers.

A brief overview of such channels can be found in Table 2.1, listing the ones provided by the physical layer, and in Table 2.2, listing the ones provided by the media access control (MAC) layer. Figure 2.1 provides a visual clue at the separation in place between the channels at different levels of the protocol stack.

Transport channel	Acronym
Broadcast channel	BCH
Downlink shared channel	DL-SCH
Paging channel	PCH
Uplink shared channel	UL-SCH
Random access channel	RACH
Sidelink broadcast channel	SL-BCH
Sidelink shared channel	SL-SCH

Table 2.1: Transport channels provided by the physical layer

Logical channel	Acronym	Type
Broadcast control channel	BCCH	control
Paging control channel	PCCH	control
Common control channel	CCCH	control
Dedicated control channel	DCCH	control
Dedicated traffic channel	DTCH	traffic
Sidelink broadcast control channel	SBCCH	control
Sidelink control channel	SCCH	control
Sidelink traffic channel	STCH	traffic

Table 2.2: Logical channels provided by the MAC layer

Different channels are used to convey messages such as the notification of an imminent transmission and the transmission itself, or the uplink scheduling request asked by the UE and the successive grant from the base station [9].

2.1.1 TDD operation

The most complex problem to solve when dealing with propagation delay is certainly the time division duplexing (TDD) mode of operation. In this scenario, each user can communicate only inside its assigned time slots.

Differently from 4G, where a predefined pattern was in place when allocating downlink and uplink allocations in a radio frame, in New Radio this is done much more flexibly using a plethora of parameters such as the periodicity of UL and DL transmissions, the number of consecutive DL and UL slots and symbols at the beginning of each pattern and more.

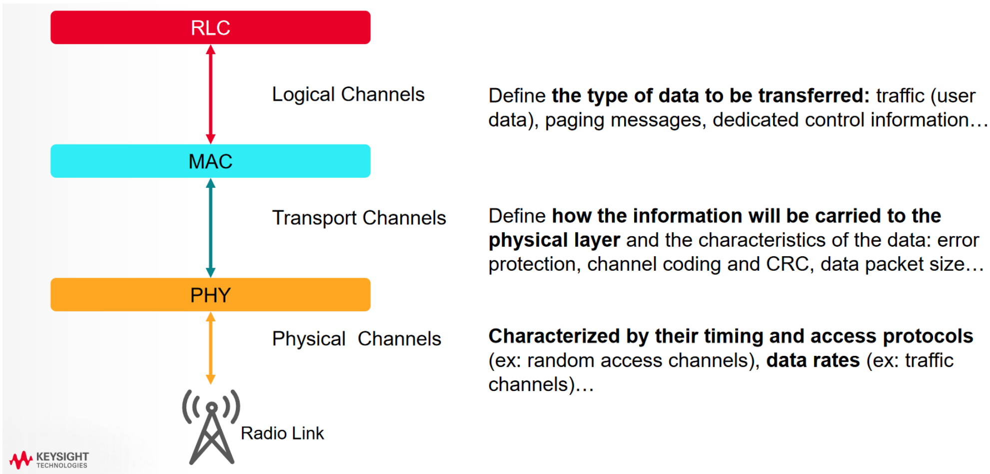

Figure 2.1: Different types of channels used in 5G, courtesy of Keysight technologies

The important key concept is that all the scheduler work has to account for the propagation delay. While in terrestrial communications guard periods of six symbols when switching from downlink communications to uplink communications is sufficient to account for the delay in a 32Km-radius cell, and timing advance commands can account for the different propagation delays experienced by users located in the center of the cell and users located in the cell edge [14], the same cannot be said for non-terrestrial use cases, where the distances that come into play are much longer, therefore any delay is orders of magnitude higher.

2.2 Accounting for propagation delay in scheduling

2.2.1 Problem description

The first encountered problem while implementing a non-terrestrial communication scenario in the simulator was the inability of the scheduler to account for the propagation delay when allocating radio resources to the connected user equipment on the ground.

The implementation of the 5G scheduler in ns-3 is designed to allocate resources less than a single subframe in advance, and since each subframe has a duration time of 1ms, the resource grant was already expired by the time it was able to reach the UE, since it was referring to a past subframe.

Example Consider a scenario with a propagation delay τ_p of 6ms. The UE sends a request for uplink resources at time t_0 since it has some data to send. In the terrestrial scheduler implementation, the gNB would receive such request at $t_0 + \tau_p = 6ms$ and, provided that other transmissions have not been scheduled yet, grant the UE the possibility to transmit in the following slot, which will start after 1ms at $t_0 + \tau_p + t_{slot} = 7ms$. However, this grant will reach the UE only after another propagation delay, therefore at $t_0 + 2\tau_p = 12ms$, when it will already be too late.

2.2.2 Proposed solution

The implemented solution assumes that the scheduler has knowledge of the propagation delay. This is a reasonable assumption since systems such as GPS already rely on a precise estimation of the delay between the user on the ground and the satellite.

The scheduling then proceeds as normal with the only difference being that the information regarding the propagation delay is used to postpone the allocated symbols.

Example Consider the same scenario of the previous example in section 2.2.1. The new implementation of the scheduler accounts for the propagation delay by allocating the first available slot after τ_p, so the time for the grant to reach the UE is accounted for, and the gNB marks the slot after $2\tau_p$ as reserved.

The last part of reserving a different slot is not as immediate. However, this mechanism needs to be in place because of the behavior depicted in Figure 2.2, where a scenario with propagation delay of 5 slots (roughly 1,2ms) is considered:

- UE sends the scheduling request to gNB at frame 1, subframe 0, slot 0.
- The scheduling request (SR) reaches the gNB after $1\tau_p$ of 5 slots and the UE is scheduled to transmit at frame 1, subframe 2, slot 2 since there is some noticeable propagation delay.
- The SR reaches the UE at frame 1, subframe 2, slot 2, and the UE can transmit right away.
- The packet reaches the gNB after another τ_p, hence the base station needs to know that it cannot schedule other transmissions to take place in this slot, otherwise interference and collisions may arise.

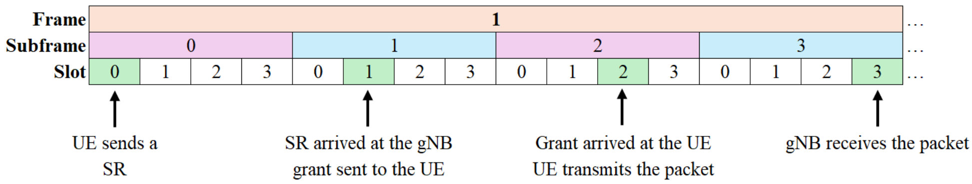

Figure 2.2: Difference between allocated slot and gNB reception

2.3 BSR timer

After implementing the solution discussed in the previous point in the ns-3 network simulator, allowing the communication to take place, some other irregularities were found regarding the periodic buffer status report timer, as followingly described.

2.3.1 Problem description

Reduced latency Figure 2.3 shows a rather peculiar trend regarding latency. The expectations were of a linear increase in latency with a slope of 3, where every packet arrived at the destination after three times the propagation delays. The reasoning behind this expectation was that every packet generated by the application should have triggered a scheduling request received by the gNB after $1\tau_p$, a grant to be emitted requiring another τ_p to reach back to the UE, and finally the transmission to be received by the gNB after the third propagation delay.

The observed pattern did not match any of the expectations. The output plots obtained from the simulation campaign showed a nonlinear saw tooth behavior presenting values that consistently stayed under the expected $3\tau_p$ threshold value (Fig. 2.3).

Periodic buffer status reports Further investigation on the subject allowed to identify numerous buffer status reports sent from the UE to the gNB at regular intervals of 10ms, even when no new packets were produced by the application.

This behavior happens because the implementation of 5G MAC layer includes a periodic BSR that the UE sends to the gNB as long as the transmission buffer contains some pending data. The details of this mechanism are documented in

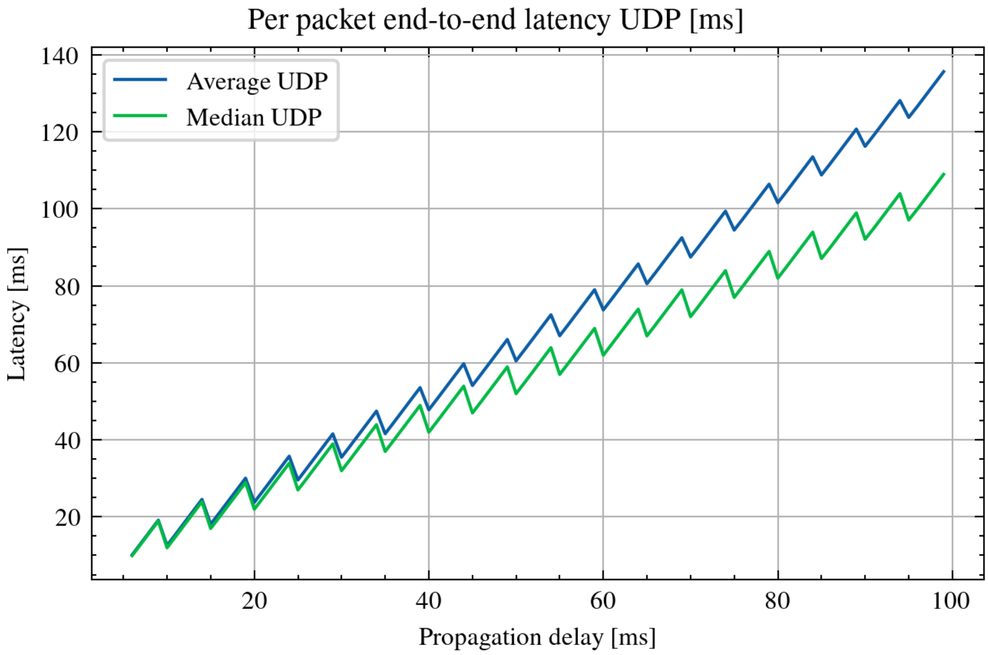

Figure 2.3: E2E latency vs. propagation delay with periodic BSR

Section 5.4.5 of the standard [3], where it is also stated that the default interval between those periodic BSR is of 10ms.

This characteristic was meant to act as a safeguard against lost scheduling requests, and does not account for the propagation delay at all. The UE therefore continues to send additional buffer status reports at regular 10ms intervals even though the previous ones were not lost, but still travelling towards the base station.

Figure 2.4 can provide a visual clue to understand this behavior. The arrival of packet p1 at time zero triggers the transmission of a first SR. Since this scenario has a propagation delay of 20ms, the SR will arrive at the g-NodeB at time 20ms. However, the BSR timer expires in 10ms by default, so during a single round-trip time the SR is repeated four times. The green arrows marked with a capital G denote all the grants issued to the UE. Only one of the four depicted will actually be used by the intended packet, effectively wasting three out of four transmission opportunities that could have been allocated to other users waiting to transmit.

The reduction in latency was observed because newly arrived packets could, in certain situations, make use of the resource grants that were meant for the

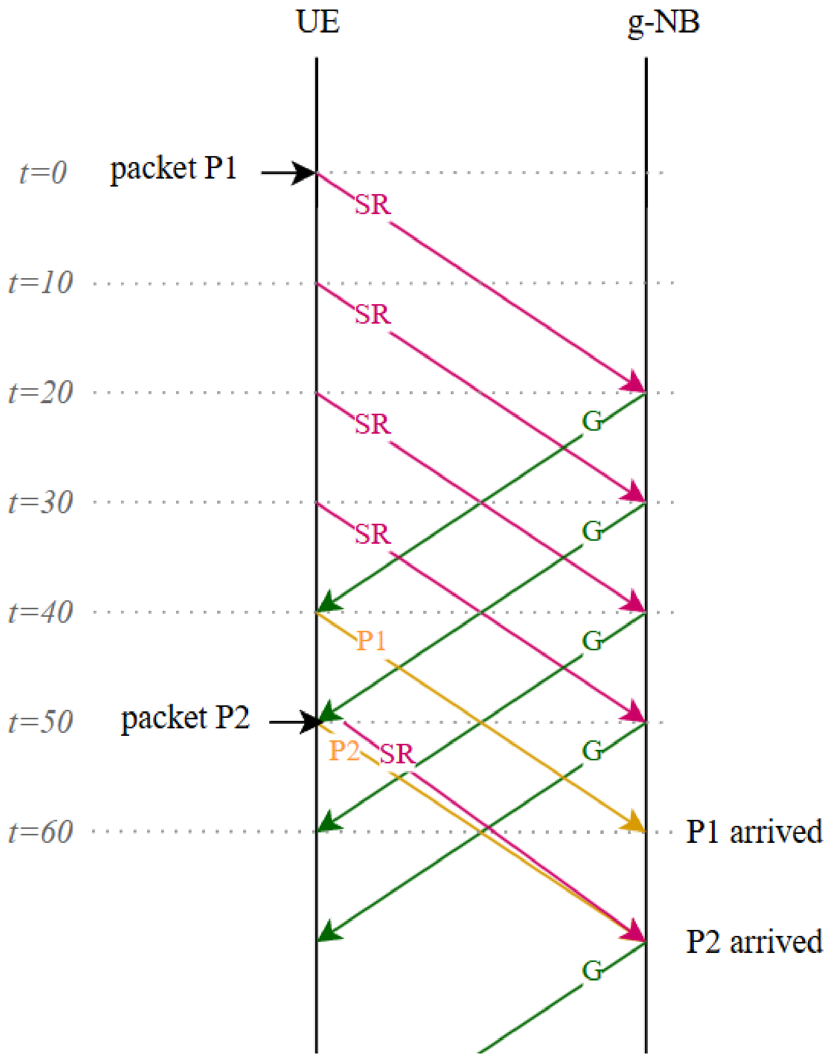

Figure 2.4: Packet diagram with periodic BSR

previous packets but have yet to arrive at the UE. Consider the arrival of the second packet (P2) in figure 2.4. It triggers another SR, but a grant is received immediately after, therefore allowing for an immediate transmission of the second packet, which will experience a latency of only a single propagation delay. This instantaneous grant is nothing but the result of the previous SR storm originated from the UE itself.

While this can present some beneficial aspects such as the reduced overall latency, it originates from a behavior that is outside the original design intentions. Moreover, this approach is particularly greedy, leading to wasted capacity.

2.3.2 Proposed solution

After the successful identification and characterization of the problem, the solution consisted of the implementation of a simple adaptive algorithm that would dynamically set the BSR timer to a value of twice the propagation delay plus a constant of 4ms that accounts for the processing times.

Area of potential improvement The implementation of the aforementioned solution, however, while working as intended, had the expected but unwelcomed effect of increasing the overall latency of the system, since all the additional grants required by previous packets and that newly generated ones could exploit were no longer available.

This unexpected problematic behavior of wasted transmission opportunities accidentally also unveiled the potential of preemptively transmitting additional scheduling requests, so that future packets could be sent with a lower latency.

If implemented correctly, the UE could adopt a predictive approach towards its traffic needs for the near future depending on the type of application, and send the gNB some scheduling requests for packets that have yet to be generated.

2.4 Inflated BSR

2.4.1 Problem description

Another problem occurred when the interval between the packets generated by the application, i.e. the packets interarrival times, was smaller than a single round-trip time.

The arrival of each packet automatically triggers the transmission of a scheduling request by the UE. However, each request is made for the whole transmission buffer. This causes a problem when many packets arrive before the gNB has the chance to respond with the appropriate resource grant.

Referring to Figure 2.5, we can see that the arrival of the first packet in the transmission buffer of the UE triggers the transmission of a SR for a single packet. The second packet P2 arriving after 10ms triggers another request, however this second request is cumulatively made for the full buffer length, regardless of the fact that the first SR is still pending. This is denoted by the writing "SRx2".

Packets P3 and P4 behave in the same way, triggering requests for the allocation of three and four additional packets respectively.

Finally, the grants begin arriving at the UE. The first 1-packet grant allows for the transmission of P1, while the second grant, requested for two packets, allows for the transmission of packets P2 and P3. The third grant, which corresponds to a previous request for 3 packets and could therefore allow the transmission of 3 packets, can now only be used by P4, wasting 2/3 of its potential capacity.

The last grant, that would allow the UE to transmit 4 consecutive packets, is completely wasted since the send buffer is found empty.

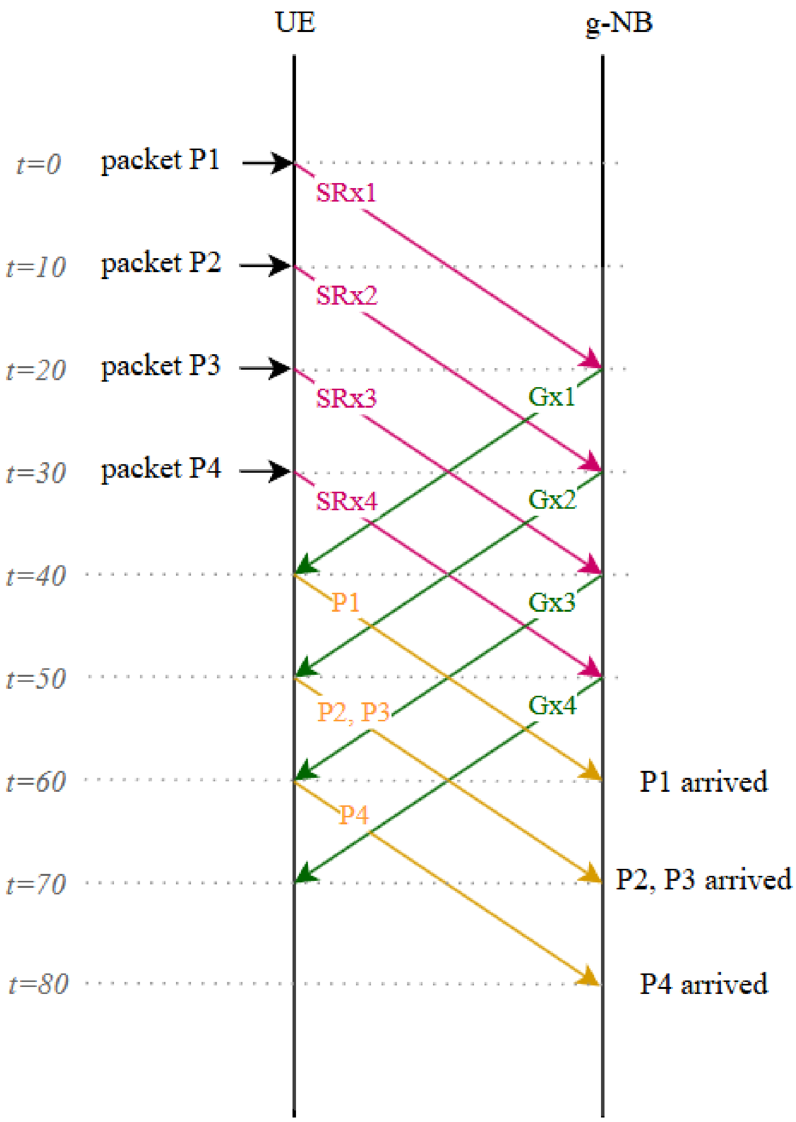

Figure 2.5: Packet diagram for interarrival times smaller than propagation delay

Repercussions The described behavior caused the physical throughput to be significantly higher than the set application source rate, and it rapidly increased with the propagation delay. Figure 2.6 shows the observed physical throughput for a source rate of 160Kb/s, indicating that a lot of radio resources are being wasted to transmit few data.

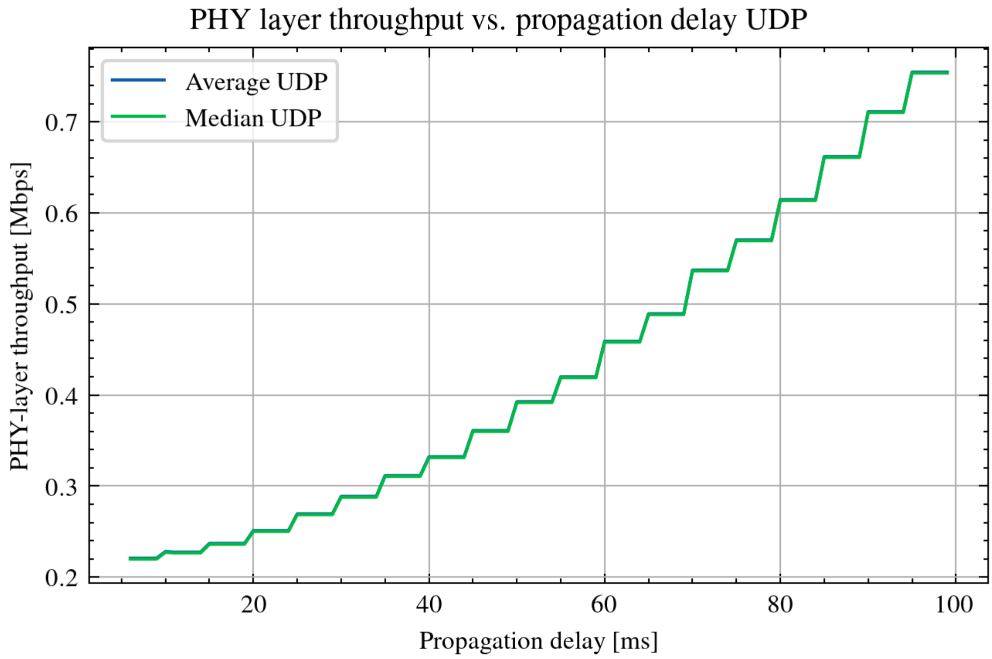

Figure 2.6: Physical throughput vs. propagation delay with periodic BSR

2.4.2 Proposed solution

Many different approaches could be undertaken in order to tackle this problem. Nonetheless, a comprehensive study detailing upsides and downsides of each of them is outside the scope of this work.

The implemented solution consists of a modification to the SR algorithm that limits the requests to only account for the newly received data, therefore disabling the cumulative behavior.

2.5 Reordering timer

TODO: The misalignment between the send interval and the propagation delay leads to the sending of fragmented packets since the grants are always a

bit bigger than the packet UE has to send. However, the gNB reordering timer is configured for terrestrial networks, so it expires before we have a chance of receiving the full picture.

3
HARQ

3.1 Concurrent processes limit

One of the problems highlighted by the 3GPP technical report [6] on the matter of non-terrestrial networks regards the maximum number of concurrent HARQ processes.

3.1.1 Problem description

The details of HARQ protocol implementation in the 5G NR standard is extensively treated in many publications such as [7]. However, for the purpose of understanding what is a HARQ process and how it affects the throughput in a non-terrestrial scenario, a brief overview of a few key concepts is enough.

HARQ working principle

Fig. 3.1 gives an overview of how HARQ processes work. Upon successful reception, depicted in the first column of Fig. 3.1, an acknowledgement (ACK) is sent back, triggering the transmission of the successive transport block (TB), which is represented in the second column. This behavior is the normal state in which transmissions are received correctly, and it keeps repeating itself until errors are detected.

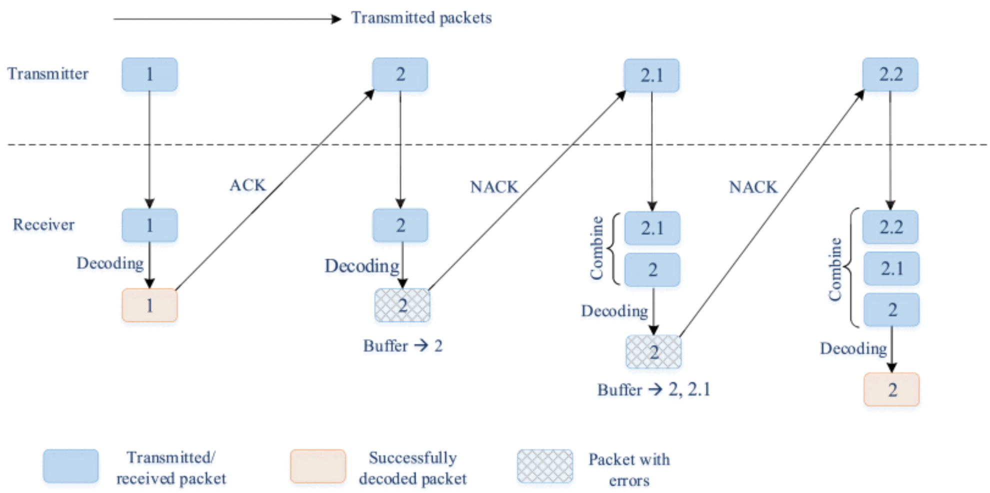

Figure 3.1: HARQ retransmission diagram [7]

Should the receiver detect errors in the received TB, represented by the greyed packet, a negative acknowledgement (NACK) is relayed back to the sender, which in turn proceeds to send some additional redundancy bits. Note that the sender does not repeat the whole TB. The receiver now proceeds to decode the transfer block using all the information that it has received so far.

This is the most important feature of HARQ protocol: it does not discard the packets affected by errors, since they can be at least used to recover some information. The erroneous packets are stored in buffers and used for joint decoding [8].

If the redundancy bits that the sender just transmitted are still not enough to allow for a correct decoding of the transfer block, or if another error is detected, a retransmission is triggered. This is shown in the last column of Fig. 3.1, where all the packets received so far contribute in correctly decoding packet 2.

Finally, if even this retransmission is affected by errors and the combination of the information received so far is not enough to complete the decoding of the TB, additional information is sent. After this fourth interaction, no further attempts are made to correct the packet [23].

The various transmissions and retransmissions being made by the protocol are called redundancy version (RV)

Stop and wait

The presented behavior means that HARQ is a stop-and-wait kind of protocol, since it is designed to wait for the arrival of previous packet's ACK before sending the new one. While this enforces the delivery of ordered packets, it also brings the downside of severely underutilizing the channel capacity, wasting resources that could potentially be used for transmission instead [4].

This limitation is overcome by the introduction of multiple concurrent processes.

Processes

A HARQ process starts when a TB is passed to the HARQ entity and finishes when the ACK relative to that same TB is received by the sender. After the ACK is correctly received, the next TB starts being processed. Considering a link with propagation delay τ_p, the minimum active time for a process is therefore $2\tau_p$, i.e. the time for the transfer block to arrive to the destination plus the time for the acknowledgement to travel back to the sender.

The 5G NR standard allows the base station to configure the maximum number of concurrent HARQ processes to be assigned to each connected user, with the default value being of 8 concurrent processes and the maximum 16 [5, 8].

Application in NTNs

Since the propagation delay of NTNs is order of magnitude larger than their terrestrial counterpart, the limited maximum number of concurrent processes lowers the maximum achievable throughput, as detailed in the following toy example.

Example Consider a scenario where each process tries to send a TB every $2\tau_p$, which is the maximum rate at which transfer blocks can be sent under the condition of waiting for the acknowledgement to arrive before starting a new transmission, to a LEO satellite orbiting at 2.000Km, therefore having $\tau_p \approx 6$ms. Assuming the best possible conditions with no need for retransmissions and assuming that the base station grants the UE to the best possible clearance of 16 concurrent processes, the total send rate is of 16 transfer blocks every 12ms. In order to target a throughput of 50Mbps, the block size must therefore be of at least

$$\frac{\textit{target throughput} \times 2\tau_p}{\textit{number of processes}} = 37,5Kb$$

Doing the same calculation for a terrestrial scenario with the gNB placed at a distance of 600m from the UE, we obtain that the minimum total block size (TBS) must be of just $12b$. Both calculations do not factor in overheads, control information, channel access requests and processing delays, but are helpful to give an idea of the disproportion between the two conditions.

While the necessary block size for the NTN case is technically possible to achieve even with 4G, it necessitates a high signal-to-noise ratio (SNR) to work properly. This constraint becomes even more conservative in the non-terrestrial case, since retransmissions adds delays in multiples of propagation delay and are therefore more costly [22].

3.1.2 Possible solutions

Increasing the processes The easiest solution would be to increase the number of maximum concurrent HARQ processes. However, this comes with some caveats mainly regarding the higher computational capabilities required and higher power consumption, that can quickly become problematic in battery-operated equipments such as smartphones. Each process also requires the presence of a buffer on both the receiver side and the sender side, so additional resources are required at the gNB, too.

Aggressive HARQ A more sophisticated approach could involve the design of an aggressive version of the HARQ protocol, where each process is allowed to send multiple packets before receiving an acknowledgement. Since there already are multiple concurrent processes, each ACK packet must already contain a field specifying the number of process it belongs to, and the information identifying the specific packet to be acknowledged within a process could be encoded inside this field.

Disable HARQ Lastly, the option of disabling HARQ completely and rely solely on automatic repeat request (ARQ) retransmissions has been proposed by 3GPP itself[19]. This, however, would come with a performance penalty since satellite links typically suffer from more severe conditions than terrestrial ones, and [24] demonstrated that a version of HARQ specifically designed for non-terrestrial networks would be beneficial.

3.1.3 Simulator configuration

While the number of concurrent HARQ processes can be configured in ns-3, it cannot exceed the value of 100. By performing some simple calculations, knowing that the SNR conditions allow for the transmissions of TBs with size of 1024B, to achieve a target throughput of 50Mbps on the best case of 6ms τ_p the necessary processes would be 74.

$$\frac{target\ throughput \times 2\tau_p}{total\ block\ size} \approx 74 processes$$

This does not account for the delays caused by retransmissions, so simulator crashes due to processes overflow are frequent while testing even the best case scenario.

4

Conclusions and Future Works

A	**B**
C	D
E	F
G	H

Table 4.1: Table example

References

[1] 3GPP. *TR 21.917 "Release 17 Description; Summary of Rel-17 Work Items" v17.0.1*. Tech. rep. January 2023.

[2] 3GPP. *TR 38.821 "Solutions for NR to support Non-Terrestrial Networks (NTN)" v16.2.0*. Tech. rep. April 2023.

[3] 3gpp. *5G; NR; Medium Access Control (MAC) protocol specification (3GPP TS 38.321 version 18.1.0 Release 18)*. Tech. rep. May 2024.

[4] 3gpp. *TR 38.214 "5G NR Physical layer procedures for data" v16.5.0, Release 16*. Tech. rep. April 2021.

[5] 3gpp. *TR 38.300 "5G NR Overall description Stage-2" v15.9.0, Release 15*. Tech. rep. April 2021.

[6] 3gpp. *TR 38.811 "Study on New Radio (NR) to support non-terrestrial networks" v15.4.0*. Tech. rep. October 2020.

[7] Ashfaq Ahmed et al. "Hybrid Automatic Repeat Request (HARQ) in Wireless Communications Systems and Standards: A Contemporary Survey". In: *IEEE Communications Surveys and Tutorials* 23.4 (2021), pp. 2711–2752. DOI: 10.1109/COMST.2021.3094401.

[8] Devopedia. *"5G NR Hybrid ARQ"*. devopedia.org. Version 6, December 19, Accessed 2024-07-14. 2021.

[9] Devopedia. *5G NR MAC*. devopedia.org. Accessed 2024-07-25. February 2021.

[10] Bruce R. Elbert. "Commercial communications satellites in the post-2020 era". In: *Advances in Communications Satellite Systems. Proceedings of the 37th International Communications Satellite Systems Conference (ICSSC-2019)*. 2019, pp. 1–16. DOI: 10.1049/cp.2019.1235.

[11] ESA. *Types of orbits*. esa.int. Accessed 2024-07-15.

[12] Marco Giordani and Michele Zorzi. "Non-Terrestrial Networks in the 6G Era: Challenges and Opportunities". In: *IEEE Network* 35.2 (2021), pp. 244–251. DOI: 10.1109/MNET.011.2000493.

[13] Marco Giordani and Michele Zorzi. "Satellite Communication at Millimeter Waves: a Key Enabler of the 6G Era". In: *2020 International Conference on Computing, Networking and Communications (ICNC)*. 2020, pp. 383–388. DOI: 10.1109/ICNC47757.2020.9049651.

[14] GSMA. *5G TDD Synchronisation Guidelines and Recommendations for the Coexistence of TDD Networks in the 3.5 GHz Range*. Tech. rep. Accessed 2024-07-21. April 2020.

[15] Marko Höyhtyä et al. "Sustainable Satellite Communications in the 6G Era: A European View for Multi-Layer Systems and Space Safety". In: *IEEE Access* PP (Jan. 2022). Licensed under CC-BY 4.0 public license, pp. 1–1. DOI: 10.1109/ACCESS.2022.3206862.

[16] Helka-Liina Maattanen et al. "5G NR Communication over GEO or LEO Satellite Systems: 3GPP RAN Higher Layer Standardization Aspects". In: *2019 IEEE Global Communications Conference (GLOBECOM)*. 2019, pp. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9014090.

[17] Andrew May. *Low Earth orbit: Definition, theory and facts*. rohde-schwarz.com. Accessed 2024-07-20.

[18] Iestyn Morgan-Jones and Pavel Loskot. *Regional Coverage Analysis of LEO Satellites with Kepler Orbits*. 2019. arXiv: 1910.10704 [physics.space-ph]. URL: https://arxiv.org/abs/1910.10704.

[19] Amitav Mukherjee. "Hybrid ARQ Schemes". In: 2020. URL: https://api.semanticscholar.org/CorpusID:219510991.

[20] Rohde-Schwarz. *Connecting the world with 5G NTN*. rohde-schwarz.com. Accessed 2024-07-20.

[21] Malik Muhammad Saad et al. "Non-Terrestrial Networks: An Overview of 3GPP Release 17 & 18". In: *IEEE Internet of Things Magazine* 7.1 (2024), pp. 20–26. DOI: 10.1109/IOTM.001.2300154.

[22] ShareTechNote. *4G/LTE - PHY processing*. sharetechnote.com. Accessed 2024-07-14.

[23] ShareTechNote. *5G/NR - HARQ*. sharetechnote.com. Accessed 2024-07-14.

[24] Alessandro Vanelli-Coralli et al. “5G and Beyond 5G Non-Terrestrial Networks: trends and research challenges”. In: *2020 IEEE 3rd 5G World Forum (5GWF)*. 2020, pp. 163–169. DOI: 10.1109/5GWF49715.2020.9221119.

[25] Dengke Wang et al. “On the Performance of Non-Terrestrial Networks to Support the Internet of Things”. In: *2022 56th Asilomar Conference on Signals, Systems, and Computers*. 2022, pp. 881–887. DOI: 10.1109/IEEECONF56349.2022.10052102.

[26] Dengke Wang et al. “The Potential of Multilayered Hierarchical Nonterrestrial Networks for 6G: A Comparative Analysis Among Networking Architectures”. In: *IEEE Vehicular Technology Magazine* 16.3 (2021), pp. 99–107. DOI: 10.1109/MVT.2021.3085168.

[27] Elias Yaacoub and Mohamed-Slim Alouini. “A Key 6G Challenge and Opportunity—Connecting the Base of the Pyramid: A Survey on Rural Connectivity”. In: *Proceedings of the IEEE* 108.4 (2020), pp. 533–582. DOI: 10.1109/JPROC.2020.2976703.

Acknowledgments