POLITECNICO DI TORINO

Corso di Laurea Magistrale
in ICT For Smart Societies

Tesi di Laurea Magistrale

Caching through aerial platforms for 6G Networks

A Detailed Simulator for Caching Performance Analysis in Milan

Relatore
Prof.ssa Michela Meo

Co-Relatrice
PhD. Greta Vallero

Candidato
Ten. Lorenzo Maria Mastrodicasa

Anno Accademico 2021-2022

Alle persone che ogni giorno hanno reso possibile la mia rinascita come persona,

All'Esercito che mi ha formato e cresciuto come uomo,

Al me che nonostante tutto ci ha creduto fino alla fine.

Summary

The staggering growth in IP Traffic, the development of ICTs (Information And Communication Technologies) and IoT (Internet of Things) devices and connections have led to an increase in the demand for new applications, and the need for a *ubiquitous*, *reliable* and *scalable* service worldwide [1], [2].
These requirements cannot be entirely fulfilled by the 5G TNs (Terrestrial Networks) due to two main inherent limitations which are the intrinsic energy costs of such an architecture [3], and its deployment in remote and inaccessible regions. Meanwhile, in addition to the aforementioned needs, the increased commercial viability of HAPS (High Altitude Platform Systems) and aerial/space technologies [4], [5], made possible by the advances in solar panel efficiency, antennas, light composite materials, cost of manufacturing and launching, has fueled the conceptualization process of the next generation of wireless communication systems (6G). Moreover, this has also led to an integration of aerial platforms in the overall network architecture, resulting in the creation of the so-called NTN (Non-Terrestrial Network) systems as an effective solution to complement TNs.
According to the 3GPP (3rd Generation Partnership Project), NTNs, emerging as complementary to the current RANs (Radio Access Networks), are part of the so-called VHetNet (Vertical Heterogeneous Network). This new architecture, where the non-terrestrial part has a crucial role, is composed of three connected layers: a satellite network, an aerial network, formed by UAVs (Unmanned Aerial Vehicles) and HAPS, and a terrestrial network [6].
This work focuses on the aerial segment, using a HAPS platform fulfilling the role of the above-mentioned layer, by equipping it with a MEC (Multi-access Edge Computing) server, thus providing it with caching capabilities. In this thesis, we assume that the HAPS is placed over the Milan municipality at an altitude of 20km and is designed to complement the TN in request handling, thanks to the offloading capabilities offered by the onboard MEC.
The HAPS offers coverage to 16 traffic zones within the municipality, whose traffic patterns are known and provided by a large Italian mobile network operator. These profiles report the value, in bits, of the traffic demand volume for 1420 BSs of Milan for two months in 2015 and they highlight districts' differences in terms

of peaks and usage. Moreover, for every zone, a sample cluster of BSs (Base Stations) is considered, comprising one macro BS with ground caching capabilities, supported by six micro cell BSs that are not able to perform any content caching. This study has the main aim of evaluating some key caching KPIs (Key Performance Indicators) such as the *hit* and *miss probabilities* and the quantity of traffic that has to be forwarded to the core network. Furthermore, it inspects how the integration of an aerial MEC changes the overall performance of some well-known caching algorithms and if some particular behaviors or patterns have to be noted when ground and non-terrestrial caches work together. For this reason, an in-depth analysis regarding the traffic profiles is provided, understanding their main features, and many possible scenarios are inspected. Results, carried out considering each of them, are obtained considering both the case of *single-beam* HAPS coverage, meaning that the HAPS offers offloading capabilities to a single district of Milan, and in the *multi-beam* context and they highlight that the considered parameters strongly vary from district to district depending on the traffic demand profile of the RAN portion under inspection.

Acknowledgements

I thank Prof. Michela Meo who allowed me to carry out this thesis work, advising and following me step by step with extraordinary positive energy. I also thank my co-supervisor, Eng. Greta Vallero for her competence, but most of all, for her patience, dedicating me a lot of time in these months and answering my questions and curiosities with great passion and kindness.

How dull it is to pause, to make an end,
To rust unburnish'd, not to shine in use!
As tho' to breathe were life!

[A. Tennyson, *Ulisses*]

List of Figures

Contents

List of Acronyms

3GPP	Third Generation Partnership Project
AI	Artificial Intelligence
AMPS	Advanced Mobile Phone System
AP	Access Point
AR	Augmented Reality
B-LFU	Block Least Frequently Used
BS	Base Station
BW	BandWidth
CAGR	Compound Annual Growth Rate
CAV	Connected and Automated Vehicle
CDMA	Code Division Multiple Access
CDN	Content Delivery Network
CPS	Cyber-Physical System
E2E	End to End
EDGE	Enhanced Data rates for GSM Evolution
eMBB	enhanced Mobile BroadBand
ESA	European Spatial Agency
ETSI	Three Letter Acronym
FCC	Federal Communications Commission
FDMA	Frequency Division Multiple Access
FIFO	First in First Out
FM	Frequency Modulation
GEO	Geostationary Earth Orbit

GSM Global System for Mobile communications

GMSK Gaussian Multiple Shift Keying

GPRS General Packet Radio Services

GPS Global Positioning System

GSMA Global System for Mobile Communications Association

HMI Human Machine Interface

HAPS High Altitude Platform System

HSDPA High Speed Downlink Packet Access

HSPA High Speed Packet Access

HSUPA High Speed Uplink Packet Access

ICT Information and Communication Technology

IoT Internet of Things

IT Information Technology

ITS Intelligent Transportation System

ITU International Telecommunication Union

LEO Low Earth Orbit

LIFO Last in First Out

LoS Line of Sight

LRU Least Recently Used

LTE Long Term Evolution

M2M Machine To Machine

MEC Multi-access Edge Computing

MEO Medium Earth Orbit

MIMO Multiple Input Multiple Output

ML Machine Learning

MMOG MultiMedia Online Gaming

MMS Multimedia Messaging Service

mMTC massive Machine Type Communications

MNO Mobile Network Operator

MR Mixed Reality

NB NarrowBand

NGSO Non GeoStationary Orbit

NM-LFU No-Memory Least Frequently Used

NR New Radio

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiplexing Access

PAPR Peak-to-Average Power Ratio

PSK Phase Shift Keying

PV PhotoVoltaics

QoE Quality of Experience

Q-SLRU Quadruply Segmented Least Recently Used

RE Renewable Energy

RTT Round Trip Time

SDG Sustainable Development Goal

SINR Signal to Interference Noise Ratio

SMBS Super Macro Base Station

SNR Signal to Noise Ratio

TDMA Time Division Multiple Access

UAV Unmanned Aerial Vehicle

UE User Equipment

UMTS Universal Mobile Telecommunications System

URLLC Ultra Reliable Low Latency Communications

V2V Vehicle to Vehicle

V2X Vehicle to Everything

VHetNet Vertical Heterogeneous Network

VoD Video on Demand

VNI Visual Networking Index

VR Virtual Reality

WAP Wireless Application Protocol

W-CDMA Wideband Code Division Multiple Access

WHO World Health Organization

Chapter 1

Introduction

This first Chapter aims at laying the foundation for modern context comprehension regarding Internet-related issues and challenges, providing the reader with all the needed notions to understand how relevant is the worldwide development of wireless networks.
Section 1.1 illustrates the fact that a ubiquitous Internet connection is not only a technological requirement but also a crucial enabler for many everyday life activities. As a confirmation of what has just been affirmed, this concept is also mentioned in one of the key points of the SDGs 2030 Agenda.
Furthermore, Section 1.2 explains the need for ICTs in modern societies, focusing on IoT-enabled applications, and related paradigms, such as the Smart City one.
Finally, 1.3 describes how all of these new phenomena have led to a sharp increase in traffic demands, thus providing modern networks with many challenges to be faced and overcome.

1.1 The Need For Worldwide Internet Connection

The adoption, between the 25th and 27th of September 2015 of the Sustainable Development Goals, in New York, marked a completely new era in international development.
In particular, between the 17 goals and 169 targets that are illustrated in the *2015 Development Agenda*, and that will guide ratifying governments policies until 2030, the 9th is dedicated to *"Build resilient infrastructure, promote, include a sustainable industrialization and foster innovation"* [23].
Furthermore, between its targets, the 9.c one calls on states to significantly strive to provide universal and affordable access to the Internet in the least developed

countries, highlighting the need for a *"Universal Access to Information and Communications Technologies"*. [24]
To confirm this, the report of the Internet Governance Forum, held between the 9th and 13th of November 2015, emphasizes the fact that an *open, free, and neutral* Internet would empower sustainable development worldwide, unlocking human capabilities, and simplifying many aspects of human life. [25]
In this sense, the *SGGs 2015 Agenda* mentions ICTs widespread also when dealing with the target 4.b, connected with the increase in the percentage of educated people, the 5.b for women's empowerment and the 17.8 when illustrating the importance of science, technology, and innovation.
In addition to this, the COVID-19 pandemic, which represented one of the worst and combined global economic, social, and health crises of the last century, highlighted the essential role ICTs have played in facing all its related challenges, starting from the advent of the so-called *Smart Working*, until the strong widespread of e-commerce and digital entertainment. [26] It is evident that the stringent containment measures that were in place all over the globe have taken to a major boost in digitalization, as social distancing has led to the increased use of digital channels. ICTs pervaded all the actors of modern society, starting from institutions, that used them for mitigating economic impacts, moving to hospitals for contact tracing strategies, schools for home education, and enterprises. In [27] ITU data show that growth rates in Internet usage have surged, with an additional 800 million people coming online in the last two years, reaching 4.9 billion users overall, or 63 percent of the global population, in 2021.
Furthermore, during 2020, 33% of EU enterprises increased both the share of people employed having remote access to the e-mail system and the proportion of employees that can use the enterprise's ICT systems. It has to be noted that this trend is a global one, as shown in Fig. 1.1 plotting the percentage of companies that, by country, have decided to increase access to their computer systems. This tendency is also confirmed by the number of people employed as IT specialists that grew by more than 50.0% during the period that goes from 2012 to 2021, having an average annual growth rate of 4.6% with peaks during the pandemics of 7.3% [28].
However, at the height of national lockdowns due to COVID-19, up to 1.6 billion children around the world were affected by school closures, causing the largest mass disruption of education in modern history, and according to [29], a joint ITU-UNICEF report still many of them did not have access to the Internet at the time, thus preventing these children from attending school lessons.
Therefore, the need for accomplishing the 9th SDG was reaffirmed stronger than ever, as reported by the 400 top representatives from the public/private sectors that participated in the sixth World Telecommunication/ICT Policy Forum (WTPF-21), held virtually from 16-18 December 2021 [30].

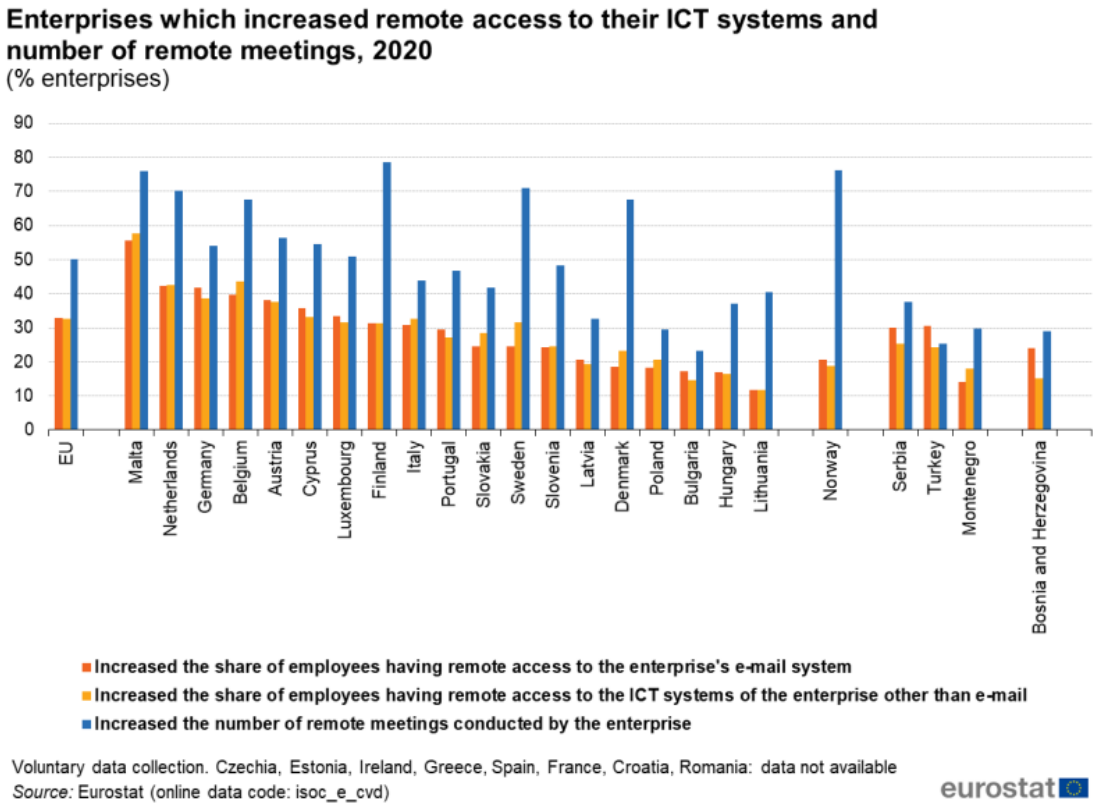


Figure 1.1: Enterprises that increased access to their ICT systems [7]

1.2 IoT Development and Smart Cities

The need for an Internet connection is not only a health, or educational concern, but something that is pervading our everyday life and fostering future needs and habits. According to the *European Commission*, [31] *" A smart city is a place where traditional networks and services are made more efficient with the use of digital solutions for the benefit of its inhabitants and business."*. This concept goes far beyond the common use of digital technologies and aims at using them to increase already present resource exploitation and reduce emissions.
This phenomenon is crucial nowadays due to the global trend towards massive urbanization, thus meaning a massive concentration of energy requests. To confirm this, the UN says that, by 2050 68% of the world population is projected to live in urban areas and that 1.5 million people are added to the global urban population every week [32]. Moreover, [33] buildings nowadays are responsible for more than 40% of the total energy consumption worldwide and 36% of greenhouse gas emissions.

Therefore, in such a connected paradigm, finding more efficient ways to light and heat buildings, having a more interactive and responsive city administration, and smarter urban transport networks is something that is not negligible anymore. The most common aims of ICTs in urban areas are [34]:

- Improve the quality of life of its citizens;
- Ensure economic growth with better employment opportunities;
- Improve the well-being of its citizens by ensuring access to social and community services;
- Establish an environmentally responsible and sustainable approach to development;
- Ensure efficient service delivery of basic services and infrastructure such as public transportation, water supply, drainage, telecommunication, and other utilities;
- Ability to address climate change and environmental issues;
- Provide an effective regulatory and local governance mechanism ensuring equitable policies.

And they are pursued through many incoming applications such as smart metering, smart grid, smart building visualizators, and intelligent transportation systems, as shown in Fig. 1.2. These applications and their possible outcomes have to be

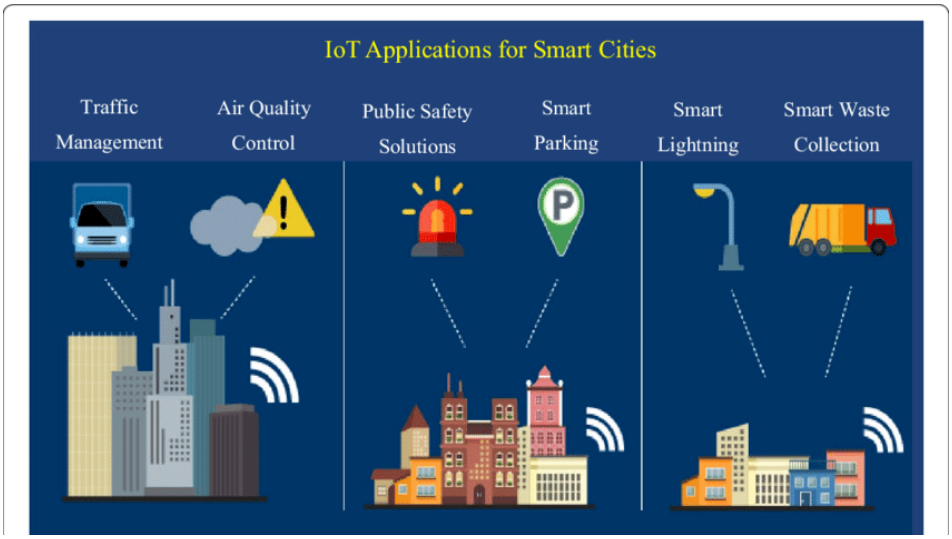


Figure 1.2: Potential Applications for Smart Cities [8]

strongly taken into account in future decades. The need for their development is confirmed in every scenario. However, one of the most impressive data is one of the road accident-related deaths that can be avoided thanks to a smart transportation system. In 2016 the WHO stated, in his *Global Status Report on Road Safety* that the number of death related to road accidents reached 1.35 million in one year, more than HIV/AIDS, tuberculosis, or diarrhoeal diseases, and that road crashes are one of the most common causes of death for children and young people between 5 and 29 worldwide. The same importance, given the modern political events, can be given to emissions and energy consumption control and reduction, which can be achieved by the use of smart monitoring systems.
All of these new services are deployable thanks to the advent of IoT devices and technologies, where the IoT is considered an umbrella term used to describe a *"growing network of objects that communicate between themselves and other internet-enabled devices over the Internet"* [35]. IoT systems are all sets of computing devices that can connect wirelessly to a network and have the ability to transmit data. Those data exchanges are fundamental, and they have to be quick, providing pieces of information that are then processed, thus giving to the final user a human-readable, meaningful result. For this reason, it is evident that IoT is the key enabler for Smart Cities to be there and, from what was affirmed before, that its worldwide development is crucial for ensuring a better life for everyone.
Given that, the future society will be completely connected, with IoT-related communications being preponderant. As a confirmation of that, it is expected that by 2025 the number of connected devices worldwide will rise to 75 billion [36].

1.3 Global Internet Trends

As seen in the previous sections, ICTs and IoT-enabled applications are nowadays fundamental and the number of connected *things* is experiencing unprecedented growth. Cisco, in his Internet Annual Report, [2], has pointed out the fact that globally, overall devices and connections are growing with a CAGR equal to 10%, thus meaning that their growth rate is higher than the one of the Internet users and population together. This result, however, was highly expected, given the analysis carried out in Sections 1.1 and 1.2 and confirms what was previously affirmed. Moreover, affirming that the share of *things* that can exploit an Internet connection, being in contact with one another and exchanging data, means to assert that the IoT is gaining global momentum, becoming one of the most booming phenomena worldwide. To confirm this, the previously mentioned white paper also highlights the fact that the terminals mix is shifting towards M2M connections, which are the fastest growing ones with a CAGR of 19%, composing half of today's total share. Moreover, by a visual inspection of Fig. 1.3 it is also possible

to see that, globally IoT-based connections are more than doubled in five years, and that also smartphone connections are growing rapidly, while PC-related ones are decreasing steadily. With the growing number of M2M enabled terminals, the term massive-IoT has been invented and has become common to describe the huge amount of sensors and devices able to be connected to the internet worldwide.

Furthermore, the Ericsson 2021 mobility report highlights the fact that, from

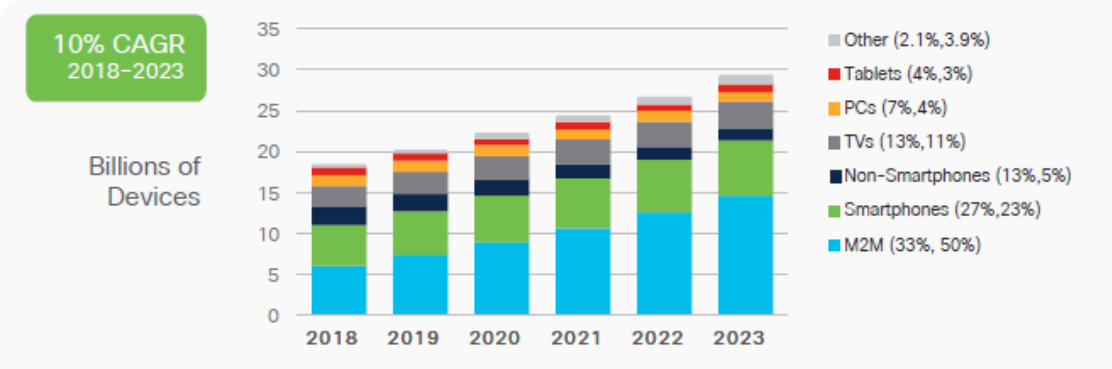


Figure 1.3: Global Device and Connection Growth [2]

2011 to 2021, the mobile traffic in terrestrial networks has increased 300-fold [37], with the total global value reaching 65EB per month and expected to grow by a factor of around 4.4 until reaching, in 2027, 288EB per month. This means that, today, the average usage per smartphone reaches the value of 11.4 GB and that in the recent future will reach the value of $41GB$ per month.

In addition to this, with the advent and expansion of 5th-generation wireless systems, the number of people migrating from 4G to 5G-enabled smartphones is increasing steadily. The 5G mobile subscriptions surpassed 1 billion in 2022 and are expected to be the most numerous ones, with an average value of 15 GB used per month by every single smartphone. Their growth is expected to be so rapid that the number of total connections will reach the value of 4.4 billion in 2027.

Lastly, according to the CISCO VNI previsions [38], AR, VR, and MR-related applications, thanks to new hardware available to individuals, and a growing body of content to consume, are expected to continue a high growth trajectory with an increment in overall traffic related to them that will be 12-fold with a 65 percent CAGR. Moreover, this work also highlights the spread in Internet gaming-related traffic, and 4K and 8K contents, which is also related to the 5G given capabilities, by saying that, by the end of 2022, the sum of all forms of IP video including file sharing, streaming, VoD, gaming, and video conferencing will account for 82 percent of the total amount of IP traffic.

Briefly, for all the aforementioned reasons the overall traffic volume is expected to grow continuously, reaching unprecedented peaks. Therefore, the overall network,

to be compliant with future expectations, and present applications must provide a flexible, continuous, reliable, and fast service to everyone.

Chapter 2

Wireless Technologies: An Overview

Having done all this complete description of the challenges that IoT-based applications are giving to the network in terms of reliability, flexibility, and quality of service, it is now time to understand if the modern network is fulfilling all the requirements for the full development of these new applications or not. To understand that, this Chapter aims at providing a brief review of the wireless networks developed technologies.
Section 2.1 illustrates the evolution of mobile technologies, from 1G to 4G, describing the major enhancements that every single generation has provided.
That said, Section 2.2 deals with the promising 5th generation main features and use cases, highlighting the most important amendments that this new technology grants with respect to the LTE one.
Furthermore, Section **??**, derives the prime challenges of 5G development, illustrating consensus and coverage-related issues in its deployment.
Finally Section 2.4 gives an insight into 6G networks, paving the way for introducing aerial platforms in the normal terrestrial-based paradigm.

2.1 The evolution path, from 1G to 4G

The first generation of mobile communication systems was first introduced in 1979 by Nippon Telephone and Telegraph Company in Japan. However, the technology was also used in the USA and its major subscribers were AMPS (Advance Mobile Phone System) in North America and NMT in Scandinavia. The 1G was based on an analog system that used FM modulation, and FDMA channel access strategy, providing users with duplex channels with one-way bandwidths of 30kHz for each conversation in the frequencies in the range of 850 MHz, allocated by the

FCC. However, every cellular system was incompatible with one another, due to their different frequency ranges. Furthermore, the network did not guarantee any inter-operator roaming, and, during its operation and due to the increase in users FCC needed to enlarge the used spectrum many times. Finally, this technology was turned off in 2009. [39]
The limitations of the described wireless architecture induced the standardization of a 2nd generation cellular mobile communication system, that was fully developed by the early 1990s. The main innovation was the use of digital mobile communication, thus marking a complete shift from the old analog communications. The development of the 2G wireless system allowed its users to have encrypted voice calls, but also SMS, and e-mail services. This cellular networks strategy was commercially launched with the GSM standard, in Finland, in 1991 and used GMSK modulation with either a TDMA strategy or an FDMA one, with BW of $200kHz$ reaching a data rate up to $9.6kbps$. Originally it was designed to operate in the 900MHz band, although for some time it was shifted towards higher ones ($1800MHz$). Other standards were the D-AMPS/IS-136, PDC used mostly in the US as a development of the 1G networks, and the cdmaOne/IS-95 which has the difference, concerning the previously mentioned ones, that it uses CDMA and not TDMA. However, all those systems still used a circuit-switching technique. Moreover, the GSM channel structure and modulation technique did not permit faster rates. Therefore, in a short time, there was a necessity for increased data rate strategies. [40]
The technologies that faced that problem were GPRS and EDGE. GPRS's importance is because it was the first applying packet switching compared to the old-fashioned circuit switching. This led to a performance increase with data rates up to 50kbps. However, the real improvement was provided with the EDGE standard, built on the existing GSM, having the same transmission technologies, but using an 8PSK modulation technique along with the GMSK, thus increasing the maximum transfer speed to 384 kbps in case of reduced coverage. These enhancements also enabled new services such as access to WAP, MMS, and access to the web.

The standardization of the 3rd generation of wireless mobile communications started as an EDGE upgrade, with the main aims of giving better data rates, having a standard network design globally, and marking a shift from voice-based to data-centric systems. IMT-2000 was done by ITU to introduce the 3rd generation system and the first commercial networks were introduced in mid-2001. In Europe it was called UMTS, being ETSI driven, while cdma2000 is the American variant. UMTS uses a W-CDMA radio access technology, offering high spectral efficiency with channels of 5MHz in bands around 2000MHz, thus meaning that new frequencies were allocated with respect to the other technologies. Moreover, the enhancements of the 3G allowed giving to users new services such as video calls.

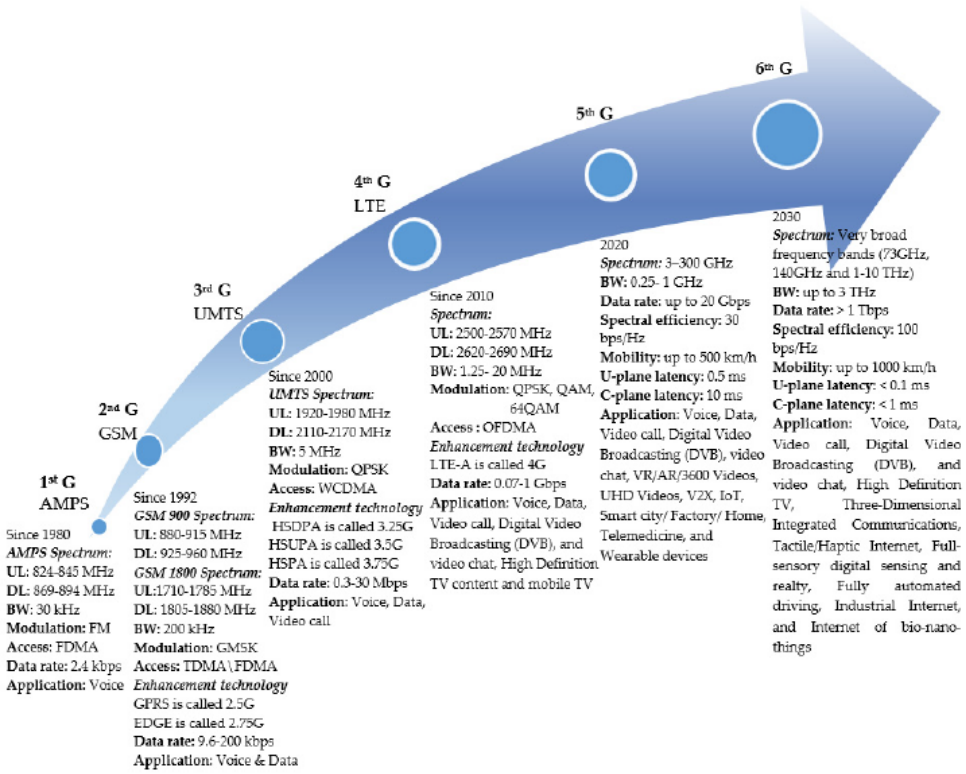


Figure 2.1: Evolution of Mobile Networks

Lastly, this architecture was enhanced by developing HSPA, which is the union of HSDPA and HSUPA, respectively used for downlinks and improved person-to-person data applications. Furthermore, in this case, also MIMO provisioning was considered, thus improving the reachable data rate to 20 Mbps for HSDPA. [41]. 4th generation communication system, well known as Long Term Evolution LTE, is the next step forward in cellular UMTS technology. The rapid increase of mobile data usage and the emergence of new applications such as streaming, mobile TV, and MMOG (Multimedia Online Gaming) pushed the data throughput and latency requirements over the already present possibilities, giving the necessity for a new specification document to allow the support to such a complex environment. To be compliant with these demands, a completely new structure was adopted using a pure IP architecture, based on packet switching. 3GPP LTE standardization group was formed in 2004 and in June 2005 Release 8 was finally frozen and then completed in December 2007. The main requirements of this work were to implement a reduced delay for connection establishment and transmission latency in the range of a few hundred milliseconds, improved spectral efficiency, simplified

network architecture, seamless mobility, reduced power consumption with respect to 3G and scalable bandwidths between 1.25MHz and 20MHz. However, going through the Release 10 to Release 13 other changes have been provided to the first standard to give an always improving end-user experience. The considered spectra are the ones in the frequency bands around 600 MHz, 700 MHz, 1.7 GHz, 2.1 GHz, 2.3 GHz, and 2.5 GHz. However, the modification that led LTE to a sharp performance increase was the use of OFDM-based radio access technology. This strategy was then suited to provide the multiple access scheme, thus having the OFDMA, orthogonal frequency division multiple access in the downlink, and SC-FDMA, single channel orthogonal frequency division multiple access in the uplink. OFDM divides the available bandwidth into several narrowband channels, with orthogonal carriers, therefore it is very resilient to selective fading. This is due to the fact that wideband signals are impaired by this phenomenon and for them, even if equalization helps, this may become very hard. Moreover, the OFDM technique improves the overall performance also against multipath, since using a narrow channel means reducing the bit duration, therefore decreasing the problem of ISI. Again this technique is also effective for interference, because the spectra of the used waveforms have zeros in correspondence with the peaks of the other carriers, providing a high degree of spectral efficiency. In this way, no mutual interference is created. This achievement is a crucial improvement concerning the previous wireless network technologies, as, differently from what happened in FDM, where guardbands were needed, thus wasting bandwidth, here it is possible to have a partial overlapping thanks to the orthogonality feature, having very low ICI. Furthermore, the OFDM signal used in LTE comprises a maximum of 2048 different sub-carriers having a spacing of 15 kHz. Within the aforementioned signal, it is also possible to choose between three types of modulation which are the QPSK, 16QAM, and 64QAM. Lastly, using OFDMA means that users are allocated by both time and frequency, therefore leading to the possibility to exploit the fact that a given user might have a better radio link quality on some specific bandwidth area of the available bandwidth, thus allocating this portion to him, while other portions to the others. Thanks to these amendments, and also to its advanced version, LTE-A, with bandwidth extensions up to 100MHz via carrier aggregation, 4G can reach speeds far greater than the ones provided with 3rd generation architecture, with peak values that may arrive up to 1 Gbps for stationary users and 150Mbps for moving users, enabling digital video broadcasting, video chat, and high definition TV and mobile TV to be there [42]. Finally, with GPP Release 12, completed in March 2015, the so-called *sidelink communication* was provided, meaning that two users may communicate without necessarily going to the eNodeB, but directly through their interfaces, called in this case *PC5*. This allows many other application fields such as discovery services (context-aware applications, social gaming, smart cities service, location enhancement application)

and also dedicated network access sharing for out-of-coverage devices. All of these features have led to the widespread of this technology, so that, GSMA Global Mobile Trends 2021 Report affirms that, by 2025, 4G will account for 57% of the global mobile customer base [43].

2.2 5th Generation, Current Times

Despite the great amendments proposed by the 4th generation cellular network, the advent of new applications such as 4K or 8K video streaming or VR and AR, together with the massive market penetration of IoT devices and the development of the concept of *Smart Cities* require greater capacity but also lower latency, and strong support for a massive number of connections. However, the long-term

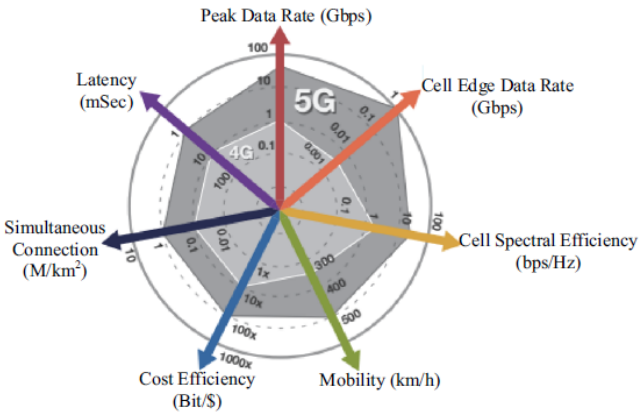


Figure 2.2: KPIs in Evolution from 4G to 5G

evolution architecture is not sufficient at all to satisfy them. Indeed, having a traditional LTE system in place, with 2x2 MIMO support, also means the possibility to give service up to $\leq 150/4$ simultaneous full HD video streaming, which is not enough for modern demands worldwide. Moreover, the aforementioned standard was designed to support originally up to 600 users per cell, thus meaning that it was not thought for massive M2M communications development and won't be able to support the CAGR these systems are experiencing now [42]. All those reasoning, have led to the conceptualization and deployment of a completely new technology, that has marked a complete shift towards the modern smart society, the 5th generation network.
This network is created with 8 main purposes, summarized by IMT-2020 when developing it, as shown in Fig. 2.2. In particular, it has to overcome the old LTE-A as regards latency, reaching less than $1ms$ as a requirement, having a better

spectrum efficiency and energy efficiency, enhancing mobility, and having unprecedented data rates. Furthermore, this architecture focuses on 3 main use cases and they are:

- eMBB concentrating on human-centric scenarios for multimedia content access. Large area coverage and hotspots are clear examples of this category. In the first case, there is the need for coverage and medium-high mobility with improved data rates, while in the last one, there is the need for high traffic capacity.
- URLLC focuses on applications that need strict latency requirements and very high throughput to be there, having low losses and continuous service. Examples of these situations are automated vehicles, wireless control of industry, and high-distance surgery.
- mMTC which is suited for IoT scenarios, where many devices are fully connected and have to exchange low amount of data, thus characterizing the so-called delay-tolerant applications, such as smart grids, e-health, and all services that can be related to the *Smart City* paradigm. Moreover, being the number of connected devices growing incredibly, as already pointed out in Section 1.2, this scenario is not negligible at all, and the attributes that have to be guaranteed are related to the need for the deployment of low-cost sensors with long-lasting batteries, that can be connected everywhere. This means that the network has to be accessible with a certain amount of reliability everywhere.

To enhance service provisioning and be compliant with these prime scenarios, many advancements are introduced.
First of all new frequency bands are used, with the usage of three main available spectra. Between them, there is the lowest one, meaning the Sub 1GHz band, the medium one, meaning the Sub 6GHz band and the highest, which is referred to as mmWave, providing ultra-high frequency radio waves that range from 30GHZ to 300GHz, used to supercharge 5G connections and give users download speeds that were unbelievable few years ago. The choice of mmWaves as a first amendment is performed both to increase the available data rates, and to overcome the reduced amount of available spectrum in frequency. Furthermore, between those ultra-high frequency bands, the most used are the ones ranging from 24 to 100 GHz [44].
The second improvement is related to the first one and is the need for small cells to be in place. Indeed, small cell deployment provides the possibility of having high data rates and also saving energy due to the short distances between the transmitter and the receiver.
Thirdly, the adoption of multiple connection techniques and the enhancement of D2D paradigms is granted, resulting in enhanced support for IoT systems.

In addition to this, SDN, NFV, and Network Slicing, which are fundamental technologies for network flexibility, are massively used in 5G wireless networks. In particular, NFV is fundamental since, thanks to that, the network functionalities are not placed on dedicated hardware but are located on software parts that can be executed by standard servers. This allows flexibility and dynamicity of the network. For the same purpose Network Slicing is strongly needed, so that a single physical network can be divided into *slices*, resulting in virtual networks that may be optimized to handle different services. This is because 5G typically includes many services, that may have different requirements, and slicing allows the creation of different subnetworks tailored to their actual needs. An example of that is shown in Fig. 2.3 taken from [9]. Lastly, also mMIMO is used and it

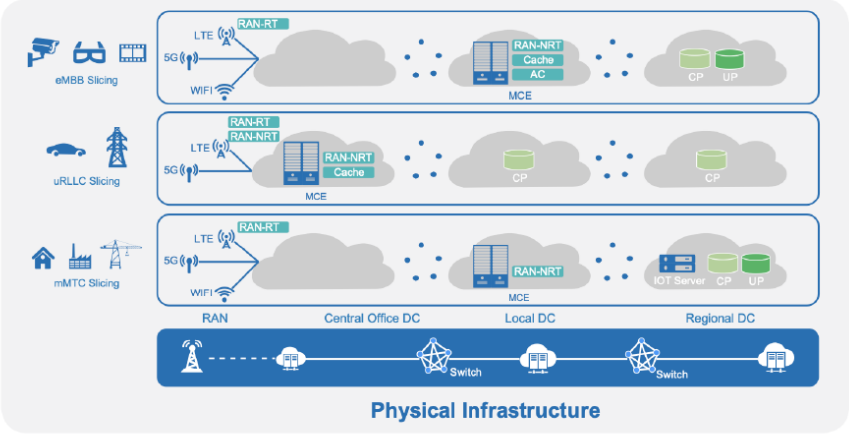


Figure 2.3: Network Slicing for Multiple Services Example [9]

is strongly beneficial in this architecture. MIMO is a concept that was applied to many wireless standards because it can significantly improve the capacity and reliability of wireless systems, and was already a key milestone in LTE evolution. However, the massive MIMO technique allows transmitting and receiving much more data streams simultaneously over the same radio channel, thus multiplying the capacity of a wireless connection without using other bands, and increasing the overall efficiency. 5G systems use mMIMO such as 8x8 array antennas, that need to be more directive, therefore ensuring better performance in terms of reliability, and being more resistant to intentional jamming. mMIMO means also using beamforming so to concentrate the signal over the physical position of the devices targeted, giving an adaptive emission as shown in Fig. 2.4

Figure 2.4: 4G vs 5G mMIMO

2.3 5G Challenges and Problems

However, even if this new generation of mobile networks seemed to be the one able to solve any issue, it faces many difficulties, both from the technical and the commercial point of view.
When dealing with 5G we already mentioned the need for higher data rates, thus taking about the use of higher frequency spectra. However, according to Friis law, this increase in frequency is directly linked with a strong free space path loss. Indeed, neglecting the shadowing phenomena, assuming that the distance between the transmitter and receiver is d and that the propagation of the electromagnetic waves occurs in empty space, (i.e., in absence of absorbing and reflecting objects), the power of the received signal decreases with the increasing distance. More precisely, if the receiver antenna is located in the so-called "far field region", the power $P_R(d)$ could be represented through:

$$P_R(d) = P_T \left(\frac{\lambda}{4\pi d}\right)^2 G_T G_R \tag{2.1}$$

Where P_T is equal to the transmitted power, G_R (G_T) is the gain of the receive (transmit) antenna, and λ is the considered wavelength. And therefore the path loss:

$$L(d) = \frac{P_T}{P_R(d)} \tag{2.2}$$

$$L_{fs}(d) = \frac{1}{G_T G_R}\left(\frac{4\pi d}{\lambda}\right)^2 = \frac{1}{G_T G_R}\left(\frac{4\pi df}{c}\right)^2 \tag{2.3}$$

Again, this mathematical law establishes that the power of the received signal is proportional to the inverse of the square of the product of distance and frequency, therefore, assuming antenna gains and distances as constants, the attenuation increases proportionally to the square of the frequency.

Even though this simple analysis already points out what is the reason for beamforming and mMIMO in 5G, in reality, the estimated quadratic dependence between propagation and distance turns out to be extremely optimistic as there is also the so-called *multipath fading*. This phenomenon explains the fact that more, delayed and attenuated replicas of the same signal arrive at the receiver, where some of them are obtained from reflections by other objects. Lastly, the dependence between path loss and distance was derived in [10] and is an exponential one, with a value n called *path loss exponent* whose typical values are in Fig. 2.5.

Enviroment	*Range of n*
Urban macrocell	3.7 – 6.5
Urban microcell	2.7 – 3.5
Office in a building (same floor)	1.6 – 3.5
Office in a building (different floors)	2 – 6
Warehouse	1.8 – 2.2
Factory	1.6 – 3.3
Flat	3

Figure 2.5: Path Loss Exponent remarkable values [10]

Moreover, the use of small cells, which has many advantages, as already explained before, is also a problem for the deployment of this technology worldwide. As the network densification process goes on, the overall costs that an MNO will undergo to develop a complete architecture are enormous, as a shorter range, implies more base stations and costs for them, as shown in Fig. 2.6 Given the fact

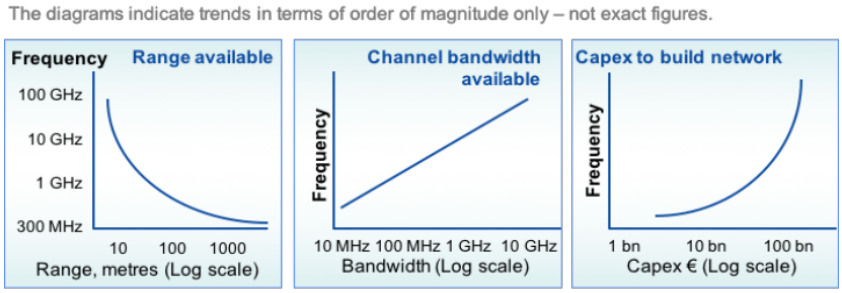


Figure 2.6: Physics and Economics in 5G [11]

that traditionally macro cells offered coverage in ranges between 2 and 15 km, and since this paradigm shift means having an average range of $200m$ [45], the overall number of deployed cells must increase a hundredfold. The total deployment cost of such a complex architecture is not only monetary but also related to the time needed to perform this shift. Moreover, MNOs are interested in investing where they need and where the return on their investments is ensured, thus meaning that there is no interest at all in providing coverage in rural areas. This trend

was already depicted for the fiber-based access in *Strategia Italiana per la Banda Ultralarga*, a report carried out in 2015 by the Presidenza Italiana del Consiglio dei Ministri, and never changed. Again Small Cells deployment, if on the one hand helps reduce the power transmitted by the entire device set, may lead to an increase in overall energy consumption due to the uncontrolled proliferation of 5G BSs, thus increasing the telecommunication networks consumption rate which is already high and equal to the 12% of the total per year [3]
Finally, another key point in the difficulty in NGA deployment is the strong reluctance of public opinion against the possible health concerns of this increase in frequency spectrum even if many works in the literature [46] have shown no scientific evidence of health effects triggered by radio wave exposure of BSs working below the maximum limits.

2.4 6th Generation, The Future

Through the years, the volume and magnitude of connected things are expected to grow further reaching unprecedented numbers. Moreover, estimations in [37] show that 5G provided requests will cover the 72% of the overall soon, meaning that traffic demand is going to increase steadily, following and if possible, overcoming the historical data-growth rate of at least a factor 10 per decade. This means that 5G will soon become insufficient in provided throughput, and more than everything, in being able to support a large number of connections. In fact, as already discussed in Section1.1 connected devices are the category that will experience the highest overall CAGR in the next future, reaching 500 billion by 2030, about 59 times larger than the expected world population which is around 8.5 billion [47], but already being three times larger in 2025. In addition to that, it has to be considered that a typical 5G communication system can support at most 50 thousand IoTs and NB-IoT systems per cell, which is already not enough for many applications. Therefore, all those factors are leading to the need for increased data rates and the necessity of a network that supports even millions of connected devices, with the consequence that a more robust architecture must be designed to realize this massive access. The main requirements and reasons for the future 6G networks to be developed are summarized in [48], [49], [12] and shown in Fig. 2.7. In particular, they are:

- Performance increase, as the traffic volume and demand for high and advanced services is going to increase rapidly, 6G needs to provide much higher data rates with respect to 5G, aiming at providing peak values of $1000 Gbps$ and almost doubling the spectral efficiency. Moreover, from the point of view of latency, the target that has been set is to reach a value that is lower than 100 µm for air delay and $1 ms$ for E2E latency, with extremely reduced jitter.

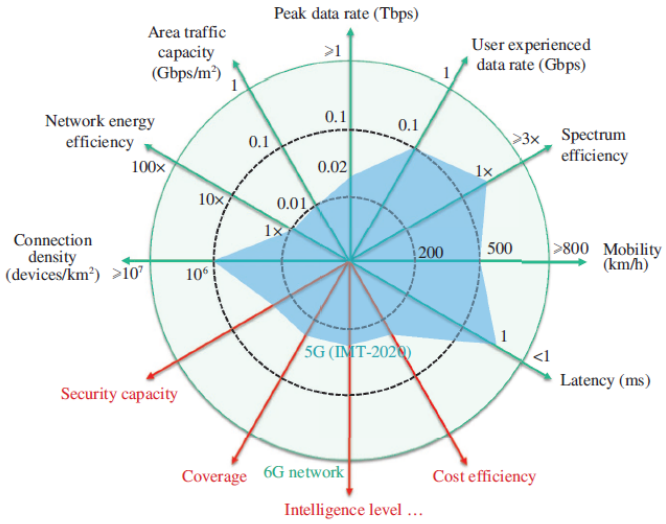


Figure 2.7: 6G Expected Performance Metrics [12]

Finally, also reliability will be a focus, aiming to improve that hundredfold from 5G with an overall error rate that has to reach the value of 10^{-7}.

- Mobility support has to be increased in the next-generation communication systems ensuring data rates to be maintained also in highly mobile devices. For this reason, larger coverage and increased supported speed, which goes much over the $500km/h$ that were enabled with the previous generation of networks are expected. In particular, ITU defines a mobility requirement for 6G to be greater than $1000km/h$ [50]

- Massive Connectivity. The overall connection density requirement is expected to improve accommodating about 10^7 devices per square kilometer [51].

- Energy efficiency, meaning that being sensors a great amount in IoT devices, all of them need energy-efficient communication to take place.

- Edge Computing, thus meaning that the new mobile networks generation should support the limited computational capacities of mobile devices by offloading some intensive computational tasks to the cloud. Therefore, extremely low latency communication is required to make all of this take place in almost real-time. With 6G a user should be able to use in real-time the computation capabilities that the network provides. To make this happen, some computing capabilities closer to the user must be given by using MEC Servers.

However, to completely fulfill and reach those requirements, some technological enhancements must be envisioned. Finally, those concrete improvements should range from system network architecture and control, to edge and ubiquitous computing and radio technologies. Some of the most powerful will be:

- Comprehensive Artificial Intelligence that should be part of the wireless network, as virtualization and slicing should be done rapidly, thus increasing performance by having a fast decision-making process that can be boosted by using ML solutions, as suggested in [52]. This will result in a network AI governance overall, with its integration carried out to automate system operation and administration. Therefore, decentralized intelligence in the cloud, meaning edge intelligence will be essential for instantly managing resources as per user requirements, thus providing the best solution anytime. Moreover, this AI-based orchestration will also accomplish predictive tasks, to arrive at the so-called zero-touch approach, where AI is able to handle slices and network functions and maintenance, granting a full self-adaptation.

- Extremely Large Bands (THz Waves), are used to increase the KPI of the given network. Data-hungry applications need support for data rates and latencies that are unreachable by only using mmWaves. The use of these frequencies has been encouraged by FCC in [53] where it is written that *"The Federal Communications Commission adopted new rules to encourage the development of new communications technologies and expedite the deployment of new services in the spectrum above 95 GHz."*, and also announcing the creation of a new category of experimental licenses for use of frequencies from 95 to 3000 GHz. If on the one hand increasing the frequency means directly conveying more data faster, a strategy needs to be found to avoid the inherent propagation shortcomings, such as the expected severe path loss, and the strong atmospheric absorption. The first solution to overcome this issue is the use of massive and ultra-massive MIMO. Other viable solutions may be the ones using intelligent reflecting surfaces to improve propagation characteristics and a user-centric approach where the cluster serving a particular UE can be dynamically determined for that given user. Combining this last approach with mMIMO the so-called cell-free mMIMO is envisioned, meaning that all APs can serve UEs cooperatively without any restriction, thus giving the user a seamless uniform system across the network. However, also new waveforms, signals, and protocols should be found as, even if the OFDM technique is still viable, it has an intrinsic disadvantage related to its high PAPR, which was already a problem in 5G, resulting in the choice of SC-FDMA for uplink in this technology.

- Evolution in Network Topology, as while normally cellular BSs are deployed in a fixed location of fixed networks, here a static network topology is costly

and useless, thus taking to a further increase in Virtualization and slicing. Furthermore, another component that is becoming not negligible at all is the use of NTN (non-terrestrial networks), which was already envisioned in the 5G Release 17. However, NTNs will be a main part of the 6G future standard, being mentioned as a key point for future wireless network development in the Huawei 2021 report [13]. Realization of such an architecture will allow wide coverage and enhanced and increased network and computing capabilities if providing them with onboard MEC.

- Wireless edge Caching as an on-demand service is characterized by asynchronous content reuse and predictable demand distribution, thus meaning that providing caching at the edge may strongly improve the overall network performance. In particular, having such a capability in some nodes or even aerial platforms will reduce the network load, thus being strongly beneficial. This positive effect is also strengthened considering the tremendous increase in traffic volume which is expected in the next future, and the aerial system usage also takes to an interference reduction. Therefore such a technique will consequently increase spectral and energy efficiency overall [54].

Given this general overview of the 6G requirements and enabling technologies, many applications are gaining momentum since this new generation of mobile networks is under conceptualization. The most important ones are the following:

- Holographic Communications, since holographic displays will be the next evolution in multimedia experience delivery. However, videos like these need a spectrum that is currently unavailable in mmWave. A hologram is a technology, firstly suggested in 1947 by Gabor [55], for capturing an object's full 3D image, thus meaning that the requirements are much stronger concerning normal images or video to reproduce them. Moreover, a user is expected to interact with this hologram, thus changing his position with respect to it. Therefore, providing an immersive 3D experience to the user is not easy at all and requires data rates that, for a human-size hologram, may arrive at 4.3 Tbps [56].

- Extended reality (XR), a mix of AR, VR, and MR, has attracted a lot of attention and opened new horizons in various fields today. To support those kinds of applications there is the need for reliable, and continuous internet provisioning with data rates exceeding $10 Gbps$ [48], thus making clear that 5G is not sufficient for them. According to [12], VR technology can be employed in several professional verticals such as education healthcare, and the military, giving a great and interactive experience, potentially reshaping our everyday life. However, a big issue in their development is the processing capacity of devices, which should be increased further to support 3D object visualization,

as modern ones will provide the user with an unsatisfactory QoE. A future VR system should satisfy many complex requirements together as it has to be cheap, have high-quality graphics, and be portable. A possible solution that was thought for those services to be in place in a few years is the one to offload some computing tasks to some MEC Servers, thus having and proposing a cloud VR architecture, as proposed in [54].

- Connected robotics and autonomous systems, industry 4.0 is a term which is gaining hype and it is a short term to describe the fourth stage of the industrial revolution aiming at significantly improving flexibility, versatility, usability, and efficiency of CPS. Robotics is strongly related to that meaning that it can be introduced as part of a future company. However, these systems need to be controlled continuously and remotely, meaning that extreme reliability and below 100 microseconds latency are required. Another example related to that is the so-called V2X or vehicle to everything. More specifically, when dealing with V2V communications and applications such as platooning, high reliability and low latencies are required. Moreover, also when thinking about remote driving, the same requirements need to be in place.

All of these services are grouped in [13] in 5 main categories. The first mentioned is the **eMBB+** which includes all applications that are creating the paradigm shift towards Metaverse, and it is the obvious continuation of 5G eMBB. It includes VR, AR, XR, and telepresence. Other services are grouped under the keyword URLLC+. This umbrella term refers to all the 4.0 industry-related devices, such as robot usage, UAVs, and all the modern and new HMIs that are being developed in the last years. The third main category is the mMTC+ one, and it the one that includes all those applications that require the interaction of thousand of devices that are very lightweight and need low-power transmission and low-cost connections. They are mostly used under the so-called Smart City use cases. The fourth group is the *Sensing* one, which is the one that relates to the so-called networked sensing, for localization purposes, while the last is the one called AI, which relates to all those systems that are ML-based.

Chapter 3

State of The Art

This Chapter illustrates the State of The Art regarding NTNs.
First of all Section 3.1 points out the fact that nonterrestrial networks, which were once considered just as a secondary actor in wireless communication technologies, are nowadays gaining more and more momentum, becoming one of the 6th generation development pillars.
Secondly, Section. 3.2 provides a complete classification of all the aerial systems that are integrated and part of the non-terrestrial segment of a future VHetNet.
Having carried out this description, Section 3.3 provides an overview of many proposed architectures in modern literature.
Moreover, Section 3.4 explains the reasons why this work is focused on HAPS, illustrating the main advantages of using them in spite of LEO satellites. This Section is also focused on giving a full categorization of them, dividing these systems into Ballons and motored platforms, and it also provides a complete description of the HAPS main use cases.

3.1 The Advent of NTN, a 6G pillar

A total focus and work on Non-Terrestrial Networks (NTN) were gained thanks to 3GPP Release 15 in 2020, with a Study Item that was centered on NTNs deployment scenarios and channel models for them. In this paper, [57] non-terrestrial networks are defined as *"Networks, or segments of networks, using an airborne or space-borne vehicle to embark a transmission equipment relay node or base station"*.
Furthermore, this Release paper posed the main expectations for non-terrestrial segments in the future. Indeed the aforementioned document affirms that they have to *"foster the roll-out of network services in un-served areas that cannot be*

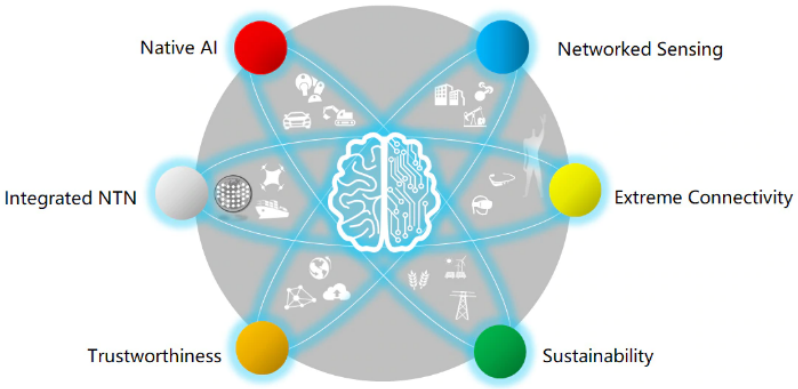


Figure 3.1: NTN as a pillar in 6G development [13]

covered by terrestrial 5G networks and underserved areas to upgrade the performance, [...] to reinforce the service reliability by providing service continuity for M2M/IoT devices or passengers on board moving platforms, [...] and enable network scalability by providing resources for data delivery towards the edge or even user terminal". From these three key points, it is already possible to understand the disruptive advantages that this new system is aimed at providing in the next future.
In addition to that, also Release 16 and Release 17 have a study item on NR NRN, to adapt the new radio access for non-terrestrial use cases. In particular, the last one focuses on enhancements for LEO and GEO-based NTN, also targeting support for HAPS networks. Moreover, it gives the foundations for NB-IoT and eMTC-based satellite access to address the massive Internet of Things. [58]
Finally, as shown in Fig. 3.1 NTNs are one of the pillars for the 6G development worldwide, thus explaining how their integration is now something crucial in modern research.

3.2 Non Terrestrial Part, Spaceborn and UAVs

A Non-Terrestrial Network has different platforms in it. Its main systems can be grouped into two main categories, which are the space-born and airborne ones. The first category is classified according to three main parameters such as altitude, beam footprint size, and orbit [59].
First of all there are GEO satellites with a circular and equatorial orbit on the

Earth at 35786 km altitude. Their orbital period is equal to the one of the earth and their footprint size ranges from 200 to 3500 km. The satellite appears motionless at a fixed position in the sky to ground observers. Included in this category are geostationary (GSO) satellites, which remain in orbit above a fixed spot on Earth. Not all geosynchronous satellites are geostationary. ESA's European Data Relay System (EDRS) program has placed satellites in GEO, where they relay information to and from non-GEO satellites and other stations that are otherwise unable to permanently transmit or receive data. This means Europe can always stay connected and online. However, some have elliptical orbits, which means they drift east and west over a fixed point on the surface during the course of a full orbit. Some have orbits that are not aligned with Earth's equator. These orbital paths are said to have degrees of inclination. Several hundred television, communications, and weather satellites all use geostationary orbits. However, deploying them is typically very expensive mainly due to their launching costs. [60] Secondly MEO Satellites, also referred to as ICO (intermediate circular orbit),

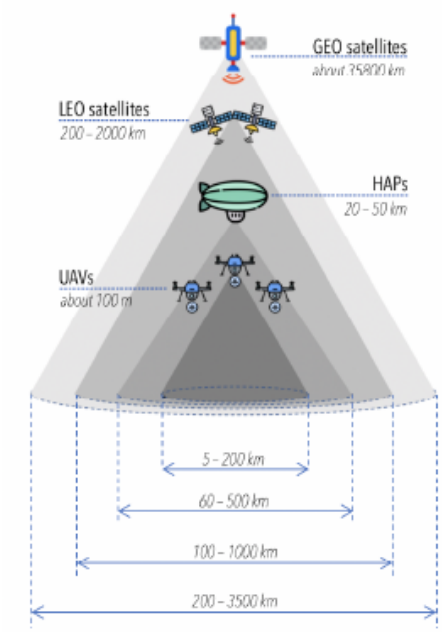


Figure 3.2: Sketch of the distribution of NTNs [14]

has a circular orbit around the earth, their orbit ranges between seven thousand and twenty-five thousand kilometers and has footprint ranges in $[100 - 1000]km$. Navigation satellites, like the kind used by car's GPS, work well at this altitude,

and normally their orbital speed is $13900km/h$, much higher than the GEO ones, as the gravitational attraction becomes greater [61].
Lastly, LEO ones occupy a region of space from about 300 kilometers to 1500 kilometers above the earth with a circular orbit. Their speed is much higher than the earth's rotation with an orbital period ranging from 40 to 120 minutes, experiencing 36 morning and night periods in 24 hours [62]. They are normally used for observations, military purposes, and weather data collection and have the same beam range as the previous ones. However, to do that a constellation of them is proposed to offer continuous coverage and communication backhaul [63]
Furthermore, the last two categories are also known as NGSO satellites due to the fact that, being their period lower than the earth rotation, they are not stationary with our planet. This feature, makes them *"less useful for tasks such as telecommunication because they move so fast across the sky and therefore require a lot of effort to track from ground stations"*. On the other hand, the airborne platforms are divided into UAS and HAPS [59].
UAVs altitude is very low, their height is normally less than 0.5 kilometers, and have a very low speed, less than $0.07km/s$, and a very low RTT that equals $1ms$. On the other hand, its coverage is strongly impaired. Its purposes cover agriculture, military, industry, search and rescue and monitoring [64].
Finally, Fig. 3.2 gives an overview of the general picture of the distribution of non-terrestrial systems placement, orbits and beam coverage.

3.3 NTNs General Architecture

Non-terrestrial systems are composed of a terrestrial terminal, acting as a UE, a non-terrestrial station, that can be either a satellite or every UAS platform that can operate as a terrestrial BS, a service link between the ground terminals and the aerial station, and a gateway with the main aim of connecting the NTN access part with the core network thanks to a feeder link. Release 16 of 3GPP well depicts this scenario, as shown in Fig. 3.3.

3.3.1 Proposed Architectures

That said, many different cases are under study and are under discussion nowadays. As an example, [59] distinguishes the NTN architectures in considering the type of access, thus classifying them as *satellite access architectures* where the non-terrestrial terminal is served by the NTN platform and *Relay-like architectures*, where the terminal and the platform communicate thanks to a relay node. Moreover, for both of them it does a further division between *transparent* and *regenerative* ones, being the first, the ones in which *"the satellite repeats the user's signal from the feeder link to the service link and vice-versa"* [65] and the latter the

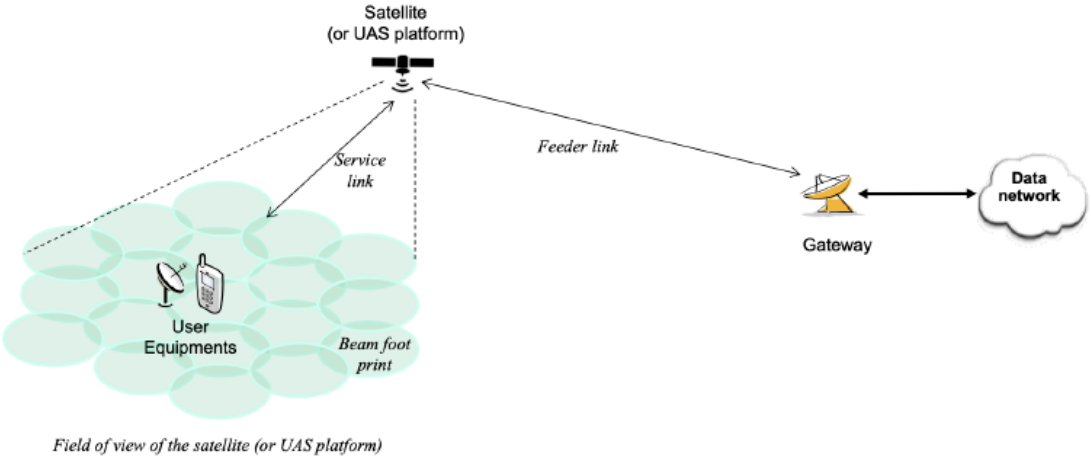


Figure 3.3: Sketch of the distribution of NTNs [15]

ones in which the satellite payload does regeneration of the signal. This division is also provided in [66] where the same categories are reported considering a single haps structure. Finally, [67] focuses on the HAPS part and the overall system configuration discussing several different cases such as a standalone HAP architecture, a multiple HAPs architecture, an integrated terrestrial/HAPs system, an integrated satellite/HAPs system and the fully integrated terrestrial/HAPs/satellites system. Lastly, [16] provides a complete overview of a fully connected, triple-layered architecture, the so-called VHetNet in 6G network, as shown in Fig. 3.4.

3.4 HAPS

This works focuses on the aerial platforms, that, according to the ITU Radio Regulations are *"Radio Stations locate on an object at an altitude of 20-50 kilometers, thus being in the stratosphere"* [18]. The HAPS concept has been explored since the early 1990s using balloons and airplanes, however, thanks to the advancements in antennas and space technologies these systems have been recently fueled.

3.4.1 HAPS Categories

The first thing we want to point out is the fact that HAPS is an umbrella term that generally refers to four main macro-categories of aerial vehicles which are *balloons*, *fixed wing*, *airships* and *hybrid approaches* [20].

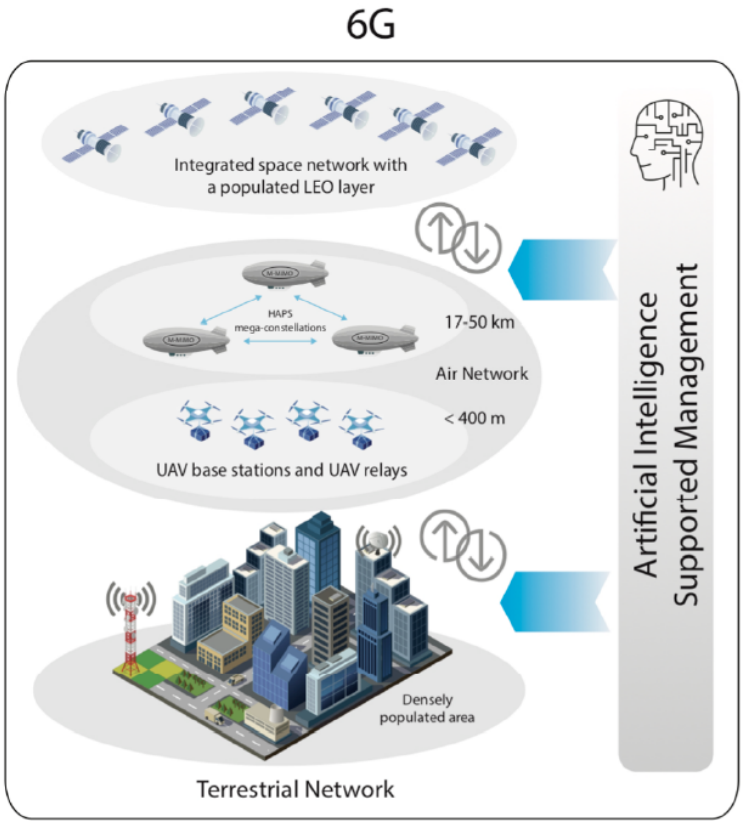
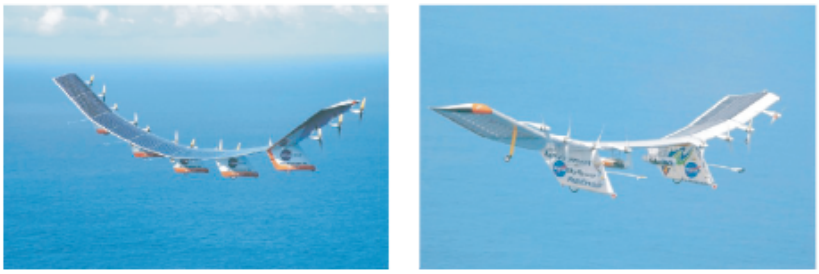


Figure 3.4: Complete VHetNet Vision [16]

Balloons are normally very small and lightweight. However, their lack of propulsion and very low payloads (tens of kilograms) are great limitations in terms of autonomy, which ranges in terms of several months, and also positioning control. On the other hand, fixed wind platforms, also referred to as *solar-powered unmanned aircraft* [17] or HALE (High Altitude Long Endurance) platforms can be precisely positioned and they have higher payload capabilities, thus meaning that their achievable operational time is much higher. They have a wingspan that reaches $70m$ and they are powered by solar cells and electric motors. Examples are the SPL, Airbus Zephyr, Softbank HAPSMobile, and Skydweller SolarImpulse. Fig. 3.5 gives some examples of them. Thales Stratobus and Altran Ecosat platforms are under the category of airships. They have great payload capabilities, arriving also to $1000kg$, and a flight duration that goes up to five years. As the previous group, they offer precise positioning control. Fig. 3.6 provides a graphical representation given by the ITU of this kind of aerial platform. Lastly, some



AeroVironment /NASA
"HELIOS" (US)
Wingspan: 75m
Payload: 50 - 100kgr

AeroVironment /NASA
"Pathfinder Plus" (US)
Wingspan: 36.9m
Payload: 50kgr

Figure 3.5: HALE Examples [17]

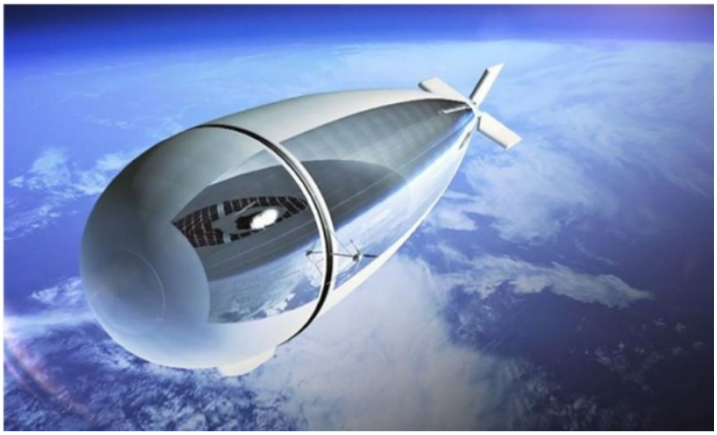


Figure 3.6: Airship Example [18]

hybrid approaches are developing, aiming at mixing aerostatic and aerodynamic principles.

3.4.2 HAPS Advantages

The reasons that lead us to focus on HAPS development are various, and among them the most important are:

- Favorable Channel Conditions, as they have a lower altitude with respect to the others ($190dB$ path loss for GEO satellites with $2GHz$ band) LEO, being placed at $20km$. Moreover, this feature, combined with the very high LoS probability gives them a low channel attenuation, and high SNR, thus allowing normal User Equipment to be able to directly connect to them without the

need for particular ground stations to be placed. [1]

- Relatively stationary position of HAPS systems makes them avoid any capacity waste related to the orbiting oven in unpopulated areas such as oceans. Moreover, its quasi-geostationary position is fundamental in reducing doppler effects already pointed out in Section 3.2. Surely doppler happens also in terrestrial networks due to the relative motion between UE and fixed BSs. However, in NTNs the LEO/MEO relative movement increases further than those frequency shifts. In [68], assuming a communication link over an LEO satellite with an altitude of 600km and a 2 GHz carrier, the doppler shift has a peak value of $48KHz$, thus being many times higher than every kind of terrestrial motion related one. Again, this attribute is also beneficial because NGSO satellites require continuous handover procedures to be there to move traffic from one LEO/MEO to another one and ensure service continuity and reliability to be in place.
- Wind Speed and Weather, the weather is mild throughout the year in the stratosphere with little change in the wind speed and also little values, that increase sharply when overcoming this part of the atmosphere, as shown in Fig. 3.7. This allows HAPS to fly with stability [69].

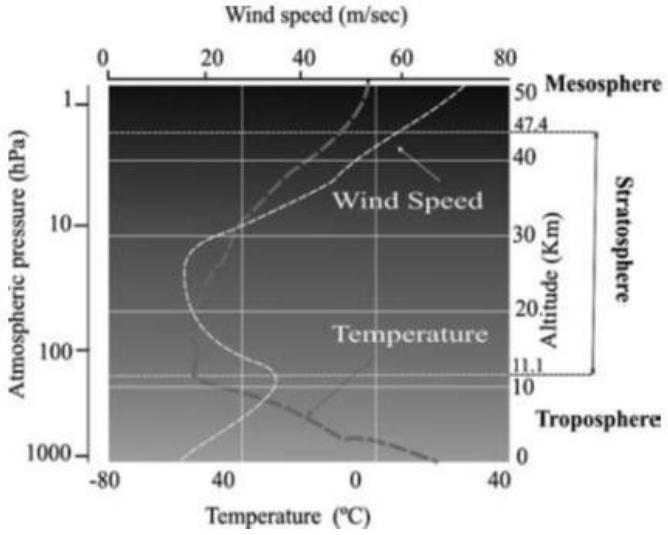


Figure 3.7: Wind Speeds at different Heights [19]

- Overall volume, a HAPS, being very large, has to possibility to have wide onboard PV panels and accumulators for energy provisioning purposes, thus drastically increasing the overall, operational time. Moreover, this feature also enables the use of massive MIMO deployment. Therefore, these aerial

stations can generate directional beams with narrow widths, thus increasing the SINR.

- Wide footprint of about 500km in radius. Therefore, an 18 HAPS system may be able to cover Greece.
- Reduced latency, due to the altitude the expected RTT ranges from $[0.13 - 0.33]ms$ thus meaning that those platforms may enable also URLLC. [70]
- Easy Deployment as aerial BSs can be deployed and be able to operate correctly in hours. On the other hand, satellite systems may require months to be in place and ready to operate. This attribute is crucial when complementing the already existing terrestrial network in case of emergencies. Moreover, the reduced cost of deployment is something that has to be accounted for.
- Higher Radius and Operative Time concerning UAVs.

3.4.3 HAPS Main Uses

Given all these great advantages in HAPS usage with respect to both UAVs and GEO/MEO/LEO constellations, the aforementioned aerial platforms, have been envisioned in many different use cases. In this work we will focus specifically on one of them, however, a general overview of the most important ones, taken from [16], is illustrated, to provide the complete range of applications of them.

Conventional HAPS Usecases

The most conventional use cases for aerial platforms are the ones related with *emergency communications and disaster recovery*, and *greenfield and white spot reduction*.

Emergency Communications and Disaster Recovery. Due to their easy and fast deployment, HAPS is well suited for emergency applications. They are an excellent candidate for disaster relief missions also because of their higher coverage concerning a normal low-altitude UAV. They may provide seamless continuous connectivity for many days, or even months. Finally, this concept and use case has already been developed respectively in Puerto Rico and Peru after disasters in 2017 and 2019 and have proven its capability to work in those harsh conditions and provide continuous and reliable connectivity in those cases [71].

Greenfield Coverage and White Spot Reduction As already pointed out in Section 2.3, the worldwide connection is still far from being achieved due to the lack of interest of MNOs in investing money in non-profitable areas. Therefore, there are still many geographical areas that today lack completely of cellular infrastructure. In Fig. 3.8 there are the developed GSM, 3G, or LTE cells. It is

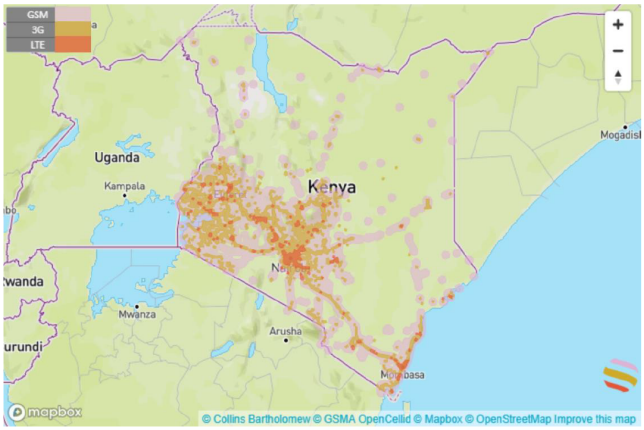


Figure 3.8: Kenya developed cells [20]

immediately noticeable the fact that most of the country lacks basic connectivity. Moreover, not only greenfield areas are there, but also much smaller uncovered zones, are diffused even in developed countries and they are called *white spots*. Those are a direct consequence of geographical obstacles or problems of propagation. The HAPS's favorable propagation conditions, due to its LoS communication, can counteract and efficiently fight those use cases. In particular [20] shows some simulations comparing the performance of the downlink spectral efficiency in a village, thus a typical white spot area, while moving with a car nearer or further from the network ground cell. Fig. 3.9 shows the obtained results suggesting that HAPS can provide wide area coverage with homogeneous performance, while the ones provided by the ground station are strongly dependent on the position of the UE.

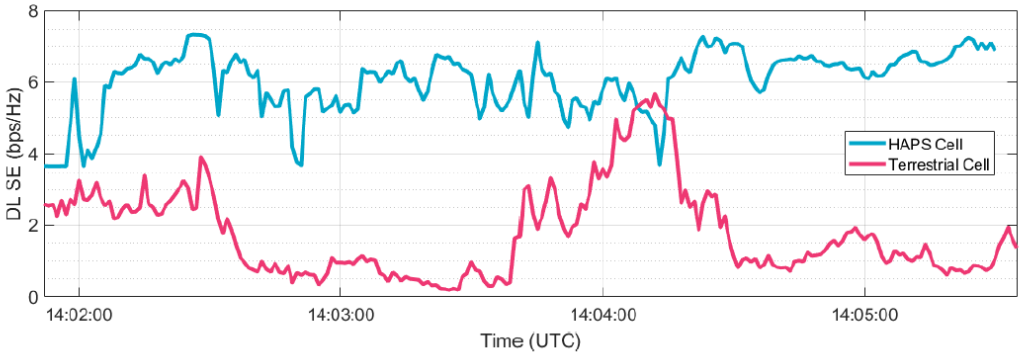


Figure 3.9: Haps vs Terrestrial Cell Spectral Efficiency [20]

HAPS SMBS

The envisioned NTN architecture with HAPS integration for future communication networks aims at providing the aerial platforms with onboard MEC capabilities, thus greatly enlarging the span of their possible applications since a macro BS is a crucial component in wireless access architecture to provide coverage and support capacity, but the increasing communication needs of metropolitan areas is making small cells deployment not sufficient to match them. Therefore, a HAPS-mounted SMBS can be a powerful solution to match the increasing traffic profiles and provide flexibility and continuous service provisioning, envisioning also the great increase in IoT-connected devices. Fig. 3.10 shows the main use cases of HAPS-SMBS in Future Networks.

HAPS-SMBS for IoT Services, the wide coverage and the fact that many

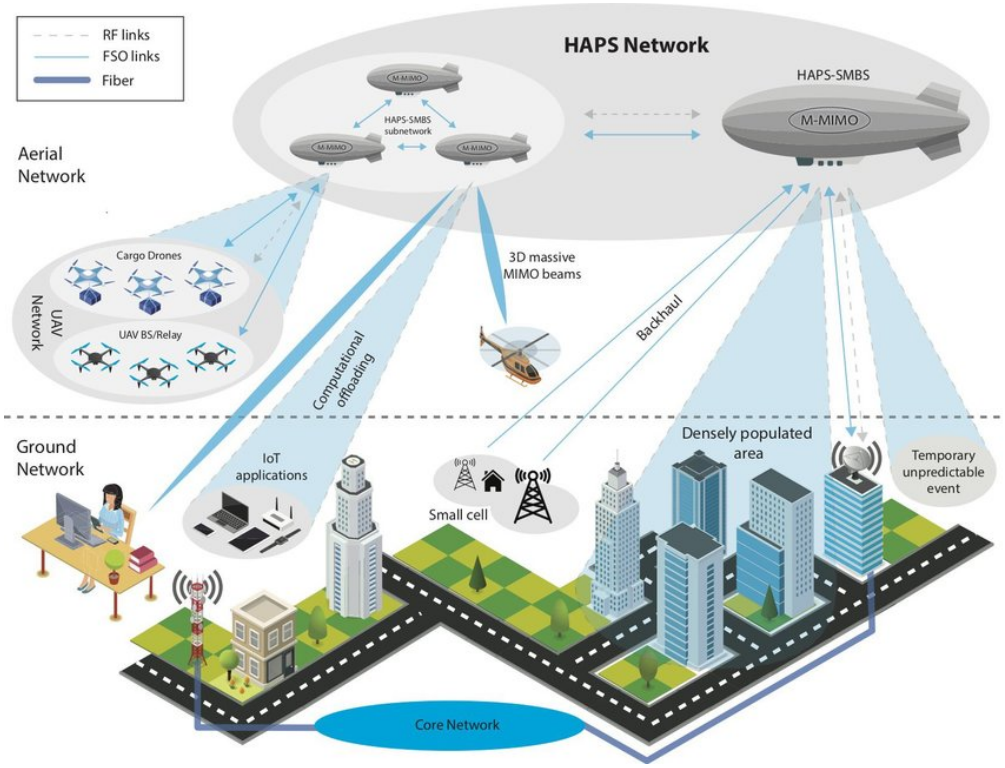


Figure 3.10: Haps Usecases in Future 6G Networks [21]

IoT-based services and applications do not need ultra-low latency provisioning (even if HAPS may be also used for these use cases) strongly enhance their use in this situations. Moreover, aerial platform coverage in white spots or greenfield areas, allows the full development of these connected devices, thus making HAPS-SMBS very useful for collecting data from IoT devices and providing reliable uplink

connections for them.
HAPS-SMBS to cover unplanned user events, as in many cases the actual terrestrial network may not be enough. Examples are flash crowds, football matches, or other particular events where there may be the need for additional support [72]. In recent times, UAVs have gained great momentum for these use cases. However, when dealing with that, the mMIMO capability of the HAPS, together with its higher operation time, are attributes that make these systems far better than simple UAVs in accomplishing those needs and making the terrestrial network more flexible.
HAPS-SMBS as aerial data center, used for computational offloading, as stated in [54], and to make computational tasks being solved by them and not by the local UEs. This may also enhance the ITS and CAV paradigm to be in place.
HAPS-SMBS for aerial network management, because UAVs have limited computational resources, they can also enhance their computational capabilities. Moreover, ML-provided HAPS may also directly manage the UAV network.
HAPS-SMBS to handle LEO satellited handoffs and provide seamless connectivity, acting as a sort of link between terrestrial gateways and LEO satellites that, being very fast result in frequent handoffs. This would remove one of the strongest barriers to the integration of LOE satellites in global wireless connectivity.

Chapter 4

Methodology

This Chapter illustrates the basic parameters used to carry out our simulations.

In particular, Section 4.1 depicts the general used scenario, giving a complete overview of the ground terrestrial network considered for every district of the Milan municipality and the HAPS on board capabilities.
Section 4.2 focuses on the way the traffic data provided from a popular Italian MNO are used, passing through updating, scaling and sampling passages to obtain the final content request number and their popularity distributions according to the different clusters in which they are generated.
Furthermore, Section 4.3 describes the caching strategies in use, highlighting their functioning, but also their main advantages and drawbacks.
Section 4.4 is instead focused on the inspected key performance indicators computed to evaluate the goodness of the performed strategies.
Finally, Section 4.5 shows the traffic management strategies that will be investigated, trying to simulate various use cases, ranging from loss-tolerant to delay-tolerant applications.

4.1 General Scenario

In this work, a portion of a future-generation LTE-A RAN is analyzed. Here, we imagine the integration, over many Milian populated areas, of a HAPS with onboard MEC capabilities, thus acting as a SMBS. The aerial platform is thought to offer coverage to 16 Traffic Zones of the previously mentioned city, having mMIMO provisioning, thus possessing a beam for every area of the Milan city that has to be covered. This situation is shown in Fig. 4.1.

These zones are considered since traffic profiles for them are known and they were provided by an Italian famous network operator. These data report the

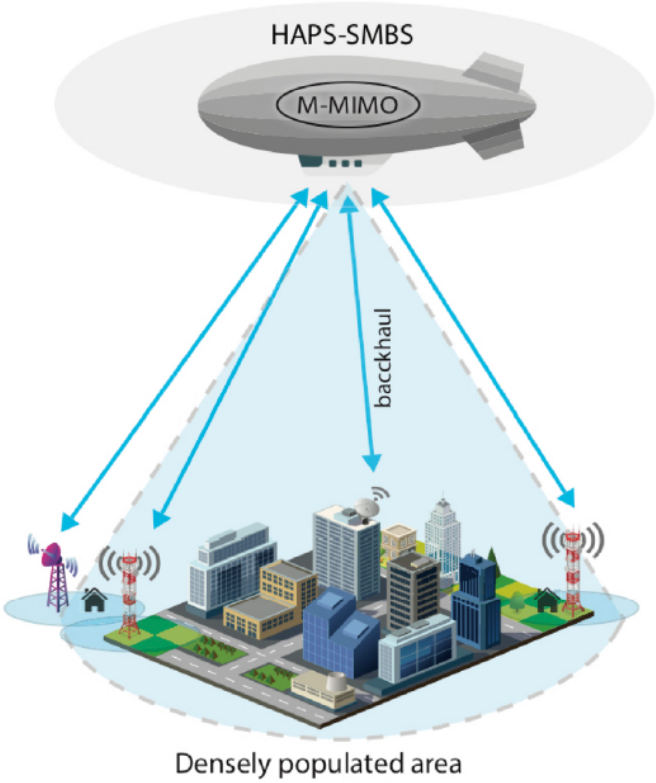


Figure 4.1: Overview of The General Scenario [22]

hourly traffic volume, in bits, for every BSs in the considered neighborhood, for more than 2 months in 2015. In our scenario, a BSs cluster is considered for every district in which the aerial platform offers coverage. The aforementioned cluster is composed of a macro BS, supported by 6 micro BSs. It has to be noted that the macro station is the only one that has caching capabilities for the system's ground segment, therefore meaning that all the cacheable content requests, are firstly offloaded to this terrestrial station. Moreover, the radio coverage of the micro and macro BSs overlaps, meaning that the small cell BSs are deployed to provide additional capacity.
In addition to that, the aerial platform is not just considered as a relay platform, but, being provided with MEC onboard server, and mMIMO capacity, it acts as an aerial data center. Therefore, some content requests can be actually offloaded to it for being cached. This means that, when a user requests contents that are locally cached either at the macro or the HAPS, they are directly delivered to

him, otherwise, they are retrieved from the cloud. Again the previously mentioned HAPS is placed at an altitude of $20km$ from the ground and it can cover an area that is greater than $40km$. This aerial system, thanks to its beams, can create cells with a radius of 2 or 3 km each. Due to these features, the entire Milan city and suburbs can be covered and a univocal correspondence between beams/cells and traffic zones is given as shown in Fig. 4.2

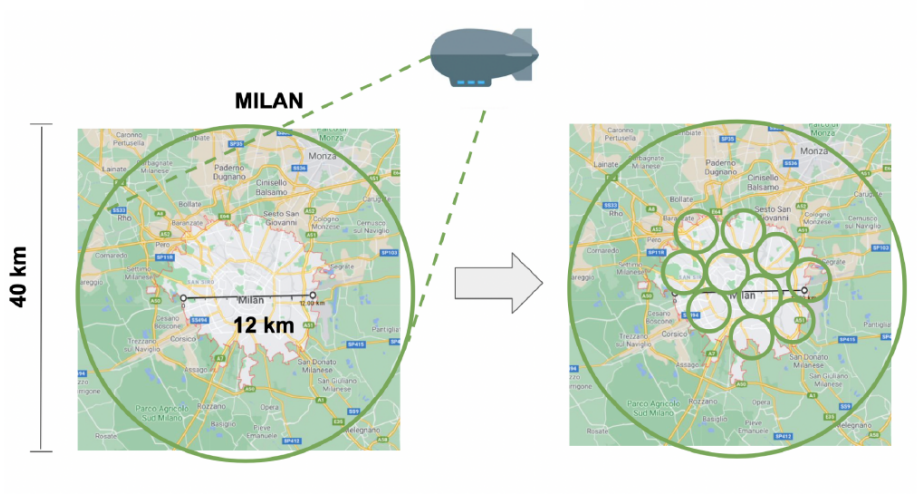


Figure 4.2: Representation of the Coverage Provided by The HAPS

4.1.1 BSs and HAPS Structure

Every cluster, composed of 1 macro BS and 6 micro BSs is connected to the power grid, but also has a PV panels series and storage capabilities, thus it is provided with an energy accumulator or a series of batteries. This means that the functioning is as follows, the cluster uses as much as possible of the RE generated by the PV panels and when the produced energy is higher than the consumed one, stores it in the batteries. When the opposite situation happens, energy is taken from the accumulator. Only in the case in which both the produced and stored energy are not enough, the power grid is used.
On the other hand, when talking about the aerial platform, it has to be considered that this system is also powered with PV panels and batteries, which are thought, for simplicity, to provide always enough energy to the non-terrestrial platform. However the HAPS has an important limitation that has to be noted, and that is the one in capacity.
This threshold is derived considering [73] that, setting the regulatory aspects of

the HAPS systems, assumes these platforms operating at a frequency between $27.9-28.2$ GHz with an efficiency spectrum, discussed in [74], equal to $4bit/s/Hz$, providing a total capacity is $1.2Gbps$. In the end, for every beam, the capacity threshold is $75Mbps/beam$.

4.2 Traffic Generation

Having understood the main aspects of the analyzed use case, content generation has to be taken into account. To do that it has to be noted that, the data provided by the MNO are the ones relating to the total traffic in the LTE network on the ground. Therefore, these time series must undergo five important steps before being used.
First of all demand profiles are sampled with a granularity of 1 hour. Then, the summation of the uplink and downlink volumes is performed. These values are respectively marked as ${T_{bs,t}}^{DL}$ and ${T_{bs,t}}^{UL}$. After that, an over-scaling operation is done since the provided profiles are outdated and need to be updated to be used. This phase is carried out in the same way as what was previously done in [75]. Thirdly, the augmented traffic undergoes an under-scaling step, since, considering it as completely cacheable traffic is not a possible operation. For this reason, CDNs related scenarios are studied and the Cisco VNI Report provided us with the correct value to perform this step. In particular, [38] says that *"72% of the overall IP traffic nowadays can be part of a CDN"*. For this reason, the total traffic is multiplied by this scaling factor as in formula 4.1.

$$ {T_{bs,t}}^{Cc} = ({T_{bs,t}}^{DL} + {T_{bs,t}}^{UL}) \cdot F \tag{4.1} $$

That can be rewritten as:

$$ {T_{bs,t}}^{Cc} = F \cdot {T_{bs,t}}^{Tot} \tag{4.2} $$

Where ${T_{bs,t}}^{Cc}$ is the cacheable traffic to be sampled, ${T_{bs,t}}^{Tot}$ is the summation of downlink and uplink volumes and F is the scaling factor $F = 0.72$. Finally, the demand profiles are sampled with a granularity of 1 hour to obtain the overall number of requests that can be cached. This is done considering the requests as equally dimensioned, with size equal to $100Mbit$. Finally, the number of content request for every time t on the given BS, which is referred to as ${N_{bs,t}}^{C}$, is computed as follows:

$$ {N_{bs,t}}^{C} = \lfloor \frac{F \cdot {T_{bs,t}}^{DL}}{S} \rfloor \tag{4.3} $$

Where S is the size of the contents. Finally, a finite library of 1000 contents is considered $\mathcal{K} = [1,2,3.....,K]$. Here each content has different popularity according to the area under investigation. Therefore, for every Milan traffic zone, a different

demand is given for every content of the aforementioned library, and modeled using a Zipf distribution, as happens in [76], [77]. The Zipf distribution has, as main

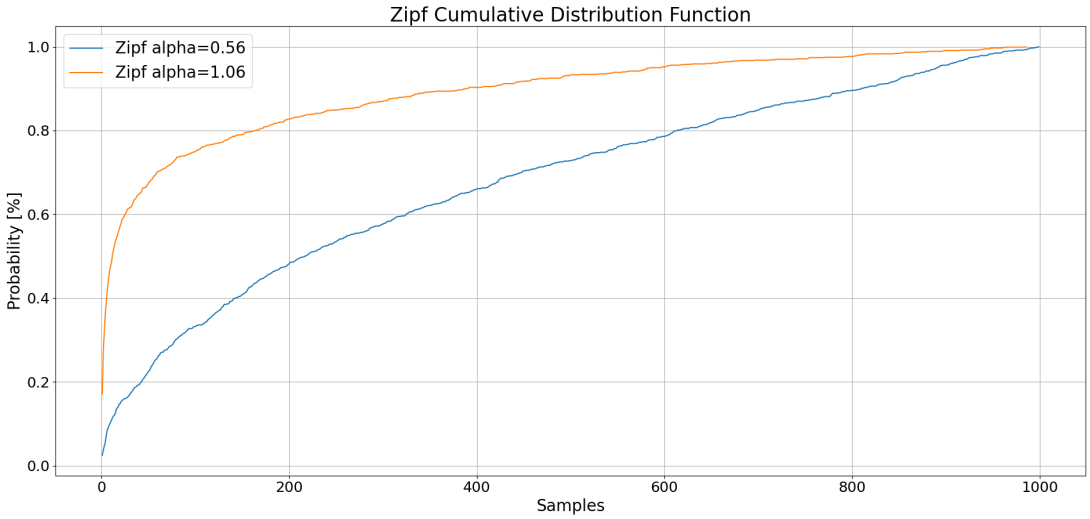


Figure 4.3: Zipf CDF with different α parameters

parameter the α value, which changes its shape, making it more peaked when high values are there. This means that, clusters that have high α parameter end up having great differences between the contents that are much requested and the least attractive ones. The opposite happens with low α values. In this simulation, the considered constant that aim at modelling these behaviors are $\alpha_L = 0.56$ and $\alpha_H = 1.06$. Their CDFs, derived from the developed python-based sampler, are shown in Fig. 4.3.

4.3 Cache Update Algorithms

After having described the entire attribute set of the main use case, the employed, well-known caching algorithms in use are illustrated and they are:

- B-LFU (Block Least Frequently Used);
- LRU (Least Recently Used);
- Q-SLRU (Quadruply Segmented LRU);
- RANDOM
- LIFO (Last In First Out);

- FIFO (First In First Out);
- NM-LFU (No-Memory LFU)

For each of them, a basic description is provided.

LIFO / FIFO

First of all, some of the most well-known algorithms are the so-called FIFO and LIFO. These strategies have a dual cache eviction policy. The first one removes the oldest content that was inserted in the cache, while the second, when a cache miss is there, evicts the last requested content. These algorithms have now lapsed as they lack both a real concept of frequency and popularity of the contents. We have chosen to implement them anyways to have a reference on the performance of the given techniques, and content assignment strategies.

B-LFU / NM-LFU

The blockwise LFU algorithm is a frequency-based algorithm, which updates the cache according to the contents that are requested the most in the past hour. It is called *block-wise* since the cache renovation is performed according to the simulation's time granularity, and not content by content. At the start of the simulation, a dictionary containing the whole library is created both at the macro and the HAPS, and their values are updated every time a content requested by a user is offloaded either to one or the other segment. At the end of every timestamp, the two caches are refurbished inserting the most requested contents. Given the fact that the simulation does the simplificative assumptions of static content library and popularity on every cluster, this strategy is intrinsically advantaged with respect to the recency-based ones, as its main drawback, which is the one of being not effective during periods in which popularity fastly changes, is not inspected.
On the other hand, the algorithm referred to as *NM-LFU* works on a content-based granularity, meaning that the dictionary of content occurrences, and thus the cache, is updated content by content. However, this algorithm is used to inspect the necessity of memory of past events for a frequency-based algorithm. Indeed, the NM prefix stands for no memory, since the algorithm forgets about the past history of content when it is evicted from the cache. This means that this algorithm ends up being completely static in case of α_H distributions, or performing more or less like a *LIFO* replacement strategy when α_L are used. This happens because as content enters the cache, it is the less requested one, thus being the more vulnerable to replacements, and when is evicted, its history is completely forgotten, as happens for the *LIFO*.

RANDOM

This algorithm performs the cache update following a completely random logic. As a matter of fact, at the end of the first timestamp, it takes a random sample from the contents and stores them in the cache. This choice is made to inspect whether an aleatoric drawn may be beneficial in some particular use cases, for example, when α_H is used.

LRU / Q-SLRU

These are the two last strategies developed in our simulations. They are based on a recency-based eviction with content granularity. LRU works by inserting every new content that results in a cache miss, or every cache hit at the head of the queue and evicting the one at its tail. Given that, this algorithm ends up performing a very aggressive policy, introducing new content directly in the most important position which is also less vulnerable to replacement. This has strong advantages when a fast change in popularity distributions and contents birth and death processes are there but is strongly impaired when flat use cases are developed, as in this scenario.
Moreover, to counterbalance its effects, a segmented alternative is provided, also called K-LRU, where K stands for the number of active queues. The *Q-SLRU* algorithm divides the cache as if it was composed of different, smaller, and equally wide segments that work together. In particular, the four provided queues are ordered from the one containing the most popular content to the one containing the fewer ones. When a cache miss happens, the new content is then inserted in the head of the less popular queue. On the other hand, when a cache hit is there, the content is promoted to the head of the adjacent queue with an increased popularity level, shifting all the others by one position backward and putting the one at the tail of the more popular queues in the head position of the adjacent, less popular cache level. Therefore, this technique is much less aggressive with respect to the previous one, ending up being less responsive in cases of fastly varying popularity, but still able to perform well in use cases where it does not change or when it varies smoothly.

4.4 Inspected KPIs

The key performance indicators used to evaluate the performance are computed for every inspected traffic zone in Milan and every developed algorithm and context. Their main aim is the one of evaluating the effectiveness of both the choice of adding the aerial cache and the technique used to perform the caching operation and content assignment. They are:

- ***Macro Miss/Hit Probability***, for every area the probability that a new request from a user results in a cache miss or hit on the ground is computed to inspect the performance of the ground cache. This index is computed as the ratio between the overall *hits* or *miss* numbers for the given content library present in the cache, referred to us *lib*%, over the summation of the aforementioned contributions. Therefore, the formula ends up being as follows:

$$M_{miss,\ \%,\ lib\%} = \frac{M_{miss,\ lib\%}}{M_{miss,\ lib\%} + M_{hit,\ lib\%}} \tag{4.4}$$

$$M_{hit,\ \%,\ lib\%} = \frac{M_{hit,\ lib\%}}{M_{miss,\ lib\%} + M_{hit,\ lib\%}} \tag{4.5}$$

- ***HAPS Miss/Hit Probability*** for every zone the possibility that an offloaded content request is handled by the HAPS and not sent to the cloud to be retrieved is inspected, together with its dual parameter, with the formula being as in these equations.

$$H_{miss,\ \%,\ lib\%} = \frac{H_{miss,\ lib\%}}{H_{miss,\ lib\%} + H_{hit,\ lib\%}} \tag{4.6}$$

$$H_{hit,\ \%,\ lib\%} = \frac{H_{hit,\ lib\%}}{H_{miss,\ lib\%} + H_{hit,\ lib\%}} \tag{4.7}$$

- ***Total Miss/Hit Probability*** for every zone the possibility that a content request either offloaded or not results in a cloud retrieval or not.

$$L_{miss,\ \%,\ lib\%} = \frac{H_{miss,\ lib\%} + M_{miss,\ lib\%}}{H_{miss,\ lib\%} + H_{hit,\ lib\%} + M_{miss,\ lib\%} + M_{hit,\ lib\%}} \tag{4.8}$$

$$L_{hit,\ \%} = \frac{H_{hit,\ lib\%} + M_{hit,\ lib\%}}{H_{miss,\ lib\%} + H_{hit,\ lib\%} + M_{miss,\ lib\%} + M_{hit,\ lib\%}} \tag{4.9}$$

- ***HAPS, Cluster and Total Miss reduction probabilities*** are obtained starting respectively from $H_{Miss,\%}$ and $M_{miss,\%}$ and they represent the percentage reduction in the HAPS, Cluster and Total miss probability that is there when the aerial platform is provided with caching capabilities or when increasing the library dimension present on it. Those indexes have the following formula:

$$R_{H_{\text{miss, [libf\%-libi\%]}}} = 100 \cdot (1 - \frac{H_{miss,\ \%,\ libf\%}}{H_{miss,\ \%,\ libi\%}}) \tag{4.10}$$

$$R_{M_{\text{miss, [libf\%-libi\%]}}} = 100 \cdot (1 - \frac{M_{miss,\ \%,\ libf\%}}{M_{miss,\ \%,\ libi\%}}) \tag{4.11}$$

$$R_{L_{\text{miss, [libf\%-libi\%]}}} = 100 \cdot (1 - \frac{L_{miss,\ \%,\ libf\%}}{L_{miss,\ \%,\ libi\%}}) \tag{4.12}$$

Where the $libi\%, libf\%$ indicate respectively the starting and final library levels. Moreover, it has to be noted that when the $libi\% = 0\%$ this can be also mentioned as the case in which there is the absence of cache on the aerial platform. These KPIs are used to give a further investigation of possible improvements related to the HAPS-SMBS provisioning.

- ***Average Hourly BH Traffic coming from the HAPS***, representing the average traffic that, hourly, has to be sent from the HAPS to the cloud to be satisfied, thus experiencing high delays. This value is the most important one when dealing with delay-sensitive applications and the one to minimize and it is computed as follows:

$$\overline{H_{BH,\ lib\%}} = \frac{H_{BH,\ lib\%}}{T_{Span}} \tag{4.13}$$

where $\overline{H_{BH}}$ is the averaged value of backhaul traffic for the given caching library, whose entirety is represented by $H_{BH,\ lib\%}$ and T_{Span} is instead the simulation period in hours.

- ***Hourly Average Total BH Traffic***, this value represents the portion of the overall traffic that is not cacheable or results in cache misses either at the macro or the aerial platform. Moreover, it has to be noted that this value is an average hourly overall traffic, thus meaning that it represents the average hourly traffic volume that either after being sent to the HAPS or resulting from a terrestrial BS miss, has to be handled by the core network. To compute it, its two main addends have to be computed. The first one was already computed before and is the ***Average Hourly BH Traffic coming from the HAPS***, while the other one is the ***Average Hourly BH Traffic coming from the Cluster*** and is computed as follows:

$$\overline{M_{BH,\ lib\%}} = \frac{M_{BH,\ lib\%}}{(T_{Span}) \cdot (\#BSs)} \tag{4.14}$$

where $\overline{M_{BH,\ lib\%}}$ is the averaged value of backhaul traffic for the provided cache library, $M_{BH,\ lib\%}$, is the total value of it, T_{Span} stands for the simulation period in hours and $\#BSs$ is the number of considered BSs for the inspected cluster.
Finally, the aforementioned KPI is then referred to as the summation of the just calculated values, thus being:

$$\overline{V_{BH,\ lib\%}} = \overline{H_{BH,\ lib\%}} + \overline{M_{BH,\ lib\%}} = \frac{H_{BH,\ lib\%}}{T_{Span}} + \frac{M_{BH,\ lib\%}}{(T_{Span}) \cdot (\#BSs)} \tag{4.15}$$

;

- ***HAPS and Total backhaul reduction indexes*** are obtained starting respectively from $\overline{H_{BH,\ lib\%}}$ and $\overline{V_{BH,\ lib\%}}$ and they represent the percentage reduction in the HAPS and Total backhaul traffic when the aerial platform is provided with caching capabilities on it and also depending by the percentage of libraries considered both for aerial and cluster cache objects. Those indexes have the following formula:

$$R_{H_{\text{BH, lib\%}}} = 100 \cdot (1 - \frac{\overline{H_{BH,\ ,\ libf\%}}}{\overline{H_{BH,\ ,\ libi\%}}}) \tag{4.16}$$

$$R_{V_{\text{BH, lib\%}}} = 100 \cdot (1 - \frac{\overline{V_{BH,\ ,\ libf\%}}}{\overline{V_{BH,\ ,\ libi\%}}}) \tag{4.17}$$

  Where the $libf\%, libi\%$ subscript highlights the quantity of cache present at the aerial platform side.

- ***Percentual Performance Decrease*** is considered and it evaluates the variation in a given index compared to the same value obtained for other cases. It can be computed similarly for every KPI under inspection. In the ***algorithm performance decrease*** particular case, the formula ends up being:

$$W_{BH,\ Alg,\ \%} = 100 \cdot (1 - \frac{\sum_{k=1}^{N}(\frac{V_{BH,\ k,\ \%}}{V_{BH,\ Alg,\ \%}})}{N}) \tag{4.18}$$

  Where N comprehends all the considered algorithms but the one for which we are computing the index, and $W_{BH,\ Alg,\ \%}$ is the worsening rate for the algorithm under inspection, given a specific library capability.
  This index is used again when dealing with the miss percentages considering different clusters or algorithms, thus meaning that, in this use case, we want to compute how better or worse an algorithm performs, as regards his related miss probability.

$$W_{miss,\ Alg,\ \%} = 1 - \frac{\sum_{c=1}^{C} M_{miss[\%],\ c,\ Alg,\ \%}}{\sum_{c=1}^{C} H_{miss[\%],\ c,\ Alg,\ \%}} \tag{4.19}$$

$$W_{miss,\ Zone,\ \%} = 1 - \frac{\sum_{a=1}^{A} M_{miss[\%],\ Zone,\ a,\ \%}}{\sum_{a=1}^{A} H_{miss[\%],\ Zone,\ a,\ \%}} \tag{4.20}$$

4.5 Traffic Management Developed Strategies

This section, which comes after the description of all of the aspects of the simulation, is dedicated to the developed traffic management strategies that will be inspected, thus simulating the behavior of different kinds of use cases tailored to various applications and requirements. The techniques in use are:

[1] MACRO Cache Priority and Content Check;

[2] HAPS Cache Priority with Content Check and Staticity Avoidance Provisioning;

[3] BS Priority with Content Check and Staticity Avoidance.

4.5.1 Strategy 1, MACRO Priority and Content Check

The first strategy, used as a general scenario, is the one already provided in [6] and works as in the following pseudocode:

Algorithm 1 Scenario 1 Strategy

```
    foreach time do:
2:      foreach BS in cluster do:
            foreach content do:
4:              if t==0 do:
                    if HAPSLim//2 not reached do:
6:                      assign content to the HAPS;
                    else:
8:                      assign it to the macro;
        else do:
10:                 if MACRO has the content:
                        assign it to the macro;
12:                 else do:
                        if HAPSLim not overcame do:
14:                         assign it to the haps cache;
                        else do:
16:                         assign it to the MACRO cache;
```

This strategy is aimed at optimizing the overall ground BS cache performance, thanks to the content check provided on that. However, it has no concerns about the HAPS Cache thus meaning that it is not optimized at all, having non-minimized traffic in backhaul that comes from the aerial platform. Therefore, it is not tailored for delay-sensitive applications, as this kind of traffic is the one

that experiences more delay and jitter as it has to go to the HAPS, and then be retrieved from the cloud. Moreover, for the way in which it works it intrinsically gives priority to the terrestrial cache, thus envisioning a ***"hierarchical caching with terrestrial priority"*** scenario.

4.5.2 Strategy 2, HAPS Priority, Content Check and Staticity Avoidance

This Scenario has been created to increase the performance regarding the **HAPS BH Traffic**, without having a complete static cache on the HAPS. In particular, this strategy performs as follows:

Algorithm 2 Scenario 2 Strategy

```
1:  foreach time do:
2:      foreach BS in cluster do:
3:          foreach content do:
4:              if t == 0:
5:                  if HAPSLim not overcame do:
6:                      assign the content to the haps;
7:                  else do:
8:                      assign it to the cluster cache;
9:              else do:
10:                 if HAPSLim not overcame and content in cache do:
11:                     assign the content to the haps;
12:                 if HAPSLim not overcame and last BS do:
13:                     assign the content to the haps;
14:                 else do:
15:                     assign it to the cluster cache;
```

Thanks to the condition on the 5^{th} line, as much content as possible are assigned to the HAPS at the first timestamp so that it can store and understand which are the most popular ones. After that, to avoid either a low traffic value or a wrong popularity detection from the HAPS, may prevent it to spot new popular contents, thus making its cache completely static, the last assignment is performed in line 13, only when the contents stored in the HAPS are so old and unpopular that their total number of requests is lower than its saturation value. This reasoning may be not beneficial in this simulation, where no popularity change in time is provided but makes this Scenario much more realistic. Finally, this situation's functioning is the opposite with respect to the previous one, thus being defined as the ***"hierarchical caching with aerial priority"*** scenario.

4.5.3 Strategy 3: BS Priority, Cooperative Caching

The last Scenario that is going to be analyzed is a sort of summary of all the other ones, trying to understand what can be the *best choice* in any case.

Algorithm 3 Scenario 3 Strategy

```
    foreach time do:
2:      foreach BS in cluster do:
            foreach content do:
4:              if t==0 do:
                    if HAPSLim//2 not reached do:
6:                      assign the content to the HAPS;
                    else do:
8:                      assign the content to the MACRO;
                else do:
10:                 if MACRO has the content:
                        assign the content to the macro;
12:                 else do:
                        if HAPSLim not overcame and content in cache do:
14:                         assign it to the haps cache;
                        elif HAPSLim not overcame and t == 0 do:
16:                         assign it to the haps cache;
                        elif HAPSLim not overcame and indexBS==last do:
18:                         assign it to the haps cache;
                        else do:
20:                         assign it to the MACRO cache;
```

This use case foresees the use of a Cluster priority on requests assignment with content check, to reduce the **Total BH Traffic**. However, in contrast with what happened in Scenario 1, here the same check is performed on the HAPS content library before assignment, thus trying to reduce the value of the most critical parameter, the **Avg HAPS BH Traffic**. Moreover, to generalize its use, the same staticity avoidance procedure carried out in Sec. 4.5.2 is given here. Finally, the algorithm performs as in the provided pseudocode. This algorithm is expected to give similar performance regarding the Cluster Cache with respect to the basic use case, having better handling of HAPS Cache, therefore increasing its overall performance. Finally, being aware of both cache stored contents this use case will be referred as to ***"cooperative caching scenario"***.

Chapter 5

Time Series Analysis

This Chapter focuses on the analysis of the provided data. Indeed, after having widely described the basic Scenario and the whole set of algorithms and use cases under study, there is the need to look to the data that we are provided with, trying to inspect some particular behaviors and general trends.
First of all, being the traffic volumes outdated, they are revised to better represent the modern demand for services. This operation is done starting from what is provided in [75].
Secondly, Section 5.1 gives a visualization of LTE BSs positioning in the city and relates inspected clusters to Milan neighborhoods. Moreover, here also the choice of the traffic zone to analyze is carried out according to different activities and expected patterns. Finally, CDFs are computed for all of them highlighting their main attributes.
Having done that, Section. 5.1.2 is carried out to compute some meaningful statistics about the chosen areas. It mainly aims at spotting their general trends, and behavior throughout the analyzed period and peak hours and days.
Sec. 5.1.3 illustrates the results when computing the rolling average and standard deviations of those time series, meaning that here we wanted to inspect if they can be considered stationary. Lastly, the same section provides also the ACF calculation for the given inputs, using two different time windows, thus trying to investigate evident periodicities both on daily and weekly periods.

5.1 Time-Series Visualization and Main Considerations

The collections under analysis come from one of the main MNOs in Italy and they refer to Milan city. In particular, they come from the deployed 4G technologies of this large operator and they report the traffic demand volume in bits, of 1420 BSs

in the different neighborhoods of the previously mentioned city with a granularity of 15 minutes. Moreover, these data have a span of 2 months in 2015, starting from the 1st of March and lasting 8 weeks, 4 days, and 23 hours, until 11 pm on the 30th of April.

5.1.1 Visualization and CDFs

In this paragraph, we aim at computing the CDF of the inspected dataset for different neighborhoods of the considered city. To do that, a first visualization of Milan districts has to be given for the 16 clusters considered. Fig. 5.1 represents the placement of the cells under investigation in the municipality areas. Moreover,

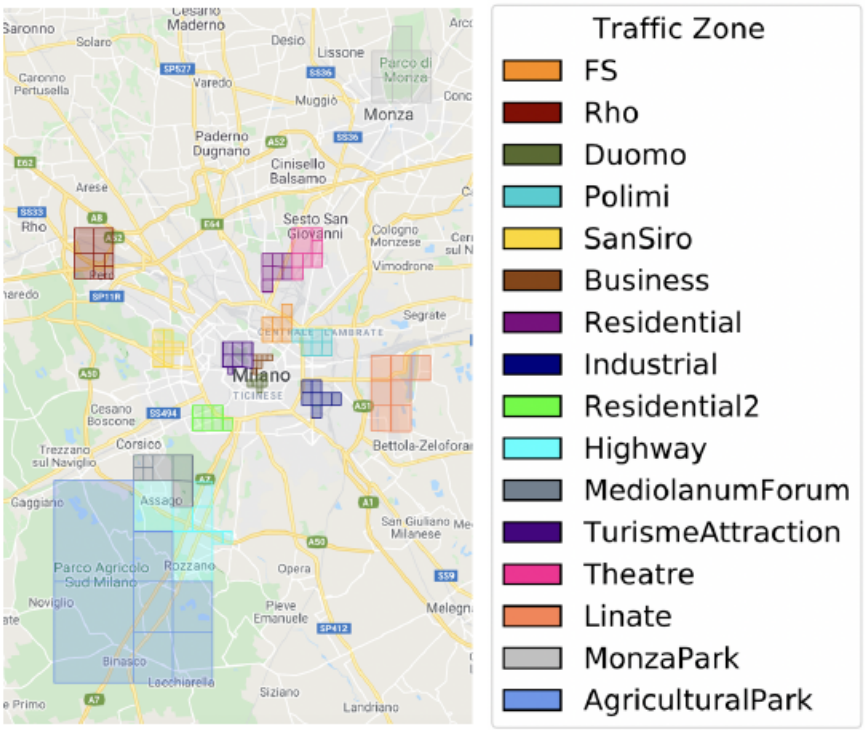


Figure 5.1: Map of the considered cells in Milan [14]

a *geomap* of the traffic zones composing Milan is provided analyzing also the neighborhoods of the city, in which these clusters are placed, as in Fig. 5.2. This plot is obtained using the *folium* library in python, with the *openstreetmap* reference, and neighborhoods data retrieved from the Milan city website [78]. Figures show that the cell data covers also many parts of the Milan outskirts, such as the *Rho*

cluster, the *Highway* and *Agricultural Park*, and *Linate* ones, with the last being placed where the airport is located. On the other hand, the *Duomo* districts and cluster coincide, as happens for the *San Siro* stadium, and *Città Studi*, referring to the *Polimi* area. Moreover, the Train Station traffic zone, which from now on will be referred to as *FS*, coincides with the *Centrale* district, while the *Residential* and *Residential2* areas are the *Niguarda* and *San Cristoforo* neighborhoods. Lastly, the *TurismeAttraction* part refers to *Parco Semione* and *Brera* districts, and the *Business* part is always inside the *Duomo* zone.

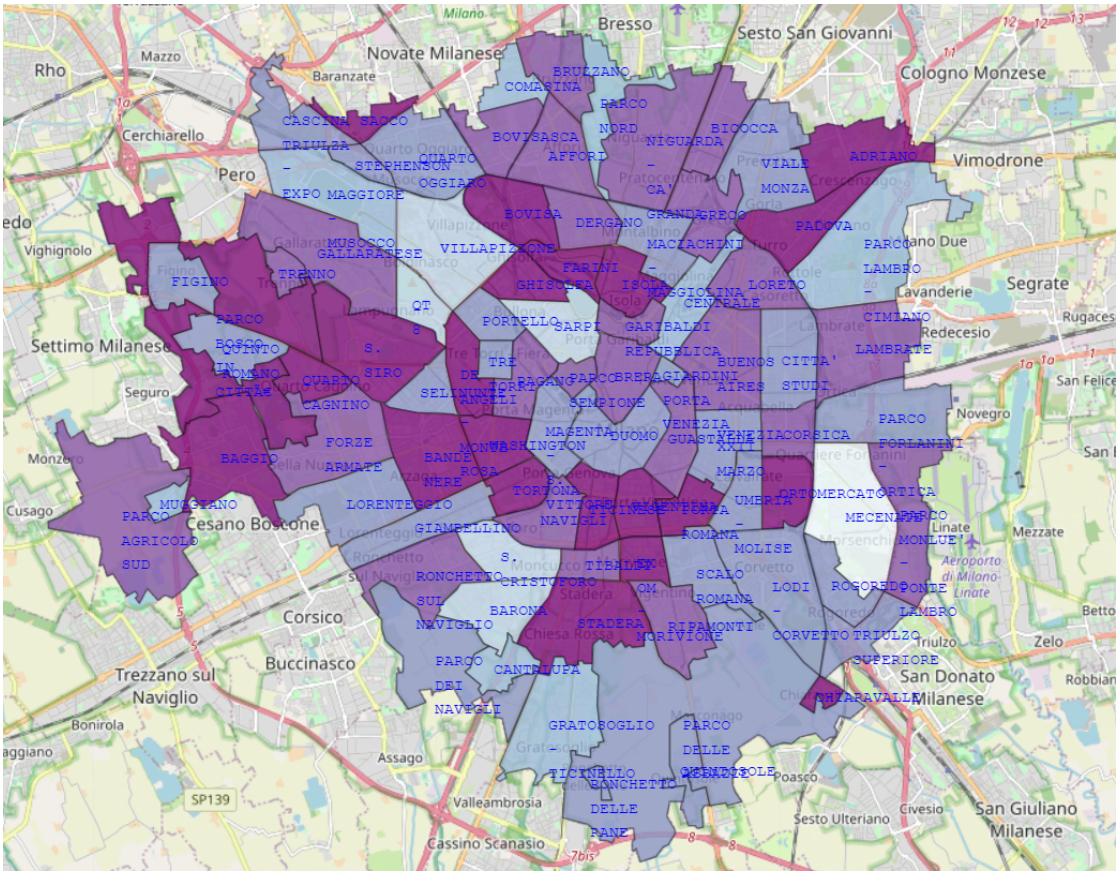


Figure 5.2: Neighborhoods in Milan

Finally, CDFs for every cluster are obtained. This step is crucial to derive the evolution of the traffic demand. The chosen traffic zones are the ones of ***FS***, ***Rho***, ***San Siro***, ***Business***, ***Residential***. This choice is made to compare completely different areas, where humans perform different actions, thus showing different habits. For this reason, we expect to have different CDFs for each of them.
Fig.5.3 depicts the results for this calculation for each of the aforementioned clusters. It has to be noted that, to improve the visual impact of the given graph, the

x-axis is plotted using a logarithmic scale, thus highlighting the differences in the various CDFs. Moreover, plotted results show what is expected. In particular, the *FS* district is the one that generally has more traffic volume than the other ones. This is due to the presence of the Milan Central Train Station, being a crucial point for people's mobility in the city. Moreover, it is interesting to highlight the intersection between the *Residential* and *Business* clusters in their curves. This is because, the first mentioned district, being the one in which people live, has more *stable* and *durable* demand profiles, therefore showing a lower slope for very low traffic values, meaning that it is very rare to have low traffic volumes compared to the business zone. This behavior is strongly expected as in the last mentioned area, the BSs are loaded during worktimes, while during the weekends they should be almost completely unloaded. Finally, *Rho* and *San Siro* are the ones that show lower volumes of traffic. In this case, it is important to consider the fact that these areas are mainly used for manifestation and fairs, thus meaning that they may have rare, high peaks in traffic profiles. This behavior is confirmed by the curves that are the only ones, particularly for the *Rho* district that has a heavy tail when moving from the 90th percentile to the 99th percentile.

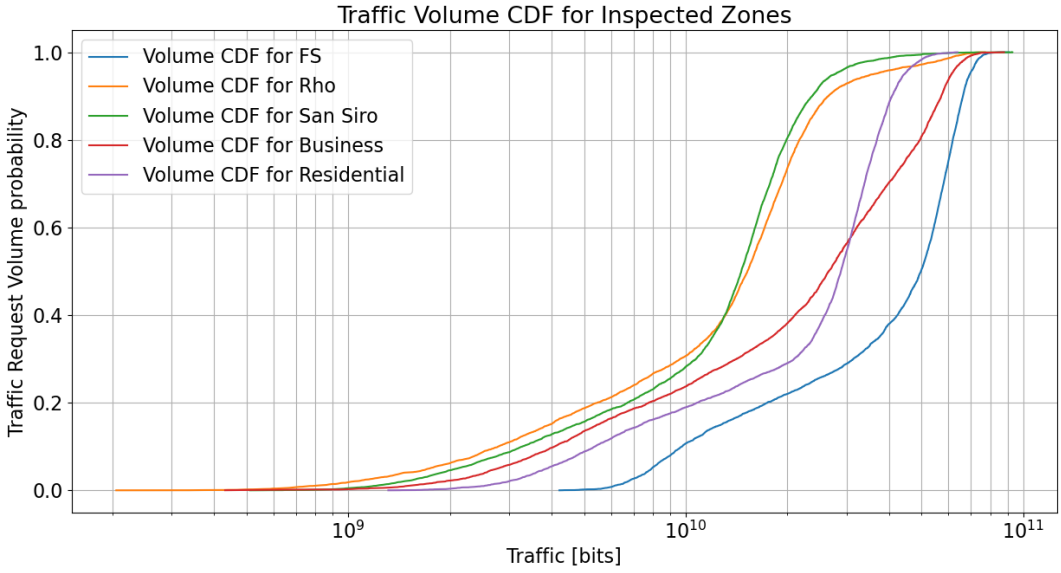


Figure 5.3: CDFs of the Chosen Districts in Milan

5.1.2 Parameters Computation

After having derived these graphs, some meaningful statistics are calculated. In particular, they are the *average*, the *standard deviation*, and the 90*th percentile*.

The computations are carried out to derive two meaningful pieces of information from this analysis which are the peak hours, and the peak days, highlighting the day-night cycle, and week one. In particular, the first step that is carried out is the one of computing the average traffic in every zone aggregating data both according to an hourly and daily base. Fig. 5.4 shows the obtained results for the first traffic aggregation in every single considered district. First of all, a general trend, which is common to all the plotted curves, is the one, strongly comprehensible, related to the sharp decrease in the traffic volume during nighttime. Starting from 1 AM, even the most loaded neighborhood has very low volumes. Moreover, it has to be noted that the *FS* district has an overall profile that is always very high, due to the intrinsic importance of this area, already discussed before. Moreover, its peak is reached at 6 PM, and this is mainly because people go there to come back from work. As a confirmation of what just said, *Business* district's volumes strongly decrease after 5 PM, having their peak at 3 PM. In contrast to that, the residential traffic profile has its maximum value when people come back home from work, at 9 PM. The same hours are the most important ones, in *San Siro*, obviously due to the football matches that attract a massive amount of people to go there. Finally, *Rho* district has its higher values during the afternoon. This is related to the fair events that are done here, attracting people during daily hours. Fig. 5.5 highlights the traffic average value for every day in the considered cluster.

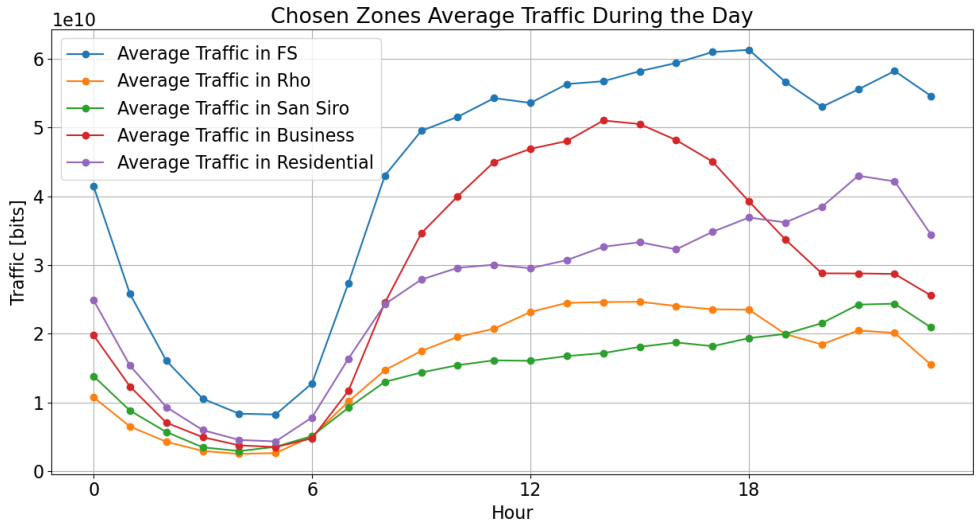


Figure 5.4: Daily Traffic Average for the Chosen Districts

This picture highlights the weekly traffic variation in the demand profile. The most general trend, visible already from a first glance at the aforementioned graph, is the weekdays-weekend alternation, with a sharp decrease in the traffic demand during

these last periods. This affirmation is easily carried out as the grid corresponds to Sundays. However, a particular district is completely in contrast with the previous affirmation and it is the one where the stadium is placed. This happens obviously since football matches are held on the weekends. To confirm this, this neighborhood has traffic peaks on Sundays when home teams are playing. High peaks are also there for the *Rho* district, highlighting the alternating events that are held there. Furthermore, another cluster has a particular behavior, related to its use, and it is the residential one, where the general alternation between weekdays and weekends is reversed, meaning that normally during Sundays the overall traffic demand is higher. Again, here the overall oscillation is much less with respect to the other ones, and this is because being heavily inhabited, it has always some traffic to handle. Finally, all the other clusters have a more *regular* behavior and cycle.

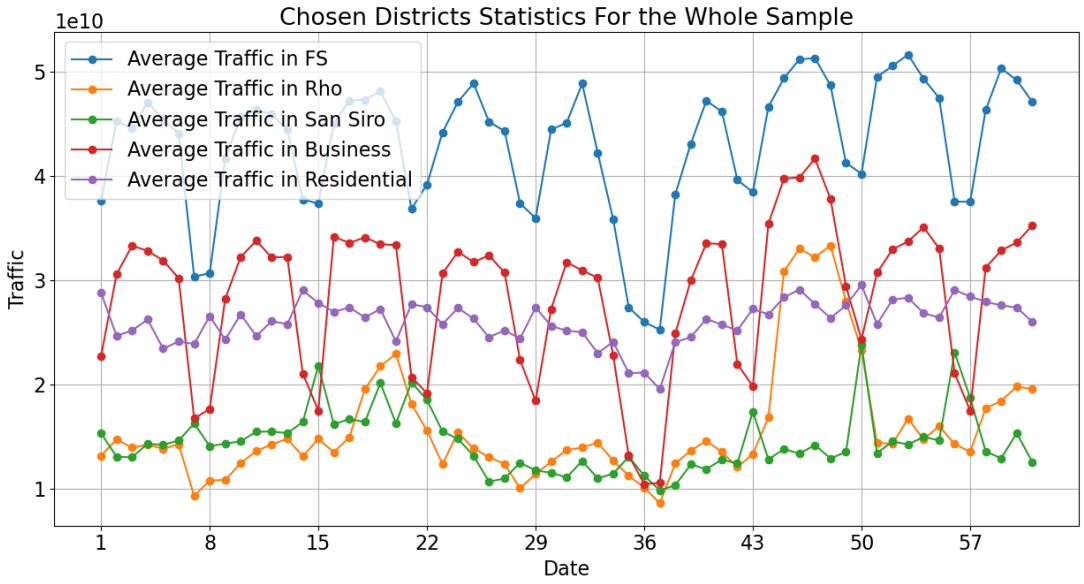


Figure 5.5: Traffic Average in the Sample Period for the Chosen Districts

The previous affirmations are also confirmed from Fig. 5.6 highlighting that *Rho* and *San Siro*, despite having low traffic demands in general, are the ones that show the highest standard deviation when an event is held there. Furthermore, it has to be noted that normally, due to the scarcity of demand, even the standard deviation here is very low. In addition to this, the *Business* and *FS* have similar trends, due to the same weekly cycle already spotted in Fig. 5.5. Lastly, the *Residential* district, as expected is the one showing the lower std value in general.

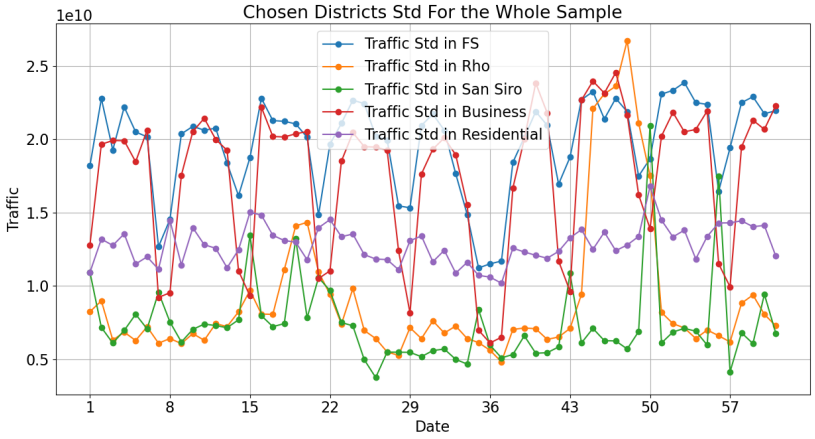


Figure 5.6: Traffic Std in the Sample Period

Finally, the 90*th* percentile is inspected. This parameter is meaningful since the CDFs have shown the fact that *Rho* and *San Siro* have heavy tails on the maximum values. For this reason, it is interesting to see that, in Fig. 5.7 the 90*th* percentile for the *Rho* districts intersects with the one from the residential neighborhood even if in the same graph obtained for the average, its mean value was far below. This is due to the first affirmation done and confirms when talking about CDFs tails.

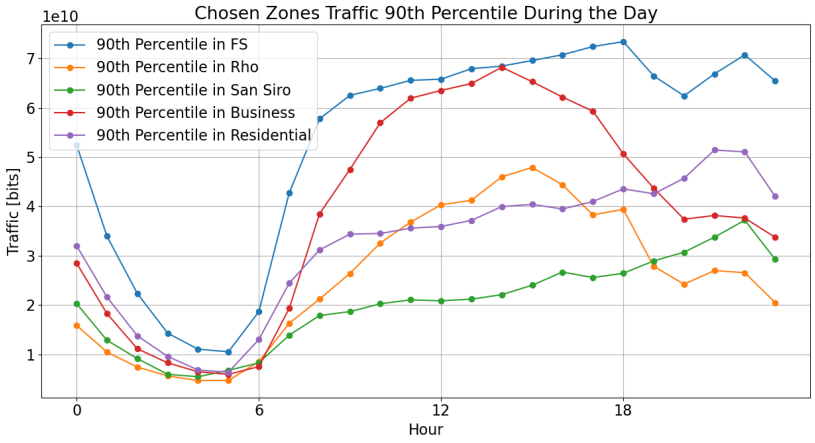


Figure 5.7: 90th Percentile of the Daily Aggregated Traffic

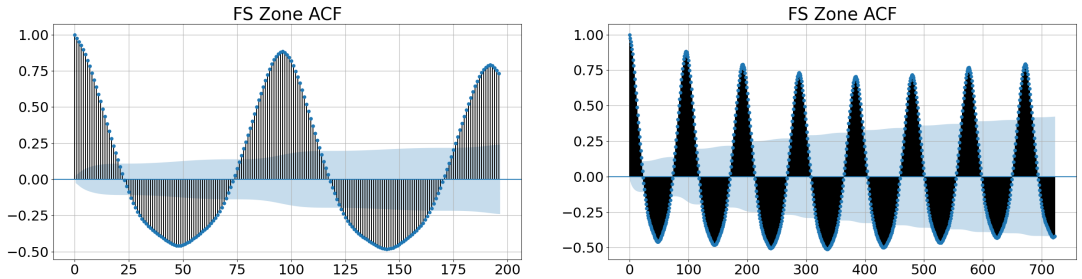
5.1.3 Statistical Properties

Lastly, some statistical properties are carried out for the time series under inspection. First of all, their stationarity is inferred by computing their rolling statistics. In particular, rolling averages and standard deviations are considered on a week-moving window. This choice is made to smooth out both daily and weekly periodicities and also local trends, thus leaving us with the global ones. Results for *FS* and other chosen zones are obtained and they are shown in Fig. 5.8. In

(a) FS series and rolling statistics (b) Residential series and rolling statistics

Figure 5.8: Clusters Time-Series and Rolling Average and Standard Deviation

particular, in this graph, it is possible to see *FS* and *Residential* areas. The choice of plotting only these two is because, generally, all the series are stationary, while they have a little decrease during the first week of April. However, this is due to the easter holidays. The choice of depicting these clusters is related to the fact that this phenomenon is less evident in the Residential one, as during the easter period people may also decide to stay home and spend their leisure time with their relatives. Secondly, the ACF is computed. This procedure is done to determine



(a) Train Station ACF for 196 samples (b) Train Station ACF for 722 samples

Figure 5.9: ACF for Chosen District, Daily, and Weekly Periodicities

eventual periodicities, and confirm the results obtained in the previous paragraphs.

In particular, the chosen windows are composed of 196 samples, thus meaning 2 days, and 722 samples, therefore considering twelve days, to depict longer periodicities. Both Figs 5.9a and 5.9b confirm what was said for the *FS* cluster, showing great daily and also weekly periodicity of data. Finally, those results are the same for every single cluster under inspection.

Chapter 6

Terrestrial Priority Scenario Single Beam Results

Having carried out such a deep analysis regarding all the aspects of our simulator is now time to analyze the obtained results for the first use case under inspection, that is the one in which the HAPS gives coverage, and thus its MEC capabilities to a single cluster in the Milan municipality. This means that its cache is entirely devoted to supporting the one developed in the macro BS that is present on the ground. The final aim of this chapter is to spot the general trends when adding a second, aerial, and thus limited in capacity, cache object to the original one provided on the ground. Moreover, this section of the work is focused on pointing out whether, and in which amount, both the traffic volume and the different contents distribution may influence the overall results when a new object is there providing offloading capabilities.
Section 6.1 illustrates the main parameters used to carry out the simulations under analysis.
Section 6.2 focuses on giving value to the provided analysis by discussing the improvements the aerial platform addition brings when performing caching in reducing the amount of traffic that may experience overwhelming delays. To do that, it computes the ***HAPS Average Backhaul Traffic*** highlighting the importance of the aerial MEC provisioning for reducig the aforementioned quantity.
Section 6.3.1 describes the results obtained on the cluster and aerial sides, thus inspecting the main factors that influence the general trends in the simulations.
Finally Section 6.3.2 derives the results regarding the ***Total Average backhaul traffic*** for each algorithm and compares them, therefore giving an overview about what are the best-performing ones, and explaining the main reasons that led them to function in that way. Furthermore, it also aims at highlighting the major differences that are found when the same analysis is carried out on different traffic

areas of the same city.

6.1 Simulation Parameters

The parameters used to carry out simulations are the following:

- FS, Rho, San Siro, Business and Residential Traffic Zones in Milan are considered;
- Zipf Distribution Parameter $\alpha_{FS,Bus,Res} = 0.56$ and $\alpha_{Rho,Siro} = 1.06$. Therefore, the districts belonging to the second group, so the ones with higher distribution value, surely experience many requests for the same content at the same time. On the contrary, the other areas experience many different content requests every time;
- Macro Cache Storable Library Percentage $D_M \in [0\% - 100\%]$, with sampled values that represent three main situations. The first one is the so-called ***Small Cache***$= 10\%$, then there is the ***Medium-Sized Cache*** $= 40\%$ case. Finally, also the ***Huge Cache Provisioning*** $= 70\%$ situation is envisioned.
- HAPS Cache Library Dimensions $D_H \in [0\% - 100\%]$. Also here it was chosen to use sampled values. However in this case they can represent different situations. The first one relates to the case in which ***No Cache*** is envisioned, meaning that the $= 0\%$ of the content library is storable in the cache. This situation is that in which the aerial system is used as a simple traffic relay. In addition to that, the same three scenarios compared to what was previously considered for the Macro Cache are analyzed when the HAPS is provided with MEC capabilities onboard. Further cases are not envisioned since increasing the storage capability over the 70% is considered as not valuable;
- Algorithms in use:
  - Block LFU (that from now on will be called *B-LFU*;
  - RANDOM;
  - LRU
  - Q-SLRU;
  - FIFO
  - LIFO
  - LFU no memory (that will be simply referred to as *NM-LFU*).

6.2 HAPS-SMBS Viability and Importance

The first thing that has to be performed to give value to our analysis is to affirm the fact that, not only is it commercially viable to enable HAPS caching but also that it is worth it from a performance point of view. That is the fundamental scope of this Section.
To fulfill this task, a direct comparison between results obtained considering or neglecting HAPS-SMBS provisioning is given. In particular, this first analysis is focused on the ***HAPS Backhaul Traffic***. This value, computed as in 4.13 indicates the volume of requests that, being offloaded to the aerial system, cannot be locally satisfied by the non-terrestrial MEC system onboard of this platform and have to be managed differently. Therefore, the traffic volume that experiences this chain of events is the most critical one, experiencing a higher RTT both compared to the requests that can be handled locally and also to the one that are unsatisfied at the terrestrial segment. This happens because these last ones, even resulting from a *cache miss* can reach the core network through wired connections instead of wireless ones, as happens for the requests that cannot be managed on the HAPS.
In addition to that, to deepen this general comprehension, also the ***haps backhaul reduction*** value is considered, and calculated as in 4.16. This choice is made to provide a precise idea of the performance improvement that HAPS-SMBS technology grants with respect to the situation in which the aerial platform is used only with relay capabilities.

6.2.1 Average HAPS Backhaul Traffic and HAPS backhaul reduction

Having described the main reasons that led us to inspect, as first indexes, the previously cited values, it is now time to illustrate the developed analysis.
The examination of the mentioned indicators starts by evaluating the first named index, considering the *B-LFU* caching strategy as the one of reference to investigate its behavior when HAPS-SMBS is foreseen and functioning.
Fig. 6.1 shows the ***Average HAPS Backhaul Traffic*** parameter fixing the value for the macro cluster library between one of the considered samples, thus for ***Small Cache***, ***Medium-Sized Cache***, and ***Huge Cache*** and then illustrating both the case in which the aerial platform is only used as a relay system for the considered cluster or the situation in which the same caching capabilities are granted to it. The KPI values are plotted for all the districts chosen in Section 6.1 using the *seaborn.barplot* python library, thus having a different bar for all the aforementioned cases.

The presented graph highlights the strong effectiveness of the performed operation and the great improvement that a HAPS-SMBS may furnish, already from

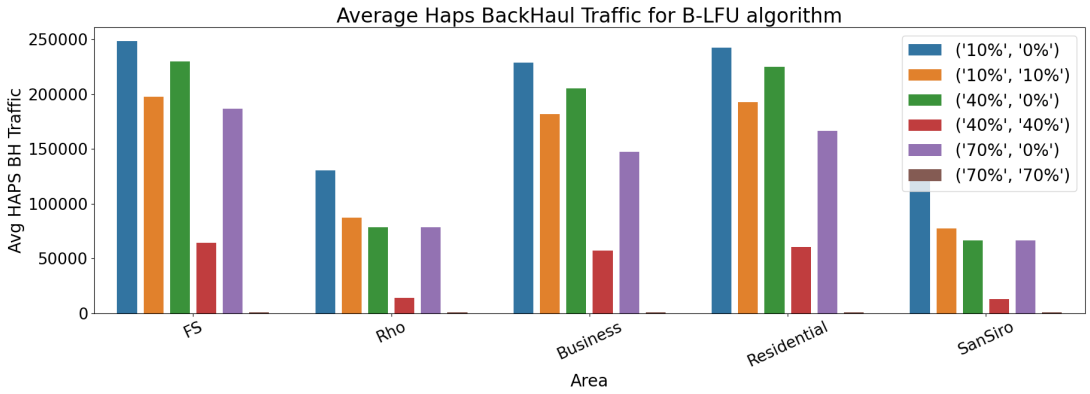
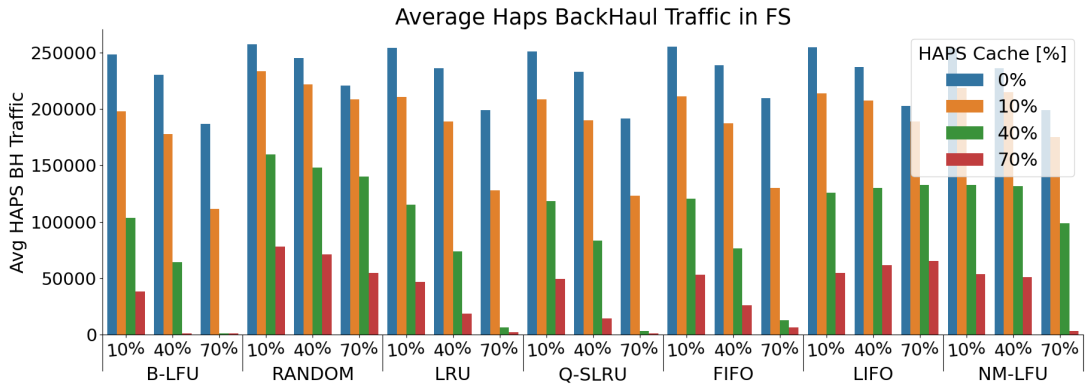


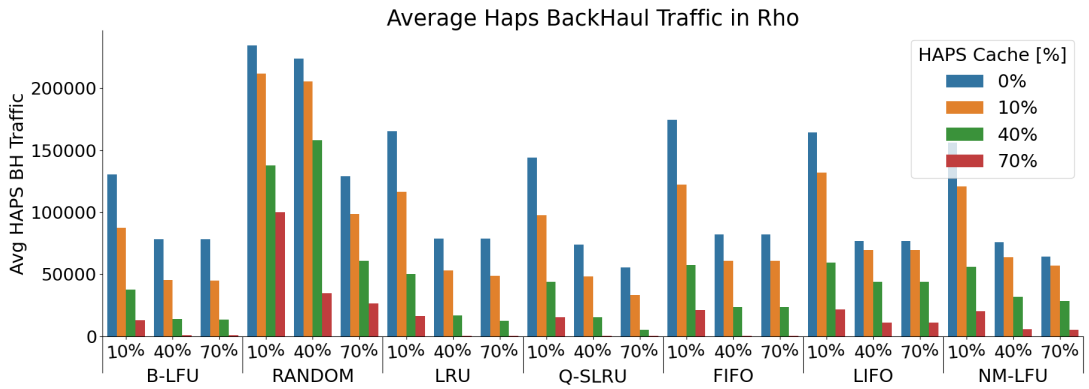
Figure 6.1: Average HAPS Backhaul Traffic with B-LFU Algorithm

the very first caching provisioning, for all the traffic zones under inspection.
Moreover, it has to be noted that this performance improvement is there not only when the HAPS would have normally been loaded with many requests from the terrestrial side, as happens when it is used as a relay platform for the *FS*, *Business* and *Residenzial* zones, but also when it would have been less loaded, as in *Rho* and *San Siro*. This phenomenon coincides with the fact that the first mentioned areas are the ones, as inspected in Section 5, that show higher average traffic volumes, while the latter ones are the most unloaded. These considerations confirm the fact that regardless of the total number of offloaded requests, and thus general traffic volumes, providing the non-terrestrial integrated platform with onboard MEC capabilities is always beneficial.
Having said that, from a deeper inspection of Fig. 6.1 it is also possible to declare that, already from the ***medium-sized cache*** provisioning, the critical traffic level is almost zeroed, while this achievement is completely fulfilled when ***huge*** sizes are there.
This highlights the fact that, even when the terrestrial cache is provided with a very performant strategies and increasingly sized libraries to handle locally BSs requests, its aerial counterpart is still able to greatly improve its performance.
This result is very encouraging and gives the driver for further analyzing this technology. In this context, it has to be said that this effect is a consequence of the *hierarchical-caching* strategy provided. Indeed, performing a presence check on the macro BS cache before the offloading operation, allows the two memory objects to be always effective, as the terrestrial end will be loaded with the most popular contents, while its non-terrestrial counterpart will be given the less popular ones, therefore being very efficient most of the time.
To carry on this analysis, all the algorithms should be inspected, thus being able to

affirm that the drawn conclusions are not only meant for the blockwise frequency-based caching strategy just discussed. To have a general overview regarding that, it is decided to use a traffic zone for every α parameter and traffic volume under investigation, considering each of them as enough representative for the whole sample. The elected areas for this visual inspection are the *FS* and *Rho* ones, considering all the possible cache library capacity combinations and the entire set of developed algorithms. Fig. 6.2 shows that the illustrated behavior is not just related to the previously tested algorithm emphasizing that the decreasing tendency shown in Fig. 6.1 is general and confirmed for all the employed caching update strategies, thus strengthening the previous discussion and affirmations.



(a) FS Haps Averaged BH Traffic [Mbits/h] behavior



(b) Rho Haps Averaged BH Traffic [Mbits/h] behavior

Figure 6.2: HAPS Averaged BH Traffic [Mbits/h] for FS and Rho and all the algorithms

Moreover, from Fig. 6.2 it is also possible to confirm another meaningful already

given insight, that is the one that the HAPS-SMBS provisioning is **always** beneficial for reducing the ***HAPS average backhaul traffic***, regardless the sample library of which the macro BS on the terrestrial side is provided. This result is crucial for our simulations.
Furthermore, from a more precise inspection of Fig. 6.2 it is also possible to have many other insights into the general performance of the provided strategies. Indeed, from Fig. 6.2b, which depicts the *Rho* traffic zone results, it is possible to see that the value of the ***HAPS average backhaul traffic*** clearly changes from algorithm to algorithm, showing the fact that not all the strategies are able to spot correctly the content popularity on their terrestrial segment, thus ending up with the necessity of exploit their wireless link, offloading the requests that their ground segment is unable to handle to their aerial counterpart. This happens because, being this scenario a *hierarchical-caching* one with content check provisioning on the terrestrial end, a high quantity of ***HAPS average backhaul traffic*** when the HAPS library is 0% means that the cluster cache did't perform well.
Following this intuition, we can confirm that *B-LFU* seems to be the most performing caching strategy, while *Random* is surely the worst, directly followed by *LIFO* and *FIFO* ones. As expected, the artlessness of these strategies is conducting them towards worse performance with respect to the most complex ones.
In addition to that, Fig. 6.2 also highlights another important trend that depends only from the α parameter given in the inspected districts. Indeed, when looking to the last two bar-groups for every algorithm in the provided plot, thus meaning the ones referring to the ***medium-sized*** and ***huge*** cache provisioning, it is possible to see an important difference in their behavior when dealing with the *FS* and *Rho* zones. In particular, for the for the latter mentioned district, it can be seen that the increase in ground capacity does not lead to a large reduction in backhaul traffic from the aerial platform, which, except for the random strategy, seems to have reached a plateau. This is due to the fact that the ground cache is already capable, since it is provided with medium capacities, of understanding which contents are most in use, remaining with very little-requested contents which, given how the scenario is managed, can be forwarded to the aerial platform thus resulting in zeroed cluster misses. Therefore, the increase in capacity to the 70% is not fully exploited both because α is large and because the low volume makes the simulations less dynamic.
A further insight of that is now provided to conclude this discussion, thanks to a quantitative analysis of these effects. To do that, the ***HAPS backhaul reduction*** index is computed for all the five clusters mentioned in Section 6.1, considering the entire set of developed algorithms and for the basic combinations already inspected in Fig. 6.1, thus meaning the ones [10%,10%], [40%,40%], and [70%,70%]. This value is evaluated to highlight the percentual reduction in the inspected metric compared to the case in which no HAPS-SMBS is envisioned in the investigated

district. The choice of using the three basic combinations previously mentioned is due to the fact that this operation is carried out with the final aim to further analyze and confirm if, even when less popular contents are considered, as happens in this scenario for the ones that are offloaded to the haps, a high or low α value has still some influence. This means that we expect still a higher performance increase, already from very small cache libraries, in areas in which the α value is high, with a plateau behavior when moving towards higher capacities. The obtained results are reported in Fig. 6.3 and perfectly confirm the aforementioned insights, with the *Rho* and *San Siro* clusters that always take advantage of a stronger reduction in both cases of ***small*** and ***medium-sized*** libraries, thus meaning that they are faster in spotting the content popularity correctly. However, when cache size increase further, they show a sort of plateau, when comparing with the other traffic zones, provided with less peaked content popularity distributions and higher traffic volumes, both because the terrestrial cache was already able to spot almost perfectly the content popularity and because of their reduced dinamicity, as previously affirmed.

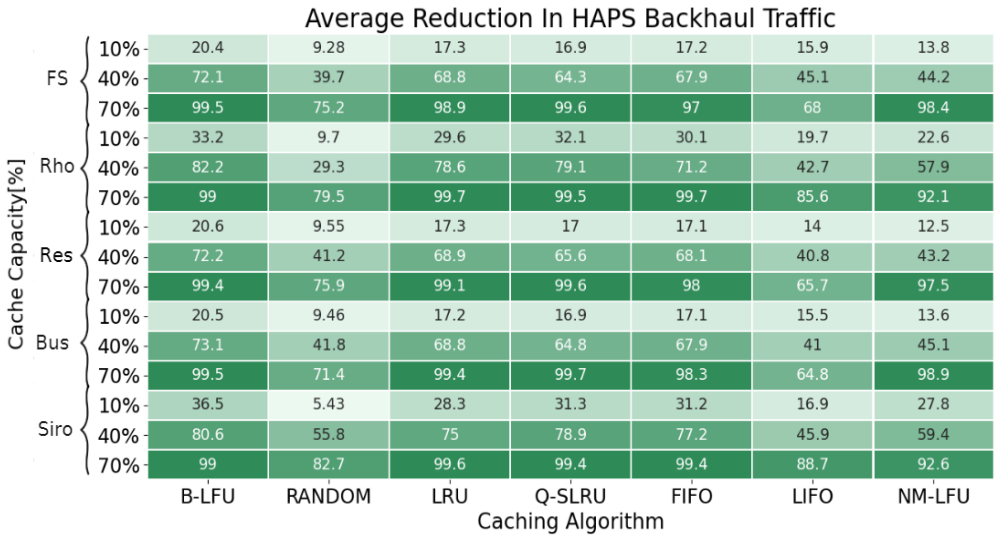


Figure 6.3: Average HAPS Backhaul Traffic Reduction

Furthermore, the plotted heatmap also highlights the fact that both RANDOM and LIFO strategies are the ones that experience the lowest performance improvement with a library capacity increase and that their behavior is somewhat very similar. This is easily explained by thinking about how the second strategy is

performing caching insertion and eviction. This technique fills its queue as much as it can and then when it is full, it simply replaces the last inserted element, thus ending up being very close to what the RANDOM algorithm does. Indeed much of its performance is directly linked to how it fills the queue. This mechanism is very similar to the one of having a random draw, the only difference is that here you can substitute the last element. However, this does not change the core of its functioning.

6.3 Further Analysis, Algorithm Comparison, and Cluster KPI Inspection

Having confirmed the viability of the HAPS-SMBS solution from the point of view of reducing the most critical metric, thus the one of the backhaul traffic generated by the aerial platform, and having spotted some general trends and behavior when dealing with this scenario, it is now time to focus on understanding how the different implemented strategies work in the overall and both on the terrestrial and aerial sides. To do that the ***Cluster*** and ***HAPS Miss Probabilities*** and the ***Total BH Traffic Volume*** values are inspected, showing the main achievements in adding the HAPS-SMBS and also the main differences between the developed caching update technologies.

6.3.1 Haps and Cluster Miss and Hit Probabilities

Even if the previous Section has provided us with a good overview of the way algorithms function overall, it still lacks a complete treatment of the developed scenario. Since now, we have had only some insights regarding how much the α parameter, and traffic volume impair, or generally influence the obtained KPIs. To carry out such analysis, the ***Cluster*** and ***Haps miss probabilities*** are chosen. This decision is taken as these KPIs, being directly related to the times the cache can spot correctly the popularity of a content are not influenced by anything else than the traffic volume and the municipality popularity distribution, thus being perfect for this study item.
The investigation is executed using the parameters specified in Section 6.1 and the first index under inspection is the terrestrial-related one. Therefore, when dealing with it, it has to be considered that, the macro BS is loaded with the most popular contents, having the macro a priority on content assignment and being the only one that is provided with a content check capability. It has to be reminded that this KPI is computed as in 4.4, meaning it is the ratio between the overall miss during the simulated use case, and the summation of hits and misses during the same period.

Figure 6.4: Cluster Miss Percentage with (10%,10%) libraries

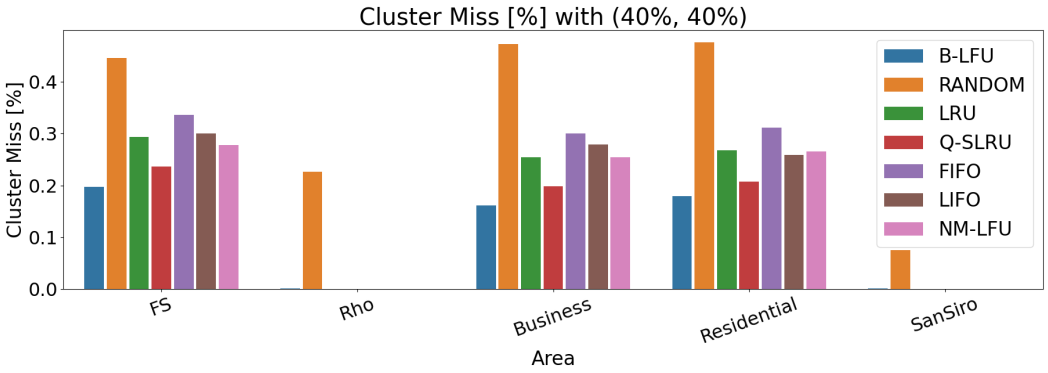


Figure 6.5: Cluster Miss Percentage with (40%,40%) libraries

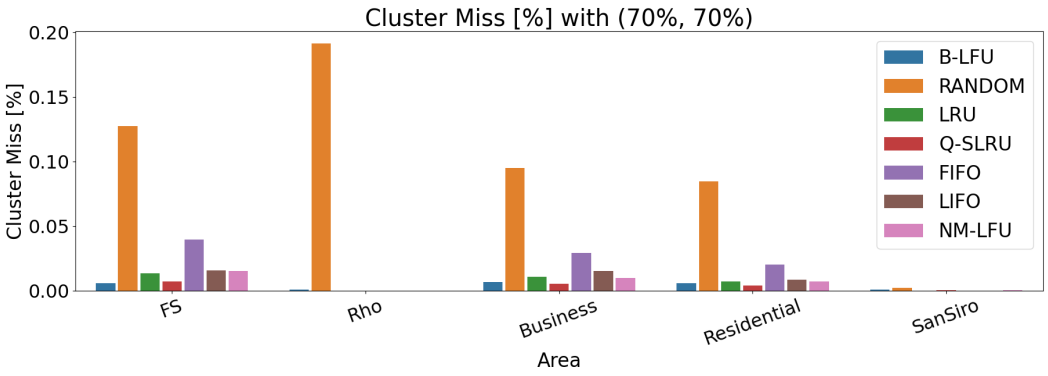


Figure 6.6: Cluster Miss Percentage with (70%,70%) libraries

The obtained results are depicted in Fig. 6.4, 6.5, and 6.6. Here it is chosen to show those figures sequentially to improve their graphic impact and grant the possibility to identify trends that only depend on the considered caching capabilities.
It is immediate to see how, even when very small libraries are there, the ***cluster miss probability*** is extremely reduced when considering districts presenting Zipf distributions with very high parameters. This result is both due to that and also to the generally scarce traffic volume in *Rho* and *San Siro* municipalities, and an insight into it was already provided in Section 6.2. However, now what was only insight is strongly and evidently justified. Indeed, in these use cases, the macro cache, already from 40% is able to perform no miss, as visible from Fig. 6.5, and the remaining contents, thus the ones that are not present in its library are few enough not to overcome the capacity limit of the haps link. Therefore, the requests that are not handled and cacheable by the cluster, are immediately sent to the aerial platform, and the added capacity when providing the macro cache with 70% libraries is hardly ever exploited, thus arriving at a complete justification of the *plateau* behavior inspected in Fig. 6.2. Therefore, we are able to affirm with certainty that increasing the distribution parameter, together with a low volume of traffic, has twofold effects. The first one is, as expected, the one of increasing the overall performance of the caching strategies, and the second is related to a plateau-like behavior that has to be considered as ***huge*** cache libraries are in fact not exploited.
On the other hand, when dealing with traffic zones in which low α is used, the increase in cache library is always used and positively correlated to the decreased cluster miss probability.
To conclude the analysis regarding the ***cluster miss probability*** we move our attention to the developed algorithms, Figs. 6.4, 6.5 and 6.6 plots confirm all the aforementioned affirmations about the*Random* strategy weaknesses, which is overcome also by the *LIFO* strategy. This happens because *LIFO* has still a small cache variability compared to the first mentioned strategy, as it is able to change the last inserted content and this behavior is beneficial overall.
Having investigated the terrestrial miss probability-related KPI is now the time to show and inspect what happens, in the same cases to the HAPS cache. When performing this description we have to remember the fact that, lacking any content check possibility, and having no priorities on requests assignment, the aerial platform is provided with *"less popular contents"*. Figs. 6.7, 6.8 and 6.9 confirm this insight. As a matter of fact, for all the provided situations, the ***haps miss percentage***, computed as in 4.6, is always higher concerning the same index related to the terrestrial cache for the same library capacity.

In addition to that, these plots also highlight another important trend, that was already present in 6.4 but that there is much more evident and that is the

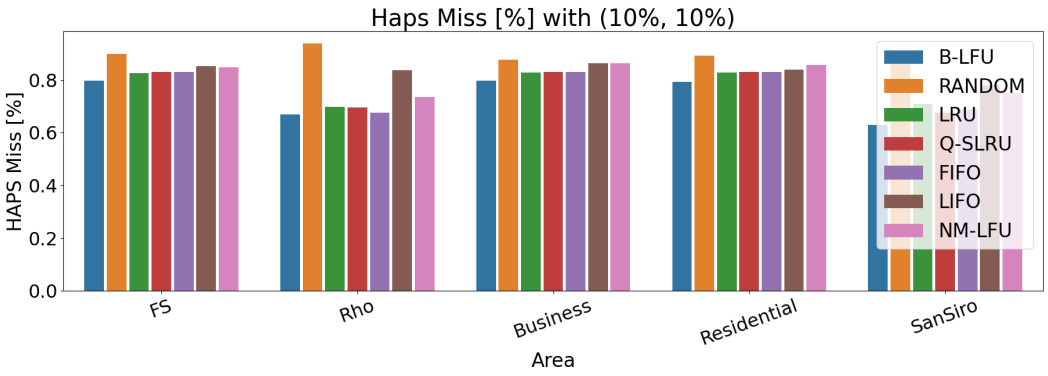


Figure 6.7: HAPS Miss Percentage with (10%,10%) libraries

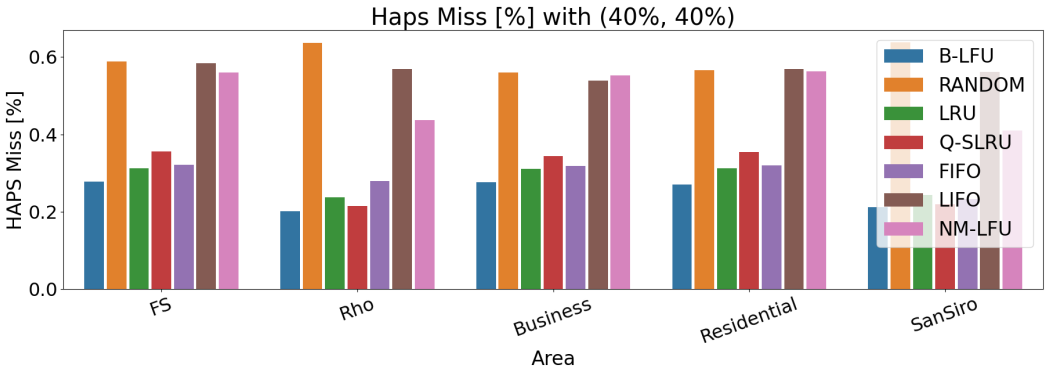


Figure 6.8: HAPS Miss Percentage with (40%,40%) libraries

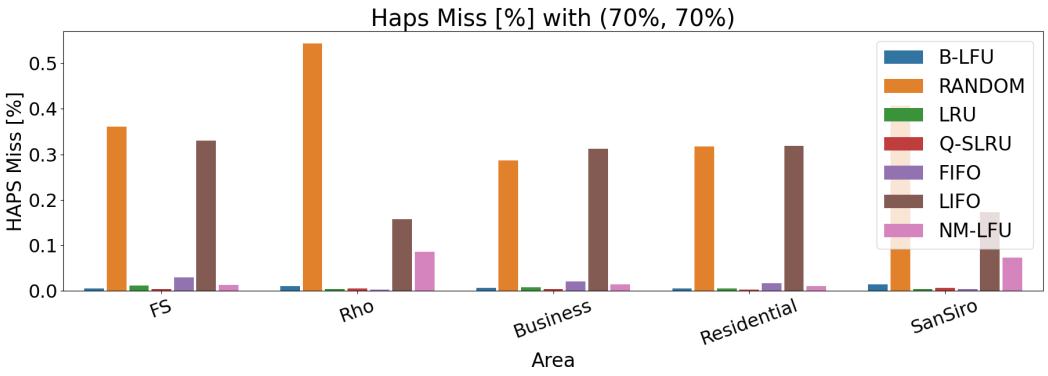


Figure 6.9: HAPS Miss Percentage with (70%,70%) libraries

one related to the fact that the recency-based algorithms end up performing as the simple *FIFO* strategy.
Moreover, comparing Figs. 6.4, 6.5 and 6.6 with Figs. 6.7, 6.8 and 6.9 it is immediately evident how, if on the terrestrial end the performance adjacency was there only when low library values are there, on the non-terrestrial counterpart this coincidence goes on until ***huge*** library are envisioned. This is related to the recency algorithms strategies of eviction and update. As a matter of fact, the coincidence on the terrestrial cache is related to the fact that, when *LRU* is provided with a very short queue, it may end up always placing at its head contents that may never be accessed again. This happens more when the Zipf distribution is almost flat, thus when the α parameter is low, and content requests are evenly distributed. When this happens, this algorithm ends up performing as the *FIFO* putting contents at its head, and then switching them back in the queue until eviction, without replacing them since they are never reaccessed due both to the small library dimension and to the flat Zipf.
However, Figs. 6.8 and 6.9 show that if on the macro, as soon as the cache library increases, the *LRU* and *Q-SLRU* start working properly, thus meaning that they can exploit their updating functioning to insert popular contents to the head of their queues, on the HAPS segment they are much less able to do that, because less popular, evenly requested contents arrive there. Furthermore, it has to be noted that in this situation, the *Q-SLRU* technique, may end up also being the most impaired one, as shown in some cases in Fig. 6.8. This happens because one content may have many requests locally in time, thus ending up being in the head of the last cache queue of this algorithm. Therefore, much time has to pass for the content to be removed, even if it is never requested anymore, thus making the cache not efficient.

6.3.2 Average Total Backhaul Traffic analysis

Finally, to conclude this Chapter we aim at providing a full comparison between the performance of the entire algorithms set the ***Average Total Backhaul Traffic Analysis*** is carried out.
Again a first visualization of the general trends for this KPI is given in Fig. 6.10, showing the *B-LFU* algorithm performance. The obtained result confirms the great improvement that this hierarchical caching strategy is able to provide, with a continuous and significant reduction of the traffic that is not handled by edge objects.

However, this plot is still focused on the behavior of a single algorithm, thus not generalizing. Moreover, without any quantitative comparison, there is not any possibility to highlight the effective trends for this index.
For this reason, the ***total backhaul reduction index*** is computed as in 4.17 for

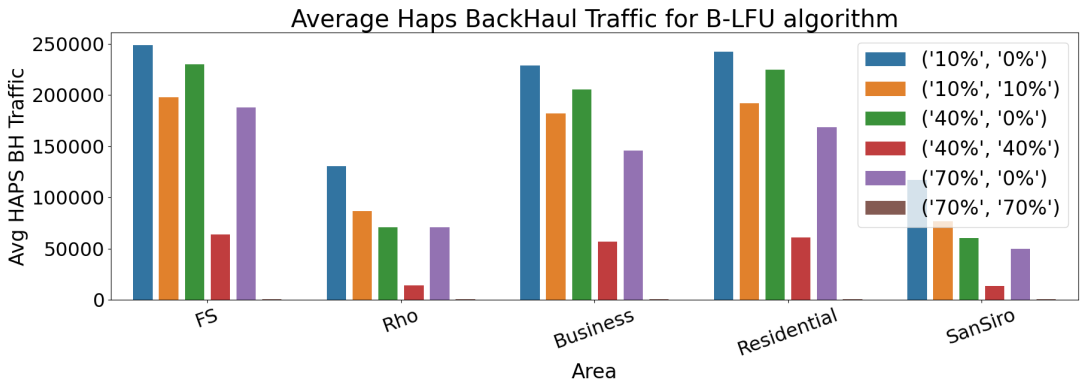


Figure 6.10: Average Total Backhaul Traffic for *B-LFU* algorithm

every cluster under inspection and algorithms developed. This aims at comparing the obtained results when the aerial system is used without caching capabilities with the ones that come out when the same capacity is provided both to the terrestrial and non-terrestrial segments. The reached outputs are plotted in a heatmap obtained using the *seaborn heatmap* python library.

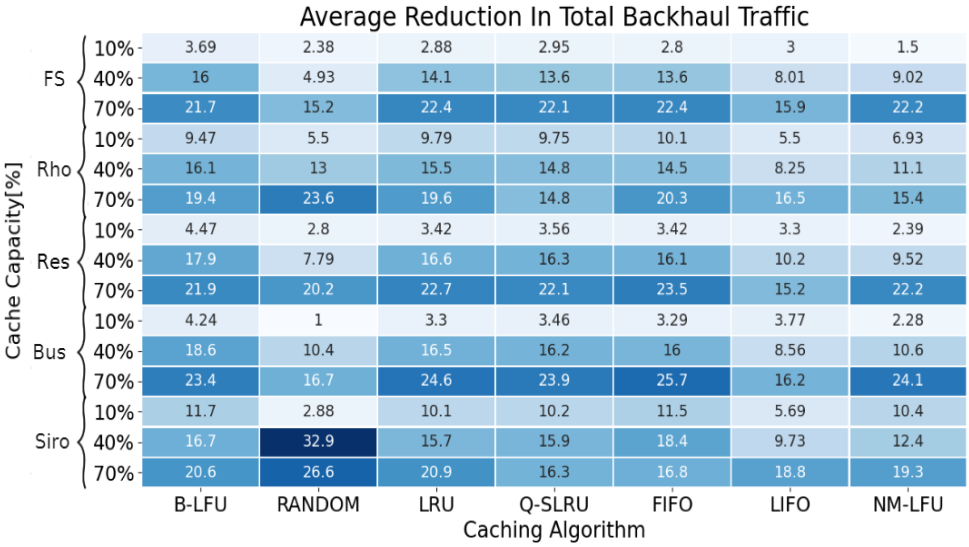


Figure 6.11: Backhaul Reduction Ratio in Every inspected District

This figure confirms the fact that adding a HAPS-SBMS is strongly beneficial

also as regards the overall performance that can be achieved in all the deployed algorithms. The overall backhaul volume, therefore the one that has to be retrieved from the core network experiences a steady decrease with respect to the normal only relay-based scenario.
Furthermore, from a visual inspection of the provided graph, it is immediately possible to see the aleatoric behavior of the RANDOM caching strategy, which is the one that has more oscillations. Indeed, a very lucky draw in *San Siro* district for the ***medium-sized*** cache use case ends up in very good performance, with the parameter computed in this case being more than ten times higher compared to the one of the ***small*** situation.
However, the fluctuation presence in this municipality and the high value obtained considering the *Random* update in the *Rho* traffic zone, are not trivial outcomes. Indeed, those two areas are the ones having $\alpha = 1.06$, thus meaning that here, there very few contents are requested many more times than all the others. It is therefore straightforward to understand that, in these situations, a *"good"* draw may be strongly beneficial, much more than when the α value is low, therefore having a more flattened content requests distribution.
In addition to this, Fig. 6.11 confirms another important, and, after this third confirmation, general trend, which is related to the fact that the backhaul reduction parameter is monotonically and sharply increasing as the cache library raises in volume for the traffic zones having low peaked distributions, while it shows a less evident behavior for the other ones, with some cases where oscillations are detectable.
This tendency is related to the alpha parameter provided in these zones and also to the low traffic volumes inspected there. That considered, it is possible to affirm the fact that aerial cache provisioning has a strong positive impact.
Having provided this general progression in the inspected KPI, and to conclude the discussion about the first scenario a total comparison between the provided algorithms is carried out by computing the quantifier that is able to figure out how a caching strategy performs compared to the other ones, thus inspecting the so-called ***percentual performance decrease***, computed as in 4.18. The obtained values are reported in Table 6.1 and they highlight the fact that the *B-LFU* algorithm is by far the best one, directly followed by the *Q-SLRU* and *LRU* algorithm.

	B-LFU	***Random***	***LRU***	***Q-SLRU***	***FIFO***	***LIFO***	***NM-LFU***
$\overline{W_{\mathbf{BH,\ Alg}}}$	−12.66%	18.83%	−3.58%	−8.73%	−0.06%	2.01%	−1.21%

Table 6.1: Worsening Rate for Every Caching Update Strategy in use

This result confirms the fact that popularity-based cache eviction techniques are much better concerning the first developed *Random*, *FIFO*, *LIFO*. Indeed, as

already specified, the first one is largely worse than the others, while the last mentioned is slightly worse than the average. Furthermore, from the two recency-based strategy values it is possible to see how the segmented strategy, thanks to its less aggressive eviction, ends up being far better.
This result was expected, as the simulation carried out in this thesis work does not envision any popularity variation in time, therefore intrinsically impairing *LRU* performance, whose attributes make it very effective in highly popularity varying scenarios. For the same reason, the result regarding *B-LFU* does not come as a surprise, confirming the goodness of the developed simulator. Again, Table 6.1 also highlights, how memory is fundamental when developing a frequency strategy. The memoryless frequency counterpart of the best strategy is impaired greatly by this weakness. In fact, it may happen that this strategy, being memoryless, will never replace contents that may have very local popularity, therefore being requested many times just for a limited period. Moreover, if it experiences an even distribution of content requests, it will end up always removing the last inserted content, considering it as the less popular. In those situations, due to its behavior in content eviction and update, it ends up performing and operating barely as the LIFO.
However, the analysis of this KPI can be deepened by inspecting its value when changing the library dimensions, thus having further insight into the strategies' general behavior. The obtained results, reported in Table 6.2 highlight that the *Q-SLRU* algorithm is less efficient with respect to the *B-LFU* when the provided library is very low, while its performance increases and barely coincides with its frequency based counterpart when the ***medium-sized library*** situation is inspected.
Finally, in the last use case, the one where ***huge*** libraries are given to the clusters, it slightly overcomes the frequency, timestep-based technique. However, it has to be noted that the difference between the two algorithms is almost negligible being equal to 0.36%.

	B-LFU	***Random***	***LRU***	***Q-SLRU***	***FIFO***	***LIFO***	***NM-LFU***
K = 10%	−14.10%	18.01%	−0.75%	−8.71%	0.88%	0.00%	−1.11%
K = 40%	−16.50%	23.28%	−6.52%	−10.92%	−2.07%	2.72%	0.05%
K = 70%	−5.23%	13.71%	−4.84%	−5.99%	−2.00%	5.02%	−3.83%

Table 6.2: Worsening Rate for Every Library and Caching Update Strategy in use

Again this table also allows us to confirm the fact that recency-based strategies are the ones that most benefit from a content storage increase, with the *LRU* performing very close to *FIFO* in the ***small-size*** case, and being then far better than the last algorithm when higher capacities are given.

Chapter 7

Aerial Priority Scenario, Single Beam Results

This chapter is aimed at comparing the results obtained for the first use case, with the second scenario depicted in Section 4.5.2.
Section 7.1 illustrates the drivers that led us to develop such a new Scenario, initially considering that the ***Average HAPS Backhaul reduction***, and thus the ***HAPS miss probability***. This is made because the values obtained for these indexes in the first situation may not be enough for some applications to be in place. Moreover, this paragraph also aims at understanding why the development of a so-called ***hierarchical-caching with HAPS priority and aerial content check*** case study was chosen, inspecting the reasons for its performance improvement.
Given this first analysis of the aforementioned KPIs, the behavior of this Scenario is deeper inspected trying to figure out at what cost has come the further reduction of the previously mentioned value. Therefore, in Section 7.2, the defections related to this content assignment choice when dealing with the ground caching performance are investigated, evaluating the ***Cluster Miss Probability***.
Finally, Section 7.2.2 does the same, and highlights the influence of the previously cited *hierarchical aerial priority* choice when dealing with the ***Total Average Backhaul Traffic***, as a conclusion of the discussion about this Scenario.

7.1 Hierarchical Caching with HAPS priority, the reasons

When analyzing this new scenario the reasons that led us to develop it must be illustrated.
First of all, if on the one hand the 1^{st} developed situation is able to reduce the

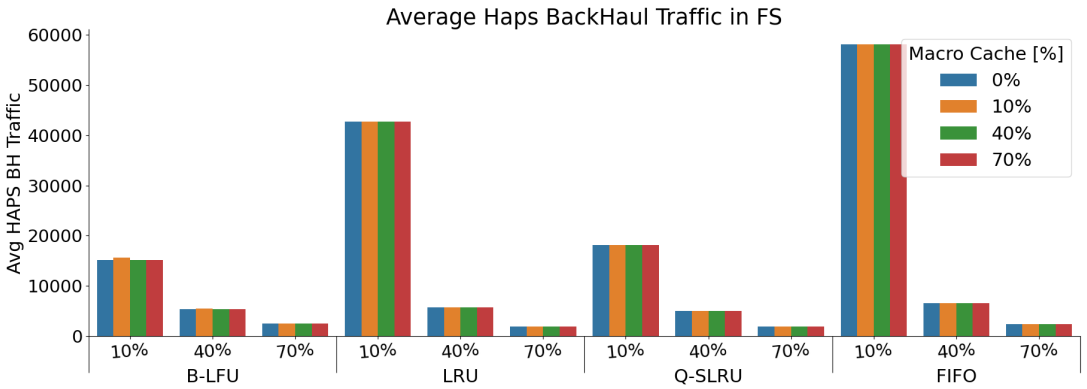
Average HAPS Backhaul traffic, thus giving meaningful results to consider really viable the HAPS-SMBS solution in future NTN integration, on the other hand, and as visible from Fig. 6.2, for the districts that are characterized by $\alpha = 0.56$ and a very high volume of traffic this value is still not negligible at all.
Indeed, also for the algorithm that performs better compared to the other ones, which is the *B-LFU* the aforementioned KPI halves only when at least ***medium-sized*** cache is envisioned in the aerial section, thus having the library couple $[D_M = 10\%, D_H = 40\%]$, while it becomes very low when both the terrestrial and non-terrestrial segments of the VHetNet under investigation, have at least ***medium-sized*** caching capabilities.
This is affirmed not to consider the less performing caching strategies, which also may end up never zeroing this value, even when ***huge*** cache provisioning is foreseen in both the considered segments. Moreover, this trend, even if less dramatic in districts where low traffic is there, and the $\alpha = 1.06$ is always there.
It is therefore for all to see the fact that, the ***hierachical strategy with macro priority and check*** was able to improve the performance of this metric, but was not suited to reduce this value to very low values from the beginning. However, this behavior is expected in many futures, delay-sensitive applications, whose requirements go, as already stated in Chapter 2, far below the one related to the only RTT contribution depicted for HAPS in Section 3.4, that is $RTT = 0.33ms$.
These were the drivers that led us to develop a new strategy to cover those situations, thus aiming at reducing as much as possible the ***Average HAPS Backhaul Traffic*** and thus also decreasing to zero the ***HAPS miss percentage***. Indeed, to confirm what was previously said, Section 6.3.1 depicted how the terrestrial cache worked far better with respect to the one owned by the aerial segment since giving priority and content-check provisioning on the ground station was directly connected to the offloading of very unpopular and evenly distributed content requests to the non-terrestrial system, which was thus left with the hardest task to perform.
For this reason it is straightforward to think that, the best and easiest solution to almost remove the most critical traffic volume, thus the one which is offloaded to the HAPS and that is not handled by this aerial system, is the one of *switching* priorities and content checking capabilities, providing those two features to the non-terrestrial cache. Moreover, when doing it, it is also considered that using the terrestrial end as the *secondary cache* means for the non-terrestrial platform, to have an ally that has no capacity limitations. For this reason, the first aim of the aerial counterpart is to spot the most popular contents, thus filling and reaching its threshold in the wireless link thanks to the offloading of sole contents that he possesses, thus zeroing its misses number. However, this would result in a non-terrestrial cache that is completely static, thus non-generalizing well in a real context, when popularity changes with time. To avoid that, a staticity avoidance

procedure is envisioned, giving the HAPS cache, if the number of hits does not fill totally the wireless link (meaning the cached contents are not so popular on the HAPS, therefore meaning that in real cases that the aerial platform is storing contents whose popularity has decreased with time), the possibility of receiving a number of totally random requests, for the sake of popularity updating.

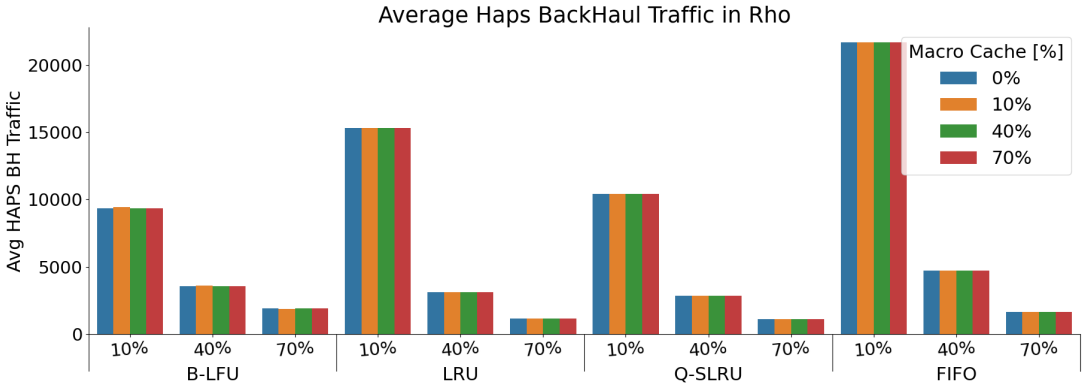
7.1.1 Average HAPS Backhaul Traffic and HAPS Miss Probability

Having illustrated the main drivers that took us to develop this analysis, it is now time to describe the main parameters used for the simulations when considering this Scenario. As a matter of fact, when dealing with the first use case, two important conclusions were drawn. The first one was about the fact that the prime factor that influenced the obtained results was the Zipf distribution parameter, together with the traffic volume, while the second one, regarding the implemented strategies, told us that the *Random*, *LIFO* and *NM-LFU* are always the worse, or, as for the last strategy, never beneficial with its non-memoryless counterpart. These considerations lead us to the decision, for this analysis, to generally consider the same parameters as in Section 6.1, but for the inspected districts, that will be *only* the *FS* and *Rho* ones, considered as enough representative for the entire sample, and the investigated algorithms, that will be the *B-LFU*, *LRU*, *Q-SLRU* and *FIFO*.
Fig. 7.1 depicts the obtained results for the simulations carried out and it is created in the same way as Fig. 6.2 was made, thus inspecting all the possible cache combinations.
However, a structural difference is there indeed, as, if on the one hand in the last mentioned figure every bar referred to a different $D_H[\%]$ and every value on the x-axis regarded the $D_M[\%]$, now the things have been switched, thus having a bar for every macro cache library and the labels referring to the HAPS-related caching capabilities. This choice is carried to make more evident the first result that we want to discuss which is evident from Fig 7.1 and that is related to the fact that HAPS performance is insensitive with respect to the cluster cache dimension. This is something directly deriving from the choice of giving priority to the HAPS platform.
Moreover, comparing the plots obtained for the 1st and this last scenario, it is evident the great difference and reduction that this use case provides as regards the investigated KPI. As an example, comparing the obtained bar for the most performing algorithm, thus the *B-LFU* for the capacity couple $[D_M = 10\%, D_H = 10\%]$ this value is almost **20 times smaller**. However, this trend is always there, for all the possible values under inspection.
Furthermore, it has to be seen that the same tendency is there also for the *Rho*

district, as evident comparing Fig. 7.1b and Fig. 6.2b where this value reduces **more than tenfold**.
In addition to that, it is also possible to affirm that, as expected, the increase in the HAPS caching library is always beneficial and helps reach values that approach the zero for the inspected KPI. However, it has to be said that, given the fact that a content request is considered, in our simulations to have a size of $100Mbits$, the results obtained, even for the worse case of the *B-LFU* algorithm highlight that, in an hour, with ***small cache-size*** on the HAPS, ≤ 15 requests end up experiencing a miss, in the *FS* district, and ≤ 9 for the *Rho* traffic zone. Those values and achievements were unbelievable in the previous Scenario, in some cases, even when ***huge-sized*** libraries were provided.



(a) FS Haps Averaged BH Traffic [Mbits] behavior



(b) Rho Haps Averaged BH Traffic [Mbits] behavior

Figure 7.1: HAPS Averaged BH Traffic [Mbits] for FS and Rho and all the algorithms

Briefly, this analysis has confirmed the goodness of the performed *switching*

operation, and the fact that, giving priority and content-check to the aerial station has worked as we expected, thus making this Scenario highly suitable for ***delay-sensitive*** applications.

7.1.2 HAPS Miss Probability

Having inspected the first important KPI for this section, and having received great results for it, it is now time to investigate another value that is strongly related to the previous one, the ***HAPS miss probability***. This index will enable us to confirm what was previously stated for the *haps backhaul traffic*. When considering this particular parameter, it has to be remarked that when no HAPS is provided the miss percentage is equal to 100%, thus being not reported in the following obtained pictures. Moreover, as regards this second use case, the miss percentage value does not change when the cluster is provided with different caching capacities, since, being the priority given to the non-terrestrial segment, its performance is not influenced by the ground station functioning.
That said, the aforementioned KPI is inspected by using the *sns heatmap* library in python, considering every single capacity combination between $D_M\%$ and $D_H\%$ for both clusters under investigation.
In particular, Fig. 7.2 reports the obtained results for the *FS* district for the two compared scenarios. The performance difference between the two is evident. The first thing that is noticeable is the discrepancy in the percentage range between the two use cases, especially when dealing with the first column of both the heatmaps, thus the one where ***small haps cache*** is used. In this context, the 1st Scenario shows values between $[59\%, 83\%]$, while for the *aerial priority* one, the range drastically decreases with values in the interval $[6\%, 25\%]$. This means that the lowest value is almost **15 times smaller** compared to the one obtained for the first use case, while the maximum one is one-fourth with respect to the *terrestrial priority* situation.

Moreover, from a deeper inspection of the provided graph, it is also possible to visualize that the maximum performance disparity between the two contexts is reached when dealing with the *Q-SLRU* algorithm when referring to the $[10\%, 10\%]$ library couple, whose respective indexes for the two use cases are equal to $H_{miss,\ \%,\ 1} = 83\%$ and $H_{miss,\ \%,\ 2} = 8\%$. Given this simple analysis, it is possible to confirm that the *aerial priority* scenario is with no doubt the one of election when delay-sensitive applications are there.
In addition to that, when looking at Fig. 7.2 it is also possible to confirm the behavior already inspected in Fig. 7.1, meaning that the ***HAPS miss probability*** does not depend from the cluster capacity.
Furthermore, the general affirmations carried out in Section 6.3.2, when performing an algorithm-wise comparison for the 1st context, continue to be valid. Indeed,

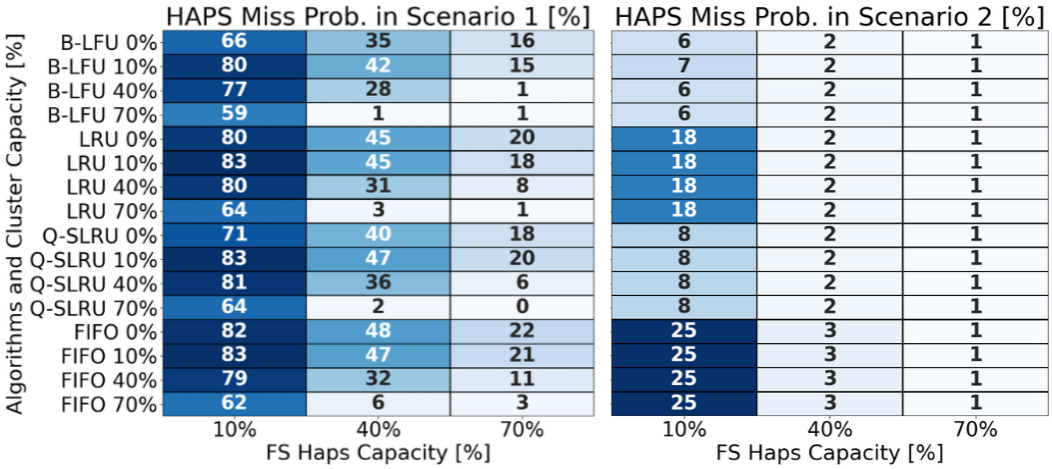


Figure 7.2: FS Miss Probability Heatmap for Scenario 1 and Scenario 2

also in this last-developed use case, it is possible to see that *B-LFU* is the most performing caching update, directly followed by the *Q-SLRU*, while *LRU* and *FIFO* are less effective. In particular, when dealing with very small library sizes, their indexes tend to be similar, as already pointed out in the aforementioned section.
Finally, another insight about the 1st developed context is also inspectable, as it is possible to see that, for that use case, the Cluster cache provisioning was causing a defection in HAPS cache performance, even if decreasing the overall backhaul traffic. This is due to the fact that, in that situation, when no ground capacity was there, the HAPS received 2700 randomly sent requests from the initial sample while, then, with the provision of ground caching capacities, the contents arriving at the aerial station are the results waste of the terrestrial system, therefore the least popular ones.
Having carried out such a complete analysis for the *FS* district, and having obtained these encouraging results it is now time to inspect the same KPI for the *Rho* district whose results are shown in Fig. 7.3. From a visual inspection of the aforementioned plot, it is immediately possible to affirm that all general trends already investigated in Fig. 7.2 are still there, with a sharp decrease in the computed index everywhere.
Moreover, it is also possible to see that those tendencies are even more evident since the highest value detectable in the *Rho* district for the *aerial priority scenario* for the miss percentage is equal to 7%. Similar values for the same algorithm in the other scenario are almost reached only when ***huge*** cache is envisioned, thus

confirming the fact that using this scenario compared to the other one is like exponentially multiplying the reduction possibilities when dealing with the aerial platform performance.

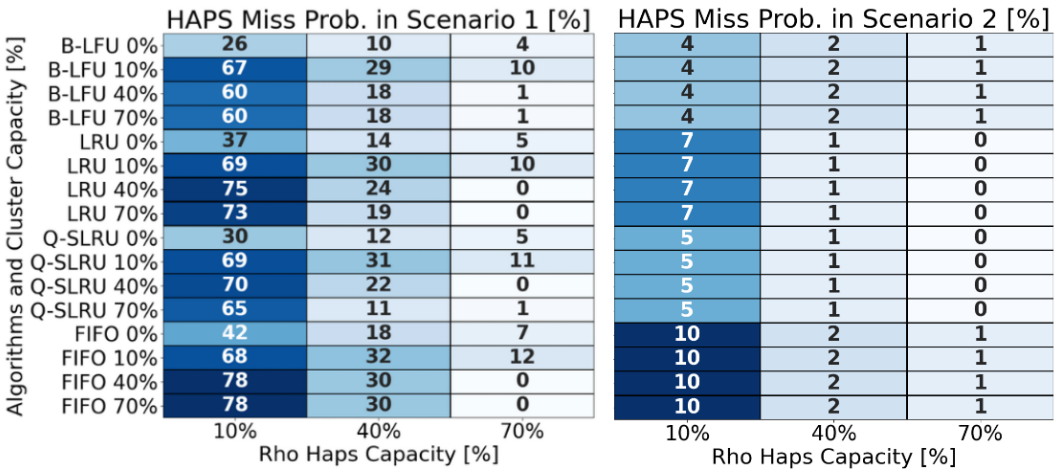


Figure 7.3: Rho Miss Probability Heatmap for Scenario 1 and Scenario 2

Again, continuing with the description of what is shown in the given graph, it is possible to state that the same trend seen in the previous figure, i.e. the one concerning the worsening of haps performance after inserting the terrestrial cache, is even more marked here. This, like the fact that all the described behaviors are much more evident in this simulation than for the one in the previous cluster, is due to the high α parameter of the Zipf distribution, which in this case, being so high, causes the difference in priority between land-based and air-based scenarios to be even more decisive.

7.2 Terrestrial Segment, the impaired part

Having deeply understood the reasons that led to a performance improvement for the non-terrestrial segment, we want to inspect the so-called ***cluster miss probability***.
This choice is carried out to compare the performance of this newly developed use case with the previous one, understanding if and in case to which extent this situation, giving priority to the aerial end, ends up impairing the performance of the terrestrial segment.
Therefore, using the parameters specified in Section 6.1 for the *FS* and *Rho* districts, again considered as figureheads of their category, a complete simulation

campaign is carried out.

7.2.1 Cluster Miss Probability

The results for this KPI and the *FS* cluster are provided in Fig. 7.4. Here, the first thing that is possible to inspect is that, looking at the graph that refers to the *terrestrial priority* context, the ground station miss probability is not influenced by the HAPS cache dimensions. This reflects what happened in the previous plot for the aerial scenario when the HAPS performance did not change when shifting the macro library towards higher values.
This result is directly related to the content offloading strategy adopted in the first situation, which, regardless of its capacity, always provides the HAPS with random samples of unpopular content. Therefore, the terrestrial cache miss percentage is not improved, nor impaired by this operation.
The same does not happen in this new use case, where the situation is the opposite, meaning that, if the algorithm developed on the aerial system works well, the cluster is provided with unpopular and evenly distributed requests, therefore experiencing a higher number of *misses*.
Furthermore, following the same comparative reasoning carried out to perform these affirmations, another consideration can be drawn. Indeed, in this *aerial priority* scenario, it is the HAPS cache provisioning that impairs the Cluster performance, as confirmed from Fig. 7.4, showing the same, but opposite, behavior already inspected in Figs. 7.3 and 7.2.

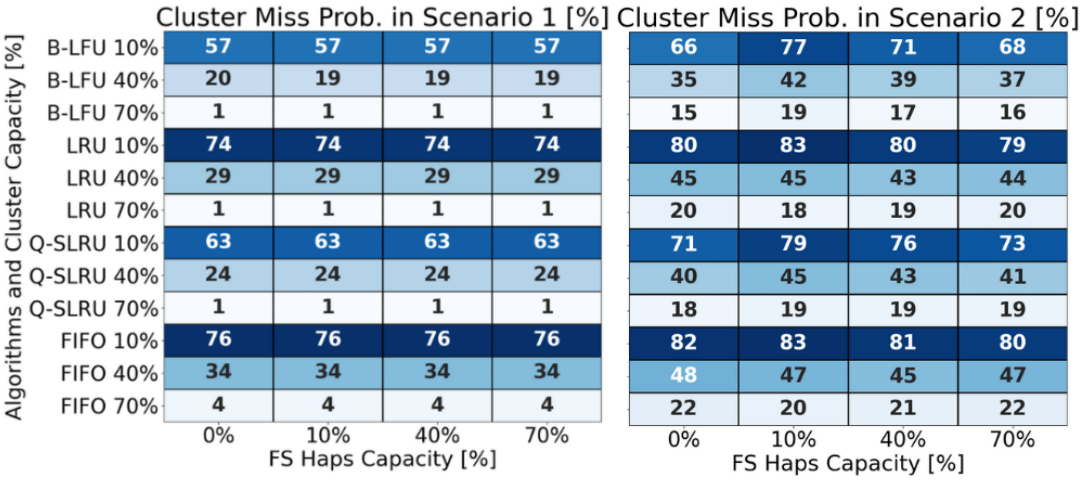


Figure 7.4: FS Cluster Miss Probability Heatmap for Scenario 1 and Scenario 2

Again, considering the maxima and minima for both the first and the second context under inspection, the performance decrease is evident, being the range for the *terrestrial-priority* situation in [1%,76%] and the one for this last one in [15%, 83%] with the maximum difference which is inspectable for the *Q-SLRU* algorithm, when [40%, 10%] capacity couple is evaluated, with a ***cluster miss probability*** which is almost doubled. In fact, this index is equal to the 24% for the first use case and 45% for the second situation.
Furthermore, to see the progression of this decrease in performance through the capacities provided to the ground station, the ***percentual performance decrease*** is computed as in 4.18, considering the inspected index for the two different scenarios and every single macro library envisioned in the simulation parameters.
Table 7.1 shows the obtained results for the aforementioned analysis and highlights the fact that, the ground station capability and the performance decrease in the ***Cluster miss probability*** are positively correlated, meaning that as the ground cache increases, the difference in the terrestrial segment performance between the two proposed scenarios further increases.
This is due to the fact that, if on the one hand the Cluster library improvement is fully exploited on very popular contents in this last *non-terrestrial priority based* use case, this incremental capability is not used as efficiently as in the past, being employed to detect and *hit* less popular and evenly distributed requests.

	B-LFU	***LRU***	***Q-SLRU***	***FIFO***
K = 10%	19.43%	8.64%	15.72%	6.75%
K = 40%	49.02%	34.46%	42.51%	26.88%
K = 70%	94.03%	95.59%	94.44%	80.25%

Table 7.1: Percentual Performance Decrease in Cluster Miss Probability for FS

Moreover, it is not trivial that *B-LFU*, which is the algorithm that always performed better, is the most impaired in this use case. This is due to the fact that being the content's popularity fixed in time and being this algorithm frequency-based, in this simulation, it is always the best performing one, thus the best in spotting the most popular content. However, if on the one hand, it is surely beneficial for the aerial platform, in this scenario this has, as a direct consequence, the fact that those very popular requests, thus the easiest to *hit* are not given to the terrestrial BS, therefore strongly decreasing its performance much more than the other algorithms.
This is confirmed also from the fact that *FIFO* which was the one that performs worse in *FS* district, is also the less impaired one, as it has many difficulties in understanding the most popular contents and it is directly followed, only when small libraries are there, by the *LRU*. This result was however something expected

after the discussion carried out in Section 6.3.2.
However, the real disadvantage of using this strategy is shown in Fig. 7.5, where the percentual impairment is total almost in every case. This is because the plotted heatmap regards the *Rho* traffic zone, which, having a much more peaked popularity distribution and even smaller traffic, makes the haps HAPS surely able to spot the most popular contents, when the priority is given to the aerial platform. Therefore, it leaves the cluster with low and equally requested traffic, as for the *FS* district, while in the first scenario it was loaded with a high number of very popular requests, thus having the best situation to perform as less misses as possible. Those are the reasons that have led to the great difference when dealing with this KPI in *Rho* neighborhood. Here the maximum impairments are there

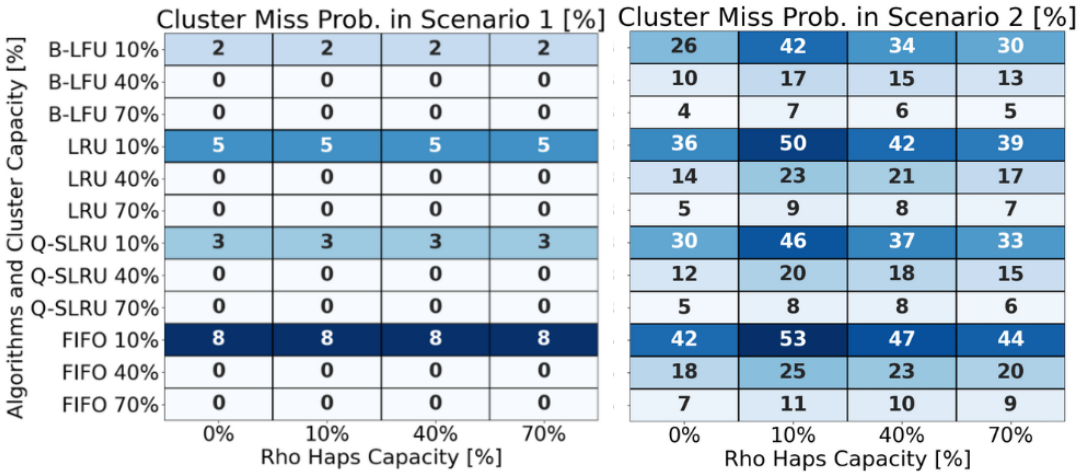


Figure 7.5: Rho Cluster Miss Probability Heatmap for Scenario 1 and Scenario 2

when considering the [10%,10%] capacity for every algorithm, but are generally present everywhere.
Finally, having carried out such a complete analysis it is possible to affirm with certainty that, the cost of having an almost perfect aerial cache for this 2nd Scenario is the one of a massive performance decrease in its terrestrial end.

7.2.2 Comparison on Total BH Traffic

To conclude this analysis, the results of this new content assignment strategy for the **Average Total BH Traffic** are compared with the ones carried out before. This step is carried out to highlight the performance of this Scenario when dealing

with the volume of requests that should normally be requested to the core network. However, even without having depicted any graph, insight can be obtained from deeply analyzing the results obtained in the previous paragraphs. Indeed, this new strategy is able to remove, almost completely the traffic that experiences the highest delay, thus the one which, being offloaded, is not processed at the edge by the HAPS, even when it is provided with onboard MEC. However, as previously highlighted, this reduction comes at the side cost of an increase in the *cluster miss probability*. Then, we have to remember that the aerial platform has an intrinsic wireless-related limit regarding the number of requests that it can handle hourly. These 3 affirmations combined together lead us to the conclusion that this strategy will end up zeroing the traffic experiencing the highest delay but much increasing the one that has to be retrieved from the core network after a ground miss event, thus loading further the core links of the network, and increasing the ***Average Total Backhaul Traffic***.
In this case this KPI will be mostly composed of the Cluster Backhaul traffic, resulting from the inefficiencies on the terrestrial segment.
The graph chosen to confirm or reject this hypothesis is a barplot obtained for every cluster and capacity under investigation. The obtained results are shown in Figs. 7.6, and 7.7 and only partially confirm the previous affirmations.

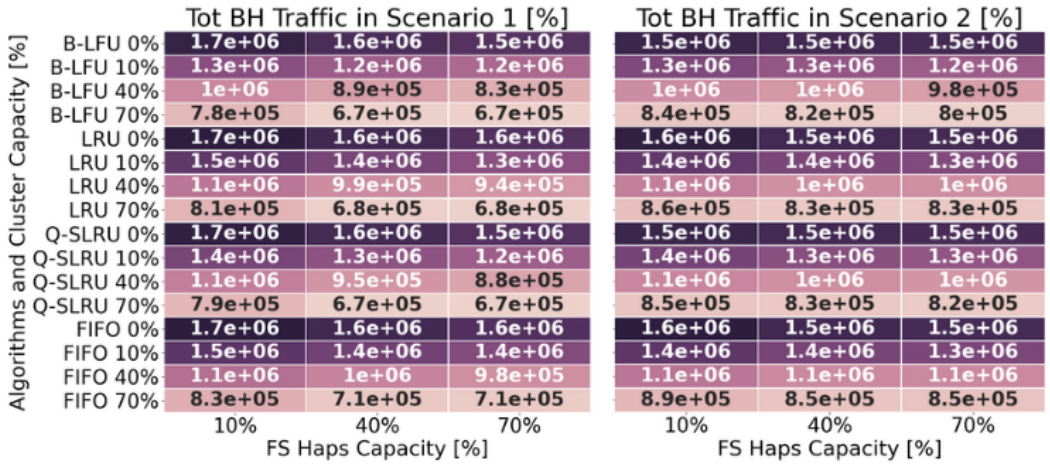


Figure 7.6: FS Avg Total BH Traffic Heatmap for Scenario 1 and Scenario 2

If on the one hand, generally, the overall backhaul traffic is less in the *ground station priority* use case, there is still a situation in which the opposite content assignment and cache-checking strategy perform better than the first mentioned one. This case is the one in which no cluster cache is envisioned, thus making

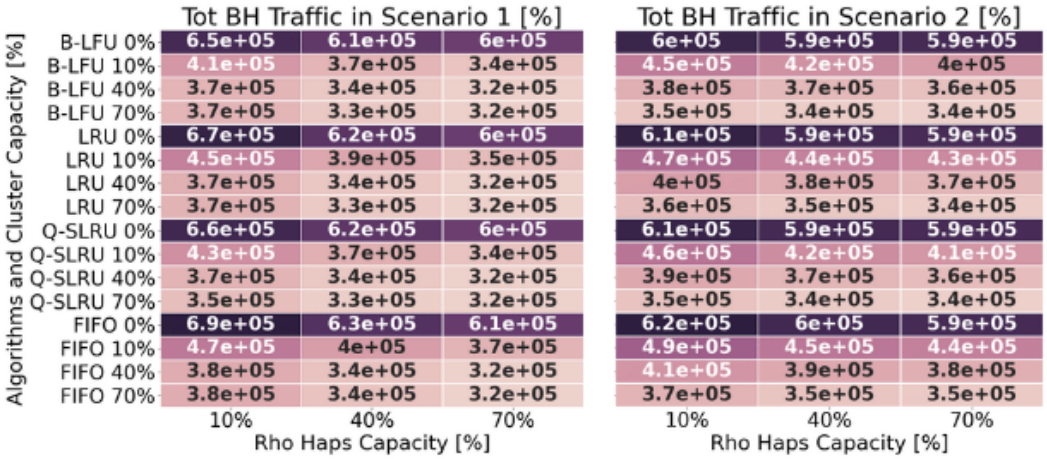


Figure 7.7: Rho Avg Total BH Traffic Heatmap for Scenario 1 and Scenario 2

this scenario also more useful where no ground library is there or even in cases in which the terrestrial station is not present or out of work for some reason.
Furthermore, a big distinction has to be carried out for the depicted figures. Indeed in Fig. 7.6, referring to the *FS* neighborhood the real increase in the performance for the 1st Scenario is there **only** when the cache dimensions start to grow a lot. This happens because the HAPS is somewhat always limited in its ability to spot correctly the contents due to the wireless limitation, thus the content-check capability is not exploited at its maximum for the second content assignment strategy. However, it is important to see that, the first real differences that cannot be neglected between the two cases in the inspected KPI are there when the ground station cache has at least a ***medium-sized*** cache provisioning, thus meaning that, in the other cases, when cache libraries lower than a medium amount are there, the ***hierarchy with HAPS priority*** is beneficial from the point of view of the most appalling traffic volume (the one in backhaul from the HAPS) and also does not impair the ***average total backhaul traffic***, thus making this strategy the one of choice.
However, when libraries start to grow, the difference is not negligible anymore, making the 1st Scenario highly preferable when there is the aim to reduce the traffic that goes to the core network.
Finally, the same cannot be said for the *Rho* related plot, meaning Fig. 7.7. There, having a higher α parameter, as soon as more caching capability is given to the terrestrial segment, the *terrestrial priority* scenario performs always better than the last developed one. This result is very important and highlights a great behavior difference due to the popularity distribution of the investigated clusters.

Indeed, in neighborhoods that show very peaked distribution, the second content assignment strategy, thus the one that gives priority to the non-terrestrial platform, is still able to zero the *haps BH traffic* having however **always** the cost of a higher load towards the core network.

Chapter 8

Cooperative Caching Scenario, A Solution?

The last Scenario that is going to be analyzed is a sort of summary of all the other ones, trying to understand what can be a sort of *best choice* in general situations when good performance overall is researched. To do that, this use case foresees the use of a Cluster priority on requests assignment with content check, aiming at reducing the ***Total BH Traffic*** as much as possible. However, in contrast with what happened in the $1^{st}Scenario$, here the same check is performed also on the HAPS content library before any requested assignment, thus trying to decrease the value of the most critical parameter, the ***Avg HAPS BH Traffic***. Moreover, to generalize its use, the same staticity avoidance procedure carried out in Sec. 4.5.2 is given here.
Finally, this algorithm is expected to give similar performance regarding the Cluster Cache with respect to the basic use case, but have better handling of HAPS Cache, thus well generalizing for many real use cases, without having the drawback of being completely static.
The analysis of this Scenario goes on as follows. Section 8.1 starts discussing the ***cluster*** and ***haps miss probabilities***. This choice is made as, having now understood how the simulator works generally, there is no need to inspect, as the first index, a general KPI as it could be the total backhaul traffic, but it is possible to directly face the behavior of this strategy segment by segment.
Section 8.2 deals instead with the ***Average HAPS BH Traffic*** and ***Average Total Backhaul Traffic*** thus deducing the *cooperative* overall performance and being able to conclude the discussion about the single beam use cases.

8.1 Comparison with other Scenarios, Miss Probabilities

The first part of this analysis focuses on the Miss Probabilities. In particular, here, many references will be done considering Secs. 4.5.1, and 4.5.2 to carry out a complete performance comparison, thus drawing meaningful conclusions.

8.1.1 Cluster Miss Probability

In this first part of the analysis, both the other scenarios have to be considered, as this new use case is given to try to replicate the same behavior of the first one, increasing the performance on the haps as the second does, without impairing much the ones on the ground station as happened in the previous content assignment strategy 4.5.2. For this reason, Fig. 7.4 has to be borne in mind when proceeding in analyzing the *cooperative caching strategy*.
The first thing that has to be highlighted is how, this cooperative scenario, performs as the one where the aerial station only in one case, which is the one of *no-haps* provisioning. This result is just a consequence of how the content assignment is performed, since these last two assignment strategies, focusing on the HAPS performance when it has some caching strategies on it, end up not exploiting totally its relay capabilities when there is no cache provisioning on it, thus impairing the ground cache performance, since it can offload fewer times the unpopular contents that it is not able to handle, indeed for both of them it is able to perform this operation moving only the traffic coming from one micro BS. However, this behavior is not the focus of our work, as now we are concerned with the behavior of all of these scenarios when both segments of our VHetNet are given caching features.
So, focusing on the results for the *FS* district, therefore considering Figs. 8.1 and 7.4, it is possible to compare this new strategy with the *terrestrial priority* one. From a visual inspection of the aforementioned graphs, it is possible to see that when the cluster library is ***small-sized***, meaning $D_M = 10\%$ there is almost no performance impairment between the two use cases under investigation. The maximum percentage distance between the two is equal to 5% and this value reduces to 0% when the HAPS Cache dimension grows further. This is because, the more the haps content library grows in capacity, the more the two scenarios tend to coincide. When the HAPS is given a content library equal to 100% the two work exactly the same, with the contents offloaded to the aerial system soon after the ones that are already present in the BS cache without basically developing any check.
However, when the ground station library starts to grow the decrease in performance between the two becomes evident. When $D_M = 40\%$, and the $D_H < 70\%$,

the difference between the corresponding values ranges on a value equal to 10%, while further inspecting the ***huge*** terrestrial cache provisioning this impairment is even greater. It has to be noted that, for the aforementioned reason, when $D_H = 70\%$ or greater, this *cooperative* strategy tends to converge to the *aerial-based* one.
In addition to that, looking at the *aerial priority* scenario, it is possible to say that, the ***cooperative caching strategy*** is *always* able to outperform its *hierarchical non-terrestrial* counterpart as regards this index. This is due to the assignment priority and content-check capability, that, being given to the terrestrial station, end up increasing the times the ground station is able to perform a hit.
Furthermore, it has to be noted that this tendency decreases in one particular use case, that is the one with $D_M = 70\%$ and $D_H = 10\%$. This phenomenon is due to two main reasons. The first is that in this situation, the macro BS does a so-called ***wise*** content offloading, thus meaning that the first motivation to give the aerial system a request is related to its presence in the non-terrestrial cache, and the second logic why the ground station offloads a content is the one related to let the haps avoid staticity when real cases, with time-varying popularity, are there.
Therefore, thanks to this content assignment technique this cooperative scenario never performs a real and massive offloading operation as happened in the 1^{st} discussed *hierarchical* situation, unless the HAPS has very high libraries inside.
Thus, when low D_H and high D_M are there the ***cluster miss probability*** is impaired, due to the lack of the possibility to offload many contents to the aerial station, while in all the other cases, the ground content check feature, makes this ***cooperative*** use case very effective.

The same trend is also visible when dealing with the *Rho* district. Here, as inspected throughout the entire previous chapter, these behaviors are even more marked, due to the high α parameter and the low traffic value. Indeed, when looking at the *Rho*-related heatmap, and considering low HAPS libraries to be provided, the inability to offload all the *unpopular contents* ends up impairing performance massively. This happens in such an evident way in this traffic zone because in the 1^{st} scenario, due to the low traffic volume, the $\alpha = 1.06$ and the possibility to perform an unconscious offloading campaign, the cluster ends up never experiencing a *ground miss* already from very small capacities, while the same is not happening here.
Finally, when the *HAPS library* starts to grow, thus meaning that the quantity of offloaded contents also increases, the two strategies barely coincide, as already depicted for the *FS* district.

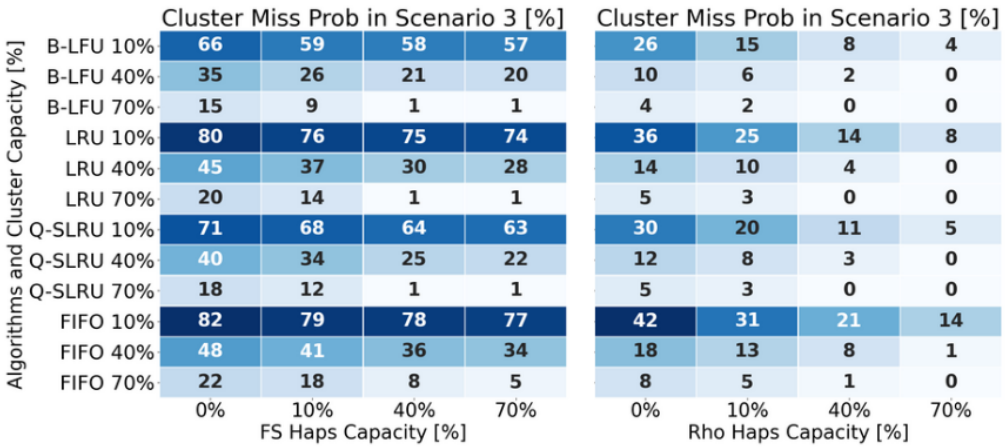


Figure 8.1: FS and Rho Cluster Miss Probability Scenario 3

8.1.2 HAPS Miss Probability

The HAPS miss probability is now considered.
First of all, looking at Fig. 8.2, it is immediately possible to see a strong performance increase in the third scenario with respect to the first developed one. This is due to the double-check provided during the content assignment, which is fundamental. However, this also happens thanks to the ***wise offloading*** procedure, which prevents the HAPS from being inopportunely loaded with too many random contents, that are now divided between the two segments according to their capacities, which gives to the simulation results much more balance.

In addition to that, this haps miss probability *reduction* is also detectable, and even higher when no caching capabilities are provided at the terrestrial end. This is because, in this extreme situation, the ***cooperative caching*** converges towards the ***hierarchical one with aerial priority***.
The same cannot be said when low haps capacities are given and $D_M \neq 0$. This phenomenon is there because, this strategy is cooperative, but still gives priority to the ground cache, thus meaning that the *most popular contents* go there, and when the aerial library is very low, the number of contents that are offloaded to the aerial station due to a ***forecasted hit*** event are not a lot, ending up being even comparable with the number of requests offloaded to perform the staticity avoidance step.
However, as confirmed by both plots in Fig. **??**, this trend vanishes as the haps gain caching capacity.

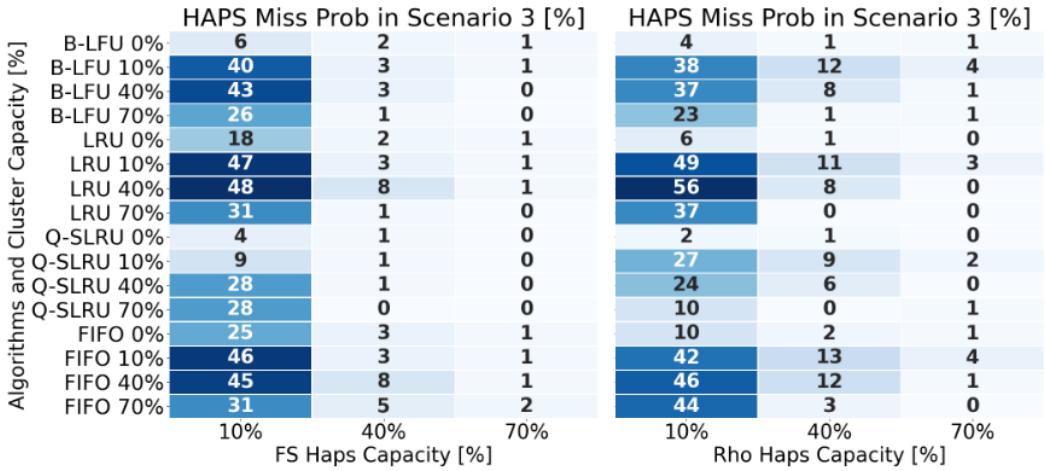


Figure 8.2: FS and Rho HAPS Miss Probability Scenario 3

8.2 Comparison on Total and HAPS BH Traffic

8.2.1 Total and HAPS BH Traffic

The **Total BH Traffic** and ***HAPS BH Traffic*** are now under inspection.
The first index that we want to analyze is the one referring to the total number of requests that have to be retrieved from the total network, regardless of the segment from which they come. This choice is made because the main aim of this Scenario was to improve the *HAPS* performance without showing an increase, as it happened in the aerial priority scenario, in this traffic value, thus sticking with volumes next to the ones seen for the ***hierarchical with ground priority*** use case.
Fig. 8.3 shows that this value is initially the same as in the reference situation, meaning when the Cluster cache has $D_M = 10\%$, the great increase in performance for the HAPS is counterbalanced by the defection detected on the Cluster side.

However, if on the one hand, while dealing with the *FS* district a general performance coincidence is there when ***small-sized*** ground cache provisioning is foreseen, in the *Rho* neighborhood, this ***cooperative*** strategy, performing as in the first developed content assignment technique, is already outperforming the aerial-based one.
In addition to this, by a deeper inspection of the provided graph, it is possible to note that, for all the other $[D_M, D_H]$ couples, this scenario is even outperforming the base case.

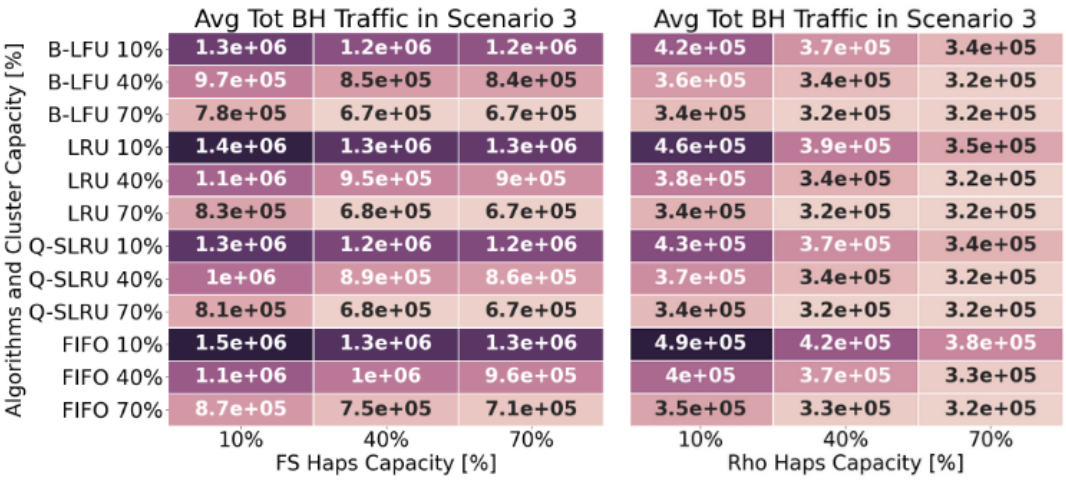


Figure 8.3: FS and Rho HAPS Miss Probability Scenario 3

This result is strongly encouraging and makes this use case very important because, when compared to the *best strategy* developed so far when dealing with this KPI, it is able not only to experience the same, or even reduce the traffic volume that has to be sent to the core network but also to do it reducing the amount of ***Average HAPS Backhaul volume*** as testified from Fig. 8.4 when compared to Fig. 6.2 with a general increase that it almost of an order of magnitude.

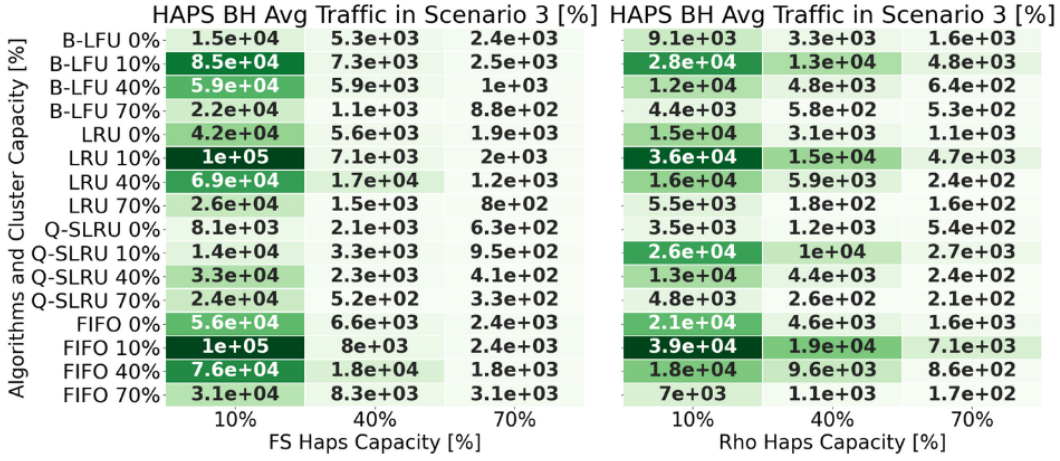


Figure 8.4: FS and Rho HAPS Miss Probability Scenario 3

Moreover, as expected, it has to be noted that, as the HAPS capacity increases,

these values tend to converge to the ones seen in Fig. 7.1, depicted for the second used scenario, thus confirming the behavior already seen and inspected in the previous section for the ***HAPS Miss Probability***.

Chapter 9

Multi-Beam Results, a General Overview

Until now this thesis work examined two main topics, namely the search for the best possible cache update strategy, and the development of various scenarios that can cater to the most diverse needs.
When doing that, we have discovered that, in any case, the most effective caching technique is the *B-LFU* one, directly followed, as regards its performance, by the *Q-SLRU* recency strategy. However, we have deeply inspected that the first mentioned library update was able, almost everywhere, to outperform its best rival. Again, we discovered that this trend is due to the fact that our simulations do not envision a variance over time in the popularity of the contents, nor their change, through the insertion of a birth and death process, thus intrinsically benefiting the frequency-based algorithm, when comparing it to the other ones.
Said that we have focused on the development of 3 main content assignment approaches. The first foreseen methodologies were envisioning ***hierarchical caching*** contexts. Between them, one was provided with *terrestial-priority* while functioning, and for the other one, the *non-terrestrial segment* was privileged.
After that, we proceeded with a deep analysis of the aforementioned techniques, therefore investigating the effectiveness difference for all the KPIs under study. Finally, we have also developed a third, ***cooperative-based*** use case, that differently from the previous ones, was granting content-check capabilities to both the deployed cache objects.
In the end, we were able to conclude that every situation could either minimize or optimize a different value between the inspected ones, thus making each of the studied use cases suitable for different applications.
However, the analysis carried out so far fails to consider that the aerial platform provides coverage only for one Milan district at a time. Indeed, in Chapter 4 we

have said that the non-terrestrial segment can activate more than just one single beam, thanks to its mMIMO provisioning. Therefore, to complete this work, the ***Multi-Beam*** simulations are carried out, considering the aerial system to provide coverage to up to 5 neighborhoods in the city.
Section 9.1 shows the parameters considered for these new simulations and the reasons that led us to choose them.
Section 9.2.1 illustrates the results obtained when more beams are activated in the first scenario highlighting how the content assignment strategy and the cluster popularity distributions directly influence also these results.
Section9.2.2 aims at comparing the obtained outcomes for the first use case with the ones that came out when the *aerial* platform is the *privileged* one.
Section 9.2.3 shows the results carried out for the last provided scenario, focusing on how the ***cooperative*** choice changes the paradigm compared to the ***hierarchical ones***.

9.1 Simulation Parameters

The parameters used to carry out our simulations are the following:

- Train Station, Rho, San Siro, Business and Residential Traffic Zones in Milan are considered, and their Zipf Distribution Parameter is assumed to be $\alpha_{FS,Bus,Res} = 0.56$ and $\alpha_{Rho,Siro} = 1.06$.

- The KPI that will be inspected specifically is the ***Averaged HAPS Back-Haul Traffic***. This choice is made considering what has been said in the introduction to this chapter. Indeed, it is possible to understand that the activation of several beams, in addition to involving more traffic to the aerial platform, means above all that the non-terrestrial cache object is shared between the different Milan traffic zones, being no more dedicated to each district. Therefore, an investigation of the aforementioned index in a complete and specific way becomes of fundamental importance. Moreover, the ***Average HAPS Backhaul Traffic*** is evaluated while varying the libraries with which the aerial cache is provided.

- For what has been said in the previous point, the sampling strategy adopted to have different specimens of the HAPS library is more punctual compared to what was done in Section 6.1, thus reducing the sampling interval to 10%, with values in $D_H \in [0\% - 100\%]$.

- The Macro Cache Storable Library Percentage is in $D_M \in [0\% - 100\%]$, with sampled values that are $D_M = 0\%$, $D_M = 10\%$, and $D_M = 40\%$. The value

$D_M = 70\%$ will be used only in some cases when it makes some results and trends more evident, but in general, it will be avoided since, as already shown in the previous chapter, to have such a high storage capacity is rare.

- The algorithm under study is the *B-LFU*. This choice is because, as previously said, the intrinsic advantage of this strategy with respect to the other ones is related to one of the simulation assumptions, that is the lack of popularity dynamicity, and its removal goes, for now, beyond the thesis scope, thus leading us to focus on the developed content assignment strategies performance more than on the given caching update techniques.

9.2 Content-Assignment Strategies and Multi-Beam Simulations

Having depicted all the simulation parameters in use, and the reasons that led us to perform some choices is now time to enter into the result analysis phase. This step will be carried out considering one scenario at a time, starting from the one that gives priority to the ground station, and ending with the cooperative technique strategy in place.
For every situation, the main differences between the single and the double-beam contexts will be highlighted and discussed, explaining the reasons for some behaviors to be there. In addition, for each case, the obtained outcomes will be correlated to the considered strategy functioning methods, but also to the attributes of the areas under study, relating their performance to the experienced traffic volume and the Zipf parameter in use.
After that, an analysis with an increasing number of clusters will be carried out. Finally, concluded our study, conclusions will be drawn for each of the developed caching contexts.

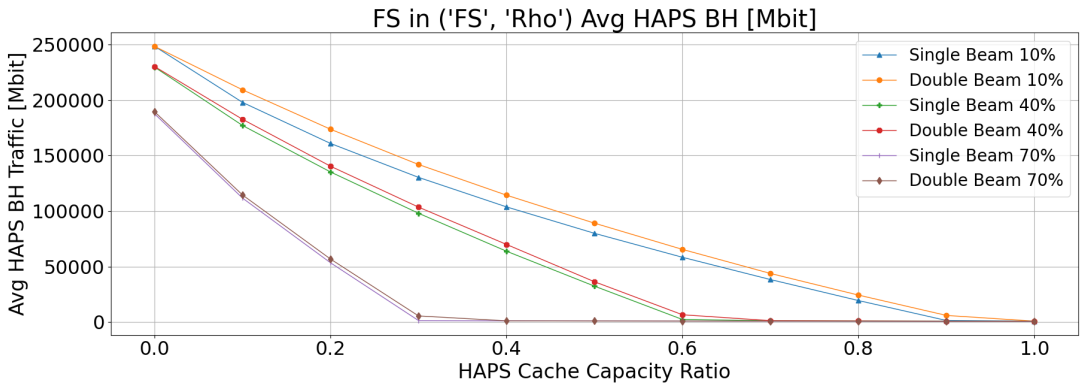
9.2.1 Terrestrial Priority Multi-Beam

In this use case, we start inspecting the variations in the ***Average Backhaul HAPS Traffic*** dealing with single and double-beam provisioning. When executing this step, we have to remember that the provided clusters have different α parameters and traffic volumes, thus strongly changing the quantity and type of traffic that will arrive at the aerial platform, after being assigned, for priority reasons to the ground station.
In particular, due to that, the offloaded requests will be more evenly distributed when compared to the ones left at the terrestrial end, and the number of them varies according to the overall demand in the given neighborhood, increasing when the number of requests is very high and having an opposite behavior when it is low.

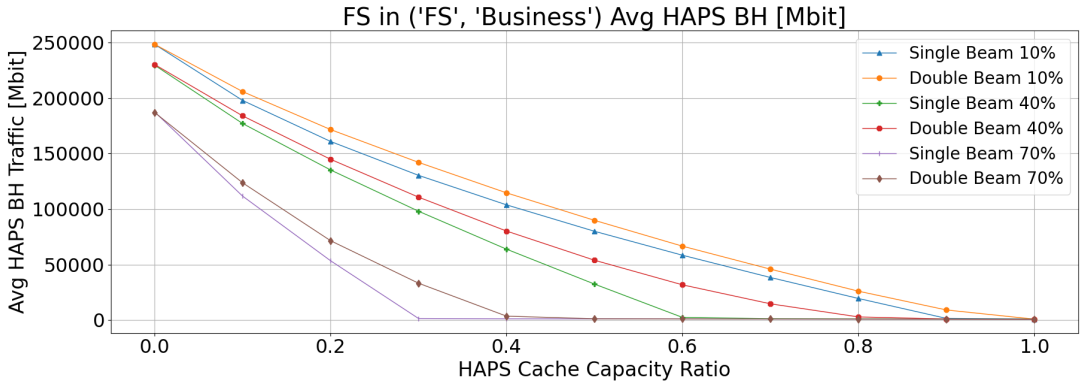
This intrinsically means that the HAPS will perform better for the most loaded clusters since it will receive much more requests from the districts that experience high traffic volumes. Therefore the aerial station ends up being more likely to consider the requests coming from loaded areas. Moreover, it is possible to affirm that, in this case, also the peaked Zipf distribution in some ways *"impairs"* the *Rho* and *San Siro* districts. This because, a high α parameter, as already shown, makes the ground station to be more effective, thus being needless to offload a big volume of requests and making these zones less present at the aerial end, having the smallest share of the common non-terrestrial cache. Therefore, when dealing with the ***Double-Beam*** case, the possible zone combinations will surely change the obtained performance. This effect vanishes as the quantity of inspected districts increases.

Double-Beam

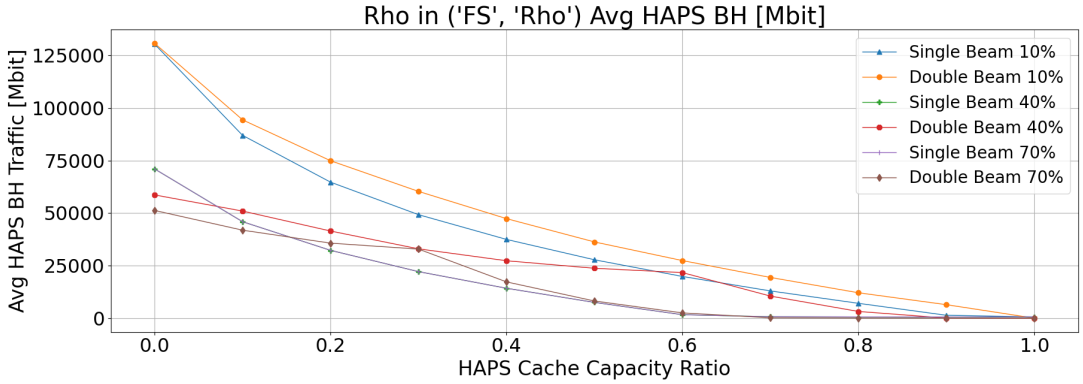
In this part, we will discuss the main outcomes of the activation of a second beam in the HAPS-SMBS platform. The first considered district is the *Train Station* one, used both when coupled with the *Business* district, and with the *Rho* neighborhood. To perform the comparison between the single and double-beam results, the backhaul traffic resulting from any of these two contexts is plotted together with varying haps cache dimensions while considering macro cache provisioning in all the three possible D_M envisioned in Section 9.1.
The obtained results are plotted in Fig. 9.1 and they highlight the fact that, as expected, the activation of the second beam impairs the overall effectiveness of the HAPS cache. Moreover, from a general look it is immediately possible to confirm the affirmations drawn at the introduction of this Section, when saying that the high traffic, low alpha parameter areas are the ones that suffer less from the beam activation, almost when coupled with the ones with very peaked distribution and low traffic volumes. To confirm that, Fig. 9.1 clearly shows that the performance reduction in the central station district is much stronger when this municipality is combined with a neighborhood that has the same popularity distribution value, and a traffic volume that is not so small, such as the one in the *Business* zone, compared to the ones seen in 9.1a when it is paired with *Rho* traffic area.
In addition to this Fig. 9.1a gives us another important insight. In fact, the predominance of the most loaded cluster increases, even more, when the ground station capacity grows, arriving at almost vanishing the impairment caused by the second beam activation when $D_M = 70\%$. This outcome is not shown at all when the train station is coupled with the *Business* zone.
Conversely, Fig. 9.1b illustrates that, when the ground cache dimension improves, the damage due to the activation of the second beam is even bigger. This is due to the fact that, in this second cluster pair, both districts always use the offloading



(a) FS and Rho Pair Results



(b) FS and Business Pair Results



(c) Rho Single and Double Beam Results when coupled with FS

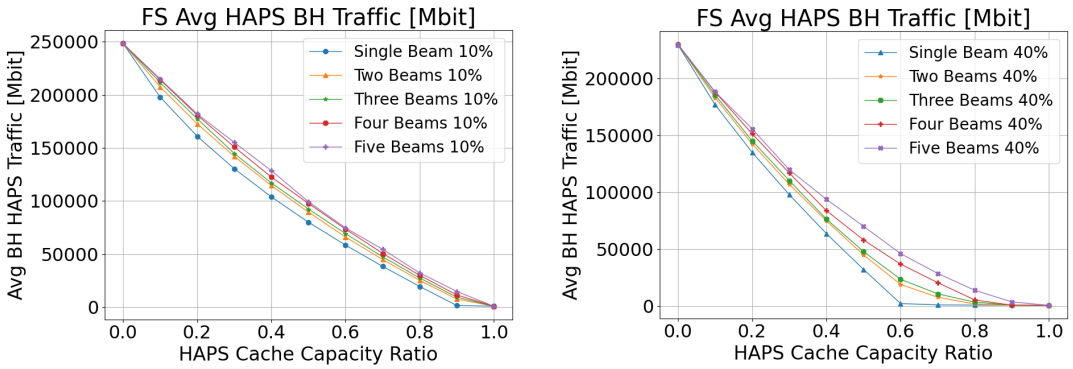
Figure 9.1: FS and Rho Single and Double Beam Results With Terrestrial Priority

capabilities they have, thus the HAPS is provided with evenly distributed, different contents coming from disparate areas. Therefore, guessing the correct one to cache is even more difficult.
Again, all the above statements can be confirmed by Fig. 9.1c, obtained for Rho Fiere when the air station also provides coverage to the central station district. In fact, in this case, the performance deterioration of the zone under analysis is continuous and much more evident than in its much more loaded counterpart.

Furthermore, also this graph outlines that this performance decrease does not disappear at all when the capacity given on the ground BS increases. This phenomenon, together with the coincidence inspected in Fig. 9.1a, is clearly due to the popularity parameter of the latter neighborhood and its low traffic volume. It is clear that, as the ground capacity increases, Rho uses even less HAPS for offloading purposes, both due to the scarcity of contents and to the very high performance that such a large cache with this peaked popularity distribution can provide. Therefore, it ends up using the aerial system so little that the cluster with which it is coupled can easily be the predominant one.
In conclusion, it is also possible to remark on something already shown in the Single Beam chapter, that is the fact that high cluster caches are not exploited at all for neighborhoods with high alpha parameters and low traffic, as confirmed by the plot 9.1, where the single beam results are the same for both $D_M = 40\%$ and $D_M = 70\%$. Finally, from a deep inspection of Fig. 9.1 it is also possible to state a general behavior of all the simulations. In particular, the obtained results outline that, even increasing D_H values is beneficial also for the double-beam use case, the performance enhancement rate experienced when the HAPS provides coverage to more districts is lower compared to the one experienced when it gives offloading capabilities to a single cluster at a time. This is because increasing the size of a shared cache certainly improves its performance, but it does not resolve the main reason why it has a lower improvement rate than in the case of a single beam, i.e. the fact that, in the first investigated use case, the additional dimension is provided to only one cluster as the cache is dedicated to it, which obviously does not happen in the situation we are now inspecting.

Multi-Beam

After having carried out the previous discussion, and having identified some particular features of this scenario, it is now time to investigate what happens when further increasing the number of covered districts, thus meaning that the HAPS cache will be shared between more and more Milan areas.
In this situation, to make the analysis insensitive concerning the previously considered *"combinations effect"*, the results obtained for the train station neighborhood will be averaged, every time, for some of the possible cluster pair mixes. Therefore,



(a) FS with $D_M = 10\%$ Results

(b) FS with $D_M = 40\%$ Results

Figure 9.2: FS Multiple Beam Results

when considering the double cluster situation, both the $(0.56, 0.56)$ and $(0.56, 1.06)$ alpha parameter possible couples will be considered. The same will be carried out when adding more districts, thus obtaining general outcomes. Fig. 9.2 shows the gained results and confirms all the intuitions illustrated in the previous paragraph.
When the number of covered districts grows, the ***Average HAPS Backhaul Traffic*** does the same, as a result of the cache-splitting activity. Furthermore, it is possible to confirm that, as the D_M value increases the performance decrease does the same. This is because, with a small ground caching capability, the terrestrial station is not able to spot all the most popular contents, therefore still giving some popular ones to the aerial system. However, with the improvement of the terrestrial cache object capacity, only very unpopular and evenly distributed requests are offloaded, therefore impairing, even more, the effectiveness of the non-terrestrial computing server and increasing the differences compared to the single beam scenario.
Finally, the percentual performance decrease is computed as in 4.18 for every deployed capacity in the *Train Station* haps and cluster. The obtained outcomes are reported in Table 9.1.

The provided outputs highlight what was possible to analyze only from a qualitative point of view until now. In particular, all the provided figures throughout this section showed that, as the aerial station cache grows, the performance deterioration increases, until $D_M + D_H > 100\%$, meaning that the summation of the cache dimensions provided at the HAPS and on the ground, station overcome 100% of the library. Therefore, until the two caches can store, even more than the total library when acting together, the performance difference between the single and multi-coverage use cases is always increasing. Finally, it is possible to say that

$\frac{D_H, D_M=10\%}{Beams}$	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
2 Beams	5.10%	6.78%	7.66%	9.88%	10.10%	12.08%	12.64%	22.70%	79.94%	0%
3 Beams	6.65%	9.59%	11.71%	12.57%	15.53%	17.63%	19.50%	32.11%	85.78%	0%
4 Beams	7.40%	10.83%	13.72%	16.00%	17.72%	19.86%	24.50%	34.82%	86.68%	0%
5 Beams	7.99%	11.46%	15,81%	18.41%	19.21%	21.87%	29.71%	39.92%	89.70%	0%

$\frac{D_H, D_M=10\%}{Beams}$	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
2 Beams	3.05%	5.67%	9.21%	12.75%	27.97%	86.64%	83.67%	26.81%	0%	0%
3 Beams	4.73%	8.45%	13.39%	21.65%	39.91%	91.21%	91.68%	81.41%	35.33%	0%
4 Beams	5.78%	9.98%	16.18%	25.48%	46.76%	93.14%	94.46%	88.29%	36.30%	0%
5 Beams	7.17%	12.28%	20.40%	31.66%	54.11%	94.88%	96.00%	93.61%	79.63%	0%

Table 9.1: Percentual Performance Decrease in Haps BH Traffic when adding a new beam

the performance deterioration is directly correlated to the HAPS cache increase, confirming what was stated in the previous paragraph. Indeed this is because the KPI improvement rate of a shared cache is always slower compared to the one of a dedicated object. The provided mathematical results perfectly highlight this behavior for the given content assignment strategy.
Indeed, when $D_M = 10\%$, the difference increases until the end, while when $D_M = 40\%$ a plateau is reached at $D_H = 70\%$.
In conclusion, it is also possible to certify what was said in the 9.2.1 paragraph. In fact, as the cluster cache improves, the performance decrease while adding a new beam is higher, due to the offloading to the aerial system of less popular contents.

9.2.2 Aerial Priority, Multi-Beam

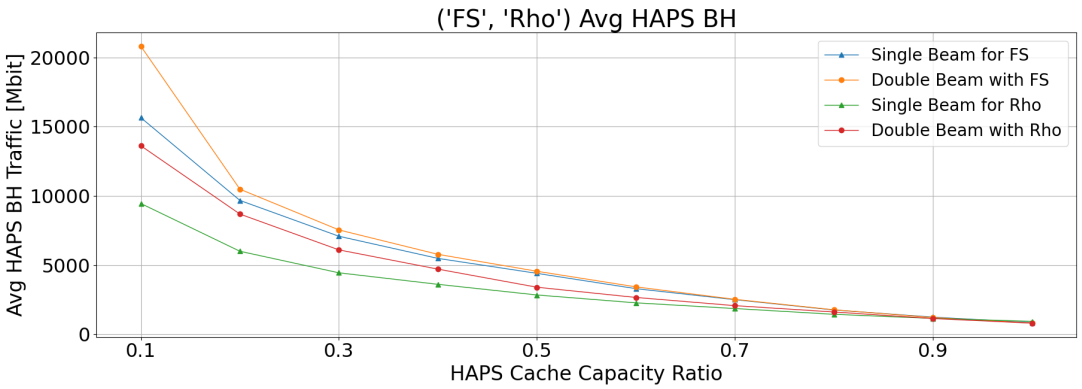
In this passage, we will repeat the analysis carried out in Section 9.2.1 considering the ***aerial priority*** content assignment technique. While performing these simulations, we have to remember the most important differences between the previous strategy considered and the one that is now under investigation.
First of all, in this use case, the privileged segment is the non-terrestrial one, meaning that, as soon as a content, already present in the aerial cache, is requested, it directly goes to it, while the less popular ones will be handled by the macro station. Therefore, in this situation, a high Zipf distribution parameter, should be beneficial, as the aerial cache will surely insert, between the most popular requests, the most demanded ones, even if the overall volume in the considered timestep, is lower. Therefore it is possible to say that a peaked content distribution acts counteracting the difference in the quantity of effective content.
Moreover, adding a new beam, mostly for the cases in which low distribution parameters are there, should not result in a high-performance decrease, particularly

when these districts are coupled with neighborhoods with low traffic volumes and high alpha values. This happens because the haps cache will spot the most popular contents coming from the districts that have high-peaked popularity distributions, which, being their Zipf parameter very high, will be few, leaving the left space for clusters that present high average traffic volumes. Therefore, in this situation, none of the districts should be strongly impaired when adding a new beam, at least when the comparison is carried out between one or two beams, and districts that show different behavior.
On the other hand, when adding new beams, a performance decrease is always expected, but the overall quantity of ***average HAPS backhaul traffic*** should still be much lower compared to the one experienced in the first case scenario.

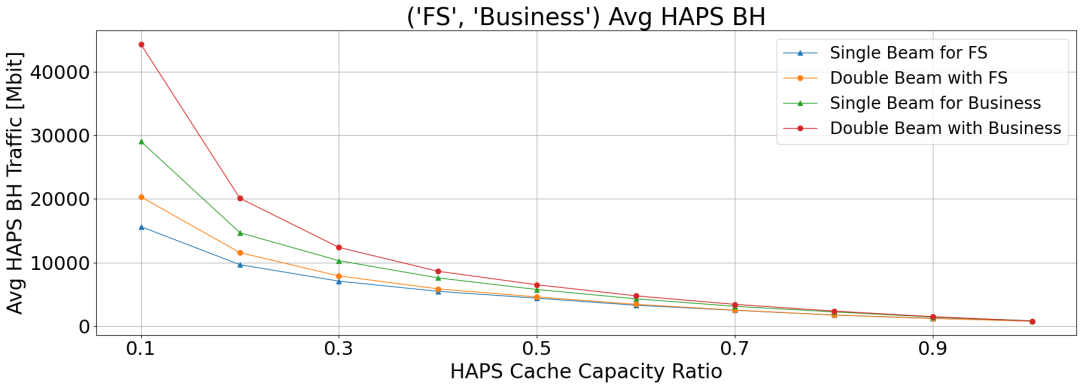
Dual-Beam

In this part of the work, the dual-beam simulation results are discussed. When dealing with them, it has to be reminded that, a change in the cluster behavior, thus meaning its average traffic volume and its popularity parameter, is the reason that influences the most the provided outcomes. Therefore, when illustrating the obtained outputs, all three major cases are taken into consideration. The first one is the one in which the train station cluster is coupled with the *Rho* municipality. Secondly, the couple between the *central station* and *Business* area is inspected. Finally, the one considering *Rho* and *San Siro* traffic zones is simulated.
Fig. 9.3 shows the obtained outputs for all the aforementioned pairs. The retrieved outcomes put in evidence the fact that the cache sharing operation in this context is strongly effective and allows to maintain high performance, even when doubling the number of covered areas.
Moreover, from a deeper inspection of all the provided graphs, some of the already discussed insights can be confirmed. Indeed, in this content assignment strategy, adding a new beam is much less influential on the overall results compared to the same situation in the first Scenario. This result has to be strongly considered when dealing with delay-sensitive applications when the ***average HAPS backhaul traffic*** must be almost zeroed.

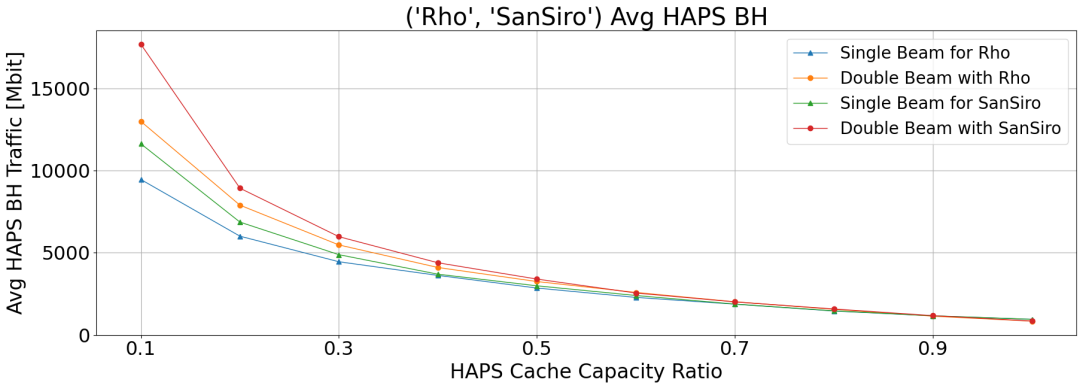
Moreover, considering the first beam couple, thus the one in Fig. 9.3a another meaningful comparison has to be carried out. Considering the same simulation for the *terrestrial priority context*, shown in Fig. 9.1, it is possible to see that the value of the inspected index is always higher compared to what happens in this use case, thus meaning that, even adding new beams, this strategy is still the best-performing one when considering this KPI.
Furthermore, we can highlight that the two values start becoming comparable only when ***huge*** ground cache provisioning is envisioned.
In addition to that, in the previous use case, the HAPS performance was highly



(a) FS and Rho Pair Results



(b) FS and Business Pair Results



(c) Rho and San Siro Pair Results

Figure 9.3: Couples Single and Double Beam Results With Aerial Priority

influenced by the macro BS capacity, which is not the case here, since, as already explained in Section 7.1.1, the aerial station values are, for this content assignment technique, completely insensitive to the cluster variations.
Another difference is there, and it is because, in this second situation, as the haps capacity increases, the results obtained when adding a new beam tend to coincide with the ones that came in output when considering the single beam coverage. The same tendency is not there in Fig. 9.3. This is due to the fact that giving priority to the non-terrestrial station, means offloading to it the most popular content. Therefore, an increase in the aerial library is much more effective here when compared to the situation shown in the previous analysis, where less popular requests were given to the satellite segment.
Finally, the results obtained when the *FS* and *Business* couple is considered, are plotted in Fig. 9.3b. From a visual investigation of this graph it is immediately possible to see that, between the two clusters, the most impaired one is the last mentioned, meaning the one where offices are placed, while the train station performs as well as in the previous case as if the couple performed very similarly to what happened before while considering the *Rho* area.
This result confirms the insight given in the introductory section of this paragraph. The *Business* traffic zone has a higher volume compared to the *Rho* municipality, thus meaning it may perform better than the first mentioned one. However, two things prevent it from performing as the *central station* zone. The first one is the fact that its average traffic is lower than the one experienced by the aforementioned district- The second reason is that, compared to the *Rho* zone, which has $\alpha = 1.06$, the office's area has a low Zipf parameter, meaning that the improvement compared to the *Rho* cluster, that it can reach due to its higher average demand, is the same that the *Rho Fiere* neighborhood can achieve from its high peaked distribution, thus ending up being equally impaired when coupled with the train station district.
Finally, Fig. 9.3c report the outputs for the last considered pair, that is the one between *Rho* and *San Siro*. Here the two clusters have, as shown in Chapter 5, the same behavior. For this reason, their performance is comparable, meaning that the HAPS cache is evenly shared between them, with the first mentioned cluster which still has some advantage, due to its traffic volume that, as shown in Fig. 5.5 is slightly higher when compared to the one experienced in the stadium area.

Multi-Beam

Having understood the main features of this new scenario when adding a second beam coverage, it is time to expand the discussion considering the so-called, *multi-beam* provisioning. In particular, as performed in Section 9.2.1 the ***average haps backhaul traffic*** is considered in the *FS* district.

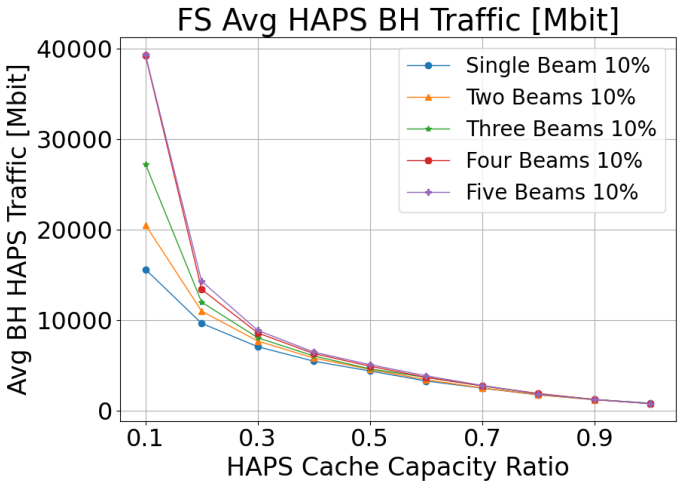


Figure 9.4: FS Avg HAPS BH Traffic with additive beams

Fig. 9.4 shows the obtained results, again highlighting the small difference when in this content assignment strategy. Moreover, it is possible also to confirm the insight, already illustrated, regarding the reduction in performance deterioration, thus their convergence, as an increase in the HAPS cache library is provided.

Finally, the percentual performance decrease is computed to give a mathematical confirmation to these statements. In particular, it is evaluated, as in the previous case, when adding new beams and when changing the haps cache value for the *central station* cluster. In this case, only one table is reported since changing the ground station library does not vary the obtained results.

$\frac{D_H}{Beams}$	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
2 Beams	23.92%	12.24%	8.22%	5.92%	3.80%	3.98%	0.96%	0.47%	0%	0%
3 Beams	42.64%	19.53%	12.28%	9.50%	5.10%	10.60%	9.04%	6.27%	0.51%	0%
4 Beams	60.14%	27.94%	17.73%	13.29%	10.70%	10.85%	9.28%	8.31%	0.06%	0%
5 Beams	60.25%	32.46%	20.36%	15.41%	13.64%	14.75%	10.54%	7.33%	0.61%	0%

Table 9.2: Percentual Performance Decrease in Haps BH Traffic when adding a new beam

The reported outcomes confirm all the intuitions proposed previously. In particular, the inspected performance deterioration has a quasi-asymptotic behavior with the increase in the size of the air cache, converging towards the values of the single-beam scenario.

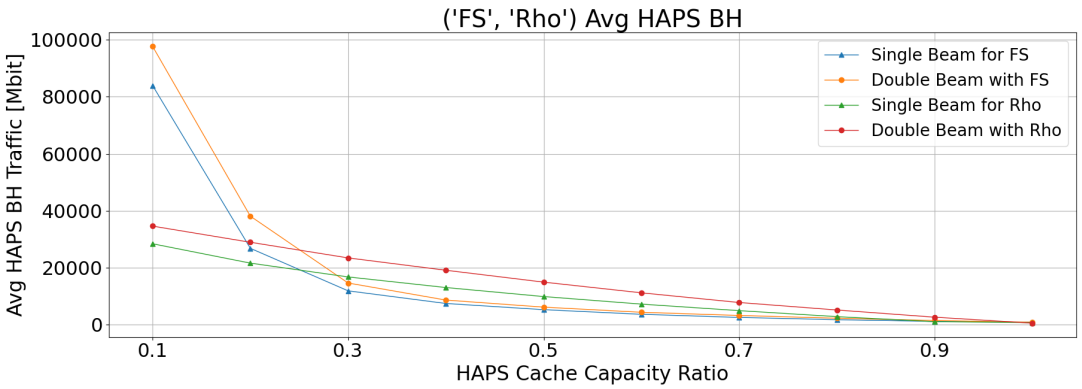
9.2.3 Cooperative Strategy, Multi-Beam

The last strategy is now under analysis, thus the one that aims at performing a sort of ***cooperative caching***. In this use case, the terrestrial station has still a priority, however, the aerial segment is provided with the content-check strategy, thus meaning that contents, before being offloaded to it, have to be present in the non-terrestrial cache. Only at the end, some requests are sent regardless of this reasoning, due to the staticity-avoidance provisioning.
In this case, the performance will change when shifting the cluster capability towards higher values, but we are expecting that the difference when doing it won't be so evident as in the first content assignment strategy.
On the other hand, as already inspected, an increase in the overall ***Average HAPS backhaul traffic*** is foreseen, since the privileged cache object is still the terrestrial one.

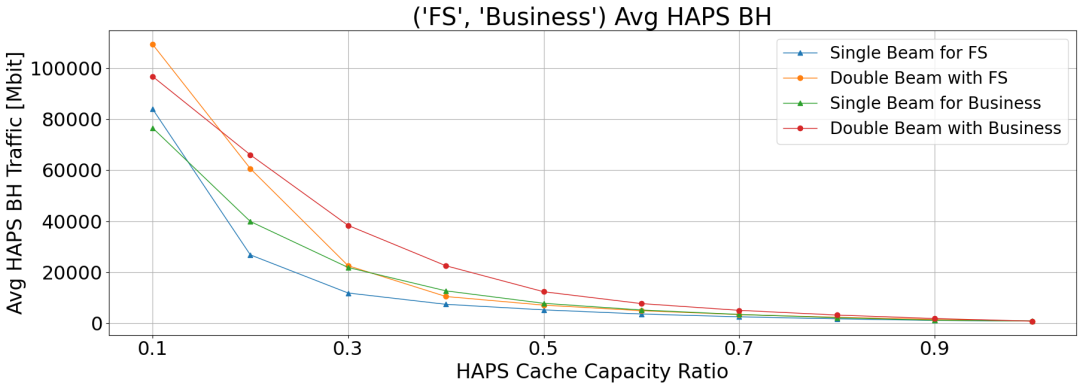
Dual-Beam

As in all the previous cases, the obtained results change when different pairs are considered, thus meaning that more than just one couple is considered. Fig. 9.5 shows the obtained results for the *FS-Rho*, *FS-Business*, and *Rho-San Siro* couples, highlighting what was previously mentioned for the 1st scenario developed. In this case, even if a haps content check is carried out, the impairment in aerial performance in districts that experience a smaller traffic volume is not counterbalanced by the presence of a peaked distribution. Moreover, when comparing this third strategy with the first one, it has to be noted that, even when the haps offers coverage to more districts and even if the increase in the ***Average backhaul Haps traffic*** is still there and it is not as negligible as in the case of *aerial priority* content assignment, still its total value is much smaller compared to the one obtained for the *terrestrial-priority* technique.

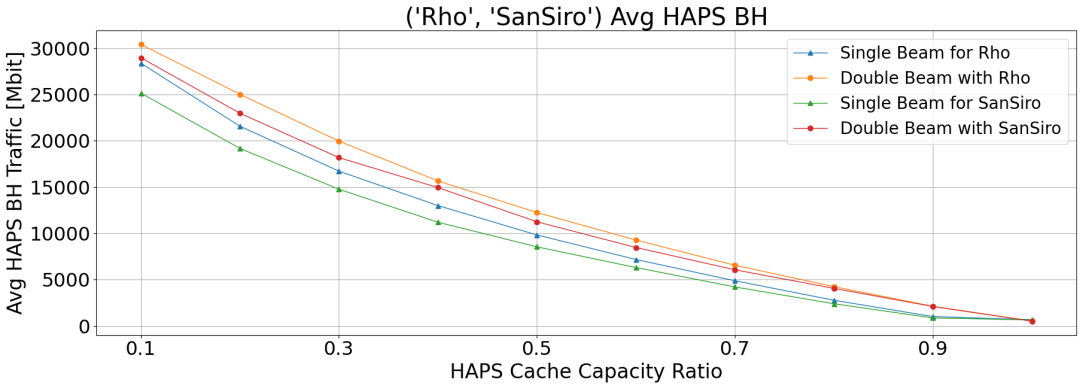
Furthermore, to confirm what just said, the *Rho* related lines in Fig. 9.5a highlight that here the deterioration is much bigger when comparing it with the one experience in the train station neighborhood.
As further confirmation, when the *central station* zone is coupled with the *Business* one, its index value increases faster than when its behavior is simulated for the other couple.
However, something has to be remarked, which is the fact that, even if the quantity of backhaul traffic has increased compared to the case in which the HAPS had priority, still an increase in its cache dimensions is highly effective, meaning that convergence towards the single case situation is there when high library provisioning is given to the non-terrestrial segment even before that $D_M + D_H = 100\%$. This is directly related to the ***wise*** content offloading strategy described in Section 4.5.3, allowing the haps to be more responsive and effective concerning the first



(a) FS and Rho Pair Results



(b) FS and Business Pair Results



(c) FS and Business Pair Results

Figure 9.5: Couples Single and Double Beam Results With Cooperative Strategy

developed hierarchical use case.
Again, it is also possible to see how, when two similar districts, such as *Rho* and *San Siro* act together, their index behavior is similar.
Finally, to conclude this section, the last consideration has to be drawn, which is the one related to the fact that, the cluster that, when the HAPS cache is very low, has higher backhaul traffic, due to the increased number of requests that it sends to the aerial station, ends always up being the less impaired, and therefore the best performing one.

Multi-Beam

Finally, the last comparison is carried out considering more beams at work.

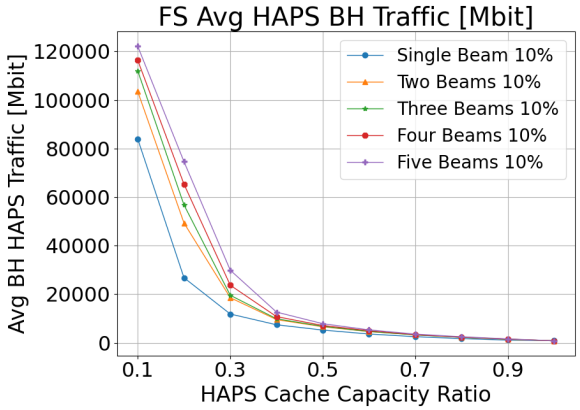


Figure 9.6: FS Avg HAPS BH Traffic with additive beams

The same, classical behavior is inspected when developing this scenario. Adding a new beam is always not beneficial for the overall performance, as happened for the first use case. However, it is important to see how, if, on the one hand, its performance decrease is still higher than the aerial caching strategy, as its overall value of average haps backhaul traffic, on the other it still has the same convergence tendency. This means that the performance deterioration decreases and reaches a plateau as soon as the haps library increases in dimensions. While discussing that, it is also important to say that this behavior happens regardless of the number of active beams, thus meaning that an increase in the aerial cache size will always result in a stabilization of the *haps backhaul traffic* to low values. The previously mentioned results are confirmed when considering the *FS* cluster with $D_M = 10\%$ as shown in Table 9.3. This last analysis highlights the correctness of all the affirmations carried out in the previous paragraph, thus making this 3rd content

assignment technique, the most *balanced* one, able to reduce the ***average backhaul traffic*** coming from the HAPS, without impairing the cluster performance a lot.

$\frac{D_H}{Beams}$	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
2 Beams	18.87%	45.71%	36.21%	22.36%	20.79%	21.73%	22.78%	24.01%	24.00%	0%
3 Beams	25.03%	51.56%	39.44%	25.01%	22.55%	22.55%	22.59%	24.03%	24.79%	0%
4 Beams	27.85%	59.00%	50.04%	31.23%	26.48%	27.42%	24.86%	26.34%	26.87%	0%
5 Beams	31.25%	64.09%	60.37%	41.40%	33.40%	32.03%	29.53%	30.73%	30.79%	0%

Table 9.3: Percentual Performance Decrease in Haps BH Traffic when adding a new beam

Chapter 10

Conclusions

In today's world, the need for an internet connection to be reliable and available anywhere in the globe is indisputable. In dealing with this topic, we have explained how this purpose is not a mere technological requirement, but an integral part of the development of modern society. This thought is confirmed by the role ICTs have played in the management of the Covid-19 pandemic and by the emergence of new applications such as those of AR, VR, and MR, which were once unthinkable. To ensure that these services do not remain mere aspirations, the integration of NTNs in the TNs classical paradigm was considered to be the most viable solution. Therefore, a real scenario comprising a HAPS offering coverage to various Milan districts is considered, and its performance is analyzed thanks to the traffic traces provided to us by an Italian MNO. We assume that it is equipped with a MEC server, which provides caching capabilities. The simulated algorithms for the MEC cache update were the B-LFU and the NM-LFU, based on a frequency popularity estimation of contents, the LRU and Q-SLRU, that work considering request recency, the FIFO, LIFO, and Random update technique. Furthermore, for all of them, three priority strategies are developed and referred to as: the ***terrestrial priority*** check and assignment, the ***non-terrestrial priority*** one, and the ***cooperative caching*** technique. For each of them, simulations considering two weeks of the aforementioned traffic profiles are performed and the analysis was divided into two main blocks. The first has dealt with the results obtainable when the aerial platform provides coverage to only one district at a time, while the second one considers the cache object to be shared amongst more areas. The main outcomes of the first analysis have been:

- Terrestrial Priority Scenario:
  - Strong reduction in the ***average hourly HAPS backhaul traffic*** compared to the case in which the aerial system is used with only relay purposes, with a percentage decrease ranging from 5.4% to 36.5% in less

performing situations, up to a zeroing of this value in very favorable cases, i.e. when large capacities are provided to the cache.

- Optimized cluster cache performance. Indeed, when clusters with high peaked distributions and low traffic volumes are considered, such as Rho and San Siro, thanks to the aerial platform support, there is a reduction of ground station miss probability to almost 0%.
- Suboptimal HAPS content provisioning, meaning that the air segment always faces the problem of being assigned unpopular content;
- Decrease in the core network load with a maximum improvement equal to 32.9%.

- Aerial Priority Assignment:
  - Immediate elimination of backhaul traffic created by the non-terrestrial station for performing algorithms and zeroing of ***HAPS miss probability***. The first value is, in some cases, *20* times lower compared to the previous situation, with averaged eightfold reduction. Already from ***small-sized*** cache dimensions, only a maximum of 15 requests are missed hourly.
  - Suboptimal Cluster performance, due to the content assignment strategy. The averaged *percentage performance decrease*, when $D_H, D_M = 10\%$. is equal to the 12.64%. This strategy indeed does not solve the problem of unpopular content, but moves them away from the non-terrestrial segment;
  - Suboptimal *total averaged hourly backhaul traffic*.

- Cooperative Caching:
  - Optimized Cluster and Haps Cache performance, with a reduced ***cluster miss-probability***, compared to the one derived in the second strategy, without the cost of having high ***average backhaul HAPS traffic volume*** as in the first.

Moreover, general trends about the inspected algorithms are derived. In particular, regarding the Random strategy, even if it is much more efficient in districts with highly peaked content popularity distributions, it remains the least performing one. Subsequently, it was found that the LIFO strategy, due to its intrinsic functioning, is very close to the previously mentioned one, filling its free spaces with the first possible contents and finally replacing only the last one accessed, therefore explaining its poor performance. Regarding the LRU algorithm, we have seen that for low cache values, it performed almost like the FIFO, since for small

capacities and more uniform distributions, it is very unlikely that a content inside its queue will be requested again, thus never activating its aggressive strategy. The LRU content replacement is also the reason that makes it less performing for the static popularity simulation in use when compared to its segmented counterpart (Q-SLRU) while making it very effective in situations where popularity varies quickly. Finally, we have seen how the B-LFU is always the most performing technique and analyzed the importance of keeping the memory of past hits, as the NM-LFU algorithm, which works the same as the B-LFU except for this feature, has decidedly negative performances.
Therefore, the following conclusions are drawn:

- The first strategy with terrestrial priority has little traffic in the core network with minimal levels of terrestrial misses. However, it does not greatly reduce the cases of content with high delay, thus being the least viable one, in case of delay-sensitive applications;
- The second architecture eliminates the backhaul traffic related to the HAPS at the price of a greater load in the backhaul network. It is the best solution for delay-sensitive applications;
- The cooperative scenario is the context that balances the characteristics of the previous two.

Finally, we proceeded with the simulations of the scenario, where the aerial platform offered coverage to various districts in the city of Milan. Due to the results obtained in the previous block, the analysis focused on the ***HAPS backhaul traffic***, considering the three use cases described above and the B-LFU algorithm. The results showed that:

- In each of the strategies analyzed, an increase in beams corresponds to an increase in backhaul traffic, therefore a reduction in performance, since the aerial cache is no longer dedicated but shared;
- The traffic volumes and popularity distributions values are highly influential in these cases, preferring areas, where there are high averaged demands and the content distribution is peaked;
- The first scenario shows a reduction and a continuous deterioration in performance with the growth of both caches until reaching a plateau, and finally a relative reduction of the performance decrease. However, this happens only when the terrestrial and non-terrestrial caches store together more than 100% of the contents in the library, which is unlikely. Consequently, it is concluded that this scenario presents an increasingly negative monotonic trend, with higher cache libraries, and covered districts;

- In the second strategy, a much smaller performance decrease was identified. This behavior, related to the aerial priority, is inspected to be monotonously decreasing with the increase in D_H. By this, we mean that, as the size of the caches increase, the differences that exist in the activation of several beams are reduced to zero. This strategy is therefore more viable compared to the previous one since, as the traffic volume sent to HAPS rises, due to the activation of more beams, it becomes increasingly important to improve its performance.

- The balanced behavior of the third scenario is confirmed. Its initial trend is similar to that of the first case, with a rising performance deterioration. However, in the end, as happens the second context, it reaches a plateau, having higher average values compared to the last cited scenario, but still stable and much smaller than the first one, making this solution still attractive also for delay-sensitive applications.

Having presented these considerations, it is possible to affirm that, further developments of this work would require the addition of a process of temporal variability of contents popularity, and also that of the birth and death of new content. This would allow a better investigation of algorithms performance.
Moreover, future studies will also include the HAPS energy consumption models to be in place, together with more modern traffic profiles.
In conclusion, this work confirms the strong viability of NTNs as a future paradigm for 6G, enabling future communication networks to be in place and a worldwide internet connection to be given throughout the globe. This achievement will help accomplish the SDG 9.c target that considers these two concepts as crucial in reducing social inequalities and fostering a world in which health is, as stated by the WHO, *"a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"* [79].

Bibliography

[1] Mohammad Mahdi Azari, Sourabh Solanki, Symeon Chatzinotas, Oltjon Kodheli, Hazem Sallouha, Achiel Colpaert, Jesus Fabian Mendoza Montoya, Sofie Pollin, Alireza Haqiqatnejad, Arsham Mostaani, Eva Lagunas, and Bjorn E. Ottersten. Evolution of non-terrestrial networks from 5g to 6g: A survey. *CoRR*, abs/2107.06881, 2021. URL https://arxiv.org/abs/2107.06881.

[2] CISCO. Cisco annual internet report (2018–2023), Mar 2020. URL https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[3] Sofie Lambert, Ward Van Heddeghem, Willem Vereecken, Bart Lannoo, Didier Colle, and Mario Pickavet. Worldwide electricity consumption of communication networks. *Opt. Express*, 20(26):B513–B524, Dec 2012. doi: 10.1364/OE.20.00B513. URL https://opg.optica.org/oe/abstract.cfm?URI=oe-20-26-B513.

[4] Starlink Statistics. Jonathan's space pages. URL https://planet4589.org/space/stats/star/starstats.html.

[5] M. S. Alam et al. High altitude platform station based super macro base station constellations. *IEEE Communications Magazine*, 59(1):103–109, 2021.

[6] Greta Vallero, Daniela Renga, and Michela Meo. Caching in the air: High altitude platform stations for urban environments. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2244–2249, 2022. doi: 10.1109/WCNC51071.2022.9771568.

[7] Eurostat. Impact of covid-19 on the use of ict in enterprises. Mar 2022. URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Impact_of_COVID-19_on_the_use_of_ICT_in_enterprises#Impact_of_COVID-19_pandemic_on_remote_access_to_enterprise_resources_and_remote_meetings.

[8] Sachin Kumar, Prayag Tiwari, and Mikhail Zymbler. Internet of things is a revolutionary approach for future technology enhancement: a review. *Journal of Big Data*, 6, 12 2019. doi: 10.1186/s40537-019-0268-2.

[9] Huawei. 5g network architecture a high-level perspective, 2016. URL `https://www-file.huawei.com/-/media/corporate/pdf/mbb/5g_nework_architecture_whitepaper_en.pdf?la=en`.

[10] Vitetta G.M. Luise M. *Teoria Dei Segnali*. Mc Graw Hill, 1612.

[11] ITUWTPF. Sixth world telecommunication/ict policy forum 2021, Dec 2021. URL `https://www.itu.int/dms_pub/itu-s/md/21/wtpf21/c/S21-WTPF21-C-0013!R1!PDF-E.pdf`.

[12] You Xiaohu, Cheng-Xiang Wang, Jie Huang, Xiqi Gao, Zaichen Zhang, Michael Wang, Yongming Huang, Chuan Zhang, Yanxiang Jiang, Jiaheng Wang, Min Zhu, Bin Sheng, Dongming Wang, Zhiwen Pan, Pengcheng Zhu, Yang Yang, Zening Liu, Zhang Ding, Xiaofeng Tao, and Ying-Chang Liang. Towards 6g wireless communication networks: vision, enabling technologies, and new paradigm shifts. *Science China Information Sciences*, 64, 01 2021. doi: 10.1007/s11432-020-2955-6.

[13] Huawey 6G Research Group. *6G: The Next Horizon: From Connected People and Things to Connected Intelligence*. Cambridge University Press, 2021.

[14] Scarpa F. An in-depth analysis of future sixth generation networks with integration of aerial platforms. Master's thesis, Politecnico di Torino, 2021.

[15] 3GPP. 3rd generation partnership project; solutions for nr to support non-terrestrial netowrks (ntn) (release 16), 2020.

[16] Gunes Karabulut Kurt, Mohammad G. Khoshkholgh, Safwan Alfattani, Ahmed Ibrahim, Tasneem S. J. Darwish, Md Sahabul Alam, Halim Yanikomeroglu, and Abbas Yongacoglu. A vision and framework for the high altitude platform station (haps) networks of the future. *IEEE Communications Surveys & Tutorials*, 23(2):729–779, 2021. doi: 10.1109/COMST.2021.3066905.

[17] S. Karapantazis and F. Pavlidou. Broadband communications via high-altitude platforms: A survey. *IEEE Communications Surveys & Tutorials*, 7(1):2–31, 2005. doi: 10.1109/COMST.2005.1423332.

[18] ITU. Haps – high-altitude platform systems, Apr 2022. URL `https://www.itu.int/en/mediacentre/backgrounders/Pages/High-altitude-platform-systems.aspx`.

[19] Kamal Harb, Abdi Abdalla, Muhidin Mohamed, and Samir Abdul-Jauwad. Haps communication in saudi arabia under dusty weather conditions. 11 2013. doi: 10.1109/MICC.2013.6805858.

[20] GSMA. High altitude platform systems, towers in the skies, 2021. URL https://www.gsma.com/futurenetworks/resources/high-altitude-platform-systems-haps-whitepaper-2021/.

[21] Md Sahabul Alam, Gunes Karabulut Kurt, Halim Yanikomeroglu, Peiying Zhu, and Ngoc Dũng Đào. High altitude platform station based super macro base station constellations. *IEEE Communications Magazine*, 59(1):103–109, 2021. doi: 10.1109/MCOM.001.2000542.

[22] Md Sahabul Alam, Gunes Karabulut Kurt, Halim Yanikomeroglu, Peiying Zhu, and Ngoc Dũng Đào. High altitude platform station based super macro base station (haps-smbs) constellations, 2020. URL https://arxiv.org/abs/2007.08747.

[23] UN. Sdg goals, 2015. URL https://sdgs.un.org/goals/goal9.

[24] UN. Sustainable development goal 9: Investing in ict access and quality education to promote lasting peace. Jun 2017. URL https://www.un.org/sustainabledevelopment/blog/2017/06/sustainable-development-goal-9-investing-in-ict-access-and-quality-education-to-p

[25] IGF. Community connectivity: Building the internet from scratch. 2016. URL https://internet-governance.fgv.br/sites/internet-governance.fgv.br/files/publicacoes/community_connectivity_-_building_the_internet_from_scratch_0.pdf.

[26] Björn Döhring, Atanas Hristov, Christoph Maier, Werner Roeger, and Anna Thum-Thysen. COVID-19 acceleration in digitalisation, aggregate productivity growth and the functional income distribution. *International Economics and Economic Policy*, 18(3):571–604, July 2021. doi: 10.1007/s10368-021-00511-. URL https://ideas.repec.org/a/kap/iecepo/v18y2021i3d10.1007_s10368-021-00511-8.html.

[27] ITU. Pandemic preparedness: Tech for covid-19 and future outbreaks. Dec 2021. URL https://www.itu.int/en/mediacentre/backgrounders/Pages/Pandemic-preparedness.aspx.

[28] Eurostat. Ict specialists in employment. Apr 2022. URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=ICT_specialists_in_employment#Number_of_ICT_specialists.

[29] UNICEF. How many children and young people have internet access at home? estimating digital connectivity during the covid-19 pandemic., 2020. URL https://www.itu.int/en/ITU-D/Statistics/Documents/publications/UNICEF/How-many-children-and-young-people-have-internet-access-at-home-2020_v2final.pdf.

[30] European Parliament. 5g deployment state of play in europe, usa and asia, April 2019. URL https://www.europarl.europa.eu/RegData/etudes/IDAN/2019/631060/IPOL_IDA(2019)631060_EN.pdf.

[31] European Commission. What are smart cities, 2020. URL https://ec.europa.eu/info/eu-regional-and-urban-development/topics/cities-and-urban-development/city-initiatives/smart-cities_en.

[32] UN. 68% of the world population projected to live in urban areas by 2050, says un, 2018. URL https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

[33] European Commission. In focus: Energy efficiency in buildings, Feb 2020. URL https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-lut-17_en.

[34] Evelin Priscila Trindade, Marcus Phoebe Farias Hinnig, Eduardo Moreira da Costa, Jamile Sabatini Marques, Rogério Cid Bastos, and Tan Yigitcanlar. Sustainable development of smart cities: a systematic review of the literature. *Journal of Open Innovation: Technology, Market, and Complexity*, 3(3), 2017. ISSN 2199-8531. doi: 10.1186/s40852-017-0063-2. URL https://www.mdpi.com/2199-8531/3/3/11.

[35] European Parliament. The internet of things, opportunities and challenges, 2015. URL https://www.europarl.europa.eu/RegData/etudes/BRIE/2015/557012/EPRS_BRI(2015)557012_EN.pdf.

[36] Tanweer Alam. A reliable communication framework and its use in internet of things (iot). 3, 05 2018.

[37] Ericsson. Ericsson mobility report 2021, Nov 2021. URL https://mb.cision.com/Main/15448/3460965/1502237.pdf.

[38] CISCO. Cisco visual networking index: Forecast and trends, 2017–2022, 2019. URL https://twiki.cern.ch/twiki/pub/HEPIX/TechwatchNetwork/HtwNetworkDocuments/white-paper-c11-741490.pdf.

[39] Bodunrin Bakare and Edidiong Bassey. A comparative study of the evolution of wireless communication technologies from the first generation (1g) to the fourth generation (4g). 12:73–84, 10 2021.

[40] A. Gupta and R. K. Jha. A survey of 5g network: Architecture and emerging technologies. *IEEE Access*, 3:1206–1232, 2015. doi: 10.1109/ACCESS.2015.2461602.

[41] A.R.S. Bahai and H. Aghvami. Network planning and optimization in the third generation wireless networks. In *First International Conference on 3G Mobile Communication Technologies*, pages 441–445, 2000. doi: 10.1049/cp:20000088.

[42] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 18(3):1617–1655, 2016. doi: 10.1109/COMST.2016.2532458.

[43] GSMA Intelligence. Global mobile trends 2021 navigating covid-19 and beyond, 2020. URL `https://data.gsmaintelligence.com/research/research/research-2020/global-mobile-trends-2021`.

[44] Wiley. A brief survey of 5g wireless networks, 2021. URL `https://www.wiley.com/learn/computerscience/pdf/engineering-5g-whitepaper.pdf`.

[45] Nisha Panwar, Shantanu Sharma, and Awadhesh Kumar Singh. A survey on 5g: The next generation of mobile communication, 2015. URL `https://arxiv.org/abs/1511.01643`.

[46] Luca Chiaraviglio, Marco Fiore, and Edouard Rossi. *5G Technology: Which Risks From the Health Perspective?* 12 2019.

[47] Samsung Research. 6g the next hyper connected experience for all, 2021. URL `https://research.samsung.com/next-generation-communications`.

[48] Yun Chen, Wenfeng Liu, Zhiang Niu, Zhongxiu Feng, Qiwei Hu, and Tao Jiang. Pervasive intelligent endogenous 6g wireless systems: Prospects, theories and key technologies. *Digital Communications and Networks*, 6 (3):312–320, 2020. ISSN 2352-8648. doi: https://doi.org/10.1016/j.dcan.2020.07.002. URL `https://www.sciencedirect.com/science/article/pii/S235286482030242X`.

[49] Madhan Raj Kanagarathinam and Krishna M. Sivalingam. *Challenges in Transport Layer Design for Terahertz Communication-Based 6G Networks*,

pages 123–136. Springer International Publishing, Cham, 2021. ISBN 978-3-030-72777-2. doi: 10.1007/978-3-030-72777-2_7. URL https://doi.org/10.1007/978-3-030-72777-2_7.

[50] Shaimaa Abdel Hakeem, Hanan Hussein, and Hyungwon Kim. Vision and research directions of 6g technologies and applications. *Journal of King Saud University - Computer and Information Sciences*, 34, 03 2022. doi: 10.1016/j.jksuci.2022.03.019. URL https://www.researchgate.net/publication/359613210_Vision_and_research_directions_of_6G_technologies_and_applications.

[51] Saddam Alraih, Ibraheem Shayea, Mehran Behjati, Rosdiadee Nordin, Nor Fadzilah Abdullah, Asma' Abu-Samah, and Dalia Nandi. Revolution or evolution? technical requirements and considerations towards 6g mobile communications. *Sensors*, 22(3), 2022. ISSN 1424-8220. doi: 10.3390/s22030762. URL https://www.mdpi.com/1424-8220/22/3/762.

[52] Muhammad Waseem Akhtar, Syed Hassan, Rizwan Ghaffar, Haejoon Jung, Sahil Garg, and M. Shamim Hossain. The shift to 6g communications: vision and requirements. *Human-centric Computing and Information Sciences*, 10, 12 2020. doi: 10.1186/s13673-020-00258-2. URL https://www.researchgate.net/publication/347711248_The_shift_to_6G_communications_vision_and_requirements.

[53] FCC. Fcc takes steps to open spectrum horizons for new services and technologies, 2019. URL https://www.fcc.gov/document/fcc-opens-spectrum-horizons-new-services-technologies.

[54] Mikko A Bernardos, Carlos J.; Uusitalo. European vision for the 6g network ecosystem, 2021.

[55] Sean F. Johnston. A cultural history of the hologram. *Leonardo*, 41:223–229, 2008.

[56] Richard Li. Network 2030 a blueprint of technology, applications and market drivers towards the year 2030 and beyond. 07 2019.

[57] 3GPP. 3rd generation partnership project; technical specification group radio access network; study on new radio (nr) to support non-terrestrial networks (release 15), 2020.

[58] 3GPP. Ntn & satellite in rel-17 & 18, Mar 2022. URL https://www.3gpp.org/news-events/partner-news/ntn-rel17.

[59] Federica Rinaldi, Helka-Liina Maattanen, Johan Torsner, Sara Pizzi, Sergey Andreev, Antonio Iera, Yevgeni Koucheryavy, and Giuseppe Araniti. Non-terrestrial networks in 5g & beyond: A survey. *IEEE Access*, 8:165178–165200, 2020. doi: 10.1109/ACCESS.2020.3022981.

[60] D.L. Oltrogge, S. Alfano, C. Law, A. Cacioni, and T.S. Kelso. A comprehensive assessment of collision likelihood in geosynchronous earth orbit. *Acta Astronautica*, 147:316–345, 2018. ISSN 0094-5765. doi: https://doi.org/10.1016/j.actaastro.2018.03.017. URL `https://www.sciencedirect.com/science/article/pii/S0094576517315011`.

[61] NASA Earth Observatory. Catalog of earth satellite orbits, 2009. URL `https://earthobservatory.nasa.gov/features/OrbitsCatalog/page1.php`.

[62] Yannick Borthomieu. 14 - satellite lithium-ion batteries. In Gianfranco Pistoia, editor, *Lithium-Ion Batteries*, pages 311–344. Elsevier, Amsterdam, 2014. ISBN 978-0-444-59513-3. doi: https://doi.org/10.1016/B978-0-444-59513-3.00014-5. URL `https://www.sciencedirect.com/science/article/pii/B9780444595133000145`.

[63] Bassel Al Homssi, Akram Al-Hourani, Ke Wang, Phillip Conder, Kandeepan Sithamparanathan, Jinho Choi, Ben Allen, and Ben Moores. Next generation mega satellite networks for access equality: Opportunities, challenges, and performance. *IEEE Communications Magazine*, 60:18–24, 04 2022. doi: 10.1109/MCOM.001.2100802.

[64] Antonio Guillen-Perez, Ana-Maria Montoya, Juan-Carlos Sanchez-Aarnoutse, and Maria-Dolores Cano. A comparative performance evaluation of routing protocols for flying ad-hoc networks in real conditions. *Applied Sciences*, 11 (10), 2021. ISSN 2076-3417. doi: 10.3390/app11104363. URL `https://www.mdpi.com/2076-3417/11/10/4363`.

[65] Marco Giordani and Michele Zorzi. Non-terrestrial networks in the 6g era: Challenges and opportunities. *IEEE Network*, 35(2):244–251, 2021. doi: 10.1109/MNET.011.2000493.

[66] Hsieh F. Nokia Bell Labs, Ghosh A. Haps: Connect the unconnected, 2020.

[67] Abbas Mohammed, Asad Mehmood, Fotini-Niovi Pavlidou, and Mihael Mohorcic. The role of high-altitude platforms (haps) in the global wireless connectivity. *Proceedings of the IEEE*, 99(11):1939–1953, 2011. doi: 10.1109/JPROC.2011.2159690.

[68] Xingqin Lin, Björn Hofström, Eric Wang, Gino Masini, Helka-Liina Maattanen, Henrik Rydén, Jonas Sedin, Magnus Stattin, Olof Liberg, Sebastian Euler, Siva Muruganathan, Stefan Eriksson G., and Talha Khan. 5g new radio evolution meets satellite communications: Opportunities, challenges, and solutions, 2019. URL https://arxiv.org/abs/1903.11219.

[69] HAPS MOBILE. Haps. URL https://www.hapsmobile.com/en/.

[70] ITU-R. Spectrum needs of high-altitude platform stations broadband links operating in the fixed service, 2018. URL https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-F.2438-2018-PDF-E.pdf.

[71] ITU. Emergency telecommunications, 2021. URL https://www.itu.int/en/mediacentre/backgrounders/Pages/emergency-telecommunications.aspx.

[72] Irem Bor-Yaliniz and Halim Yanikomeroglu. The new frontier in ran heterogeneity: Multi-tier drone-cells. *IEEE Communications Magazine*, 54(11): 48–55, 2016. doi: 10.1109/MCOM.2016.1600178CM.

[73] Dong Zhou, Sheng Gao, Ruiqi Liu, Feifei Gao, and Mohsen Guizani. Overview of development and regulatory aspects of high altitude platform system. *Intelligent and Converged Networks*, 1(1):58–78, 2020. doi: 10.23919/ICN.2020.0004.

[74] ITU. Technical and operational characteristics for the fixed service using high altitude platform stations in the bands 27.5-28.35 ghz and 31-31.3 ghz, May 2002.

[75] Daniela Renga and Michela Meo. Can high altitude platforms make 6g sustainable? *IEEE Communications Magazine*, 60:1–7, 09 2022. doi: 10.1109/MCOM.002.2101048.

[76] Min Chen, Yongfeng Qian, Yixue Hao, Yong Li, and Jeungeun Song. Data-driven computing and caching in 5g networks: Architecture and delay analysis. *IEEE Wireless Communications*, 25(1):70–75, 2018. doi: 10.1109/MWC.2018.1700216.

[77] Zhaohui Luo, Minghui Liwang, Zhijian Lin, Lianfen Huang, Xiaojiang Du, and Mohsen Guizani. Energy-efficient caching for mobile edge computing in 5g networks. *Applied Sciences*, 7:557, 05 2017. doi: 10.3390/app7060557.

[78] Comune di Milano. Territorio: Nuclei d'identità locale (nil), Nov 2022. URL https://dati.comune.milano.it/dataset/activity/ds61_infogeo_nil_localizzazione_.

[79] WHO. Constitution of the world health organization, July 1946. URL `https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1`.