NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

COMPUTER NETWORKING

MASTER THESIS

The 5G era of mobile networks: a comprehensive study of the related technologies accompanied by an experimentation framework

Panagiotis K. Boumpoukiotis

Supervisor: **Lazaros Merakos,** Professor

ATHENS

MARCH 2020

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΚΤΥΩΣΗ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Η εποχή των 5G δικτύων: μια ολοκληρωμένη έρευνα των σχετικών τεχνολογιών συνοδευόμενη από ένα πειραματικό πλαίσιο

Παναγιώτης Κ. Μπουμπουκιώτης

Επιβλέπων: **Λάζαρος Μεράκος,** Καθηγητής

ΑΘΗΝΑ

ΜΑΡΤΙΟΣ 2020

MASTER THESIS

The 5G era of mobile networks: a comprehensive study of the related technologies accompanied by an experimentation framework

Panagiotis K. Boumpoukiotis

Reg.No.: M1491

SUPERVISOR: **Lazaros Merakos,** Professor

EXAMINATION COMMITTEE: **Lazaros Merakos,** Professor
Athanasia Allonistioti, Associate Professor
Nikolaos Passas, Instructional Lab. Personnel

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Η εποχή των 5G δικτύων: μια ολοκληρωμένη έρευνα των σχετικών τεχνολογιών συνοδευόμενη από ένα πειραματικό πλαίσιο

Παναγιώτης Κ. Μπουμπουκιώτης
Α.Μ.: Μ1491

ΕΠΙΒΛΕΠΩΝ: **Λάζαρος Μεράκος,** Καθηγητής

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Λάζαρος Μεράκος,** Καθηγητής
Αθανασία Αλωνιστιώτη, Αναπληρώτρια Καθηγήτρια
Νικόλαος Πασσάς, Ε.ΔΙ.Π

ABSTRACT

The ever-increasing demand from mobile communications networks for the provision of better services and interconnection of more devices is pushing the industry's community to develop new network organization methods and technologies in order to effectively address this challenge. As the current technology has reached its limits in terms of traffic management capability, it is necessary to develop a new operating framework that can effectively respond to the new conditions created by the telecommunications market.

The 5th generation of mobile communication networks (5G) aims to solve this exact issue by developing a new operating model. This model, by thoroughly restructuring the way the network operates at all levels, forms a new ecosystem of network infrastructures and functions that enables the provision of high-level services to users, tailored to their particular needs.

The fundamental principles and key technologies that govern the operation of a new generation network throughout its entire length were extensively studied in the context of this paper. Starting with the innovations regarding the structure of 5G networks at the architectural level, the analysis extends to a bottom-up approach: from the broadcast and access levels to the network (C-RAN & MAC) to the mechanisms responsible for delivering the network's functions and services (NFV). Then, the new network-based routing and traffic management (SDN) model is introduced, and the technology for providing distinctive services to users (E2E Slicing) is presented. Furthermore, some characteristic indicators and metrics related to the standardization of the network's technologies are presented, as well as all the current developments related to the development of 5G in Europe.

Then, the data of the experiment carried out for the purposes of the paper is presented. On the one hand, this data concerns the modeling of an existing network based on the new 5G standards and, on the other hand, the evaluation of its performance based on some scenarios regarding the topology and the amount of data exchanged at any time on the network. The examination of the efficiency parameters focuses on the ability of the ONOS SDN controller to manage the traffic of the data when certain events affecting the original network structure occur.

In terms of the results of the measurements being carried out, although the positive impact of the incorporation of new technologies on the performance of mobile communications networks appears to be positive, there are still some individual open issues that need further research by members of the telecommunications community in order for the original vision of the universal operation of all mobile devices under one single umbrella not to be ultimately undermined.

SUBJECT AREA: 5th Generation Networks (5G)

KEYWORDS: 5G, Software Defined Networks, Network Functions Virtualization, Network Slicing, Cloud Radio Access Network, ONOS

ΠΕΡΙΛΗΨΗ

Οι συνεχώς αυξανόμενες απαιτήσεις από τα δίκτυα κινητών επικοινωνιών για τη παροχή καλύτερων υπηρεσιών και τη διασύνδεση όλων και περισσότερων συσκευών, ωθούν τη κοινότητα του κλάδου στην ανάπτυξη νέων μεθόδων και τεχνολογιών οργάνωσης των δικτύων προκειμένου να αντιμετωπιστεί αποτελεσματικά αυτή η πρόκληση. Δεδομένου ότι η παρούσα τεχνολογία έχει φτάσει στα όρια της από άποψη ικανότητας διαχείρισης της κίνησης, απαιτείται η ανάπτυξη ενός νέου πλαισίου λειτουργίας το οποίο θα μπορεί να ανταποκριθεί αποτελεσματικά στις νέες συνθήκες που διαμορφώνονται από τη τηλεπικοινωνιακή αγορά.

Η $5^{η}$ γενιά των δικτύων κινητών επικοινωνιών (5G) αποσκοπεί στην επίλυση ακριβώς αυτού του ζητήματος, μέσα από την ανάπτυξη ενός νέου μοντέλου λειτουργίας. Το μοντέλο αυτό αναδιαρθρώνοντας εκ βάθρων τον τρόπο λειτουργίας του δικτύου σε όλα τα επίπεδα, σχηματίζει ένα νέο οικοσύστημα δικτυακών υποδομών και λειτουργιών το οποίο επιτρέπει τη παροχή στους χρήστες υπηρεσιών υψηλού επιπέδου, προσαρμοσμένες στις εκάστοτε ανάγκες τους.

Στα πλαίσια της παρούσας εργασίας μελετήθηκαν εκτενώς οι θεμελιώδεις αρχές και οι κυριότερες τεχνολογίες που διέπουν τη λειτουργία ενός δικτύου νέας γενιάς καθ' όλο το μήκος του. Ξεκινώντας από τις καινοτομίες που αφορούν τη δομή των 5G δικτύων σε επίπεδο αρχιτεκτονικής, η ανάλυση επεκτείνεται με μία προσέγγιση από κάτω προς τα πάνω· στα επίπεδα εκπομπής και πρόσβασης στο δίκτυο (C-RAN & MAC), στους μηχανισμούς που είναι υπεύθυνοι για παροχή των λειτουργιών και υπηρεσιών του δικτύου (NFV), ενώ εν συνεχεία γίνεται αναφορά στο νέο μοντέλο δρομολόγησης και διαχείρισης της κίνησης συνολικά στο δίκτυο (SDN) και σε επόμενο στάδιο παρουσιάζεται η τεχνολογία που αφορά την ικανότητα παροχής διακριτών υπηρεσιών στους χρήστες (E2E Slicing). Ακόμα, παρουσιάζονται ορισμένοι χαρακτηριστικοί δείκτες και μετρικές που σχετίζονται με τη προτυποποίηση των τεχνολογιών του δικτύου καθώς και όλες οι τρέχουσες εξελίξεις που αφορούν την ανάπτυξη του 5G στην Ευρώπη.

Στη συνέχεια παρουσιάζονται τα δεδομένα του πειράματος που διεξήχθη για τους σκοπούς της εργασίας και αφορά αφενός τη μοντελοποίηση ενός υφιστάμενου δικτύου με βάση τα νέα πρότυπα του 5G και αφετέρου την αξιολόγηση της απόδοσης του με βάση ορισμένα σενάρια σχετικά με τη τοπολογία και το πλήθος των δεδομένων που ανταλλάσσονται κάθε στιγμή στο δίκτυο. Η εξέταση των παραμέτρων αποδοτικότητας εστιάζει στην ικανότητα του ONOS SDN Controller να διαχειρίζεται τη κίνηση των δεδομένων όταν προκύπτουν ορισμένα συμβάντα που επηρεάζουν την αρχική δομή του δικτύου.

Ως προς τα αποτελέσματα των μετρήσεων που διεξάγονται, παρόλο που φαίνεται το θετικό αντίκτυπο που θα έχει η ενσωμάτωση των νέων τεχνολογιών στην απόδοση των δικτύων κινητών επικοινωνιών, υπάρχουν ακόμα ορισμένα επιμέρους ανοικτά ζητήματα τα οποία χρήζουν περαιτέρω έρευνας από τη πλευρά των μελών της τηλεπικοινωνιακής κοινότητας ώστε να μην υποσκαφθεί τελικά το αρχικό όραμα της καθολικής λειτουργίας όλων των κινητών συσκευών υπό μία ενιαία ομπρέλα.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Δίκτυα $5^{ης}$ Γενιάς (5G)

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: 5G, Software Defined Networks, Network Functions Virtualization, Network Slicing, Cloud Radio Access Network, ONOS

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Professor Mr. Lazaros Merakos for the confidence he has shown in me, as well as for the overall contribution of his valuable knowledge to this thesis.

I would also like to extend my warmest thanks to the other members of the Examination Committee, and especially to Dr. Dimitrios Tsolkas, for the help and guidance he offered me throughout the preparation of the thesis, with constructive comments and helpful observations.

Finally, I would like to thank my family, my friends and colleagues, particularly from the National and Kapodistrian University of Athens, for the support they provided me throughout my studies.

CONTENTS

LIST OF FIGURES

LIST OF GRAPHS

LIST OF TABLES

PREFACE

This thesis was drawn up at the Department of Informatics and Telecommunications of the National and Kapodistrian University of Athens as a master thesis in the context of the M.Sc. program “Communication System and Networks” of this department.

1. INTRODUCTION

The utility of communication in people's lives has always been a crucial factor, contributing to the evolution of humanity. Today, the need for communication and information exchange has become vital, since the modern living sets very challenging requirements for the infrastructures that will support every aspect of human activity. Practically, it has been emerged the need for the support of continuous mobility and higher data rates, ability to provide custom services as well as security and privacy protection. In order to address those needs, a rapid re-shaping of the current network model was decided, leading to the creation of the 5^{th} Generation of mobile networks.

The path towards the new operating model of mobile communications networks which is under development involves a radical overhaul of how they operate at all levels. Due to the size and the importance of the entire project, it is imperative that this transition is completed covering every possible use case, while leaving room for future extensions. In order to do this, it is first of all necessary for the stakeholders to understand the background on which the new network will operate and the basic principles governing its operation.

The major scope of the present thesis is to analyze the constituents that jointly form the heart of the 5th generation mobile networks, through an extensive study of the technologies and advancements that those networks incorporate. In this view, we analyze the main components of the network that are related to the network's structure (multi-layer cells, mMIMO, D2D, UDN, MEC), the core and radio access advancements (e.g., C-RAN, E2E Slicing) as well as the management and orchestration (such as the NFV model and the SDN concept). Related Key Performance Indicators (KPI) are also presented, complemented with an overview of the pathway towards launching 5G in Europe.

The contribution is completed by simulating, operating and evaluating the transport network of the end-to-end chain that composes the 5G network. We adopt characteristics of a real network (in terms of deployment) and we utilized the SDN concept to conduct KPI measurements. To evaluate the network's performance we used ONOS in the role of the controller and assumed network nodes connected with either single links or double links, where and random link failures are enabled.

The positive impact of the experiments carried out can be reflected in the process of setting up transport networks where there is a need for guaranteed network performance in terms of bandwidth and high reliability. It was found that this is perfectly possible by the use of duplicate links in the connections between the nodes of the network, without requiring any special actions for its configuration and without an extra computational cost as regards its management. In addition, it has also been shown that the implementation of such a network can be made by using the ONOS Controller, which provides the network with the capability and flexibility required to adapt to the various conditions affecting its mode of operation.

2. THE 5G COMMUNICATION PARADIGM

2.1 Introduction

People's need for more mobility, along with the passage of much of the human activity from the natural to the digital world, have caused a massive increase in mobile data traffic, which, according to studies, is expected to multiply by 100 times in the next decade. This effect is enhanced, if one considers that apart from the classic People-to-People (P2P) type of communication, more and more Machine-to-Machine (M2M), People-to-Machine (P2M), Machine-to-People (M2P) communications are occurring, with heterogeneous features and demands (from anywhere, at anytime, to transfer anything by any terminal kind) within the framework of the Internet of Things (IoT), thus making it even more imperative to have a network infrastructure to respond effectively to these needs [1] [2] [3]. If the above-mentioned parameters include the natural constraints arising from the lack of additional available radio frequency spectrum for mobile communications and from the need to improve the sector's environmental impact (energy consumption mitigation), then a fairly difficult crossword for the network operators is composed [2].

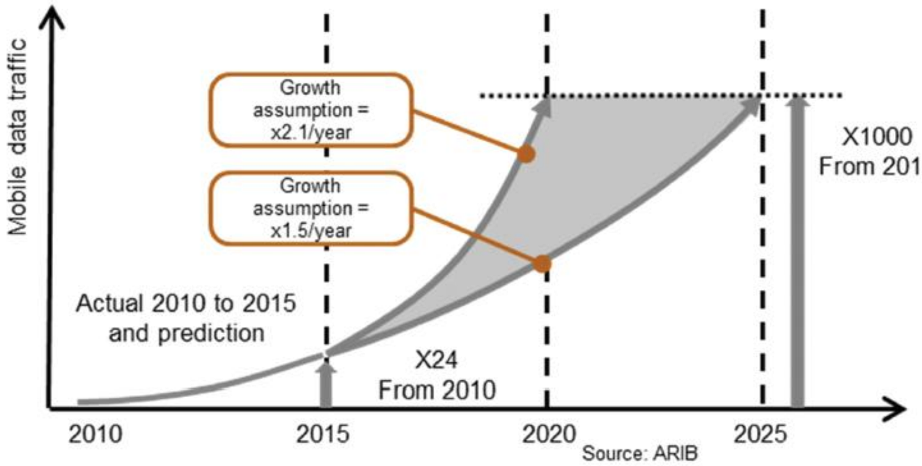


Figure 1 : Expected increase of mobile data traffic [4]

The goals of 5G, as set by the research community and vendors, include data rates of 1 - 10 Gbps, i.e. ten times more than the ones of 4G, diminished round trip latency time to 1/10 of the corresponding round trip latency time of the previous generation (1ms is expected in 5G) and a comparatively higher bandwidth capable of servicing the huge number of connected terminals expected to be in a region (IoT). In addition, network availability and coverage are expected to reach 100% while reducing the energy consumption required by both the network nodes themselves and the users' devices, at 90%, essentially materializing a green network [5].

5G, in order to effectively address these challenges, will need to redefine both the current operating models and the business and service delivery models of the whole ICT field. This will of course be caused by the transition from a traditional Client-Server concept to an Anything-as-a-Service concept. At this new concept, the variety of network terminals (mobile phones, tablets, vehicles, machines, sensors, robots, etc.) will be network nodes of an ecosystem that will produce, consume and exchange information for the purpose of providing services. Through this logic a significant part of the traffic and information processing is shifted to the edge of the network. As a result, the exchange of data between the devices for the operation of their applications does not depend solely on the availability of a centralized network infrastructure but is now based on distributed ad-hoc networks that are created on the basis of the availability of information and the proximity of network devices [5] [6].

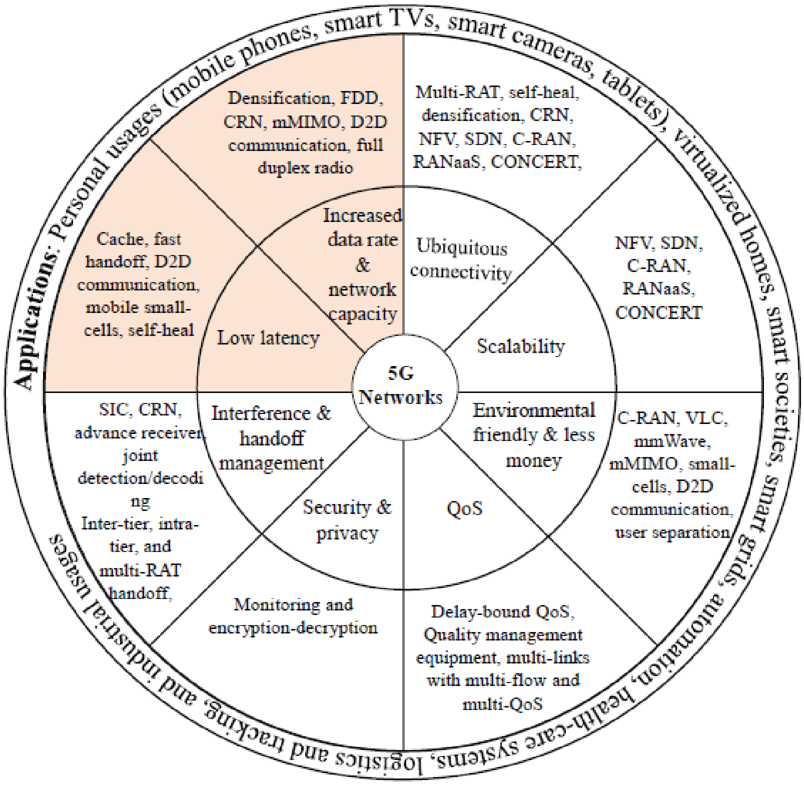


Figure 2 : Challenges and candidate solutions for the evolution of mobile networks [7]

Regarding how the network works in an abstract view due to the heterogeneity of the nodes that compose it and the huge range of use cases it should support, the appropriate operating mechanisms are required to allow each different type of equipment - node to make a flexible selection and use of those network infrastructures (both at a software and a hardware level) required to meet its requirements. This inherently heterogeneous nature of the network is another challenge for network operators, as each node may have completely different network requirements at a time. For this reason, the International Telecommunication Union (ITU) has defined different basic types of mobile networks services (enhanced Mobile Broadband – eMBB, massive Machine Type Communications – mMTC, Ultra-Reliable and Low Latency Communications - URLLC) in which terminals are classified according to their needs in terms of latency sensitivity, bandwidth, connection reliability, connection density and other parameters. Therefore, there is a new model of operation (that of differential services), where the network operator enables applications to choose at any time which elements of the network infrastructure (technologies, infrastructures, etc.) they will use in order to provide their services (network-as-a-service) and even on-demand [8]. This feature gives a user-centric character to the network where, unlike previous generations (3G, 4G) [9], network capabilities can now be dynamically readjusted and developed (adds flexibility and scalability elements) on the basis of the requirements that exist at a time [10].

Table 1 : Communicational needs of different service types [11]

Service type	User experience data rate	E2E latency	Mobility
eMBB	DL: Up to 1 Gbps UL: Up to 500 Mbps	10 ms	On demand, up to 500 km/h
mMTC	Low (typically 1 – 100 kbps)	Seconds to hours	On demand
ULC	DL: Up to 50 Mbps UL: Up to 25 Mbps	1 ms	Up to 500 km/h

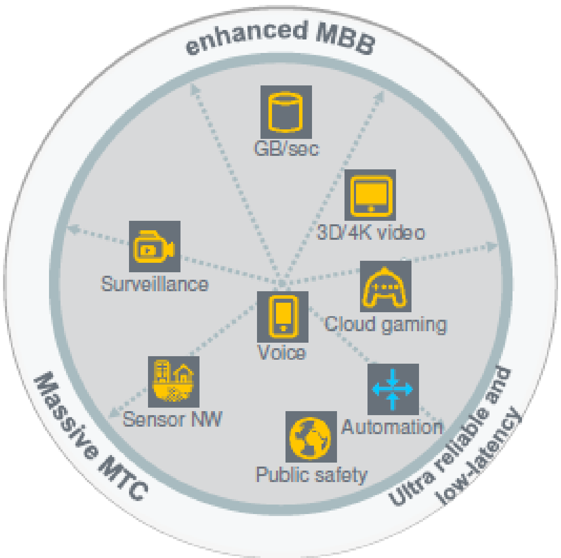


Figure 3 : Basic types of mobile network services [11]

As regards the business models of the operators, the implementation of 5G is going to profoundly change their function, since the economic benefit is shifting from the exclusive provision of infrastructure (hardware) that is today to the provision of innovative applications - services (software) of the network itself towards end-users [10]. It essentially causes a shift to a "pay-per-use" model or, in financial terms, a shift from CAPEX to OPEX [6]. In addition, it is commonly accepted that 5G will significantly contribute to the reduction of network operator costs as it maximizes the performance and the utilization of the equipment [12], while, at the same time, the use of common hardware for the operation of the network allows for the reduction of the required funds both for its acquisition and its maintenance as well as for the development of new network applications [1] [6] [13].

2.2 The evolution of wireless networks architecture in 5G

Implementation of the set goals can be achieved at the implementation level through the adoption of a new architecture model and the use of a mixture of efficient and intelligent tools and methods, such as the concepts of Software Defined Networks (SDN), Network Function Virtualization (NFV) and Network Slicing allowing network to orchestrate and operate billions of terminals with less predictable traffic patterns than in past times.

As regards the new techniques that are to be used, SDN, through the separation of Control Plane (CP) and Data Plane (DP), enables real-time, automated and low-cost customization and software (centralized) management of the network's functions (routing, mobility management) by treating them as micro-services [14] [15], essentially transforming the network's setup into a programmable infrastructure [16]. Correspondingly, NFV, using virtualization techniques, enables traditional network functions (routing, switching, etc.) to be implemented as software that can run on independent hardware, thus separating physical network equipment from network functions [17]. As regards Network Slicing, this is a collection of capable configured network functions, network applications and cloud frameworks that can be considered as a slice of the network that can only serve a particular use case [14].

In combination with the above techniques, there are structural changes at the lower levels of the network, where the way of accessing and managing the spectrum through an additional functional split (Control Plane – User Plane) and of the use of cloud techniques such as Cloud-Radio Access Networks (C-RAN), where all operating parameters can be adjusted ad-hoc according to the needs at any given time, thus contributing to the improvement of the overall 5G performance, is optimized [18]. In addition, there are changes to the air interface and the network topology by introducing new technologies in the context of 5G New Radio (5G – NR) standard, such as the Space - Beam Division Multiple Access Techniques (SDMA / BDMA) and the use of beamforming directional antennas for transmissions. Combined with the above, massive Multiple Input Multiple Output (mMIMO) techniques are applied, while, from a topological perspective, models such as Ultra Dense Networks (UDN) and Multi-Layer Cellular networks redefine network access both in terms of the protocol that is being followed and of the location of the base stations that are the gateway to the grid [5].

5G harmoniously combines all of the building blocks and the varied innovations that make it part of a common umbrella and aspires to play the role of a ubiquitous network that will integrate the full range of wireless communications (of every form), providing a high Quality of Experience (QoE) to users [6].

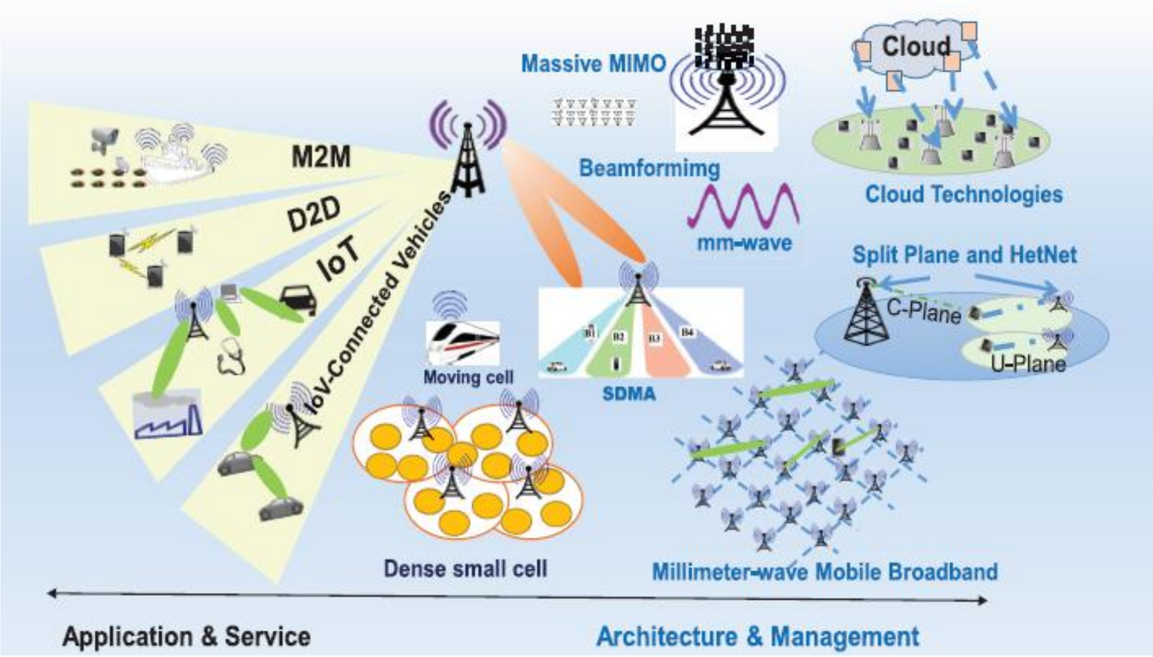


Figure 4 : Variety of techniques used in 5G networking modelKey elements of 5G architecture [5]

2.3 Towards launching 5G in Europe

2.3.1 Background

In order to complete the research and the final deployment of the 5G so as to make it commercially available to the general public, partnerships and workgroups of all the stakeholders (academia, researchers, national / regional / international regulatory authorities, standards development organizations, component manufacturers, vendors - integrators, network operators, telecommunication service providers, vertical application developers) that make up the community of development of the new generation of networks have been created at a global level (EU, USA, Brazil, Japan, China, South Korea). Their main goal is to jointly create a roadmap for completing the development of the 5G both at the operational level, in terms of technologies, and at the business level, creating a sustainable development model [19] [20].

The individual areas the 5G roadmap focuses on is promoting R&D initiatives by providing the necessary funding, introducing an open model for managing technological achievements so that knowledge is available to the community and facilitating the further development of the industry strengthening the competition and the co-operation of the participants with the aim of maximizing the exploitation of the research results and, finally, drawing up strategic plans for the use of new technologies [19].

In addition to the above mentioned aims, in case of the 5G Infrastructure Public Private Partnership (5G PPP) which is "*a joint initiative between the European Commission and European ICT industry (ICT manufacturers, telecommunications operators, service providers, SMEs and researcher Institutions)*" [5gpp.eu] the main goal of the relevant roadmap is to support the European community (industry, researchers etc) in order to conquer a leading position in the 5th generation networks sector in both research and commercial levels. Moreover, the goal is, after evaluating the contribution of the changes introduced by 5G technologies in the operation of vertical sectors, to set out a timetable for the adoption of the new model of mobile networks in Europe [20].

2.3.2 Funding for the development of 5G

As the implementation of the 5G in Europe is primarily conducted by private operators of the ICT field, it is helpful, at central government level, to develop policies that will enable it to be properly developed and supported. It is therefore useful to define a regulatory framework which, in addition to the basic rules required for the sound development of new technologies (mainly through research and investment programs), will include, inter alia, the necessary technical specifications required for the interoperability of the system with corresponding projects worldwide.

To this end, the European Commission (EC) through the European Fund for Strategic Investments (EFSI) or other financial sources, and in particular through the Horizon 2020 program, contributes to the funding of private operators to speed up the launch of 5G services while, at the same time, the same project is carried out by corresponding programs of national bodies of the EU member states. In addition, the European Space Agency (ESA) has a share in the funding of such programs [21].

2.3.3 5G implementation plan across the European Union

Based on the strategic plan already developed by the European Commission (5G Action Plan) as early as 2016, in cooperation with all stakeholders in the industry active in the 5G community, some guidelines and milestones are set in order to accelerate the development of the new generation of networks [21]. The main concern is to have a concerted effort in time for the actions required by all players of the telecommunication market so as to clarify to the operators both the 5G framework and operating rules as well as the market's need for investment in this sector. An additional goal is to eliminate the gap in the development of networks between the EU and the USA, something that had happened in previous years with the late development of the 4G in the EU member states, as opposed to the USA where the respective actions occurred faster [20] [21].

In the light of these guidelines, the following critical procedures have been defined, which should be followed and implemented on a timetable basis by all competent members of the ecosystem so as to ensure the success of the entire project:

1. Establishing a joint 5G infrastructure development plan across Europe in order to launch pilot/test applications in selected areas in 2018 [20] [21].

2. Commercial disposal of 5G services in at least one major city of an EU member state by 2020 [20] [21].

3. Wide coverage of EU's urban areas and main motorways with 5G services by 2025 [20] [21] so that 5G active connections account for 29% of total mobile phone connections in Europe [20].

4. Disposal by the relevant regulatory authorities (until 2019) of the essential portion of the radio spectrum necessary for the operation of the network and its supply, as soon as possible, with the other bands required for all of 5G services [21].

5. Promotion of pilot and test applications in large urban environments in order to assess their operation in practice and to have integrated marketing standards [21].

6. Supporting investments related to the research and development of 5G infrastructures (such as high capacity backhaul networks and multi-tier cells implementation) across Europe, with the contribution and co-operation of all players involved in the project both at public (financial community, European Investment Bank) and private level (vendors, network operators, startups, third-party investors) [21].

7. Simplification of the bureaucratic procedures required for the installation and sustainable operation of access points of small cells across European territory by creating a single regulatory plan [21].

8. Modeling of the 5G operating and development technologies under the leadership of the European Commission so as to ensure both its sustainable development and interoperability with the corresponding implementations of other members of the global community. This process includes, among other things, the establishment of collaborations between different industries (education, manufacturing, health care, smart grid, entertainment, automotive, smart cities, aerospace, threat response, cloud/IoT fog) on the establishment of operating standards for vertical applications [19] [21].

2.3.4 Trials Timeline

Ever since early 2018, pilot applications have been deployed in several European cities in order to accelerate the introduction of 5G technologies in the telecommunications market, starting in 2020 in major cities and by 2025 in urban areas and across popular transport routes [22]. Their contribution is extremely important as, through the evaluation of their operation, necessary data are collected that act as feedback in the process of completing the development of the new generation of mobile communications networks and optimizing their services. All stakeholders (vendors, operators, vertical players) are involved in this process so as to verify in practice all the theoretical capabilities of the 5G, as those have been designed in the vision of new networks [20].

At the same time, the competent authorities of most European countries, following the strategy documents drawn up by the European Commission in 2016-2017, and having conducted relevant public consultations, have already developed some national strategy plans or roadmaps with targets and key elements for the 5G launch, while in countries where the process is not complete (Cyprus, Malta, Hungary, Portugal, Croatia, Lithuania), it is expected to be completed by the end of 2019 [22].

Based on these plans, the 5G development community has planned and carried out technical tests, vertical pilots and pre-commercial trials in several stages. Following a bottom-up approach, the lower levels include small private 5G trial projects and demonstrations (category 3); in the next stage, the actions focus on vertical trial applications (category 2), while the upper layer (category 1) includes large-scale pre-commercial trials in a major cities level involving end-users and cross-border corridors culminating in a 5G run test during the UEFA EURO 2020, which is expected to be a milestone test at a European level [20] [22].

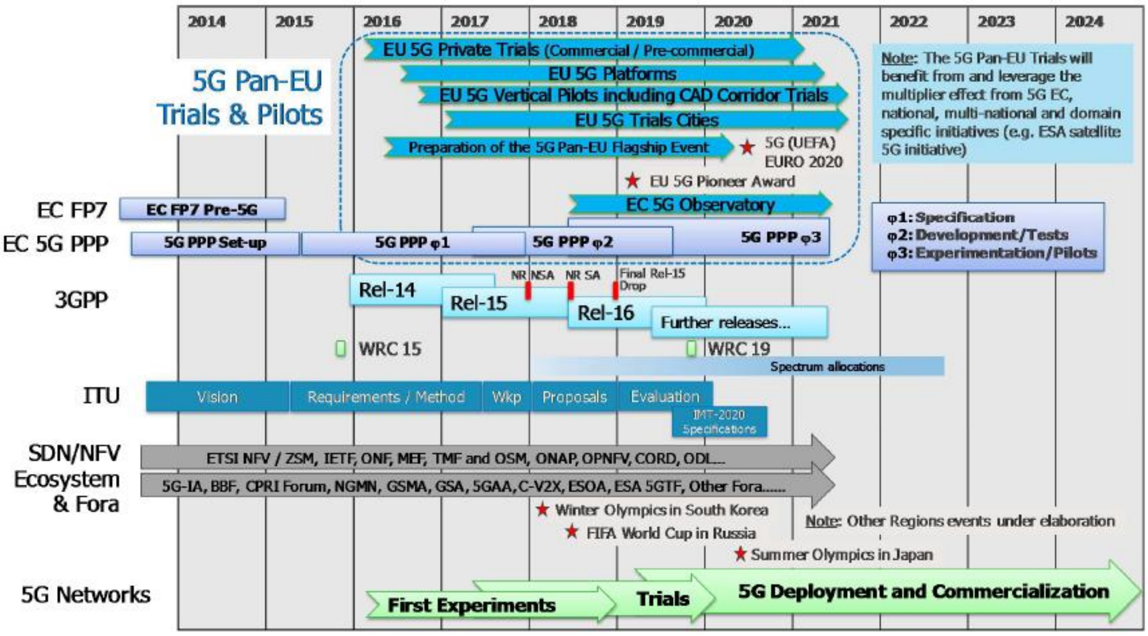


Figure 5 : 5G development plan across Europe [20]

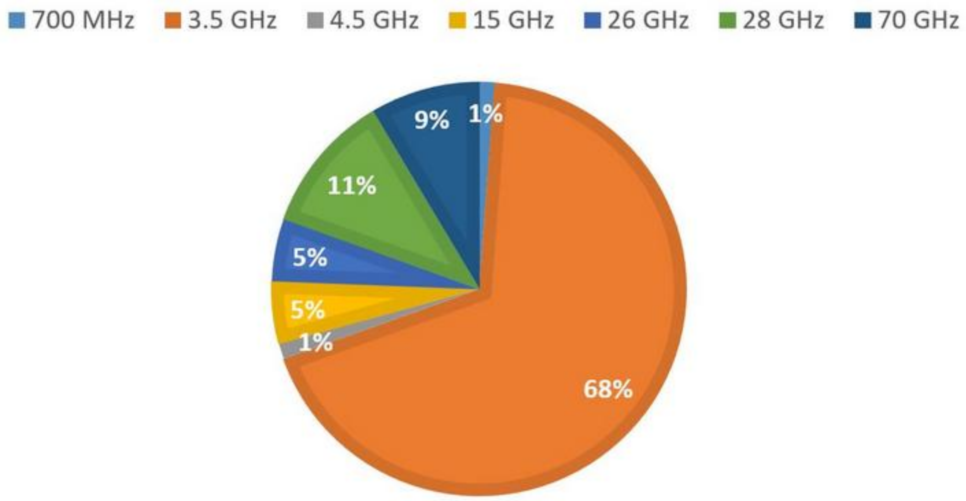
To monitor developments, the online portal *5gobservatory.eu* has been created under the auspices of the European Commission, which includes a detailed overview of progress in the 5G sector (state-of-the-art), focusing mainly on events that concern the European territory [20]. This tool contributes to the monitoring and evaluation process of the actions that take place for the implementation of 5G networks (market plans and developments, definition of major strategies, large scale trial deployments regarding infrastructure components, products and services, innovations, standardization issues, spectrum allocations, public initiatives, availability and quality evaluation of the new services) in order to ensure that the 5G Action Plan is implemented correctly and in time [23].

2.3.5 Early deployment parameters (Spectrum allocation – Standardization – NE availability)

Successful completion of 5G deployment requires some critical parameters from stakeholders, such as spectrum allocation as well as the standardization and availability of appropriate equipment in order to meet new networks' requirements [21].

As far as the first parameter is concerned, in this early phase of 5G development, it is particularly useful to clarify the scope for determining the radio spectrum bands that are to be used by the network. As it has already been mentioned in a previous chapter, the new technologies used in the 5G dictate, among other things, some particular requirements in terms of this operating parameter. The multi-level spectrum scheme used for network operation and the requirement for a single large bandwidth in the respective areas of the spectrum set the baseline frame for the settlement of the whole issue. This process is particularly necessary at this stage for vendors, network operators and developers to be properly prepared and for a reasonable period of time to perform extensive tests on network operation [20]. On this basis, the competent national authorities of the countries of the European continent should, in cooperation with the European Commission, complete the appropriate processes (spectrum release, setting auction parameters such as price and duration of the lease) for the spectrum allocation and licensing in order to avoid any obstacles during the development of the network. One such important step in this direction is the agreement reached in 2017 between all stakeholders and regulators which concerns the inherent and as large as possible harmonization of the 5G spectrum allocation between all EU countries so that there is full interoperability of services across a cross-border level, to be able to apply economies of scale methods, to have a common response to problems regarding the air interface and, ultimately, to further facilitate the network deployment [24] [21].

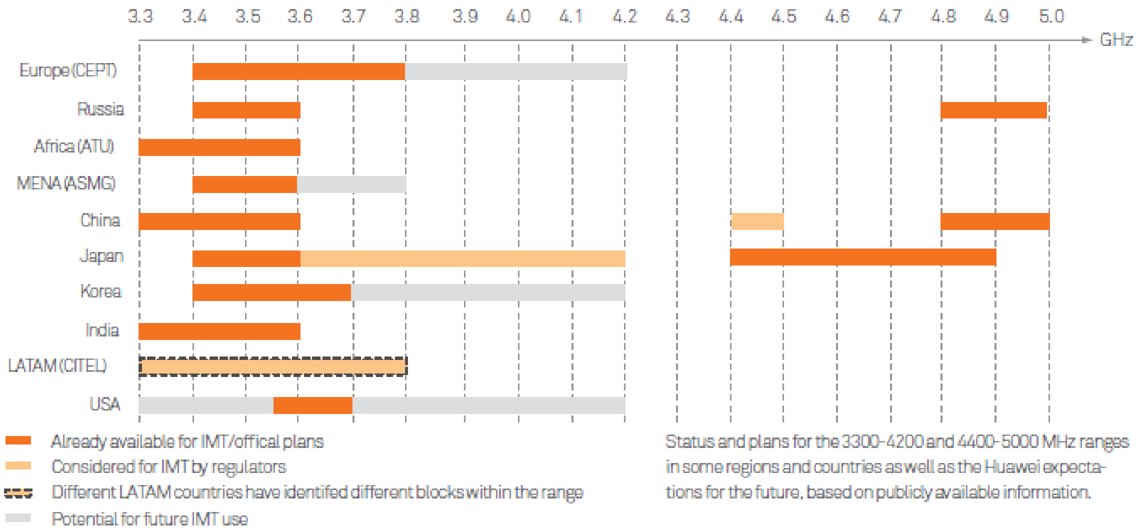
More specifically, initially, particular emphasis has been placed on addressing the relevant issues concerning bands ranging in frequencies of 700 MHz, over the 1 to 6 GHz and the upper part of the 26 GHz (5G pioneer spectrum bands). This is because of the particular interest that the community has regarding these frequencies, as these sections are expected to operate a large part of the 5G applications, at least in its early stages of operation. According to the data related to the use of the spectrum for pilot applications up to December 2018, the majority of developers focused on the use of bands around 3.5 GHz, and, at lower rates, frequencies were used in the 700 MHz range and between 26 and 28 GHz [21] [23].



Graph 1: Spectrum bands used for trial scenarios [22]

With regard to bands with frequencies in the range of 3.4-3.8 GHz, in most countries spectrum delineation procedures are in a mature phase, with auctions for operators licensing having already been announced and being at an advanced stage (Belgium, Bulgaria, Germany, Lithuania, Luxembourg, Sweden), in some cases having already been completed (Austria, Czech Republic, Finland, Hungary, Ireland, Italy, Latvia, Norway, Spain, UK) while in the rest of the countries, although there is some elementary planning, these procedures have pending (Croatia, Cyprus, Denmark, Estonia, Greece, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia). In addition to the above, it is worth highlighting the fact that in some countries such as

Belgium and Greece, where procedures have not yet been completed, trial licenses are granted to facilitate 5G Trial Deployments [20] [23] [25].

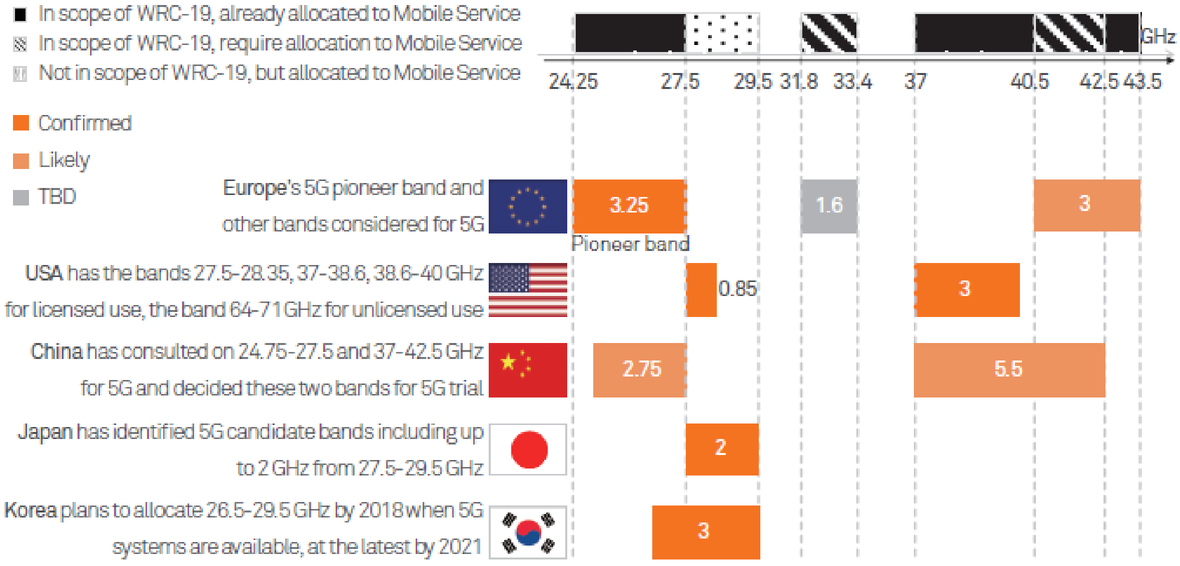


Graph 2: Maturity level of spectrum delineation procedures in range of 3,3 - 4,2 GHz and 4,4 - 5 GHz frequencies [24]

Spectrum assignment on a national basis is followed by frequency bands for 700 MHz frequencies. In this particular area, some countries have already completed the procedure (Belgium, Finland, France, Germany, Italy, Sweden) and others are expected to do so by early 2020 [20] [23].

As for frequencies exceeding 6 GHz, the timetable according to which the corresponding actions will take place has not yet been clarified, as the corresponding procedure is an important challenge for the community, since the future use case scenarios of the 5G must be envisaged as best as possible [21]. However, some countries (Ireland, Italy) innovating in this sector have already allocated bands in this area of the spectrum, thus allowing interested parties to conduct tests of their applications that will operate at these frequencies [20] [23].

Moreover, it is worth noting that some countries have advanced or are expected to do so on the basis of the plan they have made in order to make available other frequencies beyond the basic range for the purposes of the 5G (Belgium – 1,5 GHz, Bulgaria – 1,5 / 2,5 / 2,6 GHz, Czech Republic – 450 MHz, Denmark 800 / 900 MHz & 2,3 / 2,4 GHz, France - 1,5 GHz, Hungary - 1,5 GHz, Lithuania – 3,8 - 4,2 GHz, Luxembourg 900 MHz & 1,5 / 1,8 GHz, Malta – 800 MHz, Netherlands - 1,4 / 2,1 GHz, Portugal – 450 / 900 MHz & 1,5 / 1,8 / 2,1 / 2,6 GHz, Sweden – 1,5 / 2,6 GHz, UK – 1,5 / 2,6 GHz) [20] [23].



Graph 3: Spectrum assignments for 5G trials in mmWave range [24]

The other parameter that is key to globally ensuring the proper development of the 5G in terms of interoperability, availability, security, openness, scalability and flexibility of its operation is the standardization of its technologies, services and applications. This will ensure compliance with minimum network operating rules so as to provide the best possible services to end users and to further their development [21].

Given that also in this case the stakeholders who will have to commonly agree together in order to define the necessary frameworks are many in number and heterogeneous, the entire procedure is a real challenge for the industry as there should be a common component of all proposals and not scattered solutions which eventually may have conflicts between them [21]. An important step in this direction is the standardization carried out by 3GPP and its acceptance by the equipment manufacturers, which ensures the harmonization of the operation of the network on the basis of the agreed basic principles, achieving the fundamental objective of non-diversification of the parameters and functionalities of the network between different countries, which would otherwise be detrimental to the universality of the entire project [20].

It is also important for the process to be in line with the implementation timetables of the new technologies and not to hinder their development, while the standards agreed on the present technologies should cover existing use cases on the one hand and have a flexible structure on the other hand in order to ensure the coverage of future extensions and innovations [21].

Finally, an important element in this 5G development phase is the creation of compatible equipment by vendors, according to the agreed standards. The intense activity and growing interest in trial and pilot deployments in a number of vertical sectors are pushing component manufacturers to create network equipment that can cover all of the available spectrum of 5G operations at a frequency level (ranging from 700 MHz and 90 GHz), as well as at a functional characteristics level (SDN, NFV, C-RAN, mMIMO, D2D, etc.). It is noteworthy that, in this area, European manufacturers (e.g. Ericsson, Nokia) have a large shareholding of more than 51% by undertaking procurement and installation of equipment around the world (e.g. Australia, China, India, Japan, South Korea) [20].

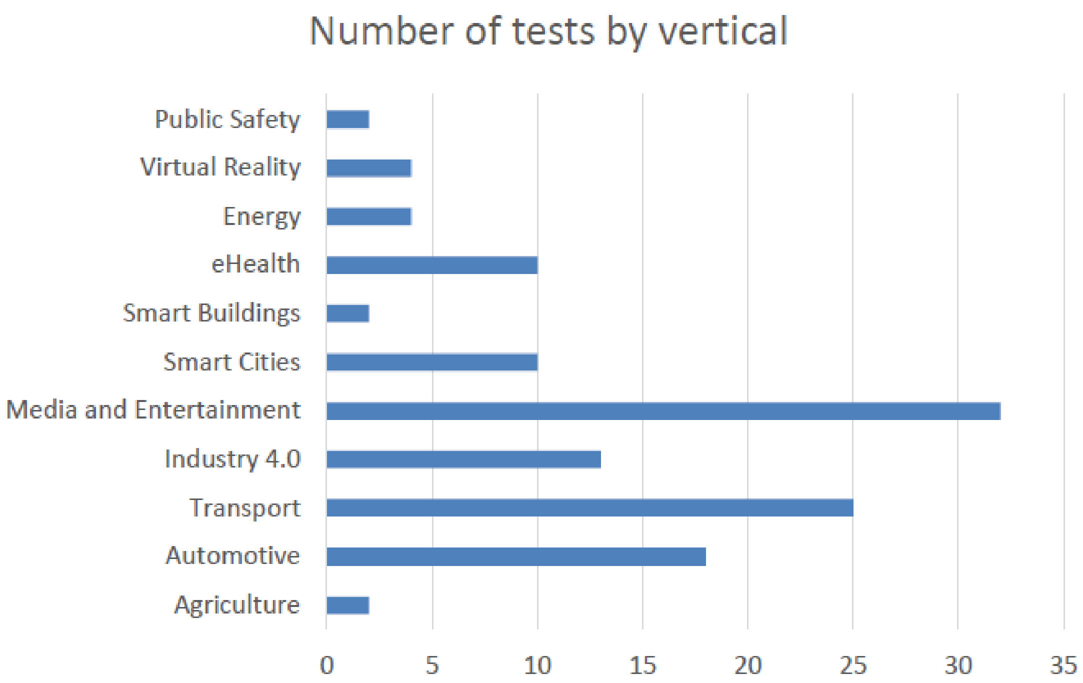
In an effort to overcome these challenges, the European Commission, having the role of coordinator, sets key targets and timetables for all members of the community and industry stakeholders to ensure timely availability of standards by the end of 2019. This will allow the test and then the commercialization of 5G services. It also calls on all parties involved to work together to create the appropriate framework to ensure the smooth functioning of networks, taking into account all possible scenarios of deliberate or accidental anomalies so as not to hinder the further development and promotion of the new generation of networks. In addition, through the promotion of cooperation, the aim is to strengthen innovation, especially in the vertical sectors where there is considerable room for the creation of intelligent applications and services [21].

2.3.6 Pre-commercial Trials

According to the original design, the 5G trials have 4 phases; the first trials began in 2017. The first phase aims at the preliminary theoretical study and evaluation of the technologies to be used in the project in order to provide the desired functionality and performance of the network in a small scale area (Technological Trial Level). The next step is to install a pilot but necessary infrastructure (if possible from multiple vendors) for a limited test of the project in a tightly confined of 1-2 sites with the involvement of real customers. (Friendly User Trial Level). The third phase of the project (which runs until mid-2019) includes, among other things, the installation of better equipment that will meet the standards developed to test the concept in a wider range, covering the needs of skilled users-testers (engineers) in 2 to 15 sites of the city (Friendly User Trial Level Cities). Finally, at the latest stage, extensive large-scale (in over 15 sites) testing is expected at a higher and more comprehensive level in terms of technologies and service provision, with the help of a large number of real customers and the use of commercial terminals in a pre-commercial base, which will lead to the operation of the final commercial solutions of the 5G (Pre-Deployment 5G Cities Level) [20].

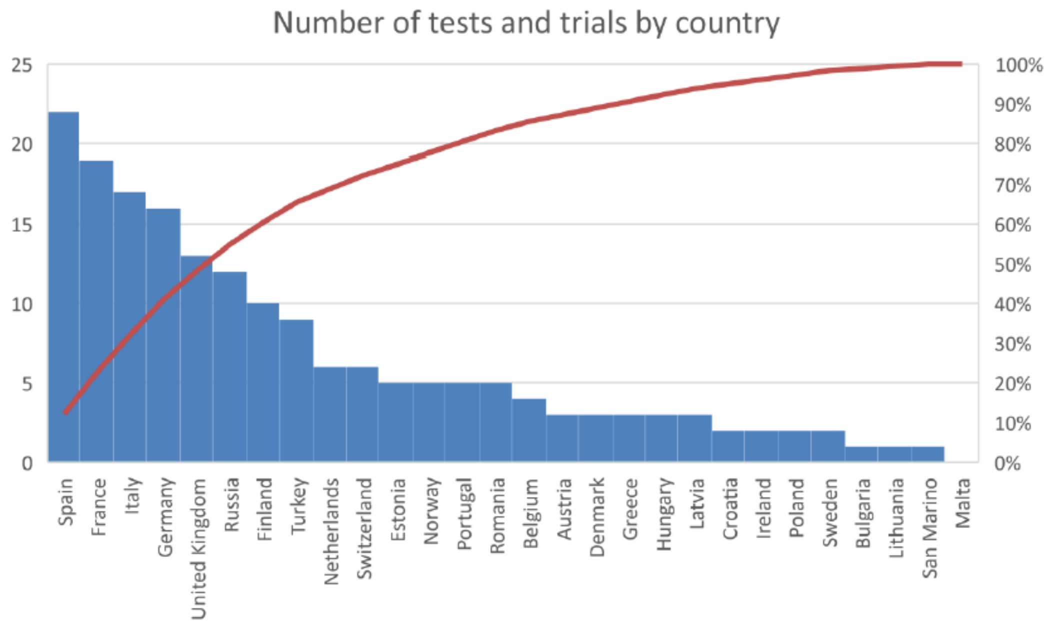
Most network operators, in order to be timely prepared to deliver 5G services, have already made their plans and have begun preparations for the commercialization of the 5G, announcing their timetables for early deployments and pilots, technical tests and commercial launches. Indicative examples include announcements by Finnish network operator *Elisa*, who announced availability of the 5G in Tampere and in the Estonian capital (Tallinn) as early as June 2018, as well as the Austrian provider *T-Mobile* that provides 5G commercial services to friendly customers since March 2019. Similar actions have also taken place in Switzerland where *Sunrise* and *Swisscom* network operators announced the commercial release of the 5G since March and April 2019 respectively, with the latter covering 100 parts of 50 different regions in the country, targeting to a population coverage of 90% by the end of the year. Similar developments have also taken place outside of Europe, with *Verizon* in the USA, which, since September 2018, has piloted the operation of the 5G using standard devices and *AT&T* with similar actions almost at the same period; *Sprint* and *T-Mobile* followed shortly after. No reports will be received for comparable efforts that have been made in Japan, China, South Korea and Qatar marking the new era beginning in the world of telecommunications [22].

According to the up until now available data (Q1/2019), extensive testing has been performed in all of the 28 countries of the EU, mainly at a private level. The number of trials amounts to 147, while, including the respective trials in Russia, San Marino, Norway, Turkey and Switzerland, their number reaches 180. Almost one-third of these (58) relate to purely technical tests, while the rest are related to service trials in different vertical sectors. The sector with the largest share of trials is so far "Media and Entertainment", followed by "Transport", "Automotive" and "Industry 4.0", thus reflecting the strong interest of the community for applications that will contribute to the development of these sectors [22].



Graph 4: 5G applications tests per vertical domain [23]

At the level of the EU member states, the countries where most of the 5G applications tests have been implemented, according to the official figures of the European Commission, are Spain, France, Italy and Germany, accounting for 40% of all the corresponding actions that have taken place in Europe [22].



Graph 5: Total 5G applications tests per EU country [23]

With regard to the 5G Action Plan targeting the development of 5G pilot applications in at least one major city of each EU member state (category-1 trials) by 2020, 38 cities have been proposed in which 63 5G pilot deployments will take place within the 5G Infrastructure Public Private Partnership framework, covering almost the entire range of vertical sectors [22].

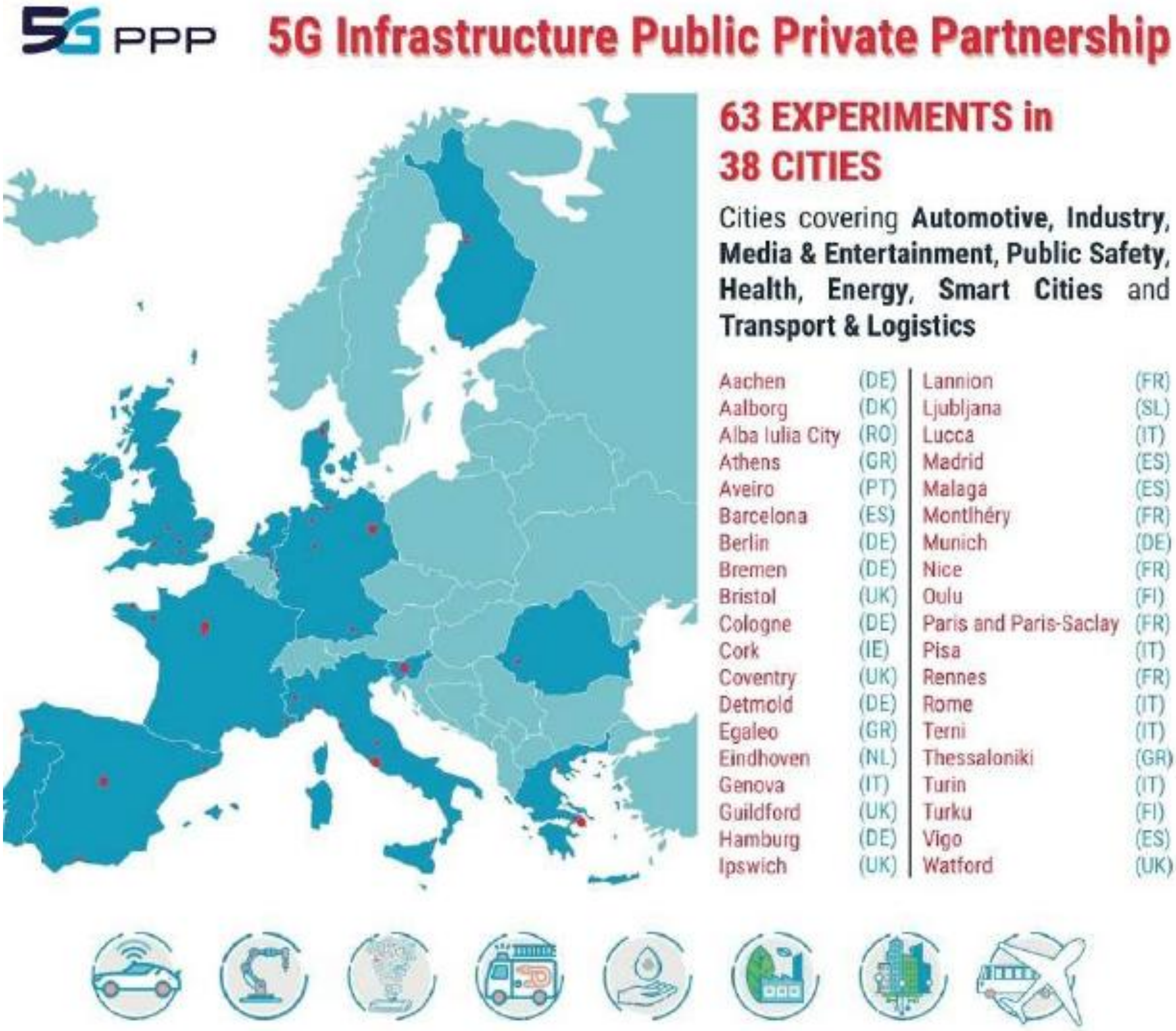


Figure 6 : 5G-PPP cross-domain trials across EU cities [20]

In addition to the above private initiatives, the public sector is making a remarkable contribution to promoting the entire project by providing support services for conducting the necessary tests. In this direction, in selected cities in EU member countries, facilitations with the provision of required resources are provided from local providers to those who want to test 5G (technological equipment, vertical use cases environment, test environment with real customers), as well as general support for the entire procedure of conducting the tests, in exchange for the faster and more complete operation of the new generation of networks in the region [22]. Besides, the cities selected in this early test period are also expected to be used during the initial 5G deployment phase in a full commercial deployment (official start of 5G deployment) and, if this is not feasible, 5G at normal operating rates will take place in adjacent areas [20].

This category includes cities such as Ghent (Belgium), Tallinn (Estonia), Espoo, Ouku (Finland), 5 cities in France, Berlin (Germany), Kalamata, Patras, Trikala, Zografou (Greece), Budapest, Zalaegerszeg (Hungary), Bari, L'Aquila, Matera, Milan, Prato, Turin (Italia), Amsterdam (Netherlands), Gliwice, Łódź, Kraków, Warszawa (Poland), Aveiro (Portugal), Barcelona, Madrid, Malaga (Spain), Umea, Kista/Stockholm (Sweden), Bristol and London in United Kingdom [22].

These tests, which are being developed in large-scale environments, are expected to provide useful feedback for the operation of the larger 5G ecosystem (different from the one generated by private initiatives) since the data gathered to assess the operation of the network will include data as representative as possible of the real operating conditions (real users, large scale area, etc.) in relation to the corresponding private projects. Areas such as access interfaces, transport, smart grids, technology platforms and open data availability can be studied thoroughly and extremely useful conclusions can be drawn for the further development of the network [22].

Along with testing at a private or at city level, some large scale trials of transnational nature are being carried out aimed at the development of applications that fall in vertical sectors of Transport and Automotive, under the Cooperative Connected and Automated Mobility framework. These tests are called "digital cross-border corridors" due to their running along roads that connect different member states and are particularly important both for the development of the new generation of networks in Europe in terms of infrastructure as well as for the development of applications in the fields of Intelligent Transportation Systems, Autonomous Cars and Cooperative Vehicles, which are expected to become highly-known sectors in the 5G ecosystem [20] [22].

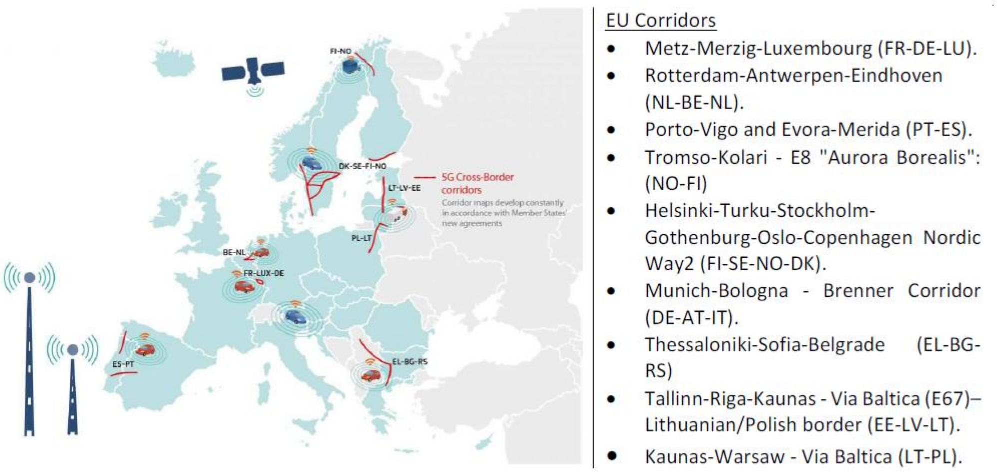


Figure 7 : EU 5G transnational trial developments (corridors) [20]

In this context, a wide range of tests are carried out with the cooperation of different countries and the involvement of stakeholders, which make up a heterogeneous set that includes network equipment manufacturers, network operators, automotive industries, transport engineers, security engineers and regulatory authorities. The framework for future systems development is co-developed in these sectors and key parameters are considered to further implement the 5G [20].

2.3.7 Notable Examples of 5G Cities

Efforts made in various European cities deserve a special mention; each one of them, in its own way, aims at a sort of universal digital transformation of the mode of operation and governance of the region, following the vision of Smart City. The aim is to provide citizens with more quality services, to increase citizen participation in administration, to enhance transparency, to encourage innovation and personal initiatives, to save resources and to strengthen competitiveness, which is a key driver of growth in modern times. Although each implementation focuses and is implemented with the corresponding priorities set by local communities in each region, the final footprint is expected to be of fundamental importance for the future operation, organization and interaction of cities [22].

In cities such as Berlin, Barcelona and UK regions such as Orkney, Shropshire and Somerset, emphasis has been put on creating the appropriate background required for the universal 5G availability across cities and the provision of innovative services in order to achieve their "transformation" from conventional to smart [22] [23]. Especially in the case of Berlin, the installation of macro-cell sites is being implemented at a rapid pace through the installation of mmWaves base stations at traffic lights, thus ensuring high data rates across the city, an element that is extremely useful for future applications [20].

At the 5G Infrastructure Public Private Partnership level, some private trial projects, which although developed in smaller-scale environments, are of particular interest in extremely useful case families. It should be stressed that many of them are already in a particularly mature stage of development and, in cooperation with NOs and other industry players, they are expected to soon become commercially operational.

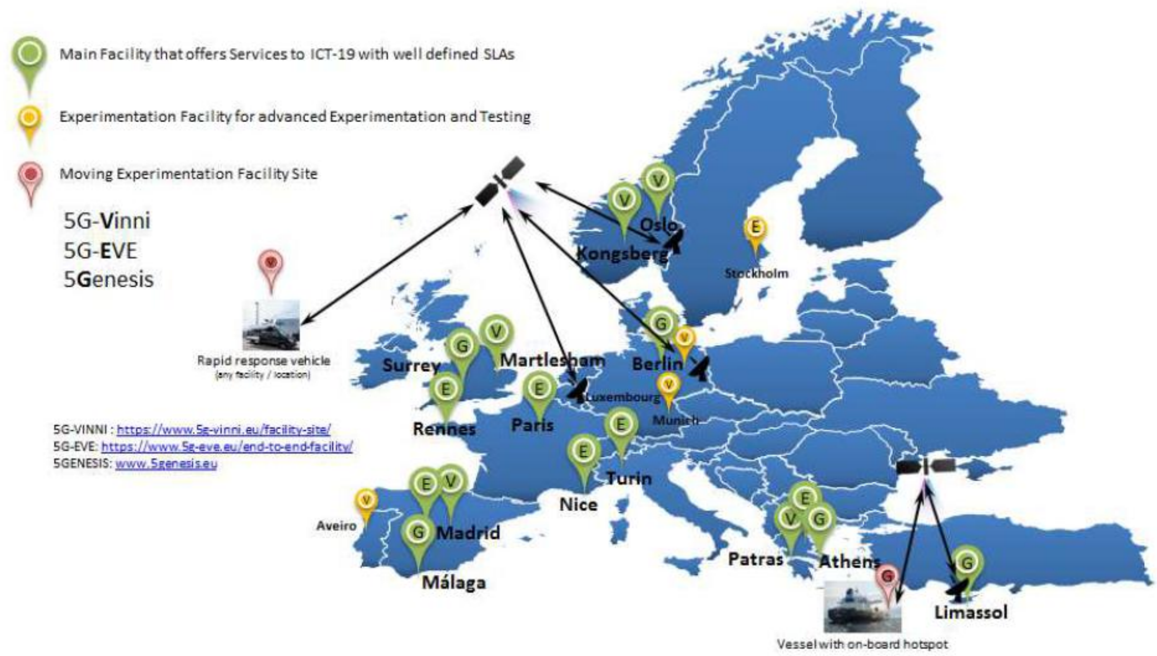


Figure 8 : Geographical imprint of small scale EU 5G trials [20]

This category includes the case of Finland, where cities such as Tampere, Oulu and certain Helsinki regions a comprehensive 5G network is available for large scale testing (The 5G Test Network Finland). With the support of the national authorities of the country and the local telco operator, the network provides the necessary infrastructure and management services required by the stakeholders (researchers, developers) to test and run pilot applications in various vertical sectors (autonomous vehicles, eHealth, public safety and energy management). A typical example is the city of Oulu where the emphasis has been put to improve the transport sector through applications that focus on modeling traffic on the city's streets and the data generated by it [20].

The Spanish national initiatives in the development of the 5G ecosystem are of great interest. Within the framework of the national networks development program, a number of innovative projects involving various vertical sectors are being implemented. Typical examples are efforts such as the one by Vodafone in collaboration with a local Hospital Clinic which is related to remote surgery, the development of applications for V2X communications between cars and bikes for collision avoidance by Telefonica and SEAT, as well as a research by Orange and Nissan regarding the study of the impact of the adoption of 5G services in the field of logistics and industry. In addition to the aforementioned private initiatives, similar actions are also being undertaken by local and national authorities, such as the initiative to install Intelligent Transportation Systems

(ITS) on a mountain road within a local park so as to operate a driver-less bus under real time conditions, as well as the promotion of actions related to technological/economical studies for the development of the 5G in industrial and provincial cities [20].

Apart from Barcelona and Berlin, where emphasis has been put on providing 5G with maximum coverage, similar actions are also being implemented in Italian cities like Turin, where TIM has installed small cell antennas in many parts of the city to carry out a variety of tests. For example, tests are being conducted in the wider area on autonomous driving, the use of drones to monitor environmental conditions, the exploitation of robotics applications in the industry and, more generally, the integration of IoT techniques into the city's operation and governance. Similarly, in cities such as Bari, Matera, Prato, L'Aquila and Milan, the joint venture of TIM and Vodafone aims to offer full 5G capabilities to the available radio-spectrum that they have committed, focusing on the provision of smart-services for almost every vertical sectors. Similar actions are also taking place in Rome, where emphasis is put on the provision of increased tourist experience services through the use of augmented and virtual reality techniques [20].

Pilot testing of 5G technologies is also carried out by providers such as Orange, Bouygues Telecom and SFR in French cities such as Lille, Douai, Marseille, Bordeaux, Nantes, Toulouse and areas such as Orange Gardens, the Linas-Monthlery Circuit and in the Paris Orange Opera megastore with a gradual increase in sites in each city where the 5G services are available for use by real customers [20].

Extensive testing is also being carried out in several cities in the Netherlands, where emphasis is put primarily on the assessment of applications of specific vertical sectors such as mobility engineering, heath, climate and cultivation rather than the testing of the 5G technologies themselves. For example, in the region of Groningen, a consortium of operators (such as Vodafone and KPN), network components manufacturers, telco scientists and other stakeholders are experimenting on these areas in order to deliver high-quality services to future 5G users, while the collaborative work with the ESA in creating innovative applications with the contribution of space technology is also of great interest. In addition, in collaboration with leading companies such as ABB, Accenture and SPIE, 5G applications are being tested on intelligent administration of tasks in industrial environments using Artificial Intelligence (AI) techniques, robotic approaches and sensor networks. An example is the ability to monitor and manage Shell refinery operations through innovative 5G services. In addition, the activities regarding the field of smart cultivation, such as the one that is carried out in Drenthe concerning the use of drones for crop monitoring and the one in the sector of self-driving in the Helmond-Eidhoven region, are also of great interest [20].

The case of Greece is also worth mentioning, where in cities such as Trikala and Athens significant efforts are being made to develop innovative applications and services within the 5G test period framework. Initially, in the city of Trikala, which has a tradition of developing IT services as part of its transition to the digital era, an innovative project is being implemented that aims to directly improve the quality of life of the residents and entrepreneurs operating in the city through the services that are to be provided to them. In this direction, a pilot project is being implemented that includes innovative solutions that, as part of the Smart City concept, operate as a ubiquitous entity including IoT sensors for waste management, an automated city lighting management system, environmental management services and an integrated application that includes a number of controllers and sensors for smart farming solutions [20]. In addition, in the city of Patras, the goal of the corresponding program is to integrate all 5G sections in a unified manner so that new services are provided to citizens in order to help them solve city management problems or other local management issues [23]. Correspondingly, at selected points in Athens, such as the campus of the national research center NCSR Demokritos in Ag. Paraskevi and the area near Egaleo's local football stadium, end-to-end 5G trials are conducted to study the operation of the network in both indoor and outdoor environments under conditions of high density of users/devices per m^2. In conducting the experiments are involved parties coming from the spectrum of the 5G networks community, under the Athens.5Glink concept [20].

In the same direction as other European countries, UK cities such as Bristol and Bath are developing innovative pilot applications that focus on reconfiguring the tourism services provided by using Augmented & Virtual Reality techniques so visitors can better navigate to tourist attractions. Similar projects are being carried out by a group of stakeholders in various cities in the country such as Worcester, Kidderminster, Liverpool, Millbrook, Cumbria, Northumberland, North Yorkshire, Inverness-shire, Perthshire, Monmouthshire and West Midlands respectively, covering both the technological test part in order to achieve maximum coverage of 5G services and high speed connectivity (even along the rail network of specific areas), as well as a huge range of use case families and vertical sectors such as autonomous vehicles, industry automation, cultivation, smart health services and public safety [20].

Apart from the small and medium sized private trials and the large scale pilot schemes taking place at the level of big cities, the European-level milestone test is expected to take place during UEFA EURO 2020. The provision of 5G services in 12 different European cities (Amsterdam, Baku, Bilbao, Bucharest, Budapest, Copenhagen, Dublin, Glasgow, London, Munich, Rome, Saint Petersburg) co-organizing the football tournament is an ideal starting point for 5G society and the stakeholders in order to start testing availability, reliability, robustness, network security and the ability of infrastructure and applications to meet the demands of real users under particularly demanding conditions (e.g. high density of users in an area, burst traffic, need for real-time services) achieving at the same time high QoE and QoS. The implementation of the entire project requires the participation and cooperation of stakeholders, such as local authorities, network components industry, network operators, service providers, the organizational authority (UEFA) and stadium facility administrators while, at the level of funding, the contribution of private and national-local authorities is essential to its success [20].

At the deployment level, each host city requires an agreement with at least one network operator so as to ensure the availability of the experimental 5G services, and the regulatory authorities are supposed to complete the spectrum allocation procedures (mainly in the bands between 3.4 - 3.8 GHz and around 26 GHz band) so that applications can run in full performance and coverage level. In each city, either different 5G services can be developed depending on the needs and priorities of each local community, or similar services by following a common policy [20].

2.3.8 Commercial launch time schedule

According to the timetables and the roadmaps that each EU country has put forward, 5G services delivery is expected to start in 2019 in a limited number of countries and not to their maximum coverage (e.g. Austria, Finland, Malta, Poland, Portugal, Switzerland, United Kingdom), while most of the member states are planning to launch 5G networking in 2020-2021, as shown in the table below [22].

Table 2 : 5G commercial launch time schedule plan per EU country [22]

Country	Scheduled launch date	Operator
Austria	03/2019 01/2020 -	T-Mobile Austria A1 Telekom Austria Hutchison 3G Austria
Belgium	2020 Q4 2020 & H1 2021 2021	Proximus Orange Belgium Telenet Belgium
Bulgaria	2020 2020 Q4 2020 & H1 2021	A1 Telenor Vivacom
Croatia	- 2020 -	A1 Croatia T-Hrvatski Telekom Tele2 Croatia
Cyprus	- - -	MTN Cyta PrimeTel
Denmark	2020 2020 - -	TDC Telenor Denmark Telia Denmark 3 (Hi3G)
Estonia	2019 2020 2020	Telia Estonia Tele2 Estonia Elisa Estonia
Finland	06/2018 H1 2019 H1 2019	Elisa Finland DNA Telia Finland
France	2020 2020 2020 2020	Orange France SFR Bouygues Telecom Free Mobile
Germany	2020 - H2 2019	Deutsche Telecom Telefonica Germany Vodafone Germany
Greece	- - -	Cosmote Vodafone Greece Wind
Hungary	- - -	Telecom (T-Mobile Hungary) Telenor Hungary Vodafone Hungary
Ireland	Q4 2019 End 2019 H2 2019 -	Vodafone Ireland EIR Three Ireland Virgin Mobile
Italy	H2 2019 H2 2019 2019 -	TIM Wind Tre Vodafone Italy Iliad
Latvia	- - 2019	Tele2 Latvia Bite LMT
Lithuania	- - 2020	Bite Tele2 Telia (Omnitel)
Luxembourg	2020	POST

	2020 2020	Tango Orange
Malta	2019 - -	Melita Vodafone Malta GO
Netherlands	2020 2020 2020	T-Mobile Vodafone Ziggo KPN
Poland	12/2018 H2 2019 - -	T-Mobile Orange Poland Plus Play
Portugal	2019 - End 2019	Altice (MEO) NOS Vodafone
Romania	- - -	Orange Telekom Vodafone
Slovakia	- - -	Orange T-Mobile O2
Slovenia	2020 – 2022 - -	Telekom Slovenije A1 Telemach
Spain	2020 2020 H2 2019 -	Orange Spain Telefonica (Movistar) Vodafone Spain Yoigo
Sweden	2020 2020 2020	Telia Tele2 Telenor
Switzerland	03/2019S 2019 April 2019	Sunrise Salt Swisscom
United Kingdom	H2 2019 End – 2019 H2 2019 H2 2019	EE Telefonica UK Vodafone UK Three UK

2.3.9 Infrastructure Equipment

At the level of physical equipment capable of operating according to the 5G standards, developments are rapid, with vendors announcing the availability of new compatible products one after the other. In terms of network equipment (routers, switches, access points, base stations, etc.), companies such as Ericsson, Huawei, Nokia, Samsung and ZTE already have products for installing the required infrastructure from this early phase of the 5G. For example, Ericsson claims to have equipment to respond to the new Air Interface and the corresponding access and switching techniques even for indoor environments while also focusing on solutions for a smooth transition from 4G technology to 5G, such as the spectrum sharing technique. Huawei offers end-to-end solutions covering the use of frequencies around 3GHz and those relating to the mmWave band, using technical active antenna elements. Nokia is also a major contributor to both providing cloud-based solutions for all network operations (in Core, Transport & Radio Access level) and providing innovative RAN components with significant benefits to network operators (halving the size of MIMO antenna units and significantly reducing the power consumption required to operate them). Samsung's achievements in this field include the development of mm-Wave components for every possible transmission framework (fixed, indoor, outdoor, etc.). This company is also actively involved in providing integrated 5G solutions. The activity of ZTE in the construction of 5G components for the entire spectrum of technologies and concepts that will operate under the 5G (massive MIMO, D2D, UDN, differentiated services model, etc.) is also noteworthy, while, at the same time, it is offering solutions for the transition from 4G to 5G using advanced technologies. The magnitude of its contribution to this sector is particularly important if one considers that a large number of 5G pilot trials in countries such as China, Japan, Spain and Austria are being implemented with its products [22].

2.3.10 Chipsets & Devices

Chipsets manufacturers, following the technical specifications defined by the 5G development community, create suitable products that are able to meet market demands. Companies such as Qualcomm, Samsung, Intel, Hi-Silicon (Huawei) and Mediatek create chipsets for communication on 5G frequencies, emphasizing, during these early stages, on the areas of the spectrum of around 6 GHz and on the corresponding frequencies of the mm-waves bands with the aim of achieving the greatest possible throughput [22].

Table 3 : 5G chipsets availability from different vendors [22]

Vendor	Chipset	Availability	Spectrum	Throughput
Qualcomm	X50	End of 2018	Sub 6 GHz & 28GHz	5 Gbps
	X55	S1 2019	Sub 6 GHz & mm-waves	7 / 3 Gbps DL / UL
Intel	XMM 8160	H2 2019	Sub6 GHz & mm-waves	6 Gbps
Samsung	Exynos 5100	End of 2018	Sub 6 GHz & mm-waves	2 Gbps (sub 6 GHz) / 6 Gbps (mmWaves)
Hi-Silicon (Huawei)	Balong 5G01	End of 2018	Sub 6 GHz & mm-waves	2.3 Gbps
	Balong 5000	01/2019	Sub 6 GHz & mm-waves	4.6 Gbps
Mediatek	Helio M70	2019 - 2020	Sub 6 GHz & mm-waves	5 Gbps

Supporting mm-waves communications is a small challenge for vendors because of the sensitivity to the obstacles that such communications have. It is common for a device to have difficulty maintaining its connectivity levels due to the blocking of the field of the antenna from small obstacles such as a hand holding a mobile device. One possible solution is to use multiple chipsets per device so that a communication channel is available at all times. Yet another important feature is the integration level for all previous generation (2G, 3G, 4G) bands of networks in order to ensure full interoperability and compatibility [22].

At a device level, development procedures involve more phases and are more complex than chipsets since they require the availability of the latter to complete. Thus, static devices (e.g. routers, access points) are initially developed to perform early tests on the 5G ecosystem, followed by the production of mobile terminals (e.g. smartphones, tablets). In this area, Huawei, Samsung and Motorola have so far led developments as, according to information reportedly leaking from time to time to the press (such as specifications of prototypes), they are already in advanced stages of development, approaching the point of promoting products for pre/commercial use. Corresponding developments are also being made by other manufacturers, illustrating the momentum of the development of devices for the 5G ecosystem [22].

Table 4 : Availability of 5G compatible devices [22]

Vendor	Product	Notes
HTC	5G Hub	In partnership with Sprint for the provision of a 5G mobile hotspot
Huawei	5G CPE Pro	Supports both sub-6 GHz and mmWave bands
	Mate 30 Series	Should be available in H2 2019
	Mate X	-
Intel	2-in-1 PC 5G Prototype	Supports the 28 GHz band
LG	V50 ThinQ Smartphone	Available in May 2019 for Sprint' subscribers in USA
Motorola	Z3	Smartphone with 5G connectivity available by additional module
Netgear	Nighthawk 5G Mobile Hotspot	Supports mmWave bands
OnePlus	n/a	Planning to launch a model in 2019
OPPO	FindX 5G	Should be available in 2019
Samsung	Outdoor CPE	Indoor / Outdoor home router with 2x2 MIMO and 32 elements per antenna
	Galaxy F Foldable	Foldable Smartphone supporting 5G
	S10 5G	Should be available in USA in 2019
Sony	Xperia X23	-
TCL	Alcatel 7	Planned for H1 2019
VIVO	Nex	Planned for H1 2019
Xiaomi	Mi Mix 3 5G	Planned for H1 2019
ZTE	Axon 10 Pro 5G	5G "smart terminals" planned to launch in 2019

3. KEY ELEMENTS OF THE 5G COMMUNICATION PARADIGM

This section outlines the most important innovations and their functionality that will collectively make up the huge 5th generation ecosystem of mobile communications networks and will offer powerful benefits to all its members. For each individual contrivance, in addition to how it works, the benefits of its use, as well as any challenges related to its implementation, are analyzed.

3.1.1 Multi-layer cellular network architecture

Based on studies of the use of cellular network services, it is concluded that, in the majority of time, users remain within buildings (pseudo-moving users) and a much smaller percentage move outside thereof (real-moving users). Based on the classical architecture of the network, each device has access to the rest of the network through the Base Station (BS) located in the center of the cell (macro cell or microcell) that covers the area where it is located. However, when a mobile terminal is inside a building, it is very likely, due to natural laws, to have signal quality losses at the moment when available RF spectrum has been spent and a not inconsiderable amount of electricity for that purpose. Of course, this also has an impact on the users' QoE since high data rates can not be achieved in order to meet the minimum requirements of each application [2] [26].

One of the solutions that have been proposed for solving this problem is the adoption of a multi-level cell architecture (from LTE-A), wherein different cell types of users for the geographic and population coverage are used, or a combination thereof. This model comprises three cell categories, the Macrocells for covering suburban areas with a lower user density, the Microcells for covering urban areas with a higher user density and Femtocells with low transmission power for use inside buildings [2] [26] [27]. Picocells, that also operate with low emission power and are mainly used in public hot spots to cover small areas within the range of a few tens of meters, can be added to these categories. Finally, by making a better categorization, both femtocells and picocells are smallcells since they have similar characteristics and are treated in almost the same way by the rest of the network [28].

Table 5: Different cell types characteristics [7]

Cell type	Range	Users
Femtocell	10 – 20 m	A few users
Picocell	200 m	20 – 40
Microcell	2 Km	> 100
Macrocell	30 – 35 Km	Many

[7]The operation of a femtocell is based on the installation of small range, small cost, low energy consumption BSs (home BSs or evolved NodesB - eNBs) within buildings that automatically connect to mobile devices when they enter into their range, provided they are allowed to do so as registered users (closed model with restricted access) or have free access (open model). This creates a new cell (the femtocell) with a small operating radius but with maximal signal quality, capable of providing a high level of connectivity services as users and BS are now very close and no obstacles significantly interfere between them. The connection of the eNB of the femtocell with the backbone may be done either over IP-based broadband connections (Very high bit rate Digital Subscriber Line - VDSL) or via a dedicated RF link [26].

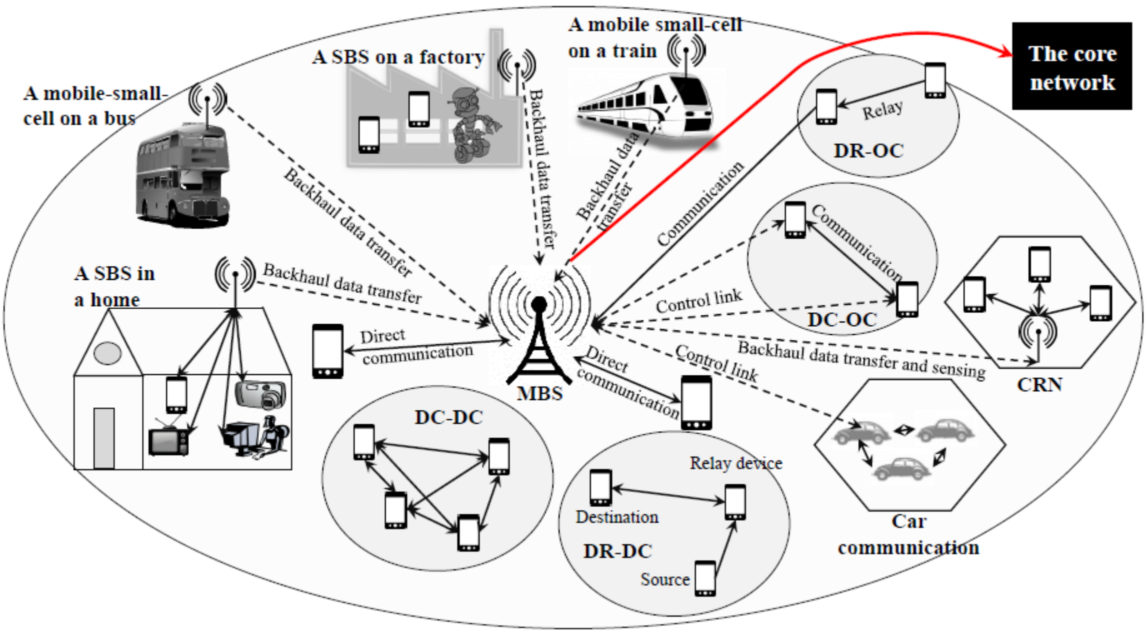


Figure 9 : The multi-level cell architecture of 5G Networks [7]

An additional addition has also been proposed, where cell types include Mobile Femtocell (MF) (a variant of Femtocell) for use in moving networks (see below), i.e. in networks in transportation means such as trains, buses, cars [27]. A similar logic of operation is followed since, also in this case, all of the Mobile Femtocell users are treated as a single unit by the rest of the network with which it interacts with the only difference from the femtocell model that, in this case, the connection with the core network changes dynamically [2].

According to the association between base stations and mobile terminals, the criterion for finding the best possible choice used in previous generations of wireless networks, that is to say, the BS-device correlation based only on optimal signal strength or quality, cannot remain the same since now the BSs' transceivers have different specifications and the architecture of the cellular network is more complex (multitier). At this point, it is worth noting that in 5G, among other things, an individual target is also the offloading of macrocells so the data transfer is made via smallcells where the transmitter and the receiver are closer and higher connection speeds can be achieved more easily. Thus, a criterion for the association of devices with the BSs could have as a top priority the connection to the nearest smallcell which will have less traffic. Taking into account all the parameters, this can be translated through an appropriate algorithm (Range Expansion - RE) again in measurable terms (weighted signal strength & signal quality) for immediate use by mobile devices. Another technique that can be used for the same purpose is Dual Connectivity (DC) in which a device is simultaneously connected to BSs of different cell levels, for example to BSs of a macrocell (signaling & data exchange) and to a number of BSs that belong to smallcells (for data exchange only). In this case, better traffic distribution is achieved in the network, while the network is more robust [28].

The adoption of such a model where mobile terminals and BSs are within close distances to each other has multiple advantages for both users and network operators. With its adoption, better population and geographic coverage, improved signal quality (increase of Signal-to-Interference-plus-Noise-Ratio - SINR) and simultaneous reduction of overhead for signaling, increased capacity, easier handoff, increased network scalability and longer battery life of mobile terminals (due to the shorter distance of the transmitter and the receiver) are achieved, when, at the same time, the required network operating costs are reduced (10x costs of obtaining and maintaining femtocells in relation to macro/microcells, restriction of electricity consumption). Additionally, the implementation of such a model is completely backwards compatible and still allows for better management and reuse of the radio spectrum due to the fact that it is possible for more users to share the same available spectrum in many small areas. In addition to the direct benefits associated with the servicing of users covered by femtocells, the improvement of the services provided to users covered by macrocells or microcells is also expected, as the number of terminals to be serviced is reduced [7] [26] [27].

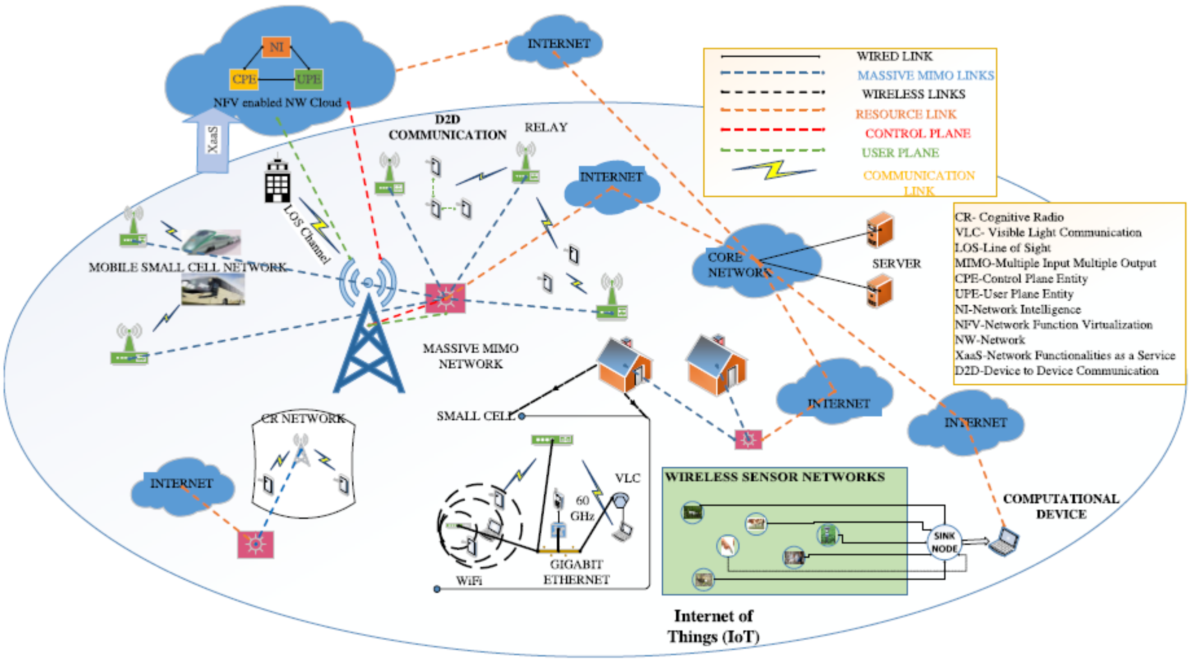


Figure 10 : An abstract view of the 5G heterogeneous networking architecture [1]

As can be understood from the above, all these features make up a Heterogeneous Network (HetNet) whose implementation poses some challenges that are to be addressed in order for it to function effectively. Some of these have to do with resource allocation in order to avoid cross-tier interferences, synchronization during handoff between different cell types (cross-tier handoff) or even between neighboring femtocells, in order to ensure the interconnection of eNBs with the backbone network over the Internet Protocol (IP), as well as overcoming certain restrictions at the medium access layer. As regards the latter category, it shall be decided by the providers if access to femtocells will be of a closed character (i.e. specific users will have exclusive access), of an open character (where macrocell users that were found in the range of operation of a femtocell will also have access) or a hybrid operating model will be followed as it affects other parameters (cross-tier, handoff, QoE, security, capacity) [7] [26] [28]. Still, consideration should be given to whether or not to move a femtocell from one location to another (similar to transferring a Wi-Fi Access Point). However, given that femtocells operate in the field of the licensed spectrum, transferring it to another site may interfere with the network of other providers whose macrocells may operate at the same frequency as the transferred femtocell [26].

3.1.2 Massive Multi-Input Multi-Output (mMIMO)

Another solution that can be applied alone or in combination with the above mentioned cellular architecture is the usage of massive Multi-Input Multi-Output (mMIMO) technology, namely the installation of multiple antenna units on all BSs, [29] when at the same time mobile terminals are also equipped with a certain number of accordingly antenna elements [1]. In this way, it is always possible for the wireless network to simultaneously transmit multiple signals (basically in multicast function but also broadcast is feasible) to multiple mobile terminals that are using the same channel, exploiting the multipath propagation transmissions in a best way (massive Multi-User MIMO, massive MU-MIMO) [30].

The utilization of multipath propagation is based on new medium access techniques that used for spatial multiplexing (transmit on different paths) and the ability of the BS by means of algorithms to know, at any time, the position of the mobile terminals in the environment and the state of both the uplink and downlink of the channel [29] [30]. These new techniques that allow spatial modulation in order to share the resources of the system are Beam Division Multiple Access (BDMA) and Filter Bank Multi Carrier [1]. Thus, the antenna arrays can allocate the available antenna beams in such a way that each mobile terminal, that preferably has LOS communication with the base station, corresponds to an orthogonal beam sending and receiving data exclusively for that device [1] [2].

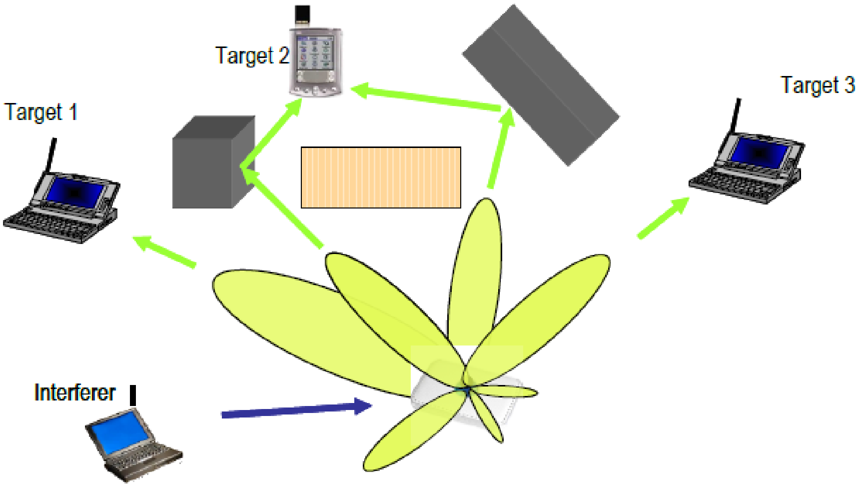


Figure 11 : Use of multiple antenna elements for spatial multiplexing implementation

Unlike the conventional MIMO used in LTE-A and allows simultaneous operation of up to 8 antenna ports per BS [31], in mMIMO, tens or even hundreds of antenna elements can simultaneously be used to each BS, maximizing the benefits of simultaneous multi-mobile terminal services from only one antenna array (with multiple transceiver units), transmitting in the same RF simultaneously (at the same time) [1].

The architecture of a network that embraces mMIMO technology requires the installation of large scale antenna arrays at BS positions (cells centers) as well as the appropriate positioning of some antenna arrays that are interconnected (via optical fiber) at the ends of the cells in order to cooperatively cover the entire region and a complete spatial awareness of the cell. Moreover, installing antenna arrays at various other hot spots such as exterior surfaces or terraces of large buildings enables indoor mobile terminals to have a high connectivity level with the rest of the network without having to directly connect to the BSs of the macrocells where, due to their location, they may not have a good connection quality. Instead, the devices are connected to the closest Access Points (APs) that are located inside the buildings, which, in turn, are directly connected (with cable connections) to the antenna arrays installed outside the buildings where, after having been carefully installed in appropriate locations, have Line Of Sight (LOS) connectivity features with the BSs of the rest of the network [1] [2].

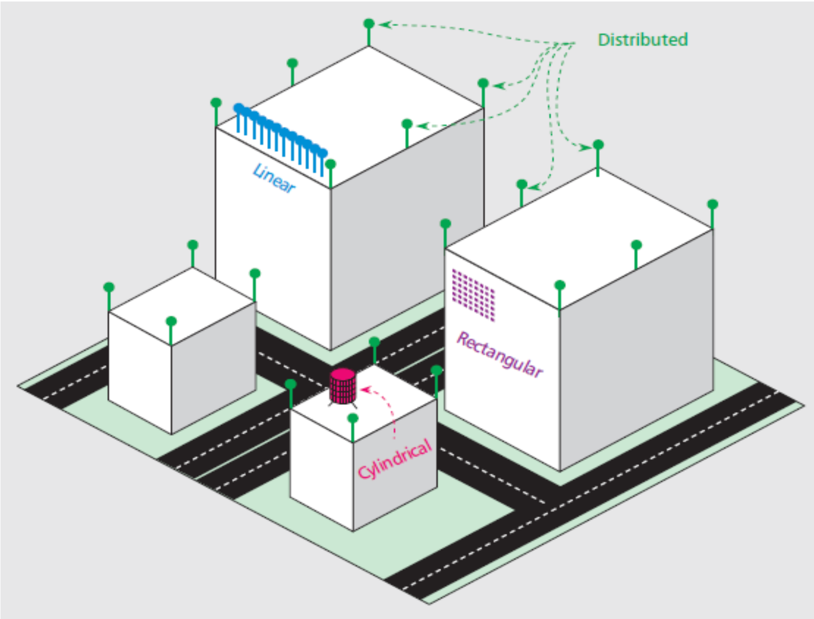


Figure 12 : Massive MIMO deployment scenarios [29]

Additionally, it is worth noting that due to the architecture that allows the existence of multiple antenna units on the mobile terminals, it is potentially possible to collaborate with each other and with the arrays at the BSs in order to cumulatively build powerful virtual large antenna arrays of a mMIMO network with even more capabilities [2].

Although this architecture has a relatively increased short-run cost for upgrading existing BSs, in the long run it allows users to enjoy improved average throughput per cell and higher data rate even for very small areas, it increases the capacity and the coverage of the network using at the same time all the available time and frequency and adapting, whenever required, the network's features make by making it easier for administrators to efficiently manage the spectrum and power saving [2] [29]. The advantages of this technology are also the reduced interference with other systems, because beams are directional and thus avoiding overlaps [29] and, even if something like this is observed, a flexible re-adjustment of the antenna elements operating in the area in function with the number of mobile terminals served during the same period, it is capable of correcting the problem [30].

In addition to the advantages mentioned above, some disadvantages of this technique, such as the increased complexity of managing such a network [12], the synchronization of devices with the BSs during handovers, the relatively limited possibility of network expansion and the difficulty that exists for the BS to be aware of the downlink, since in this case the use of pilots at link is not very efficient, should be overcome. Also, because of very narrow beams, there is an increased risk of failure of a link, making it necessary to have fallback links via dual connectivity techniques, for example [3]. In addition to this, the dissimilarity of the operating characteristics of the various mobile terminals (e.g. different number of antenna elements per mobile device type) that inevitably affects the operation of BSs should be taken into consideration, as it is necessary to find a "common component" for the operation of their antenna units in order to provide high-quality services in an efficient way [29].

3.1.3 Device-to-Device Communications (D2D)

Accelerated development of new media-rich applications (such as 3D & HD video, Virtual Reality, 3D holography, 3D printing) stimulates network researchers to find new unconventional techniques that, along with existing technologies, can meet the need for ever higher data rates and network capacity. On the side, the evolvement of context-aware services and applications that need location information share between near devices requires new models for communication between neighbors [32].

Based on this reasoning, in addition to the ability of mobile terminals in a cellular network to communicate through cell base stations, 5G now opens the way for direct Device-to-Device (D2D) communications in the same licensed spectrum which conventional communications in mobile networks are also made [32]. In other words, it is possible, along with the classical operation of networks, to exchange data exclusively on devices located in nearby areas, without any mediation of the fixed infrastructure [33]. However, the application of such an idea is not exclusively confined to direct D2D communications between two or more devices, but can be further extended by making mixed usage of stable network infrastructures, creating in that case an ad hoc mesh network [32].

In this way, a two-layer network is essentially created, where one level is characterized by the traditional network of macrocells (in which mobile terminals communicate with each other through the BSs) and the other level comprises a device-level network where communication is done either directly between the devices or there is mediation of another device in some part of the entire coupling (mesh network) for the communication of a device with the rest of the network [1].

Following this architecture, there are four subtypes of D2D communications that constitute the corresponding layer. The first type named *Device relaying with operator control link establishment* (DR-OC) refers to cases where a device is at the end of the cell, resulting in bad signal quality. In this case, nearby devices are mediated as relays to eventually communicate between the remote device and the BS. The second type of communications of this type is *Direct D2D communication with operator controlled link establishment* (DC-OC) where mobile devices that want to exchange information communicate directly with each other and not through a BS. In both of the above-mentioned types, the installation of the connection is controlled by the operator (Operator Control) taking into account relevant parameters communicated to him/her through the devices involved each time [7] [32].

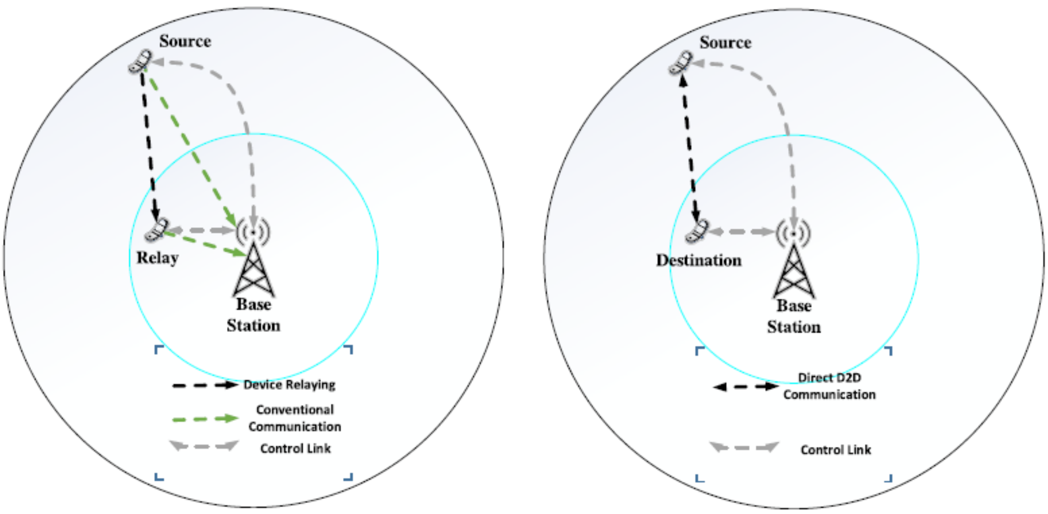


Figure 13 : Device relaying and direct device-to-device communications with base station controlled link formation [1]

In the above mentioned list there are also included subtypes of D2D communications in which there is not exist a third party (e.g. network operator) to play the coordinator role in the procedure of link connection setup. Instead of that, the synchronization parameters of the link arise from the participating members (Device Control), in a way that ensures limited interferences with other devices at both layers of the network (macrocell layer and D2D layer). One of these subtypes named *Device relaying with device controlled link establishment (DR-DC)* and refers to the case where nearby devices transmit signals to other nearby devices by mediating intermediate devices that act as relays. Finally, this category also includes *Direct D2D communication with device controlled link establishment* (DC-DC), where all involved devices directly exchange information with each other without having any BSs involved in the process. In these cases, the finding of neighboring devices is required to be made by the devices themselves, something that can be done through broadcasting messages sent periodically by the devices for this purpose [32].

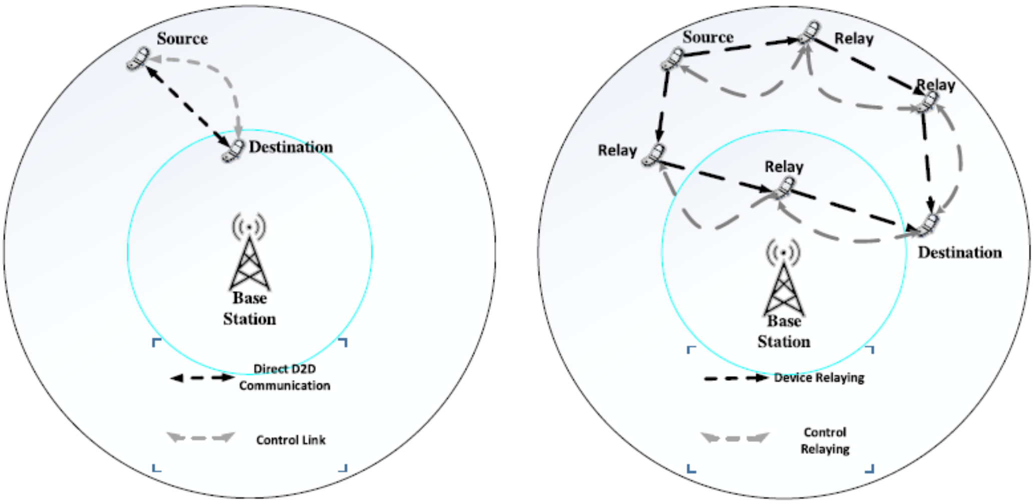


Figure 14 : Device relaying and direct device-to-device communications with device controlled link formation [1]

The adoption of D2D techniques in classic cellular networks can greatly help to improve services delivery on the basis of the dominant feature of this method, the co-operation/mutual support of the network devices themselves by also contributing to their transmission instead of being simple data consumers. This translates into improving network capabilities in terms of reliability and capacity, allowing traffic offloads in busy local areas (such as airports, train stations, stadiums, malls) and mesh/ad-hoc networks [32]. At the same time, these techniques help to increase devices mobility even in areas that are not fully covered by the classic transmission network (increase transmission range) [1], contributes to the development and operation of cloud services, strengthens efforts for resource sharing in terms of content, power and spectrum and finally allow the use of mobile terminals in case of natural disasters where the infrastructure of the cellular network can be totally destroyed or overloaded [32].

The challenges arising from this technology are summarized in the parameters of limiting interference, applying delay-sensitive processing according to the kind of traffic, resource allocating, finding a way to price services of this kind, securing communications secrecy (in terms of confidentiality, integrity, availability), maintaining reliability at high levels and, above all, whether or not there is central coordination from macrocells base stations [1] [7]. Also, with regard to the management of interference and interoperability of such systems, it should be decided whether such communications will use the same or different frequencies of the licensed spectrum as traditional networks [32].

3.1.4 Ultra Dense Networks (UDN)

It has been observed that another factor that affects features of the mobile network such as data rate, capacity, spectrum exploitation and energy consumption is the density and the location of the base stations. Following this reasoning, the Ultra-Dense Networks (UDN) technique has been developed that aims at densely placing access points in such a way that there is at least one BS per user located as close to him/her [28]. In figures, such a network can be mapped as $\geq 10^3$ cells/km^2 [28] or as a UDN access node every 10m in outdoor environment [15], highlighting its truly high density.

As far as their mode of operation is concerned, ultra dense networks are implemented by exploiting the advantages of the multi-tier cellular architecture model, where multiple smallcells that surround a user can coexist in an area. The smallcells broadcast nodes generally cover small areas and have low energy consumption, while, when not in use, they enter a state of inactivity to limit possible interference [28].

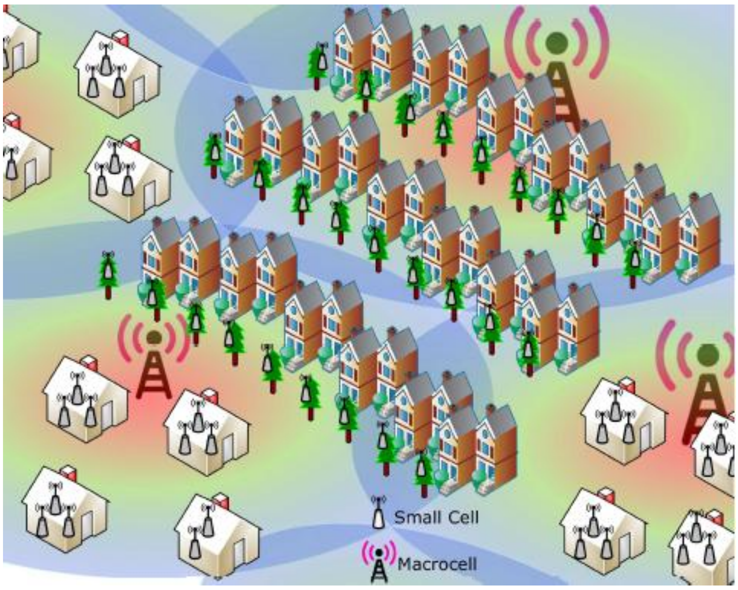


Figure 15 : An UDN depiction adopting multi-tier cellular architecture [28]

The standard for deciding on the association of devices and BSs requires a more complex mechanism in a UDN, as the available number of nearby access nodes is expected to be large in an area. One method that has been proposed is Multiple Association, according to which when a threshold in terms of average signal strength is met, one device will connect to many neighboring BSs of smallcells, sharing traffic to the backbone network from different links.

The implementation of such a network can be based on two models, taking into account specific parameters of the operating environment for what will ultimately be followed. In the first model, the densification of the BSs is made at a horizontal level (horizontal densification model), i.e. through the installation of more smallcells in outdoor areas (along the side of roads, building surfaces, public hotspots) while, in the second model (vertical densification model), the strategy of multiplying access nodes in indoor areas (multi-storey buildings, offices, hotels, apartment buildings) by creating more femtocells is followed. Naturally, each model differs from one another in performance terms since, for example, it is possible to reuse the spectrum or to connect to the backhaul in each respective operating environment. As far as the coordination of these BSs is concerned, it is also achieved in two ways, either centralized where one entity (massive MIMO BS, cloud machine) assumes a coordinating role amongst the others for broadcasts or distributed where different BSs work together for the best result (D2D links) [28].

To achieve load sharing between macrocells and smallcells, better usage of the radio spectrum, increasing the capacity of the network, improving the throughput and the contribution offered to the implementation of the IoT concept are some of the benefits that characterize the UDN concept [1] [28] [33]. On the other hand, some issues that need more investigation are about avoiding bottlenecks at smallcells backhaul due to the physical limits in the air-interface capacity of these cells, the efficient management of its complexity, scalability and the finding of an intelligent method for interference mitigation and frequency reuse across the different layers of the multi-tier network [28].

3.1.5 Multi-level spectrum scheme

Following the concept of differentiated services, the available spectrum for communications in 5G networks is divided into multiple bands in a wide range between 700 MHz and 100 GHz. In order to support all innovative functions introduced by the 5G and to achieve maximum system performance, it is necessary to use different parts of the spectrum per use case scenario family (eMBB, mMTC, URLLC, CC, mIoT, NO) so as to cover the distinct needs of each application [24].

In this context, the 5G is expected to use bands in the lower portions of the spectrum at frequencies below 2 GHz, at frequencies of 2 to 8 GHz on a middle level and, finally, at frequencies between 24 GHz and 100 GHz (cmWaves and mmWaves) in the higher portions so that both full geographical coverage and bandwidth supply of unified 500 MHz up to 1-2 GHz can be achieved in network applications in order to achieve high data rates. These frequency bands, although belonging to different spectrum categories covering areas ranging from the unlicensed sub-6 GHz spectrum up to the mmWave spectrum (30-100 GHz), will be managed in uniformly by the corresponding layers of the grid pushed by SDN and NFV concepts that can provide the appropriate background to support such functions [24] [34].

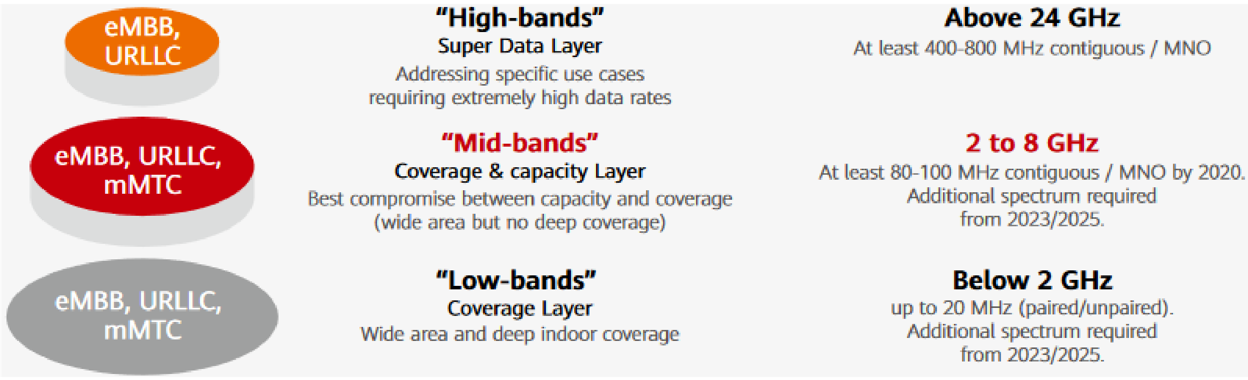


Figure 16: Basic scheme of spectrum allocation model for each 5G use case scenario family [24]

3.1.6 Mobile Edge Computing (MEC)

The review of the 5G network operating model in terms of optimizing the user's end experience sets new rules in the equation and requires the development of new methods and techniques in order to achieve this fundamental goal. This has led to the idea of transferring network functions (computing, networking, storage) as close as possible to their creation and demand source, ie to end-users or in networking conditions at the edge of the network [19].

Based on this logic, small-scale data centers that support cloud computing services are placed in the Mobile Edge Computing (MEC) model as close to as possible to the edge of the network (ideally to the base stations and, in any case, within the RAN) allowing the provision of demanding services at terminals with the least possible burden both on themselves and on the network [19] [35]. This translates into minimizing needs in terms of computational power, bandwidth, storage, energy consumption, restricting latency and failures and more generally network charges with tasks that can be implemented at the edge of the network and unload the rest of the network (core and transport) [35].

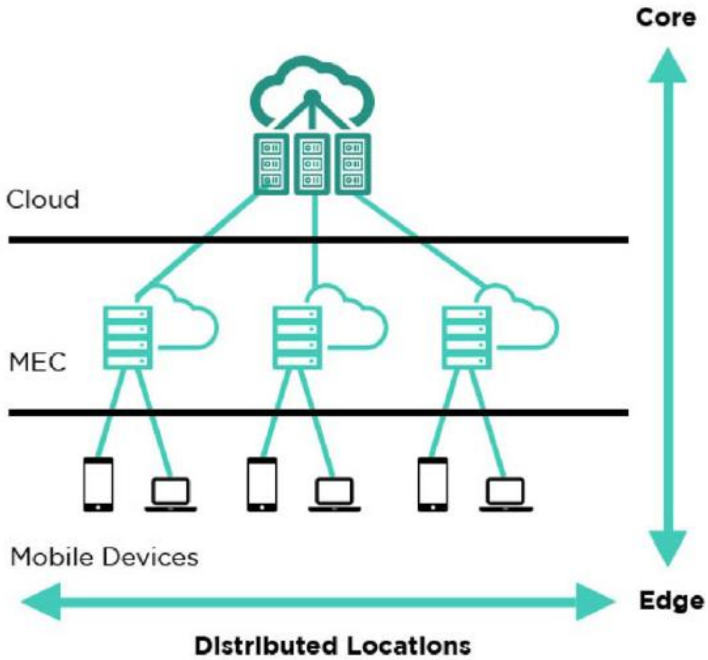


Figure 17 : MEC as an intermediate layer between cloud and mobile terminals [35]

The benefits of this technique are many when considering the variety of different communication scenarios and devices within the context of IoT that coexist in the 5G. For example, besides reducing service times and maximizing QoS, serving M2M scenarios mostly with local resources also strengthens the network's security in overall by creating a protection film due to the limitation of requests for servicing from resources that exist deeper into the network. In addition, the proximity of demand and data processing points makes it easier to handle and exploit large data for the benefit of demanding applications that use augmented reality and artificial intelligence techniques. At the same time, context and location aware applications are supported to the utmost extent with this logic, since they can receive the required information faster by the network while being aware at all times of the time required for their availability. This is possible because many of the components of the network from source to point of use are not involved in the procedure [35].

Vertical segments such as e-Health, Smart Building, Smart Grid, V2X Communications, Ocean Monitoring, Transport, Virtual Reality and, more generally, sensor networks that process data in a limited range are use case families that can benefit from faster data processing and exchange of network resources, avoiding obstacles (resources synchronization, scalability in distributed environments, bottlenecks, etc.) that under other circumstances impede their implementation. In order to make MEC's contribution to the 5G even more apparent, even the implementation of the SDN can be facilitated by strengthening the robustness of the network in cases where the ratio of packet loss is high by preventing the centralized network control [35].

3.1.7 Self-Organizing Networks (SON)

The idea behind Self-Organizing Networks (SON) focuses on automating the Operation & Maintenance (O&M) services required in every cellular network, which can bring significant benefits to providers and network administrators by simplifying the corresponding actions to benefit the quality of service for users and the reduction of costs for the operation of the network [12].

Although the idea is not new and efforts for its implementation have been made since the LTE standardization era [36], this has only been done in part. The goal is to make 5G an inherent part of the ecosystem in order to provide optimal services and make maximum use of network resources at all times at both higher and lower levels. Achieving this vision has increased chances of being successful in 5G, taking impetus from the extensive use of cloud resources in all subsections of the network (Core / Transport / Access Network) via NFV, SDN and C-RAT projects [12].

Its function is based on tracking the network's status and traffic's patterns that are prevailing in the network at all times in order to provide feedback for further forecasts that will help the network readjust cloud resources as soon as possible. Thus, in conjunction with the corresponding subsections of the network (e.g. C-RAN), the necessary actions can be performed in a fully automated way and as proactively as possible, based on the dynamic configuration of the network according to real time needs. For example, when there is increased traffic in an area due to an event (football match), then the MCE of C-RAN will act accordingly (by adjusting any parameters required) in order to ensure maximum network performance. Corresponding actions can also take place when there is some hardware damage, where, until its final restoration, the work it performed can be substituted by other resources [12] [37].

The need to develop SON is clear when considering the architecture and basic principles of 5G cellular systems. The multi-tier cell model, alongside the beamforming, MIMO, UDN and mmWaves concepts, increase the complexity of network management, requiring greater time and money for proper operation [38] [39].

Some typical cases where the SON concept can be useful are spectrum management and sharing, user association, multi-RAT optimization and directional cell search in mmWaves networks [40]. Especially as regards the management of connectivity to mmWaves ultra-dense networks, where the probability of communication problem due to interference of a user's obstacle or movement is increased, the contribution of SON can be decisive. Thus, if for some reason the beam that has been selected by the BS in order to serve a terminal during the initialization of the link fails to perform its task, then the network will have to promptly react to this challenge. It is obvious that beam-training plans are prone to all of these parameters which increase the complexity of the problem, ultimately causing delays in the whole link installation process. In this case, SON can act as a coordinator in the problem-solving plan by helping the handover process through alternative beams from other available BSs (which the terminal is already aware of from the link establishment phase). Another solution is to use location information (location and orientation) from the terminals' sensors that will be included in the control messages exchanged between the terminals and the network and feed the SON in order for it to give more complete feedback to training models and to make tracking of UEs faster [39].

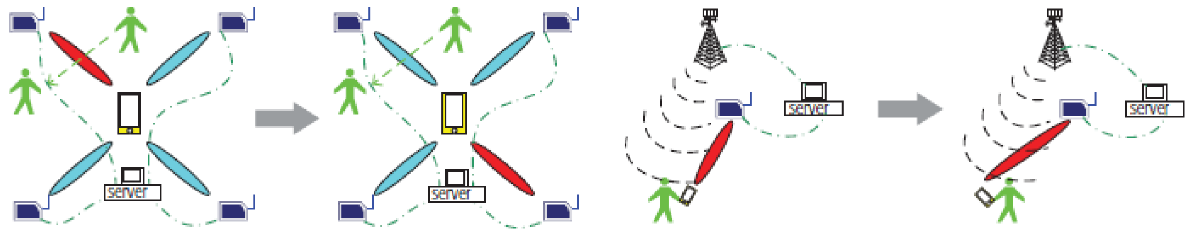


Figure 18 : Automated beam selection & directional transmission based on real time conditions [39]

SON should have a completely proactive self-organized functionality on the axes of self-planning, self-configuration, self-healing, self-optimization, self-coordination and self-awareness of topology so as to be able to control and properly adapt the network by releasing workforce from these processes [12] [39]. Developments in the field of machine learning and big-data analytics can be beneficial in giving a significant boost to the whole project, since they facilitate the creation and evaluation of trends in network traffic with a view to making proactive decisions [39] .

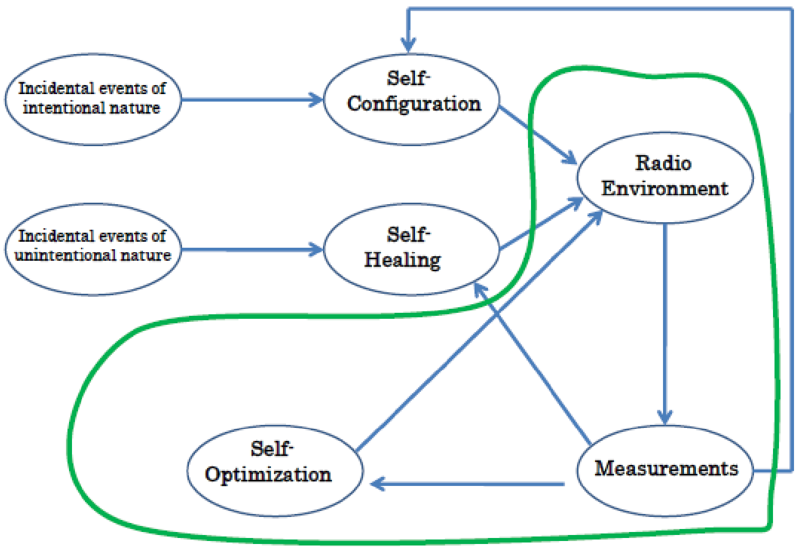


Figure 19 : SON operating model cycle [36]

3.1.8 Additional Techniques & Operations

In addition to the above, the contribution of some other techniques and functions that can add extra functionality and advanced features to 5G is possible, allowing even better service to end users and providing multiple benefits to providers.

3.1.8.1 Proactive Caching

Proactive Caching means the ex-ante temporary storage of popular information at the edge of the network (BSs, devices) for faster user service during high traffic periods [28]. This requires the timely forecasting of the information that needs to be stored at the edge of the network and can be done in off-peak times, taking into account statistical patterns for studying the popularity of the information and their correlations with other relevant information as well as among the users who consume them. It is also suggested to store information on users' devices based on their social attributes as network users (i.e. measuring whether neighboring users are influenced in using the same information depending on the location and speed at which they are usually routed to the network). For example, it is very likely that fans of a team on the same field search for the same information at almost the same time. It is worth emphasizing that this easily applicable today given that users devices (smartphones, tablets) have advanced functionality and can manage this additional function. Moreover, the study of users interactions based on their profiles in social media can contribute and this, in its turn, can contribute in the export of patterns for proactive caching prediction since it is common for “friends” on social media to have similar preferences, tastes and habits. In addition, the data study based on geotagging can provide important information to networks operators for the network points where heavy traffic is expected. Finally, the conclusions on the movement of information from the log files of the nodes of the network using Big Data analysis and Machine Learning techniques may decisively provide feedback to the prediction patterns by rendering the caching process more efficient. Thus, the synthesis of these patterns can lead to a fairly accurate model for correlations (social or at a level of spatial proximity) between users and information [41].

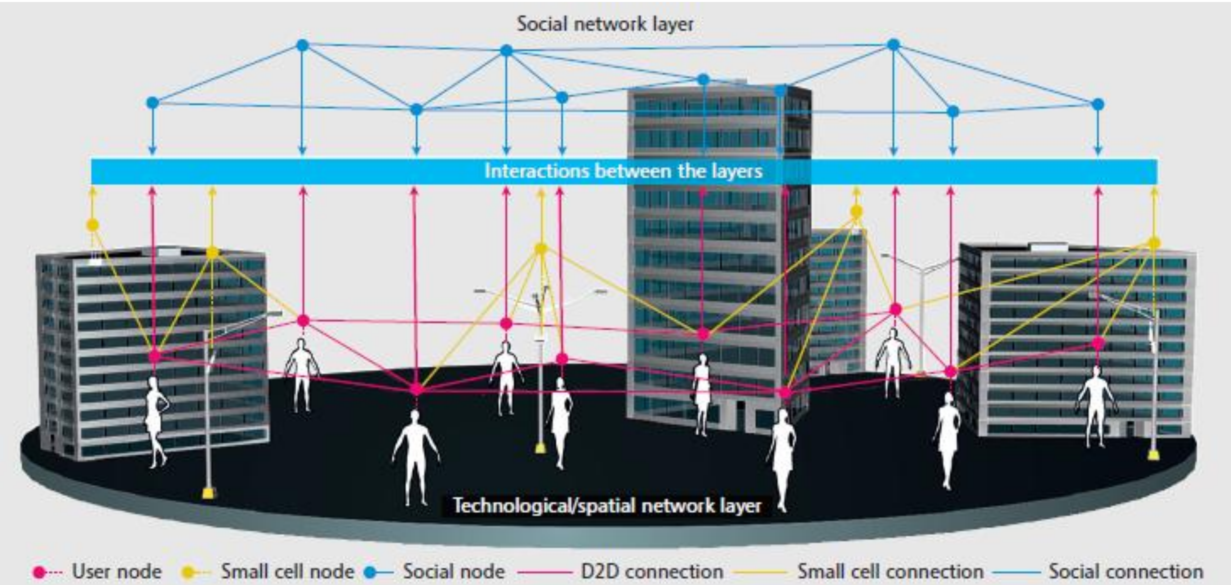


Figure 20 : The analysis of user behavior in the network as a compass in the creation of proactive caching patterns [41]

The benefits of implementing this method are multiple and primarily involve traffic offloading and better utilization of network and infrastructure resources, especially when there are high peak demands, as the network can be better prepared to cope with increased demand [41]. Moreover, in combination with other techniques used in 5G (D2D, UDN, mMIMO), it can further benefit the network’s operation by providing better services to users in terms of QoS [28]. Security issues can be considered as pending issues pertaining to this technology, especially when caching is done on user devices, [28] optimizing scheduling techniques to make the appropriate caching at the appropriate point in the network at the right time in order to avoid cache misses, as well as finding adequate mechanisms for collective load balancing and context sharing in terms of mobility management [41].

3.1.8.2 Ultra-Reliable Communication (URC)

An advanced operation mode that did not exist in previous generations of cellular networks and is expected to be integrated into 5G is the Ultra-Reliable Communication (URC) and is about guaranteed uninterrupted connectivity provision throughout the time required for the application to function. This feature may seem particularly important in a range of commercial applications (and not military, as these are treated as a completely different business model) whose availability is particularly critical for their proper operation. Such examples are the sector of cloud applications that are now widely used and are increasingly replacing the corresponding traditional ones, communications between vehicles (Vehicle-to-Vehicle – V2V) as well as sensor networks that handle critical operations and infrastructures (emergency conditions). In order to achieve this, many parameters (acceptable latency, jitter, data rate, error rate) must be taken into account and a suitable high-level connectivity model (higher levels of the OSI stack) for each respective use case should be designed, based on specific requirements that render its operation optimal (optimized QoE). In general, this can be accomplished by appropriately rethinking this methodology used for encoding and transmitting the segments of the data (in terms of header + payload encoding - decoding - certification). In brief, it may comprise differentiated operation modes for each phase (setup, data transfer, termination) and communication type (transmission of critical data without certification, or transmission of verified huge payload data) which are dynamically selected by the system depending on the communication criticality level.

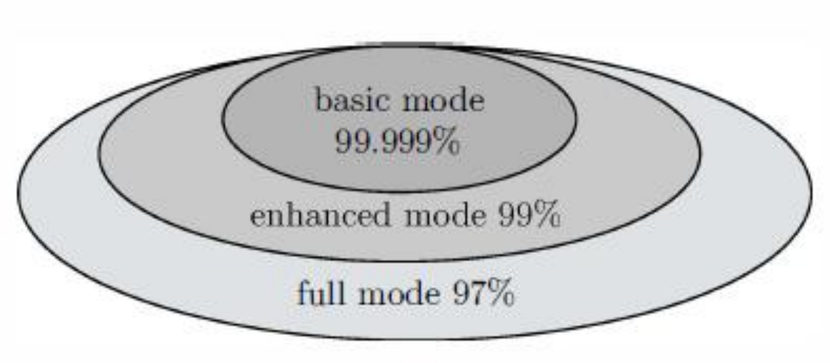


Figure 21 : Main modes of URC services [42]

Although the advantages of such a method are easily made, the challenges that come with it are summarized in the need to meet different requirements (stringent latency requirements for limited resources versus guaranteed rates with high probability for long periods) ending in a common operating mode. In addition, problems that have to do with physical parameters (interference, signal weakness) and with resource sharing limits, protocols mismatches and failures in infrastructure shall be addressed [42].

3.1.8.3 Multi-Radio Access Technologies (Multi-RAT)

Another key goal of the new generation of wireless networks is the application of Multi-Radio Access Technologies (Multi-RAT), i.e. the joint cooperation of all individual wireless access types (traditional cellular, Wi-Fi, mmWaves access) in order to give an integrated access system to the network with a single, simplified and more efficient operation than the existing scattered situation. The need for aggregate access is essentially dictated by the high 5G connectivity requirements so as to be able to cover the operation of a huge range of services to be provided over that. Thus, this ensures high quality, smooth and uninterrupted user experience regardless of whether the user, due to his/her position, is connected through different access types each time, thus enjoying the benefits of all existing individual access methods.

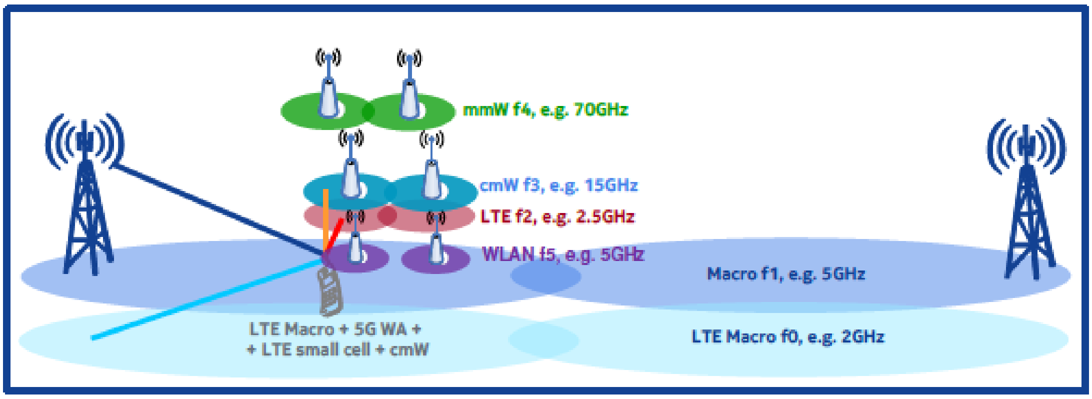


Figure 22 : Multi-RAT use case involving various radio interfaces [28]

Following the principle of differentiated services in 5G, it is necessary to create a system that can flexibly select the appropriate type of interconnection (RAT type, a single type or a combination thereof) depending on the needs of the applications served each time. At the same time, it will be able to uniformly manage multiple connectivity.

One idea in order to implement this scenario is to use a centralized Control Plane (C-Plane) that manages all resources and possible ways of connection (in terms of signaling, admission control, load balancing, etc.) in a unified way and a User-Plane (U-Plane) that includes functions for routing traffic between different RATs, possible data fragmentation, duplicate packets management, packets routing of the same flow following a different RAT type, packets aggregation where the implementation of QoS policies is required (following SDN paradigm). All these functions can be implemented in either cloud environments or dedicated hardware following distributed models, according to each time needs [11]. Of course, the application of this technique requires overcoming certain difficulties, such as to cover all possible use cases, compatibility with Cloud-RAN, interoperability and management of interference between access methods and even different ranges of the spectrum (licensed - unlicensed, mmWaves - cmWaves, etc.) [11], the simultaneous connection of devices at different RAT types with all that this implies (data segmentation, duplicate packets, routing packets of the same flow following different RAT type, packets aggregation) and finally the support of access of architectural networks of older generations (backward compatibility) [28].

3.1.8.4 Moving Networks

The concept of Moving Networks can also be included in the 5G ecosystem, which goes one step further as regards the concept of mobility, thus extending the existing communication concept. Such a network is a set of nodes located within a high speed mean of transport (train, car, bus) and communicating either with each other or with the rest of the network via an antenna element located on the outer surface of the moving mean in order to exclusively serve the users in its interior [43]. In fact, a moving cell is formed (more precisely moving Femtocell [2]) with nomadic nodes into the cellular system [44]. In such a case, it is desirable to provide high-quality connectivity and QoE services to users by overcoming any limitations arising from the speed at which they move. One idea to achieve this is to install Moving Relay Nodes (MRNs) outside the vehicles, which, on the one hand, are strongly connected to the BSs of the crossing areas (as transponders) and, on the other hand, by mediating them to the moving devices in order to be connected to the rest of the network. MRNs' installation in vehicles essentially means communicating with the rest of the network through larger, stronger and more antenna units than those that would be used individually from the user's devices, since the space and power consumption limitations on vehicles are not as stifling as on small portable devices. Moreover, in this way there are no losses that would exist for the signal penetration inside the movable vehicle (penetration loss), allowing the use of more advanced transmission techniques. Also, a strong advantage of this method for the network is the reduced handovers and the need for less signaling overhead, as, along the cells traversed, a handover is performed only for the MRN of the moving cell instead of one for all the mobile terminals. The challenges of this technique include the development of advanced transceivers to reduce interference, especially in urban dense networks, as well as the creation of new protocols suitable for managing cells mobility [43].

4. RADIO ACCESS NETWORK ADVANCEMENTS

4.1 Physical Layer & Advanced Air Interface Framework

A new era is beginning on 5G networks about the air interface, resulting from the new mobile network architecture (5G – NR) that includes a vast array of different communications. Moreover, due to the high density of terminals expected to exist on the network, specialized methods of transmission and access to the medium are required to provide high-quality services [5].

As regards the characteristics of the waveforms to be used, two versions have been proposed to meet the needs of all distinct types of network communication. In the first version, it is proposed to have a unique waveform which will be adapted according to the requirements of the service it serves and the band available, while the second version proposes the coexistence of multiple different waveforms [44].

In the case of using multiple waveforms, on their available options, other than those already in use, **mmWaves Networks** are used for millimeter band frequencies in order to transmit mobile signals alongside the other broadcasts. Besides, significant parameters related to the capacity of a wireless telecommunication system are the finest spectral allocation and the bandwidth of a channel. So far, most wireless networks use frequencies in the 300 MHz and 3 GHz frequencies due to the favorable physical properties of waves in those areas that are characterized by reliable transmission over long distances [5]. The basic feature of this type of transmission is the achievement of very high data rates (due to the wider allowed bandwidth) using even unexploited RF segments in the 3 - 300 GHz range (excluding some segments in the 57 - 64 GHz and 162 - 200 GHz ranges which are not suitable for communications because of their physical characteristics) [5] [28] [45], but at the expense of high sensitivity of transmissions of such signals resulting from natural laws (higher path loss, blocking issues, vulnerable in meteorological conditions) [1]. Consequently, the requirement for LOS transmissions and for advanced beam-training models (since their spatial division multiple access protocols in 5G are followed) [39] which this type of communication has limits to a certain extent their wide use and clearly indicates that the application of this communication method can be done mainly in open spaces (without any obstacles) in which there is a dense network of broadcasting stations (UDN, massive MIMO) and beamforming is easily applicable [1] [28]. Alternatively, a hybrid model can be implemented in which only payload is transmitted to the mm-wave bands and the channel signaling and control information is transmitted to the classical frequencies [5]. In addition, the need to continually minimize interference and effectively manage a complex network, such as this one, render the existence of appropriate technical solutions for their treatment necessary (SDN, C-RAN, SON, etc.) [45].

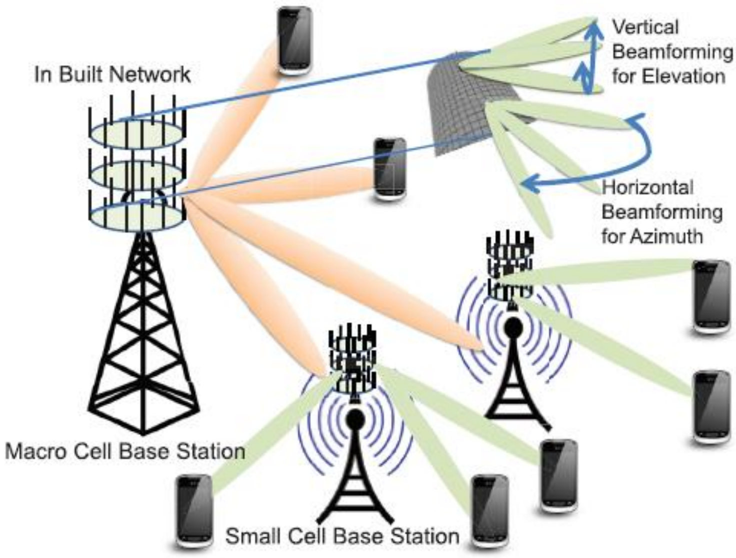


Figure 23 : Cellular transmissions using adaptive beamforming antennas & mMIMO techniques [5]

The transmission of signal using **Adaptive Beamforming Directional Antennas** both on the side of the BSs and on the side of the terminals, combined with the adoption of a SDMA model for medium access (MAC layer), are promising techniques for the operation of a dense network that uses mm-wave communications. Following this protocol of operation, more efficient transmissions from a mMIMO network are allowed, as directional transmissions or receptions on narrow beams can be implemented with the appropriate configuration of the small antenna units of the grid.

Due to the ability of the antennas to distinguish the signals in useful or interleaved in each direction, they can adjust the power and direction of their signals accordingly, each time, in order to obtain maximum gain from the mobile terminal served by the respective beam, thus following its course in the area. At the same time, interference in neighboring regions is moderated, since less powerful or even zero signal power is transmitted towards the points where the target terminal is not being served and therefore no energy focus is being emitted. Thus, high SNR, better utilization of available spectrum, energy savings and increased capacity of the network is achieved [46]. It is worth mentioning that it is challenging for the research community to create antenna training protocols in order to find the best angles for directional transceiving in both sides [5].

Regarding adaptive beamforming, there are several variants thereof, since it can be applied to different types of antenna arrays, with different parameters relating to the management of beams (weight, steering, control). In addition, it is possible to implement beamforming over analog baseband or over digital baseband modes or in a mixed (hybrid) mode. Depending on the application environment and the network operating parameters, the configuration option that outperforms the others based on the specific functions may be followed [5].

Given the progress made in implementing the adaptive beamforming model on transceivers, the road to **Full Duplex Communications** between transmitters and receivers, using the same frequency in the same time, is now open. In combination with other technologies incorporated in 5G (mMIMO, UDN, multi-tier cell architecture, C-RAN), any disadvantages of this type of communication (self interferences and interferences between neighbors) can significantly be mitigated and mobile networks can greatly benefit. In this case, the degree of spectrum utilization and network capacity is almost doubling, while the interference management and the limitation of the hidden node problem are possible [5].

4.2 Multiple Access Layer Scheme (MAC Scheme)

The application of new techniques and innovations linked to the functional features of 5G inevitably require physical modifications which, in turn, make it necessary to adapt the MAC layer to these functions.

The basic feature of the MAC layer in 5G is the adoption of a different template for sharing the medium between the different terminals. Instead of standard techniques that are based on dividing the medium by frequency (FDMA), time (TDMA) or coding (CDMA), in that case a technique capable of supporting the mMIMO and the beamforming architecture of the network must be used [5]. The Spatial Division Multiple Access (SDMA) technique, i.e. the ability to reuse the same frequencies / timeslots / code patterns in different area zones with a given distance, is able to support the operation requirements of the new generation networks. This technique exploits the fact that the adaptive antenna elements of base stations have the ability to distinguish signals received from different angles with respect to their axis (different directions), thus separating the different terminals that they need to serve [5] [46].

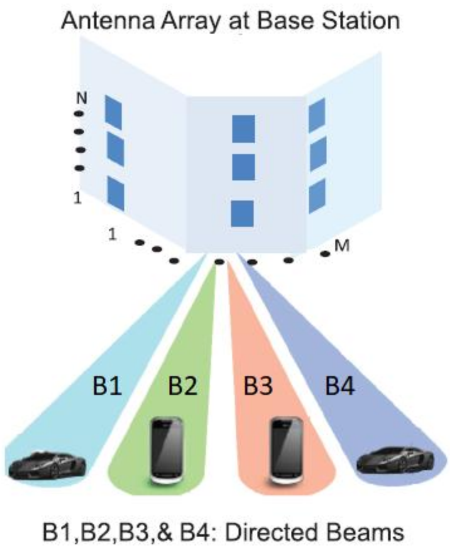


Figure 24 : SDMA scheme [5]

For the SDMA to function efficiently, it is important for the calculations made (based on the channel's characteristics) in order to find the transmission directions to be as close as possible to the optimum result. Otherwise, proper beamforming will not occur and the network performance will be greatly reduced due to interference and poor signal quality [5].

As regards management of simultaneous access to the common medium at protocols level, various techniques that can be implemented to meet the requirements of beamforming have been proposed by the community. For instance, some protocols that can be used to decide whether to broadcast or not broadcast towards a direction in the channel, respectively, are the Channel Time Allocation (CTA), the usage of Request To Send / Clear To Send (RTS/CTS) messages in a suitable variation and the Directional Network Allocation Vector (DNAV) table [5].

It is worth noting that CTA is based on TDMA and uses Super Frames (time slots combinations) in order to exploit spatial reuse. Practically, this is accomplished by exchanging information for the internal channel of each Super Frame. In addition, as regards the DNAV protocol, although it is a basic element of a directional MAC protocol, it does not deal with classical instances of inefficient operation, such as the hidden node problem and under utilization of the medium.

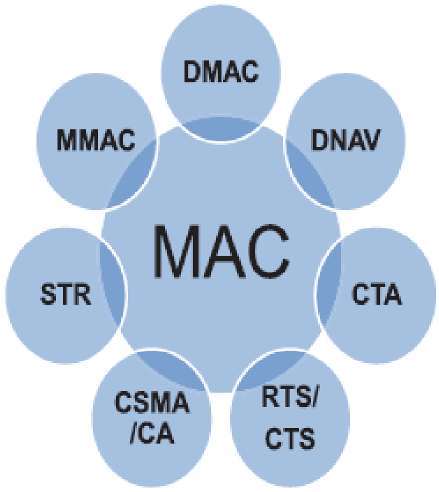


Figure 25 : SDMA add-on protocols in order to implement beamforming [5]

In addition to the above techniques, data from Multihop MAC (MMAC) protocol which unifies neighbor recognition functionality in every direction is needed can be used, and even elements from Simultaneous Transmission and Reception (STR) protocols can help to configure the protocol for sharing the medium. Finally, variants of the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol, properly adapted for directional carrier sensing, can also be used [5].

4.3 Radio Access Network (5G New RAN)

To support the entire range of communications that shall be included in 5G (P2x, M2x, V2x, x2x) and therefore all the connectivity techniques and methods associated with them (multi-tier cells, UDN, MIMO, D2D, MEC, multi-RAT, Network Slicing, etc.), Radio Access Network (RAN) is required to meet some basic specifications. It should also go hand in hand with the overall 5G vision that includes unified access for multiple technologies with different features, under a single umbrella.

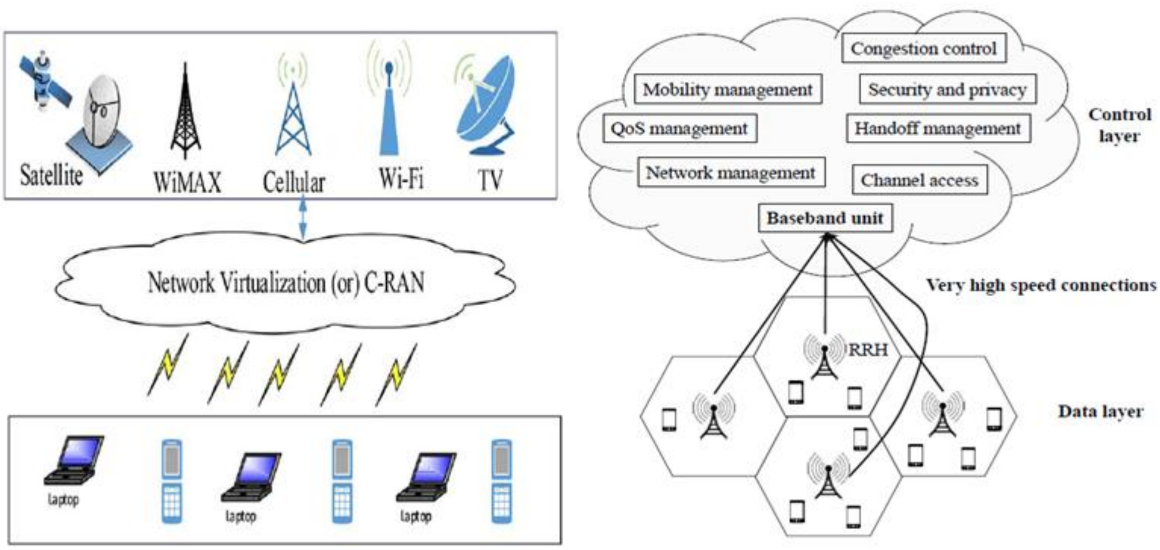


Figure 26 : 5G RAN block diagram following cloud-based architecture [47]

In this direction, it should have the ability to manage and modify the UP functions separately from CP functions as, otherwise, support for the operation of all distinct types of service and of operation of the network that will coexist in 5G becomes, in the best of cases, an especially complicated procedure. Accordingly, any modifications to RAN operation should be made quickly and with minimal cost, following a flexible framework that practically sets the existence of software-based entities as a one-way street, instead of the classical structure followed in previous generations of networks. In addition, the surge in traffic that is expected to grow constantly increases the pressure as regards the optimal utilization of natural resources (available spectrum) and network infrastructure (hardware), which can be implemented through the virtualization and centralization of processing scenarios. Finally, provision should also be made for future extensions and support for older modes of operation should not be omitted [48].

From a more practical point of view, RAT is a key part of the mobile network ecosystem and will inevitably have to evolve to keep up with the development of the other parts of the 5G ecosystem [37]. In this direction, the ideas of resources virtualization and Control Plane (CP) functions split from User Plane (UP) functions can also be applied to the (RAN) layer, thus facilitating the network's management by adding flexibility and scalability features that allow the network to dynamically adapt to changing traffic conditions and to worthily respond to user requirements [49] .

4.3.1 C-RAN Basic Architecture

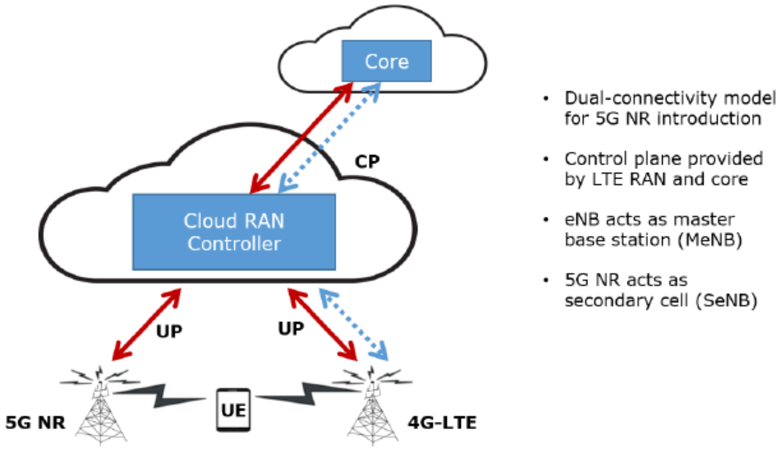


Figure 27 : Abstract view of Cloud-RAN basic operation [37]

In a first implementation step, as regards the logical nodes layer, the 5G New RAN, according to the proposals made, will be comprised of BSs that will separately manage the connectivity of terminals based on the 5G (gNBs) protocols as well as of respective BSs which will only follow the operation logic of the LTE (eNBs) [18] [37]. Thus, a terminal can be served with independent bearers from both types of nodes [37] while, in a next stage of standardization of 5G, it will be decided whether gNBs will fully integrate the operation of eNBs and the latter will cease to exist. However, BSs, regardless of their generation, they will be interconnected through the Xn interface in order to exchange necessary information for their cooperative operation while, at the core network level, the corresponding UP and CP protocols defined in CRAN will be followed for better coordination [18].

4.3.2 Main Principles of Operation

In an abstract point of view, the upgrading of RAN has been proposed to be implemented through a simultaneous double split in its operation model. The first leg concerns the separation of the Control Plane from the User Plane according to the SDN (vertical split) model, and the second leg relates to the separation of processing functions and their placement in centralized or decentralized units (horizontal split) [50].

Regarding the vertical split, the alignment of the operation with the SDN concept methodology determines that the Control Plane (CP) is responsible for the network intelligence while the User Plane (UP) is responsible only for the packet forwarding procedures. Thus, the CP in which the appropriate algorithms for setting policies and rules that guide the packets to the correct network directions can be implemented independently of the UP in a centralized way and over plain hardware, benefiting from the virtualization concept. Correspondingly, the UP which is responsible for implementing the rules provided by the CP at a lower level (i.e. redirecting the movement in the right direction) is implemented in distributed logic at the required points. Communication between them is made via standardized interfaces to ensure proper operation and satisfaction of requirements [5] [50].

As regards the horizontal split, the central idea is based on the separation of the base stations into two distinct entities, whose functions can be deployed separately, depending on the role they play in the network. Thus, instead of baseband & packet processing functionalities being implemented in the same physical entity in which radio functionalities are implemented, they are functionally separated and allowed to be implemented independently [51].

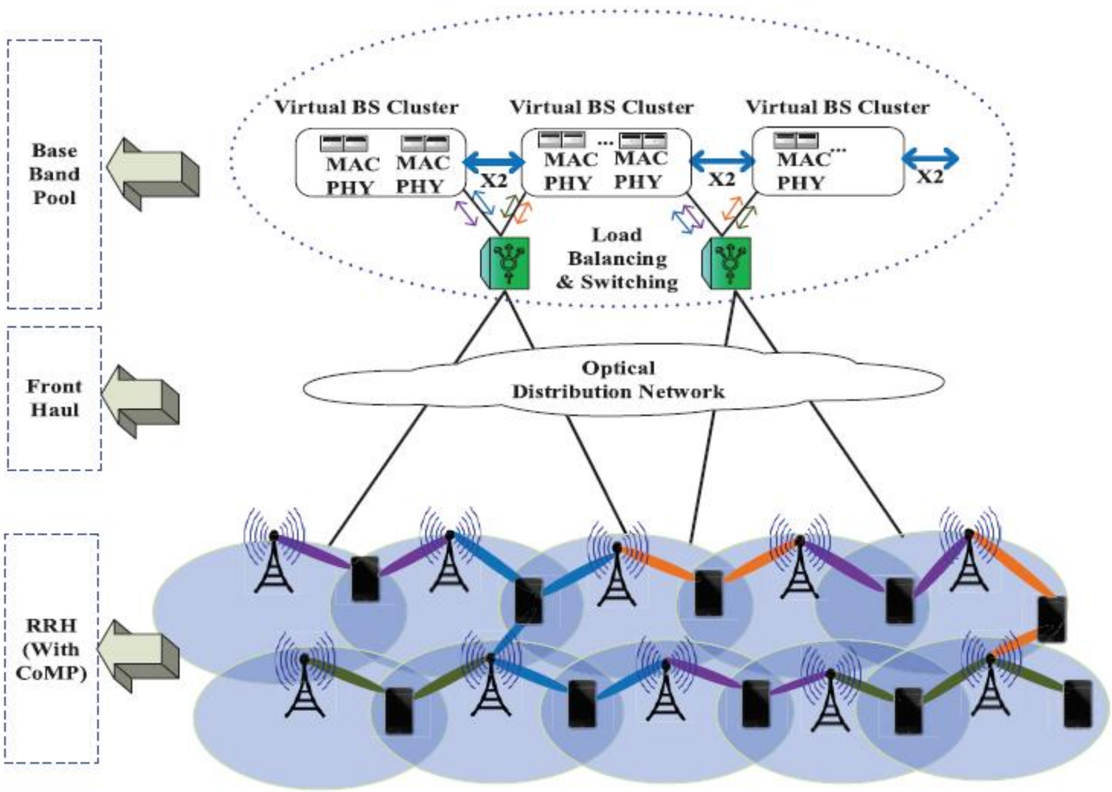


Figure 28 : 5G C-RAN architecture [5]

Going one step further, the first functions are the Baseband Unit (BBU) and are mostly related to back-end & control operations of RAT layer (such as network management, admission & congestion control, radio resource management, inter-cell handover, cell selection, encryption, upper MAC functions). These functions are mostly Non-Real Time (NRT) which means that there is a tolerance as regards the delays of a few milliseconds. Therefore, they can be collected and executed remotely from the physical infrastructure of the antennas in a centralized scheme (Central Units (CUs) located in central offices or small datacenters) as software instances over plain hardware using virtualization techniques, grouping the corresponding functionality of many BSs. By pooling multiple BBUs into larger structures, a BBU pool is created, which can be shared among different cell sites, and provides the ability to uniquely manage resources and work together in order for the RAN to benefit from the advantages of cloud computing [1] [37]. At a BBU pool level, the communication between cluster resources is made through protocol X2, as shown in the above figure [52].

The rest functions associated with Real-Time (RT) signal transmission and reception processes at lower level (lower MAC and PHY layers such as frequency conversions, amplifications, filtering, radio scheduling, link adaptation, signal transforms from digital-to-analog and analog-to-digital modes, coding, power control) make up Remote Radio Heads (RRH) and include front-end functionality. They are essentially Distributed Units (DUs) providing elementary radio signal coverage and are locally implemented where the network antennas are located (access points). It is clear from the work they perform that they have a very limited tolerance of delays, in order for them to have a better performance, and are best implemented using dedicated equipment. Cooperative Multipoint Processing (CoMP) techniques among RRHs are also applied at this level in order to maximize the utilization of the network's architecture, contributing to an increase in total capacity [1] [5] [37] [52] [53].

As regards the interconnection between BBUs and RRHs for the exchange of necessary information, this is done through the Optical Transmission Network (OTN) and the respective interface (Common Public Radio Interface - CPRI) ensuring seamless communication [51] [53] [54].

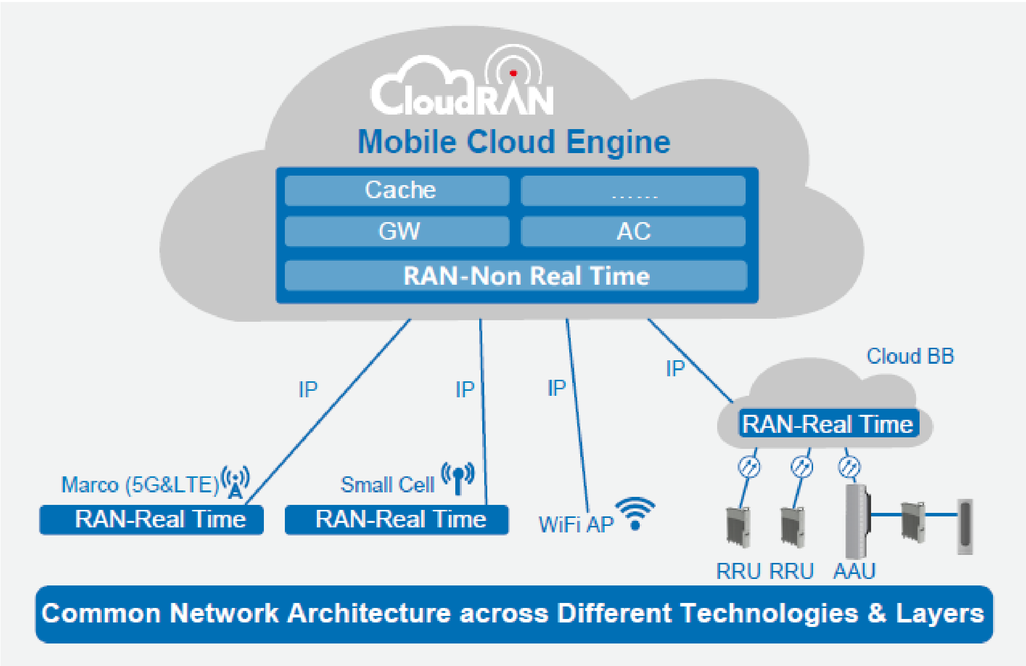


Figure 29 : The central controller unit of C-RAN which is responsive for the management & orchestration procedures of the network [8]

At a C-RAN central control and coordination level, the Mobile Cloud Engine (MCE) is responsible for the implementation of the corresponding NRT functions as well as for the general control plane and the efficient management of all existing resources that coexist in the heterogeneous 5G networks. Moreover, at this point in the network, cache and gateways functionalities are implemented as well as any others deemed necessary to support all the different access techniques that can coexist in the network [8].

4.3.3 Basic Scheme of Functional Split (Horizontal Split Options)

It should be stressed that there are several alternatives as regards the exact point at which the split of the protocol stack will be performed, i.e. exactly what parts of the network's operation (concerning both CP and UP functionalities) will be executed in CUs and which ones in DUs respectively. Thus, either a complete migration of all baseband processing functionalities to the CUs can be followed by leaving only the Radio Frequency (RF) part of the stack on the DUs, or more sub-functions can be left in the DUs, other than the radio functionalities [37].

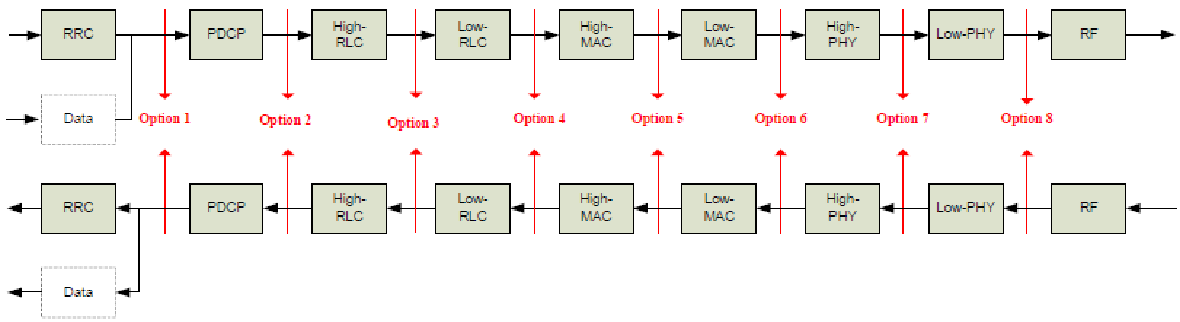


Figure 30 : Uplink and downlink horizontal split options [55]

Depending on the point of separation, the respective burden of the underlying interface (backhaul / midhaul / fronthaul) for the interconnection between two planes (in terms of delay) is affected, as well as the possible benefits produced by the cooperative scheduling (CoMP) under a powerful central coordinator. This essentially highlights the existence of a trade-off between the overall performance of the system and the performance on the fronthaul link concerning the interconnection among CUs and DUs [37].

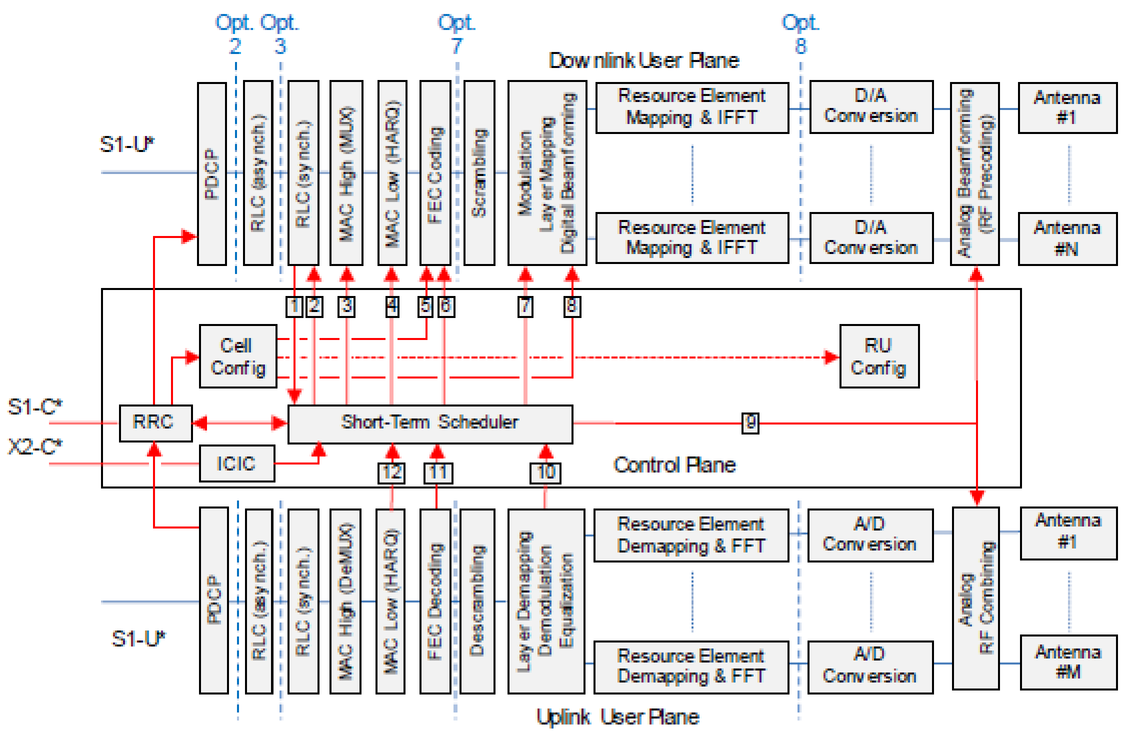


Figure 31 : Interconnections between the separated UP and CP in the RAN [50]

The main goal of this separation for the network to benefit as a whole from the centralization of functions in order to better coordinate the transmissions, as well as to allocate them to CUs or DUs according to the ad-hoc performance principles of the provided services [50]. Finding a happy medium for the stack division that will maximize benefits and minimize negative impacts depends on various factors such as the environment in which it is applied (OTN in the application area, computing power of CUs) and on business models followed by each provider. For this reason, the precise point of intersection cannot be predetermined, but it should be decided by ad-hoc area by the network administrators in order to better meet the requirements [37].

4.3.4 Topology Options (Vertical Spit Principles)

Depending on the requirements of the terminals served by specific network components within the differentiated services model, different deployment scenarios can be implemented [8] [53]. As far as RAN is concerned, this means that different variants can be configured for the placement of infrastructures supporting BBU functionality (MCEs) on the network depending on the tolerance level to network response delays that each respective served service (slice) has [37]. The factor that is essentially determining the choice of the corresponding architecture is the cost incurred in the front-haul during BBU-RRH communication and its impact on the quality of each respective service [53].

Based on the above, it is possible to place the CUs in datacenters (DCs) that are either close to the edge of the network - RRHs (simulating the classical architecture in which everything was implemented in the BSs area) or in more or less remote devices [37].

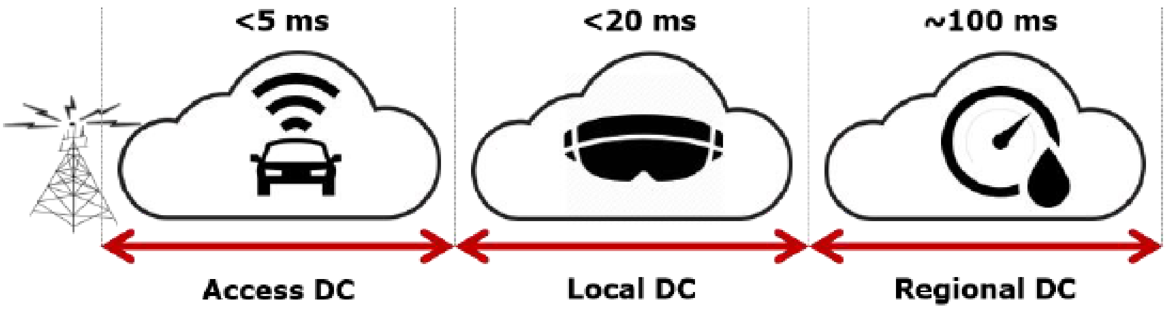


Figure 32 : CUs placement options depending on each use case scenario needs [37]

For example, when applications are too delay-sensitive and the tolerable delay is less than 1 ms (such as assisted driving), then it is convenient for the RAN-level information processing entity to be closest to the end-user (or to the RRH at the infrastructure level). Correspondingly, as the tolerance for service delays increases, MCE can be removed from the end user. Thus, for less delay-sensitive services where a 20 ms delay is not as critical for the QoE (like virtual reality applications), then the CU can be placed in local-type facilities in order to have a happy medium between cost performance and serving delay. Finally, for applications that are not particularly demanding in terms of delays (like sensors), the implementation of the corresponding functions can be done in MCEs whose physical location is even more remote from the edge of the network and is done at a regional level [37].

It follows from the above that as many implementations are made mainly in regional DCs and secondarily in local DCs, the range of resource and interference coordination increases, as well as resource savings which, in turn, causes a reduction in network operating costs [53]. In this sense, mixed models can also be followed in which the computer resources are located in separate DC-layers from the storage resources in an attempt to better balance the benefits of the cloud architecture and the service delay resulting thereof [8].

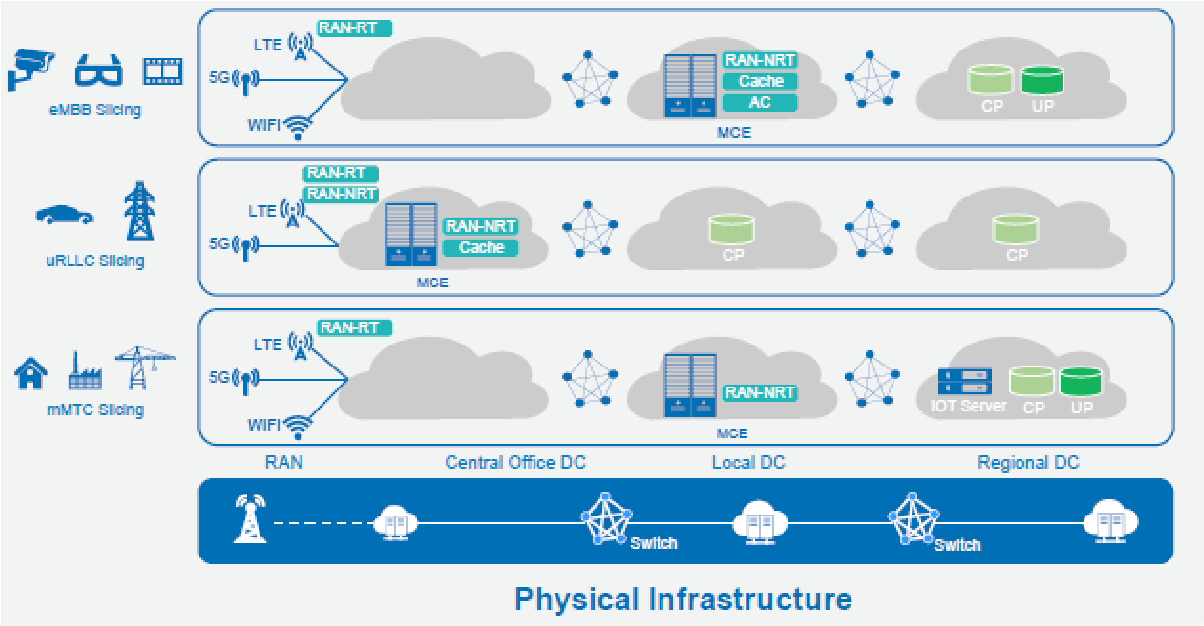


Figure 33 : Differentiated network topology according to different service requirements [8]

4.3.5 C-RAN Benefits

Benefits are expected from the vertical split in terms of easier network management, since its operation at a CP level is made by following specific operating templates independently of manufactures of network equipment. Thus, the network can be uniformly "managed" at this level, providing universal service quality across its entirety. In addition, the flexibility of network development, maintenance and upgrading is increased since changes in control functions do not necessarily entail changes in equipment forwarding, since they are implemented completely independently while saving time and money [50].

In the general case, the benefits of aggregating BBUs to a smaller number of forceful physical entities (BBU pools) with high computational power when only RRHs (which execute light functionalities) are scattered on the network greatly help network operators achieve the performance goals of the new networks. The ability to share and dynamically allocate computing resources among BBUs enables the network to on-demand manage and adjust its operating parameters, bypassing the static distribution of resources applied to older generations and imposed severe restrictions [54]. This results in better load balancing across the entire network infrastructure as well as upgrading of equipment to 100%, which automatically translates into the increase of performance of the network (throughput, capacity, delay, data rates, etc.) [5] [51].

The fact that baseband functionality of multiple BSs is grouped into larger, centralized structures that operate in a centralized way automatically gives network operators the ability to reduce the cost of procurement and operation of the equipment (purchase, maintenance, expansion, modification, power consumption), since the equipment is now concentrated on certain points and not scattered throughout the network, while expensive dedicated hardware is replaced by cheaper common servers which, using virtualization techniques, implement all the necessary functions of the access network [50] [52] [54].

Furthermore, the unified scheduling of transmissions allows for better signal quality in the cells since better transmission coordination can be achieved, resulting in a noticeable reduction of interference and better multiplexing. Optimal use of the available spectrum is therefore apparent, irrespective of the topology of access and transport networks [5] [37]. This fact contributes decisively to the proper implementation of UDNs that have particularly increased spectrum and interference management requirements [37].

The two abovementioned parameters of the C-RAN (functionalities centralization & unified scheduling) are features that enable the network at all times to be aware of the situation throughout its operating range. This feature is particularly useful for the automation of operation, administration and management (OA&M) functionalities which can be achieved within the Self-Organizing Networks (SON) project. Thus, if the radio access network can (through the analysis of traffic patterns) perceive the ongoing trends in real time (in terms of traffic congestion, capacity, etc.), then it can (through appropriate mechanisms) adjust and appropriately configure its operating characteristics in order to have optimal performance every time. As a result, there is no need to solve human resources, money and time issues. The conventional method would require multiple human resources, money and time [37].

At the same time, device mobility is enhanced by shifting the switching point to a lower level, requiring less administrative effort and allowing the network to respond faster to similar events. In this case, when a device changes an area in the network, its service goes on uninterrupted to the back-end by the same functions and simply connects to the network from a different access point [44]. This feature is extremely important to support the multi-connectivity and multi-tier cellular architecture of 5G [37] [52].

The simplified current architecture at RAN accelerates the expansion and upgrade processes of the network, since it allows separate development of the BBUs from the RRH functions, especially since the former are implemented as software instances. This is even easier to do by reprogramming the corresponding VMs [5]. This feature also offers a new fail-safe to the network, since implementation on a cloud-based environment allows a faster recovery of the system in the event of physical disruption of the infrastructure [8]. This new access network structure also results in a reduction in the time required for the development of a new service or the evolution of an existing one, ultimately facilitating service developers to scale-up the network functionality [5].

From a less technical point of view, C-RAN contributes to the development of new business models since the provider can lease from the third parties the necessary resources or even the baseband functions (cloud-as-a-service) required for the operation of the network [52]. Through appropriate APIs it can handle these resources in order to implement the necessary services for its network. As regards equipment itself, greater independence from vendors is allowed since COTS materials are used and interoperability factors between core network and radio access network are improved [44].

4.3.6 C-RAN Challenges

In order to achieve the vertical split, it will be necessary to specify exactly what RAN's Network Functions will be an element of the CP and which ones are on the UP's side respectively. The process is very complex, since the lower the functions in the radio protocol stack are, the closer they are interconnected and a possible bad partition can detrimentally affect their performance. For this reason, extensive testing is required to ensure their global operation and communication between them, for each vendor providing either CP or UP functionality. In addition, a point that requires special care in order not to hamper the flexible operation of the model as regards the introduction of new network functions is the requirement for corresponding extension of the interfaces that are responsible for the universal interconnection between the two planes [50].

Another challenge is the creation of the OTN so that it can adequately support the BBUs - RRHs interconnection, according to protocols that regulate this type of communication (CRPI). The requirements, in terms of bandwidth, latency, jitter, synchronization and security, are particularly rigorous in order to guarantee the 24/7 protocol performance [54] and the overhead caused by this communication does not work to the detriment of the overall performance of the system. Also, the whole process should be flexible and should support collaborations between different topologies and implementations that may involve a large number of nodes (BBUs & RRHs) [7] [19] [52].

5. NETWORK FUNCTIONS VIRTUALIZATION (NFV)

5.1 Introduction

Supporting a wide range of distinct communications (with separate operating requirements) that are to coexist in the 5G ecosystem is a big bet for network operators. In addition to the huge number of communications that will occur on the network, the heterogeneity of the various functions required to provide services is another parameter to be taken into account by designers [56].

The traditional model of Core Network that has been followed over the past years by the providers, must undoubtedly be harmonized with the new requirements. Namely, the purchase, operation and maintenance of dedicated physical equipment for every different network service model, can obviously no longer cope efficiently with these expectations. Every network service implementation is essentially strictly correlated with the network resources in which is deployed. Apart from the prohibitive cost (in terms of CAPEX and OPEX) that the implementation of this model would have in the future network for providers, as each new type of service would need to follow the same procedure from the beginning, it would be particularly inefficient, as it would not provide the network with the necessary flexibility degrees needed to directly adapt its functions, depending on the requirements that are now dynamically and not statically configured, as before. This necessarily led to a long life cycle of existing setups of the network, not allowing, on the one hand, the network to readjust its on-demand operating parameters in order to provide better services and, on the other hand, discouraging the development of new services as the process is very time-consuming [17] [57]. If one considers that the flexibility of readjusting the underlying functions of a network service is an integral part of the context/location/time/service-aware applications that are becoming more and more popular and are expected to play a leading role in the near future [56], then the need to change this mode of operation becomes imperative.

Given this problem in network operation, it was deemed appropriate to have a flexible design of Network Services (NSs) & Network Functions (NFs) by combining the appropriate functionalities required each time, practically disregarding the positioning within the network of the respective hardware components in which the functions are carried out as pure software [17] [56] [58]. The Network Functions Virtualization (NFV) concept, which is based on this concept, essentially transits from a closed, vertically integrated model, in which each function of a service runs exclusively on a specialized appliance (which is even positioned on specific network points) to a virtualized model where all network functions can be implemented on demand as a software instance (softwarization) that runs on virtualized layer (virtualization), namely in Virtual Machines (VMs) [14] [59] [60]. This means that functions run on commercial off-the-shelf (COTS) equipment located in a datacenter of the operator or in cloud as (NF-as-a-service) or in distributed locations or at any other point deemed appropriate for optimal network performance [58] [60]. Also, it is feasible to follow mixed scenarios where a part of the function set of a service running on virtualized components, while the rest of them staying on physical resources as Physical Network Functions (PNFs) [13]. This option is very helpful for a smooth shift from traditional vertical model to NFV [17].

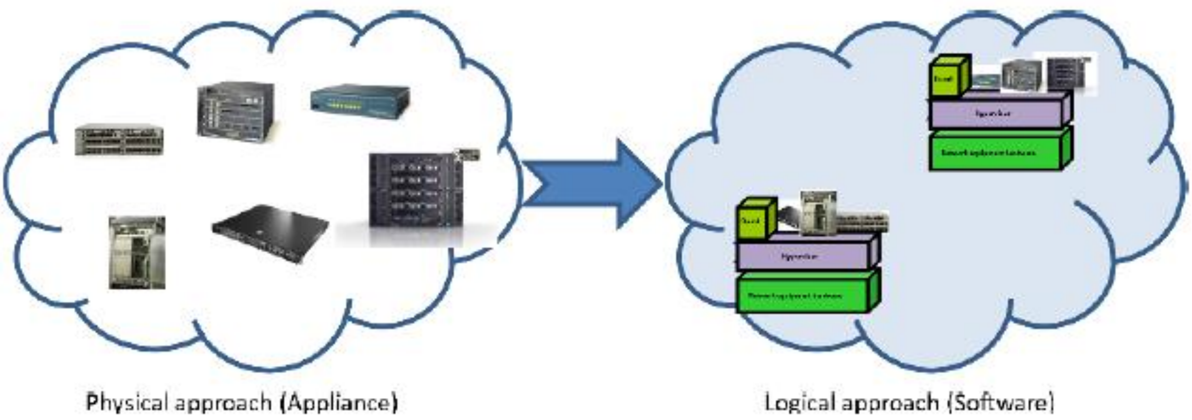


Figure 34 : Transition from a closed vertically integrated model to a flexible virtualized model for the functions implementation [60]

Based on the new model, functionality (software) is completely separated from physical implementation (hardware) by creating Virtual Network Functions (VNFs) that can be independently implemented and combined with each other with a fairly high degree of freedom in order to create the final Network Service (e.g. firewall). In this way, increased network's adaptability capabilities are provided for each requirement (dynamic scalability), its overall management and maintenance is facilitated, as well as the development of new services while, at the same time, the utilization of the equipment is maximized and the corresponding costs for energy, equipment and man-hours are minimized [17] [57] [58].

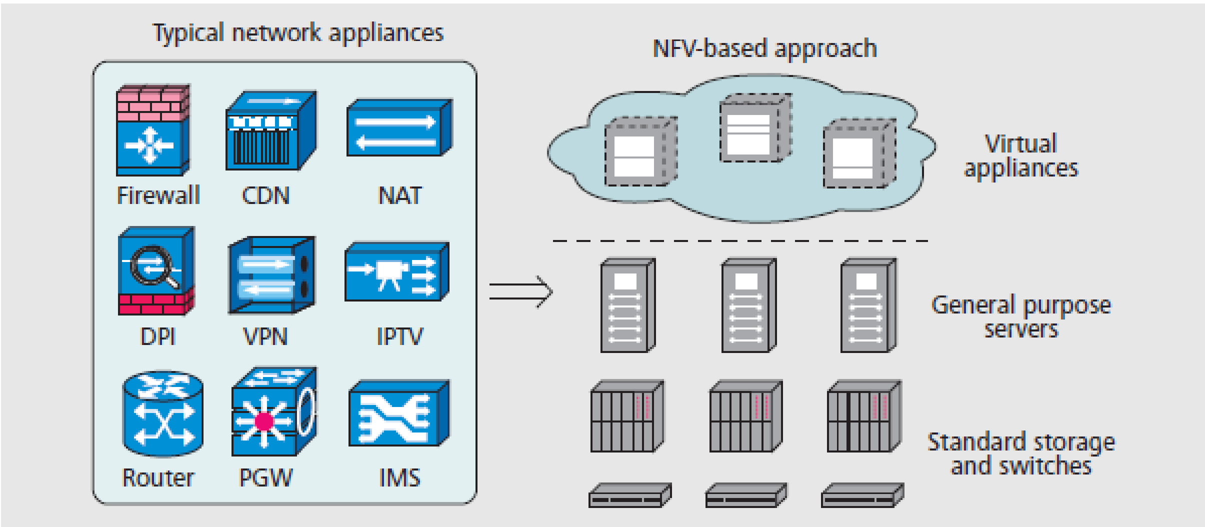


Figure 35 : Separation of network functionality from the physical hardware [59]

Some indicative examples of network activities that are able to espouse the softwarized & virtualized model and run as VNFs contain functionality of routing elements (routers, switches), network nodes (gateways, base stations, Network address translation - NAT - servers, Dynamic Host Configuration Protocol - DHCP - servers), signaling elements (Session Border Controllers – SBCs for Next Generation Networks - NGN), traffic controlling and monitoring elements (QoS and QoE measurements, diagnostics), service assurance entities (Service Level Agreement - SLA), security devices (firewalls, virus scanners) and virtual home environments [17] [58].

5.2 NFV Model Framework

The wide range of actions and their interdependencies in terms of functionality included in the NFV concept includes various network entities and operations (Network Services Physical and/or Virtual Functions/Links, Virtual Forwarding Graphs, Functions Orchestrator) that can be divided into three levels for better representation and organization of the model.

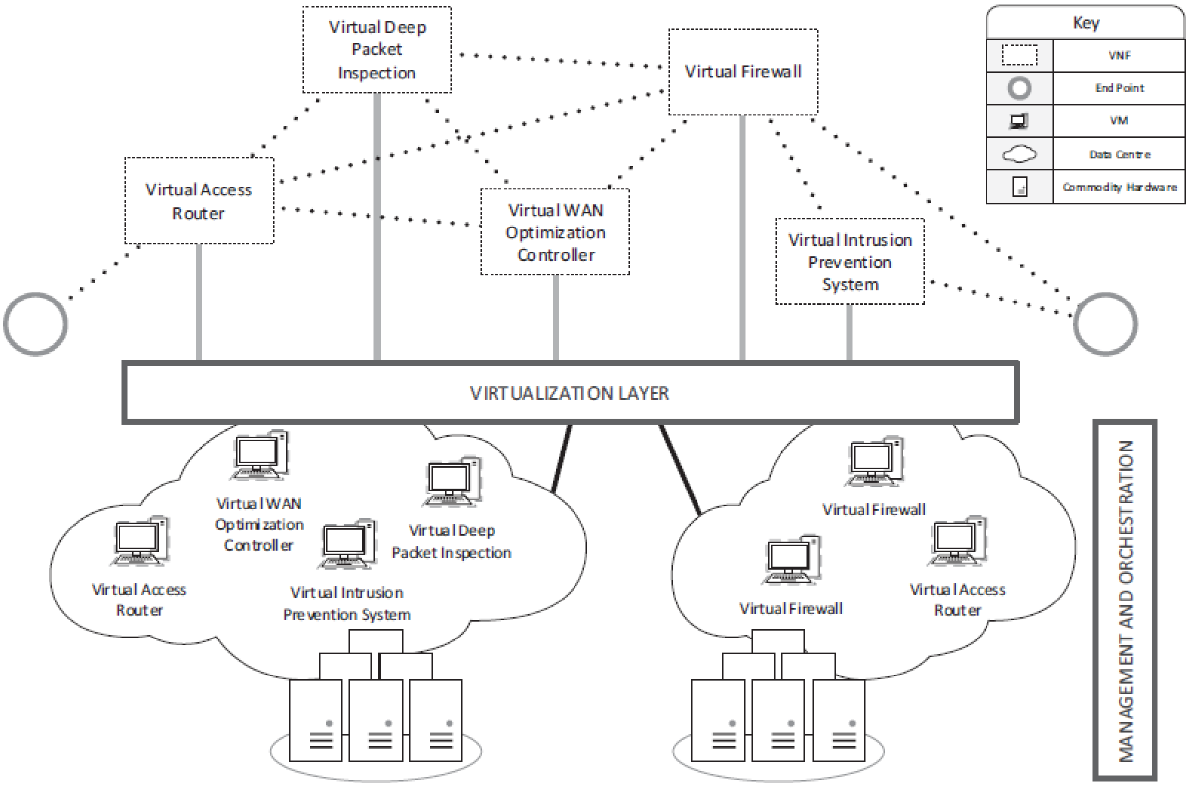


Figure 36 : Basic structure of NFV model [61]

The first section concerns the *Network Function Virtualization Infrastructure (NFVI)* which covers the fields of appliance and software that required for building and execution of VNFs. In an abstract view these resources contains computer, storage and network entities while in terms of physical hardware they are commodity servers, storage volumes and network elements (links, ports, networks, subnets, addresses, forwarding rules etc) for the connections between all of them [17] [13]. In virtual mode the above mentioned computational and storage resources represented as Virtual Machines (VMs) as well as the network nodes which essentially are software components in VMs running network algorithms (e.g. routing). Also, virtual network elements act as logical network items for the interconnections between all of these resources (virtual or not). The virtualization of the resources implemented through a software layer which dissociates the functionality from the subjacent physical equipment, in such a way that it is not perceived by the end-user if is served from virtual or physical resource [17] [58]. Hypervisors based on cloud technology, such as OpenStack and Open vSwitch are ideal for the task of virtualization [59]. It is worth noting that in a Network Service deployment both hardware and software may follow either virtual or classic model, in order to fulfill the duties of fully-virtualized network functions or semi-virtualized network functions respectively [17].

The second part of the NFV architecture is about *Virtual Network Functions and Services.* A VNF is not nothing more than a NF that instead of deployed on dedicated physical network equipment, it is deployed on virtual platform. As in the traditional model where a network function is a set of appropriate defined instructions that produce a specified functionality, the same applies to virtual model. In this new model, each VNF is typically associated to one VM which implements its functionality, but as a VNF may consists of multiple inherent parts with each of them deployed in separate VM, it is possible the implementation could take effect over numerous VMs. Similarly, a Virtual Network Service is a composition of several VNFs (virtual deployment) that in their wholeness form a new operation [13] [17] [58].

The last part of the framework named *NFV Management and Orchestration (NFV MANO)* and has principal role in the whole operation since the entity of this field is the director of the NFV concept. More specifically, NFV MANO is responsible for the management, monitoring, configuration, provisioning and orchestration of VNFs use as well as both of physical and virtual resources of the system (compute / storage / network elements). Furthermore, it involves databases in order to store metadata which are about the operation and the lifetime of the functions, services and the infrastructure of the network that managed. Especially for VNFs, management includes operations for creation, termination, authorization, scaling and update issues. Greater emphasis is given in NFV orchestration and in developing of suitable interfaces for the interconnection between all the units of the system either they are virtual or not [13] [17] [58].

5.3 Business Model

Based on the structure and principles of the NFV concept, a possible organization of the entities involved and their relationships within a modular business model is as follows: [17]

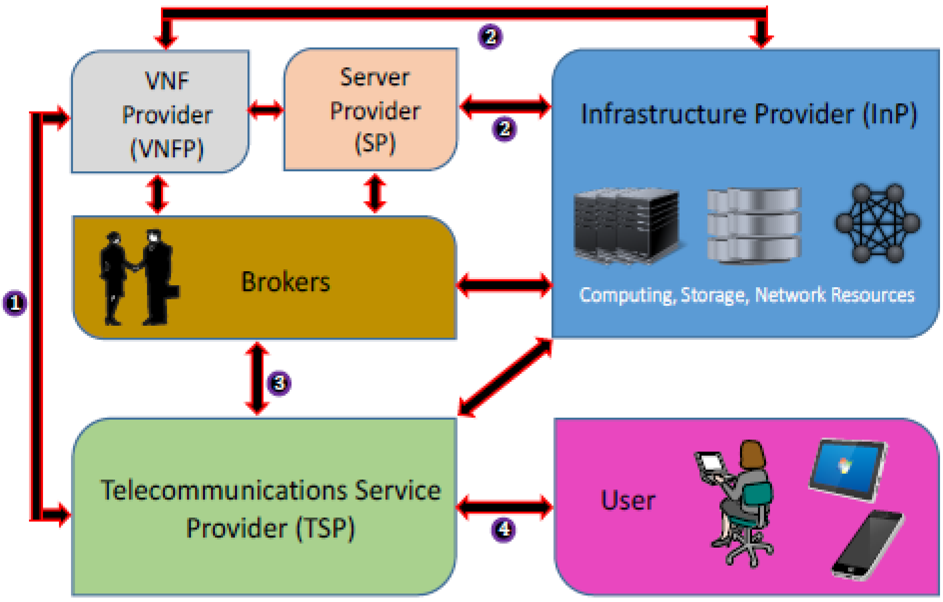


Figure 37 : NFV activity pattern [17]

– **Infrastructure Provider (InP)**: this entity includes physical hardware and related installation, maintenance and upgrade related activities and can be provided as a single service to any network service provider that does not have proprietary equipment for VNFs deployment. It essentially has a cloud environment (such as those provided by Amazon as computing resources), parts of which can be leased to multiple providers who use them through appropriate APIs. In addition, if an InP's resources are insufficient to operate an NF, the cooperation of departments between different InPs is possible in order to meet the increased needs that arise.

- **Telecommunication Service Provider (TSP)**: as understood by their name, TSPs provide telecommunication services to end users. In order to do this, they implement the necessary VNFs (individual or combinations thereof) that eventually run either in proprietary equipment (in such case InP and TSP is a unique bundled entity) or in resources that have been binded by InPs for a corresponding use. Within the context of partnerships, TSPs are feasible to sublet the resources they have binded from InPs to other providers in order to create more complex structures.

- **VNF Providers (VNFPs)**: the role of this entity is to provide programming-layer implementations of VNFs directly to TSPs or InPs which in turn sublet VNFs together with underlying equipment (as a ready cloud service) to TSPs. These implementations are hardware independent and some of them can be implemented by TSP itself (custom VNFs).

- **Server Providers (SPs)**: are the vendors of common servers that needed from InPs and/or TSPs in order to run VNFs. It is feasible for a SP to be both VNFP but, in any case, both software and hardware should not have dependencies and should be distributed and operate autonomously even with competing VNFs or servers accordingly.

- **Brokers**: an entity of this form has the role of an intermediary for the implementation of interconnections between different entities. It is useful when multi-VNFPs and/or SPs are to work together with TSPs and/or InPs to create complex VNFs in order to provide complex end-to-end Network Services. Correspondingly, it can be used to find IPnS when demanded by TSPs, acting as an intermediary in charge of finding resources and negotiating co-operation.

- **End User**: is the end consumer of the services provided by telecommunication providers (TSPs), who has the ability to use a variety of services from a range made available from providers.

5.4 Examples of Application

Although (theoretically) all of the NFs and NSs can also be virtualized by following specific standards, here are some indicative examples that can be implemented in the NFVI:

- **Virtualization of Mobile Core Network**: the virtualization of the entire core network can be made by significantly reducing the complexity of this part of the network that uses a large number of proprietary equipment often situated in inappropriate locations from a topological perspective. This favors the adaptability of the network to dynamically changing traffic conditions and allows a better allocation of virtualized resources at the points of the network that require greater emphasis. As a result, network performance is maximized and all types of communication (such as voice calls and others under development, like M2M, etc.) benefit instead of being underestimated over others. Furthermore, there can be benefit from grouping network functions into fewer entities based on workload and connectivity needs between them, avoiding delays and bottlenecks. The application of NFV to this part of the network requires, first and foremost, to ensure the reliability of the NFV model and its efficient operation in IoT environments where there is high bearer consumption [17] [57] [59] [61].

- **Virtualization of Mobile Base Station Network**: the implementation of all the functions performed on BSs as VNFs that runs in VMs is expected to give great impetus to the development of mobile communications networks since it greatly shrinks the times and costs required for their operation. In addition, it allows the sharing of BSs between different providers so that everyone benefits by achieving, on the one hand, a maximal geographic coverage (service area) and by sharing, on the other hand, the costs of operating, maintaining and upgrading the stations. However, inter alia, the BS's competencies include physical layer (OSI layer 1) functionalities, the virtualization of which is a challenge for the research community, and it is why the application of the NFV technique initially only to upper-level functions is proposed [57] [61].

- **Virtualization of Home Environment**: the execution of network functions in customer premise equipment (CPE) can be avoided through the use of VNFs. Thus, performing the necessary functions is now done in datacenters and not in special equipment on the users' side, essentially negating its existence. Instead, all that the provider needs to do is to offer end-users simplified and low-cost devices for basic physical connectivity (OSI layer 2 functionalities), as the functions of the above layers are executed on VMs. The advantages in this case are many, since this mode of operation has the effect of simplifying the network architecture, reducing installation, maintenance, renewal, extension, and troubleshooting costs of household equipment and, the most important, providing better QoE since the network can directly and effectively manage multi-dimensional user requirements. In addition, the provision of complex NSs that require communication of a series of NFs (function chains) and the introduction of new services is facilitated, since no change in physical equipment is required in order to support them; only possible changes to the software that are made on a centralized level over the entire network. The challenge in this case is to ensure the same performance as with the traditional technique and to investigate possible privacy issues that may arise [17] [59] [61].

- **Virtualization of Content Delivery Networks**: the application of NFV does not require unnecessary expenses as regards cache nodes, which is extensively used to provide quality multimedia content. In NFV, a flexible management of sources and of the quality of the managed information traveling on the network, which can be adapted to suit the needs through appropriate interfaces, is possible [61].

- **Fixed Access Network Functions Virtualization**: coexistence and co-operation between the NFVI and traditional infrastructures is also possible through the use of existing equipment as a part or sub-part of the virtual environment to which specific functions are assigned [61].

- **NFV as a Service**: the ability of all elements of an NFV platform to be provided by a third external entity as a bundled service (Outsourced NFV) [61].

5.5 Advantages

As it has already been briefly mentioned, the implementation of the NFV model in communications networks, taking advantage of the revolutionary benefits of virtualization techniques, brings about significant changes that benefit both providers and end-users. As far as providers are concerned, the benefits arise mainly from large-scale savings, which are the result of three factors. The first is to minimize the need for specific equipment purchases by specific vendors and the recruitment of suitably trained human resources for its operation; the second one is to better allocate and exploit the full potential of cheaper commodity hardware and the third is related to de facto reduction of energy consumption resulting from the supply of the entire equipment. This practically means that instead of spending money on purchasing and maintaining multi-cost dedicated components, the providers invest in common hardware in datacenters or cloud implementations where many VMs that implement different network operations can be hosted [17] [59].

Yet another factor that strengthens the trend of lowering costs and increasing network performance is the ability to flexibly update, expand or upgrade network functionalities in a minimum of time and without affecting other operations, if not necessary, as changes occur on a software level to the suitable point (VM) and not to each BS as it would be required in traditional networks [17]. The fact that, in addition to elaborate cases, the addition of a new functionality requires the development of only corresponding software that will run on existing hardware and not on a hardware extension reduces the time it takes to market new functionalities and, at the same time, gives great potential for the testing and verification of new and innovative operations at minimal cost in an efficient way [59].

In addition to economic benefits, the software - hardware separation architecture and the NFV's functionality are characterized by an openness framework that provides increased degrees of freedom in the installation, operation, maintenance and management of the network in terms of flexibility and scalability. These features offer vast network capabilities such as independent physical infrastructure and software development, the creation of VNFs in as small a portion as possible in order to make them easy to reuse, the customization of VNFs interfaces in order to provide the desired services in the best possible way in a dynamic dimension based on the current needs and not statically, as was the case with the classical architecture of mobile networks [57] [59].

All of the above benefits to the providers have ultimately an impact on the end-user of the network since he/she can enjoy higher-quality services that best meet his/her needs, by being constantly updated or upgraded, and all this by paying a lower price as the significant reduction of the costs of the network is now feasible [59].

5.6 Challenges and Open Issues

The key to the successful implementation of the NFV model is the further study of some open issues concerning the operation of the NFV concept in the next generation networks. The key elements of this issue are the efficiency and reliability of VNFs, the placement of VNFs at the appropriate points of the network in relation to the level of decoupling from the underlying hardware, the interoperability between them and between traditional implementations, the provision of guarantees for security issues, the evaluation and verification of their functionality and the ability to manage & orchestrate the large-scale heterogeneity of the network [57] [59].

As regards the efficiency of NFVs, it is also essential to consistently provide a high quality of service comparable to traditional systems operating in special purpose equipment so that the consumers of network services are not aware of the environment in which they are implemented. Studies have shown that how the sharing of resources (processors, memory, caches) in cloud services can lead the system to have significant fluctuations as regards its performance (in terms of throughput, latency), challenging the limitation of this instability. This requires, on the one hand, the exploitation of the underlying hardware to the fullest and, on the other hand, the transition of functions to the virtual environment, taking into account all the parameters required to entirely do that without any malfunctions. This transition may even include some modifications to the current form of the function code in order to efficiently work in a virtual infrastructure [57] [59]. Based on the above, it is easy to see that ensuring network performance is a complex process and, in order to be integrated, it is necessary to set standards that should be followed by all the NFV entities involved, a feature that requires more effort from designers [17].

Although one of NFV's main goals is the flexible operation of the network, it is not as easy for all NFs to operate as efficiently as virtual NFs as they were operating in the traditional model. For this reason, in some cases, it is necessary to use data from specific hardware so that there is no reduction in the VNF performance. It is clear that this is at the expense of flexibility and portability aspects, since there is a small deviation from the theoretical NFV operating model that requires the use of simple common hardware and hardware independency for all the NFs. It is therefore useful to define the appropriate standards that define the tradeoff between flexibility, portability and performance of VNF to render hardware specification to such an extent that it does not prevent the network from being flexible and portable [17] [58].

What should be adequately ensured regarding the reliability of the system is its operation even when there is a high traffic load and when extensions or modifications are required on a software level. That is, there should be a given elasticity and tolerance of the system in terms of errors that may occur in these demanding conditions as well as the possibility of recovery from them. The same level of requirements exists in cases where a failure occurs in the shared underlying hardware. In this case, it is necessary to locate and isolate the defective point immediately so that the smooth operation of the other VNFs is not affected. On the same pattern, an open issue for research remains the process of transition from the traditional to the new model or the transitions between different virtual deployment platforms where different requirements need to be matched and appropriate indicators must be defined in order to assess the reliability and stability of the system [59].

Another factor that influences the system's performance and requires a more extensive study is finding the right point within the network for the placement of the virtual resources so that it is possible to serve the users with the best possible service quality, the easy interconnection between VNFs in order to provide complex NSs (VNF chains), as well as the addition of new functions on demand without affecting other operating parameters. This factor becomes particularly critical when network functions are involved with very specific constraints, which are also affected by network topology. For example, if a network function was designed and operated based on the classic model closer to the end of the network and, based on the NFV concept, the implementation is eventually made in a provider's core datacenter, it is likely to affect its profitability from a potential traffic congestion on some intermediate path of the route. Also, when it comes to VNF chains where interconnections are required between functions, account should also be taken of possible delays caused by intermediate entities (Virtual Switch, Network Interface Controller) that are used for their connection. Finding the best position to place the equipment is a very complicated process and requires a lot of effort as it is of a flowing nature, given that the characteristics of the problem are constantly changing. If it is not addressed efficiently and there is a poor distribution of functions in the network, then the system may be unable to efficiently manage VNFs, eventually creating a vicious circle since the problem that is being attempted to be resolved is created once again.

Confidentiality, integrity and availability are the fundamental aspects of network communications and securing them is always a challenge for all those involved in the communications networks sector. Potential attack threats to network security are expected to be significantly more in the virtualized environment than in the classic one because of the network's architecture. Resource (CPU, memory) and network infrastructure (vSwitches, Controllers) sharing between different VMs, the interaction through Application Programming Interfaces (APIs) and the existence of hypervisors multiply the system’s vulnerabilities to potential attacks that are intended to damage its security, rendering its protection procedure particularly demanding. The ways of sharing the virtual environment should be further studied in order to provide an abstract-level separation in separate domains in each of which there may be separate threads with clear rules for issues such as computing resource allocation and rights (protected access) based on which all procedures and entities will comply with and will operate [57].

Another demanding issue that is remaining for the research community is the effective management and orchestration of the VNFs. It is particularly important that these procedures are simple, flexible and as automated as possible in order to respond appropriately to the needs of the network. Therefore, specific standards should be agreed upon and followed in order to appropriately reflect the flow of processes needed for the interconnection between NF images and virtual infrastructure via the appropriate interfaces [57]. It is equally important to have backwards compatibility with legacy systems so that the transition from the traditional model to the new one can be a smooth one [17].

Apart from the above, the possibilities for outsourcing certain functions to third-party service providers should be explored extensively, a scenario theoretically feasible based on NFV's operating mode and architecture. However, the matter of selection of the appropriate VNFs that can be provided in this way without affecting the overall quality of the services provided remains open [59].

Lastly, emphasis should be put on creating the appropriate northbound and southbound APIs that should cover the full range of possible interactions between VNFs and controllers on the one hand and the underlying hardware on the other, respectively. This procedure requires precision in order to define all the necessary interfaces according to the defined frameworks and standardization according to the NFV core principles, a particularly challenging element for researchers [17].

6. SOFTWARE DEFINED NETWORKING (SDN)

6.1 Introduction

The multidimensional structure that 5G has due to the large number of technologies it incorporates in conjunction with the ever-increasing amount of information expected to be traded by the various types of network terminals significantly increases the complexity of management and efficient operation of the network. Just to think about the density and the number of BSs that will work in the context of the new air interface and of UDNs and multi-tier cellular concepts is enough to understand the magnitude of the need for an advanced and ultra-efficient framework, which will be able to handle the coordination of all network devices in order to achieve smooth data transfer to the network.

Traditionally, data networks consist of a variety of decentralized and dedicated devices (routers, switches and other network middleboxes) that are responsible for the routing and the proper traffic manipulation between network nodes in order to serve the demand for communication. Their operation is based on complex protocols and policies configured and suitably defined by network administrators so that their operation responds as closely as possible to all possible demand scenarios [16]. These devices are provided by a small set of vendors that provide the final product to network operators in a black box format, i.e. a closed model where each device can have specific hardware, an individual operating system and a distinct control interface is followed [62].

The configuration of the network devices in order to implement the desired operation is required by network administrators using appropriate network commands that are often vendor specific. Based on the closed mode model, the programming procedure of the same type of network device that implements a common functionality (e.g. a router which implements standard forwarding functions) may be completely different when it comes from a different manufacturer [62]. This process is likely to need to be performed by someone physically present at each of the involved devices, separately, and, considering the network topology upon which the equipment can be found scattered in remote locations, then one can understand the size of the entire process. The same actions should be taken every time some sort of readjustment to the operation of network devices is required so that the network can meet every possible scenario that may occur, such as damage to the network's infrastructure, errors at the implementation level, changes in traffic patterns, security issues or entrance of new applications [16] [62].

If one adds to the above the fact that the network devices include, under a single structure (bundled as black box), both the processes of the control plane that are liable for handling network traffic and those of the data plane that implement traffic forwarding to the correct destinations based on rules provided by the control plane, then the entire process of efficient management and fine-grained configuration of the network is a huge challenge for network operators [16] [62].

The traditional CP and DP coexistence tactic seemed able to provide a satisfactory return on the network during the initial planning phase of its operation, giving some degree of flexibility to the entire project that, at that time, seemed to be enough. However, as it later turned out, this model was rather cumbersome and complicated and prevented the easy development of the network, since it had many and powerful restrictions [62].

Due to the size of the network and the importance it has now acquired in the financial (and not only) life, it is easy to understand that any misguided or delayed decision at the level of traffic management which can, in turn, adversely affect the overall performance of the system may has a significant direct or indirect impact on society [16]. Therefore, the requirement for flexible management that is capable of following trends in network development and growth is now necessary more than ever [62].

6.2 A New Networking Paradigm

The SDN concept is an attempt to address all the problems and constraints that arise from the networks' operation and management mode, following the traditional model. By introducing a new network scheme based on a different architecture and rules, it aims to remove the constraints of the existing operation mode, to simplify management procedures and to promote further development and innovation in this area [16] [62]. Its basic idea is to manage and orchestrate all network devices via centralized controllers, basically concentrating the brainpower of all network devices in a powerful node and letting scattered devices run a lower level command (layer-2 in OSI stack) based on high-level policies and rules provided by them [47].

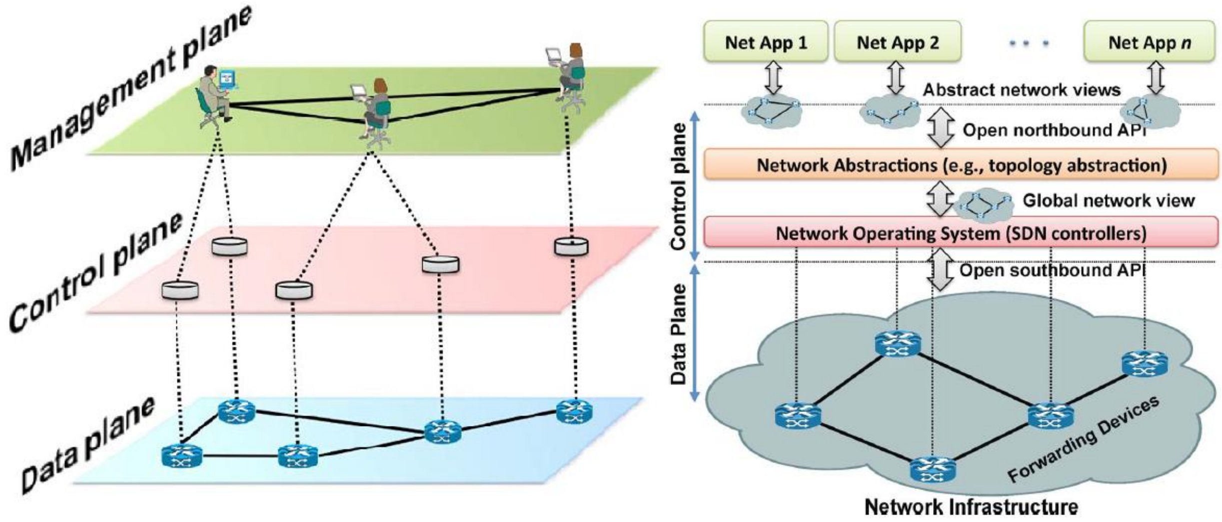


Figure 38 : SDN structure paradigm [62]

In this direction, the structural changes that introduce this new model of operation are based on some central axes that govern the operation of the network and are as follows:

The first concerns the separation of the Control Plane (CP) from the Data Plane (DP), i.e. the independence of the decision-making mechanism for the traffic management of the low-level packet forwarding procedure which, in traditional networks, has a closely related relationship and is implemented at each network node [62].

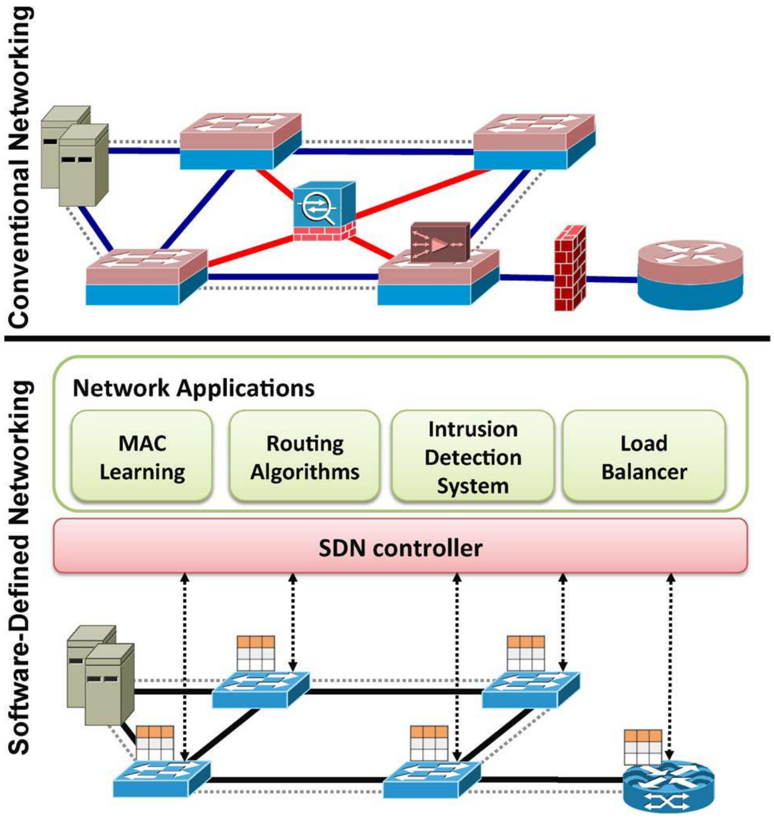


Figure 39 : Differences between traditional and SDN networking models [62]

In the second step, the implementation of the control functionality moves from the various network devices in which they were implemented (single, together with the forwarding functionality) and is concentrated in centralized independent entities (SDN Controllers) or generally in functioning platforms suitable for this software (Network Operating Systems - NOSs) which are able to instruct the network-layer operation (routing, forwarding, dropping). NOSs can deploy the Controller functionality on common hardware and their basic purpose is to provide the necessary tools and abstractions of the network infrastructure that are needed for remotely programming and managing the forwarding devices in order to follow the desired policies [62]. This automatically means that the specific functionality is removed from the network devices and they are converted to simple packet forwarding devices (virtual or physical programmable switches implementing only the data plane), which receive commands (in terms of policies / rules) for the way which they will carry out this specific action by means of appropriate open interfaces (southbound APIs) [16].

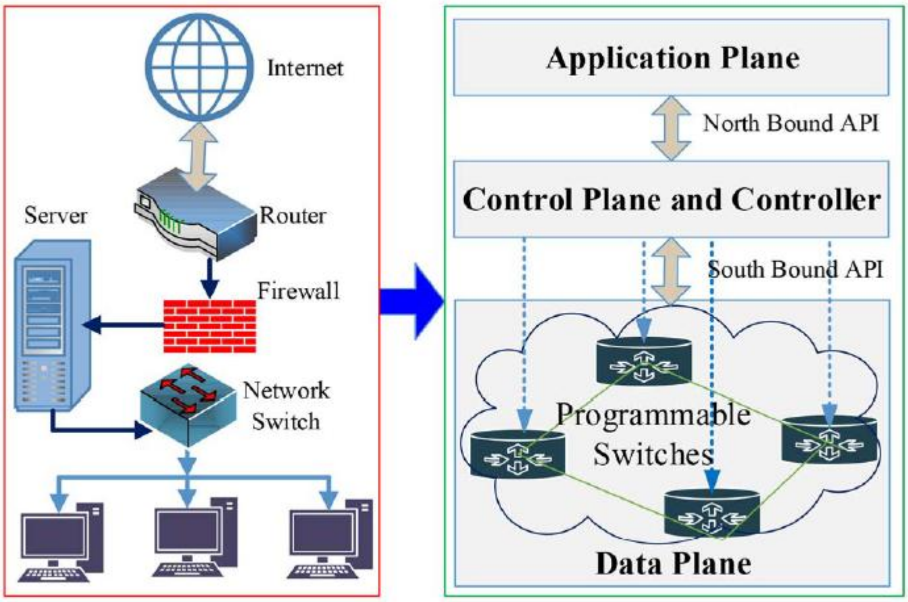


Figure 40 : Simplification of network architecture and operation [47]

Another axis of the SDN model indicates that the definition of each plain or advanced network service (such as routing, monitoring, traffic engineering, firewall implementation, network virtualization, intrusion detection, mobility management) is performed at a higher level (management / application plane) [47] through a separate high-level application (SDN Application) that runs over the NOSs and interacts with the SDN Controller via suitable interfaces (Northbound APIs) [62]. SDN Controller (NOS) is responsible to “translate” the policies into simple forwarding commands which are ready for implementation by the data plane devices. This is essentially the key to increasing network management flexibility, as this gives developers the opportunity to manage network resources in order to implement network operations with the same ease that they handle the resources of a computing system (CPU, memory, storage, etc.) [16]

In addition, it is important to emphasize that there is also a difference as regards the logic behind the data forwarding. Instead of the destination based practice that followed in the past, SDN introduces flow based techniques for the construction of the forwarding decisions. Flow is defined as a sequence of packets with specific common features and properties that is transferred to the network. Thus, based on this technique, decisions on data forwarding are taken in such a way that the requirements and criteria set by each flow are met as much as possible and not just by the destination of the data. This makes it possible to achieve flexible network traffic routing based on complex criteria (such as traffic engineering, access control and security aspects) that could not be applied based on the destination-based model [16] [17].

6.3 SDN as Part of 5G Ecosystem (relationships between SDN, NFV, C-RAN)

In the operating architecture of the SDN, an important feature is the use of virtualization techniques both in the case of controller deployment and in the deployment of network devices (e.g. virtual switch) using COTS equipment. This mode is a key pillar to achieve the sharing of abstract resources between network services, the dynamic adaptation of the network to users’ needs and, ultimately, the optimal network efficiency since its available resources are fully exploited [47]. Accordingly, the NFV is aimed at the softwarization of the NFs using plain equipment, in order to break the close model that requires implementation over special hardware. Thus, through virtual techniques, the functionality itself is separated from its implementation environment, giving huge room for flexibility to the network [17].

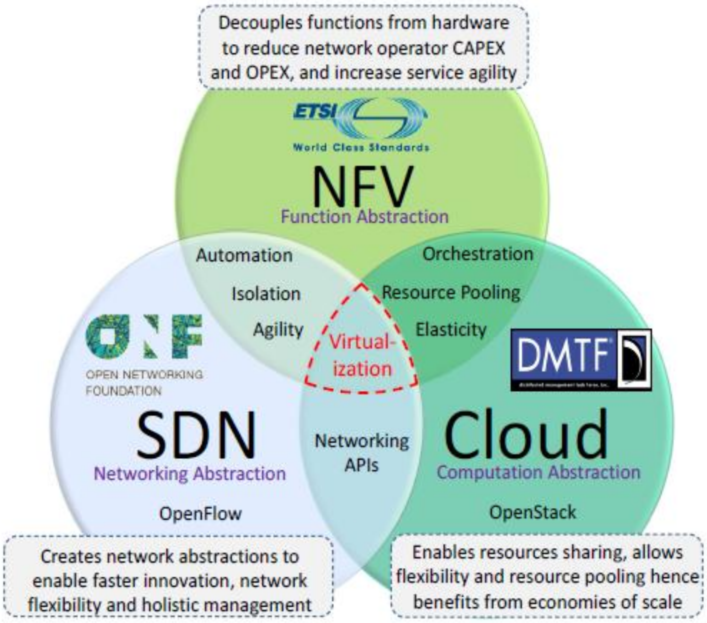


Figure 41 : Associations between NFV, SDN and Cloud concepts in 5G [17]

In this direction it can be assumed that SDN and NFV techniques complement each other, in the sense that the SDN provides the NFV with the appropriate background (in terms of quality of connectivity, automation, policy management) for network functions implementation and chaining. On the other hand, top-level applications of the SDN or the Controller can be implemented on VMs and handled as VNFs, enjoying the benefits of flexible and reliable implementation in virtual environments [17]. In fact, Network Virtualization is achieved in a large percentage by combining these two concepts [53].

Although the SDN and the NFV have parallel paths, they are not identical as they aim to provide innovative solutions for different aspects of the optimization of networks performance. The SDN attempts to better allocate and organize network components via a single management platform, while the NFV aims to simplify the implementation of network services by eliminating hardware-level constraints. The SDN essentially makes a networking-based approach to the issue that requires the adoption of a new network structure, at the same time that the NFV approaches the issue from a function-based view that can be applied to traditional networks without requiring changes in their architecture. It also follows that the SDN primarily affects networking vendors and, secondarily, mobile network providers, while for the NFV, the opposite is true [17].

Similar relationships exist between cloud computing and the above-mentioned concepts, as the cloud computing concept approaches the issue of improving network services based on the computing efficiency of networks. By allowing adaptable resource pooling and sharing, it facilitates both the SDN and the NFV in the implementation of their own individual objectives by providing them with the necessary background [17].

In addition to the relation between the SDN and the NFV, there are also important associations with the Cloud-Radio Access Network (C-RAN), which is partly based on some core principles of the SDN. The decoupling of CP/UP at the RAN level is similar to what is being done in the transport and core level following the concept of SDN. Thus, alongside the other split at the access networks level (horizontal split), C-RAN follows a model in which control plane functions and data/user plane functions are implemented independently at a central level (on Control Access Controllers for CP and UP part separately - CAC-C and CAC-U), remote from antenna sites. In this way, they fully harmonize with the SDN's operating pattern and benefit from its functional features, which can provide additional benefits to the RAN's operation, such as efficient mobility handling, improvements on tunnel-based approaches and a better handover management [50].

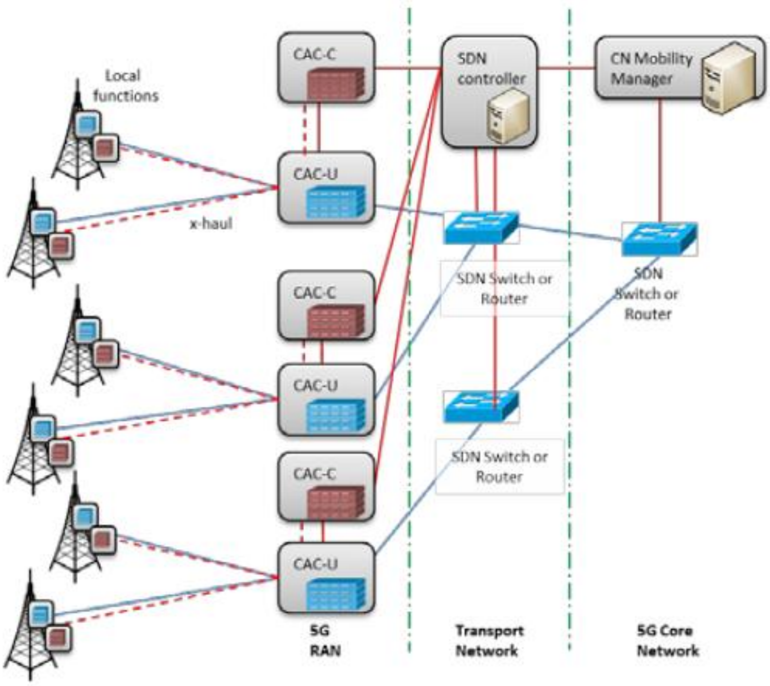


Figure 42 : End-to-end 5G network architecture based on SDN implementation [50]

6.4 SDN Controller and Interfaces

By analyzing the structure of the SDN concept, the SDN controller, which has the role of the key administrator of the network, is placed in the center of the structure and acts as the heart of the system. Correspondingly, the interfaces, which constitute communication links for the communication of the rest of the network elements with the controller, can be characterized as the system's arteries, which transfer the necessary information to the other parts of the system.

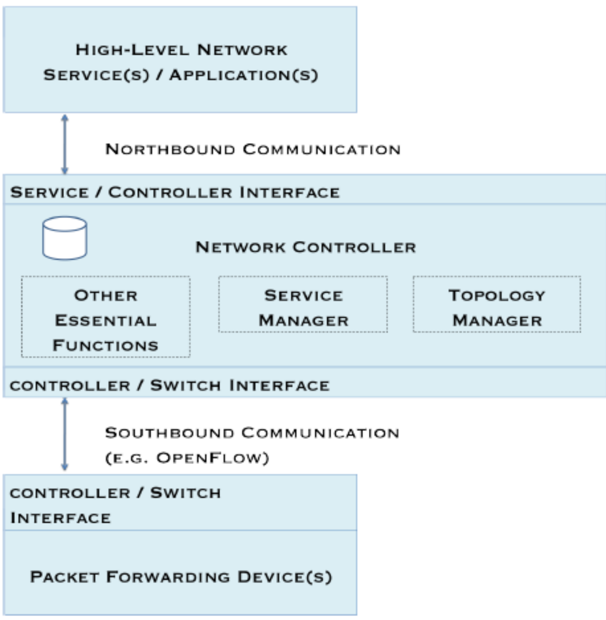


Figure 43 : Abstract view of SDN structure [16]

6.4.1 SDN Controllers

As mentioned above, the controller is the SDN's orchestrator and is responsible for coordinating network forwarding devices through appropriate rules and policies. It concentrates on controlling the efficiency of the control panel, which determines the routing of packets to the network. Through respective interfaces, it communicates, on one side, with the top-layer applications (northbound APIs) and, on the other, with the data plane devices (southbound APIs). In addition to these and in cases where a distributed model with multiple controllers is used for the design of the controller, there are also corresponding interfaces (eastbound/westbound APIs) for the communication between them [62].

There are two main design options for the controller implementation: centralized or distributed. In the first case, a physically centralized controller is used to control all forwarding devices, but this approach suffers from the classic single point failure problem, while in the second one such a problem does not occur [16]. Also, a single controller may not be able to manage a very large number of switches that may exist on some networks. Examples of such implementations (NOX-MT, Maestro, Floodlight, Trema, Ryu NOS, Meridian, ProgrammableFlow, Rosemary) developed for small business networks and data centers use multithread and multicore architectures in order to achieve a satisfactory level of performance [62].

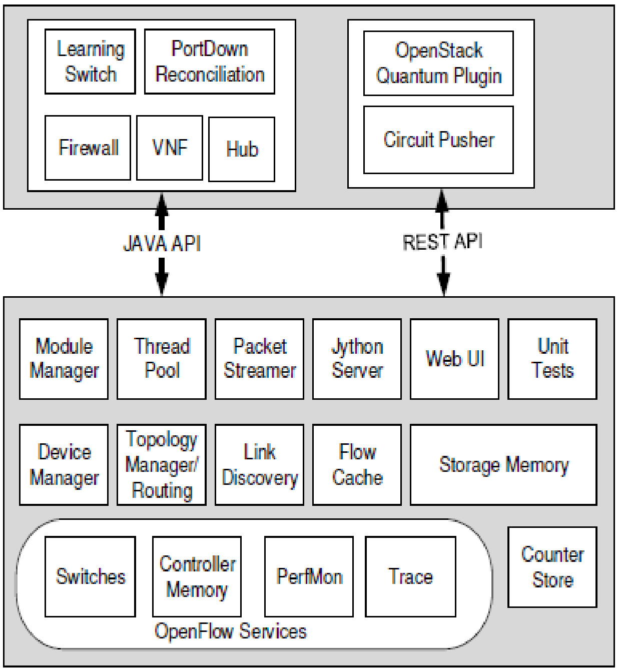


Figure 44 : SDN controller paradigm based on OpenFlow protocol [16]

Solutions developed based on the distributed model can be separated into logically distributed controller deployments that retain the advantages of centralization by operating in a distributed way on the physical layer (HyperFlow, DIFANE) as well as into physically distributed deployments where the control platform is deployed on a cluster of multiple servers (Onix, BalanceFlow, ONOS). Each approach has positive and negative aspects, mainly in terms of the need for addition maintenance and the overhead that is generated for management purposes. In addition to the above basic design models, the hierarchical model has also been proposed. In the hierarchical model, there are two different levels of hierarchically structured controllers. In the first layer, there are local controllers that handle certain switches and run some local applications, while on the second level there is a master centralized controller that runs non-local applications and orchestrates local controllers (Kandoo). The disadvantage in this case is that there is no complete awareness of the status of the network by local controllers [63].

As regards the control model to be followed, this can either be *reactive*, where each time the switches need to make a decision shall be addressed to the controller (by sending the first packet of each flow in order to receive an update on the policies and rules that are associated) or *proactive*, where, via other methods, communication with the controller is avoided (only in extreme cases) and solutions of hierarchal dispersing rules to switches are followed (DIFANE). In the second case, the whole process may be slightly accelerated compared to the first as no delays are caused each time a new decision is to be taken [64].

One of the most important and most popular software platforms for the SDN control layer is OpenDaylight, which is an abstract service layer (Service Layer Abstraction – SLA) that allows the management of fully heterogeneous multi-vendor networks, supporting at the same time several southbound protocols except OpenFlow. Indicatively, some other protocols it supports are OVSDB, NETCONF, PCEP, SNMP, BGP and Lisp Flow Mapping. It is highly flexible as regards its operation and provides network administrators with the appropriate foundation for creating network applications that manage data plane devices through software [62].

6.4.2 Southbound Interfaces

Southbound interfaces (APIs) are the link for communication between the control and data plane and their importance is key to the network (after the split in the single structure of the two planes) as, through them, all the necessary information between controllers and forwarding devices for the scheduling and management of the latter is exchanged. The forwarding devices may be either physical or virtual, following the principles of SDN and NFV [64].

It is important that there is perfect collaboration between the two planes of the network, and so certain standards should be followed in order to ensure complete compatibility and interoperability between controllers and switches of different vendors that can coexist in the network. For this reason, the interfaces should have openness characteristics so that they can be properly adapted to each implementation [62].

6.4.3 Northbound Interfaces

The Northbound APIs abstracts the deployment environment of the control plane and has essentially the form of a software ecosystem that will provide the appropriate infrastructure to the application/management layer in order to facilitate the development of top-level network applications [62]. The fact that applications should be allowed to apply different policies and rules to the same flow creates additional complexity in the task and requires additional features [16].

Unlike the southbound interface, there has been no specific standard for the interconnection of this type, but only a few fragmented and customized suggestions based mainly on the implementations and features of the controllers and of the operating systems thereof (Floodlight, Trema, Onix, NOX, OpenDaylight, SFNet, yanc, PANE) with limited and ad-hoc functionality that can not address the issue holistically. For now, it constitutes an open issue for the community [16] [62].

6.4.4 OpenFlow Protocol

The most prominent example of such an interface is OpenFlow, which is fully integrated with SDN's architecture and modeling model, has been standardized by the Open Networking Foundation (ONF) and is more popular than other solutions (POF, ForCES, OVSDB, OpFlex, OpenState, ROFL, HAL, PAD) [62].

The protocol function is based on the flow tables in the switches that contain flow entries for the management and the proper routing of the packet flows. The flow entries are records that are determined by the policies built in the controller and indicate to each switch the actions they must take in order to route each flow to the network. These flow entries are composed of three segments, which are the match fields, the activity counters and a set of actions/instructions, and are analyzed as follows:

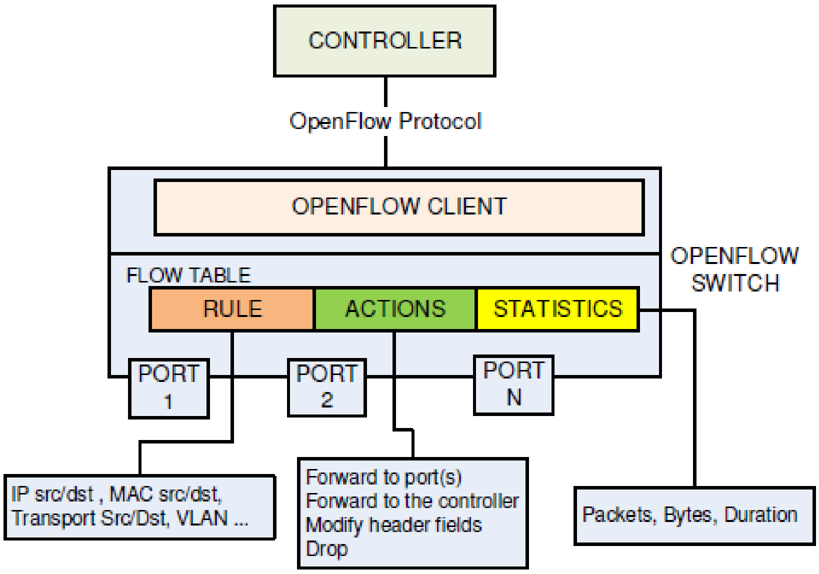


Figure 45 : Operation of a forwarding device with administration from a SDN controller [16]

The match fields contain rules and elements used to match the packets of inbound flows and mostly contain information from the header fields of the packets, the entering port of the switch or other metadata. The activity counters are used to collect statistical data from individual flows passing through forwarding devices, data which can be related to the range of packets that were received, the size and the length of the flow. The set of actions indicate the handling of the packets of a stream if there is a matching (based on the match field) with a policy/rule, as the latter has been configured by the controller. Practically, in this field the action(s) that should be performed on matching packets is (are) defined [16] [64].

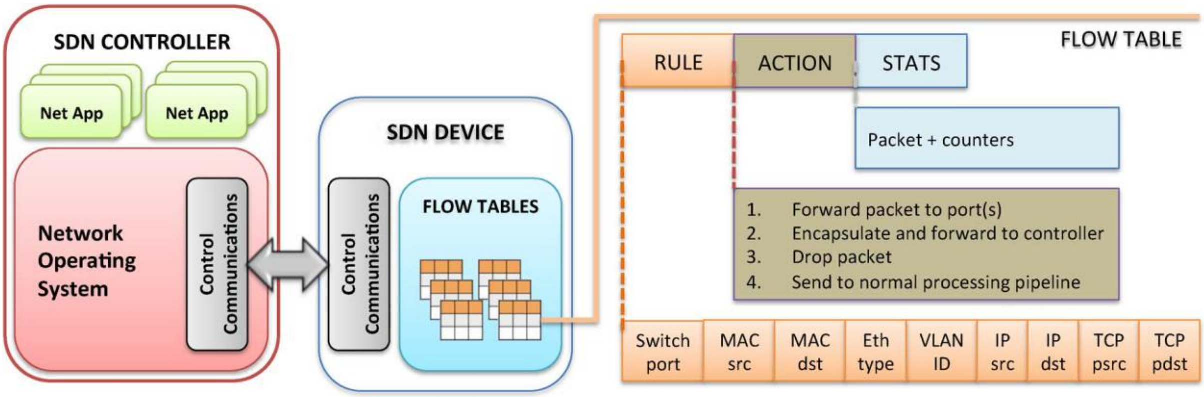


Figure 46 : SDN operated devices following OpenFlow protocol [62]

The process followed when a packet reaches the switch following the OpenFlow protocol initially involves extracting the content from its header field in order to check for matching with the match fields of the table. If there is a match with a rule, then the corresponding actions defined for this event (e.g. forward the packets to one specific port) are performed. If there is not a match, the procedure to be followed depends on the flow-miss policy followed for these cases each time. Some possible options for the packet for which a match could not be found is to drop, to check for matching with a next flow table [16] or to be sent through a secure channel to the controller in order for the latter to judge what should be done in this case, in accordance with the policies established [64].

For the interaction between the controller and the switches of the network, certain message types that are required for the operation of the protocol have been set. Among others, included are messages for flow entry modification, packet handling, statistics collection, parameters configuration, error handling and connection establishment which are exchanged through the secure channel (SSL/TCP).

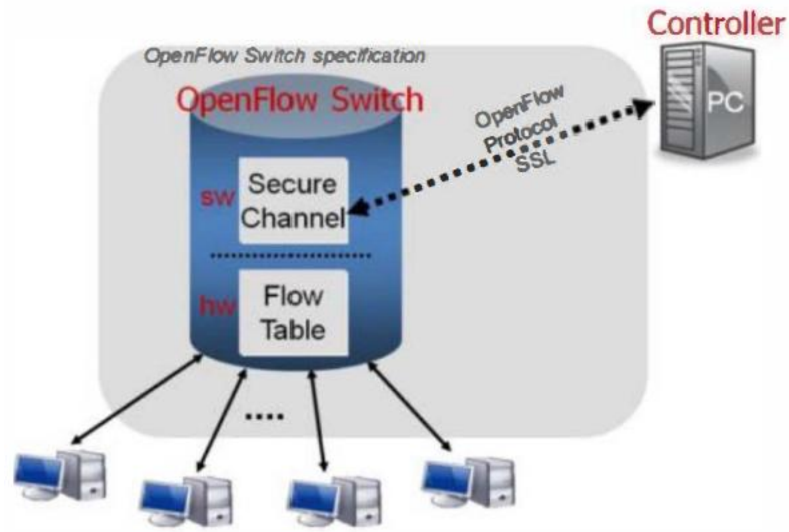


Figure 47 : Interconnection between SDN controller and switch [64]

The OpenFlow switches can be of two types, either purely based on OpenFlow or hybrid, while supporting traditional routing protocols. Most are hybrid so that there is also backward compatibility, especially in the early stages of SDN implementation [64].

6.5 Beneficial Features of SDN

The multiple innovations and new techniques introduced by the SDN in the network organization and operation sector provide tremendous opportunities for providers and managers, facilitating the coordination of network services and resources. At the same time, they solve many problems that could not be tackled on the basis of the old architecture, creating in many cases additional functionality that greatly benefits the network [47].

The SDN's contribution to flexible network management is manifold and is driven by a number of factors arising directly from the structure and mode of operation set by this new model. Initially, splitting the CP from the DP simplifies the application of policies by following simplified procedures that facilitate proper network configuration at any time [62]. The same can be assumed to be due to the organization of the network as abstract layers where implementation of network services is possible via high-level applications which, in turn, plan the network without the need for direct interaction. In addition, the universal handling of underlying networking elements, regardless of the vendor, using management layer applications and Controller mediation, harmonizes the chaos that exists as regards their management, using uniform mechanisms in order to create and enforce policies on countless networking devices scattered across the entire length and width of the network [62]. It should be noted that the SDN can be the link for the unified operation of all technologies surrounding 5G, since providing an appropriate management interface for heterogeneous networks achieved via virtualization techniques allows a simplified orchestration and their homogeneous operation [47].

As far as the scalability factor is concerned, the SDN allows the autonomous development of the two main parts of the network (CP/DP) without one affecting the other and requiring cascading changes [62]. This will make it easier to introduce new or update existing policies in order for the operation of new network services (such as traffic engineering, security enhancement and bandwidth management) to be supported by the network, designed to meet the needs of network end-users. In further analysis, this means that any upgrades needed to be made on the network will be made as software updates. Moreover, the vendor independency aspect allows the use of open standards (open source) for the implementation of the control plane functionality, enabling providers to autonomously develop their networks, favoring continuous development and innovation in the field of networks [16] [62] [47].

From the advantages of the SDN, the economic benefit resulting from its application should of course not be omitted. First of all, the integrated management of the equipment that is feasible through centralized controllers, the extension or addition of software functions and the non reliance on closed network operating policies implemented by vendors are some of the main factors that enable network operators to significantly reduce their capital expenditures and operating expenses. In addition to the above, the use of plain hardware for the control plane deployment, combined with the simplification of decentralized devices, in order to perform only elementary low-level operations, is crucial to minimizing the cost of purchasing and maintaining the equipment [62].

In the direction of reducing costs for network operation, efforts to reduce their energy footprint (green networking) can also be added to the list. Their size and the need for availability of their services on a 24/7 basis are the main reasons for the high energy consumption that arises. However, the SDN’s simplified mode of operation and its architecture features that enable virtualized resources to be dynamically adjusted and flexible programming make it possible to apply methods for power saving. Using a variety of techniques, an optimal equipment operation process can be appropriately orchestrated at the same time as a process of limiting energy consumption without affecting their intended operation and efficiency. Examples of such techniques are the use of sleep-awake techniques and appropriate rerouting through active nodes when there is low traffic on the network, the implementation of efficient load balancing in order to distribute flows among all available controllers, the use of flows monitoring techniques in order to detect elephant flows and the optimal placement of resources to achieve better performance [47].

Apart from the SDN’s individual benefits in organizing infrastructure operations, its most important aspect is the improvement of the network’s performance and of the QoE’s performance offered to end-users, while providing performance guarantees in terms of QoS. First of all, the simplified the network devices programming via high-level programming languages ensures that there is less likely to be a failure in the definition of policies during the setup process, in comparison with the process in the traditional networks that required the use of low-level and vendor dedicated commands [62]. Also, the controller continuously monitors the network’s status and can act in an automated way in situations that differentiate the network status by adjusting the operating parameters so that the policies, as defined, are always applied. Thus, there is always absolute control of data flows and, combined with the fact that resources can be committed in advance, the high availability and dependability of the network can always be ensured even in the event of equipment failure. The fact that it is perfectly possible to switch between wireless transmission techniques in the SDN also contributes in that matter [47]. Moreover, the centralization of the control operation of the whole net in one (or more) powerful Controller with global knowledge of the network state should not be omitted. This feature allows, on the one hand, the maximal exploitation of all resources [47] and, on the other hand, the creation of more advanced and innovative networking services that are able to enhance the total productivity of the network devices [62].

It would be a mistake not to mention the SDN's contribution to the network security sector, as it can de facto exert enormous influence. Given the seriousness and magnitude of this issue, its development is made in a separate subsection below.

6.6 Security Aspects of the SDN

6.6.1 Providing Efficient Secure on Networks

The assurance of the smooth operation of the network and the adhesion to the basic rules on confidentiality, integrity and availability of data traffic can be divided into two main categories as regards its treatment in the SDN. The first concerns the protection of the network itself as infrastructure and a single entity, while the second concerns the provision of additional security services at a higher level [65].

More specifically, the first category includes issues such as threat detection and reparation, definition of security configurations, as well as verification security levels throughout the network. The opportunities for real-time monitoring of network status and the fact that dynamic configuration of the entire network is allowed in the SDN via high-level applications, which are orchestrated by the central controller, favor the general management of the network and facilitate the implementation of protection measures from potential threats. In this way, the network can immediately detect potential attacks or deviations from its planned operation and react quickly and targeted by isolating or eliminating the threat [65], eventually operating as an Intrusion Detection System (IDS) and an Intrusion Prevention System (IPS) [47].

For example, a proposal to address Denial-of-Service (DoS) threats is to monitor flows (using traffic anomaly detection mechanisms) throughout their traffic within the network and, if something suspicious is recognized, a dynamic redirection away from the target point (Defense4All) shall be made [65]. Accordingly, in order to address Intrusion Attacks, there are control techniques for suspicious flows through specific statistical measurements at a packet level (TIPS, SDN-IPS, Snort-SDN) or forwarding flows to multiple security components for tight checking (CloudWatcher) [47]. Meanwhile, the network's response to the threat can be implemented via a set of APIs and modules that allows the trapping of suspicious data flows to appropriate honeypots or use of emergency alarms and dynamic quarantine for them (FRESCO) [65].

Various solutions that have been proposed highlight the importance and capabilities of the Controller who, as the only trusted entity, can define flexible and dynamic security rules and policies (based on the services provided and not just on network topology) that control access to the network (at link-layer) by limiting it to non-trusted entities (SANE, Flexible Policies) [65].

Corresponding suggestions have also been made as regards controls on network security verification and certification. This issue is especially critical when there are more than one controllers whose rules can overlap/interfere with each other or can be overriden by creating black holes or loops that endanger the smooth operation of the network. In this direction, solutions have been developed to detect misconfigurations by comparing rules by pairs (FlowChecker), proper network operation control using dynamically generated test flows (NICE), technique for real-time verification of policies and rules at the time of their entry into the network (VeriFlow), generalization/expansion of the rules for real-time conflict control (FortNOX), dynamic validation of rules based on fixed network rules in order to avoid re-labeling (check for hidden "set" and "goto" commands) (FLOVER) and to create fast failover rules based on the network's existing policies for backup operation (FatTire) [65].

As regards the second category, it concerns additional network security services that are dynamically generated and selectively activated when it is felt necessary. In the domain of anonymization, it has been suggested that ISPs assign temporary IP addresses and identifiers on flows (like Network Address Translation – NAT – operation) so as to prevent third-party tracking and correlation (AnonyFlow). This proposal is feasible in the SDN as the controller can coordinate different policies per user along a path that includes many switches without affecting its performance. Alternatively, virtual IP addresses that are constantly changing can be randomly assigned to network hosts and the controller translates them to the real IP, while at the same time forbidding direct access (OF-RHM) [65].

Another useful feature that can be provided thanks to the SDN is out-tasking of network security management to third party entities. These entities may be certified security service providers that can efficiently manage the entire process primarily on behalf of small business networks that, due to a lack of resources, are unable to implement serious security mechanisms [65].

In addition, a solution is proposed for the business networks in which the "Bring Your Own Device" model is followed by the employees and concerns the secure data offloading of critical data which is necessary for their further processing, sharing and storage. By providing a strictly controlled privacy environment that the controller can support by using distinct policies per service, it is finally possible for users to safely implement this mode of operation (ECOS) [65].

6.6.2 Challenges in SDN Security

In addition to the new capabilities introduced by the SDN in order to ensure the network's operation, it also creates new challenges related to the same issue that are directly connected to its intrinsic characteristics [47] [65]. The availability and the reliability are fundamental feature for the smooth and efficient operation of the network and should have a top priority [62].

The most important issue is to ensure the SDN model itself, based on which modern networks will operate and, therefore, the impact of a possible attack that could affect one of its most important activities (routing and forwarding) could be devastating [65]. The vulnerabilities associated with this part of the network are related with the relevant software that implements network services and can affect various parts of the network, such as controllers, switches, links and virtual machines [47].

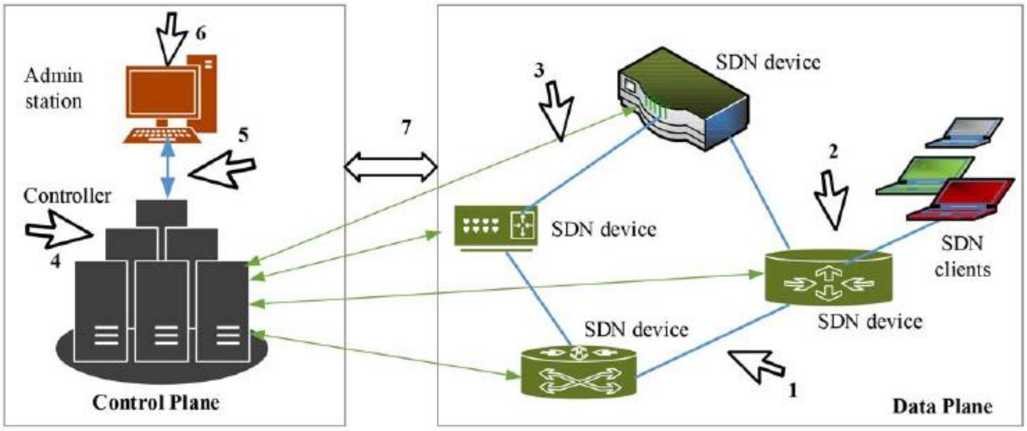


Figure 48 : Possible security vulnerabilities of a SDN based network [47]

Such attacks include, but are not limited to, spoofing attacks, intrusion attacks, anomaly attacks or DoS attacks that can cause particularly serious problems such as service unavailability, data loss and modification, unauthorized change of rules, bug entry and the introduction of a malicious code in top-level applications [47].

The main causes for the existence of these attacks have to do with matters relating to its operation architecture and are directly linked thereof. Control plane's saturation probabilities and buffers of the switches in combination with the failure of strong mechanisms for traffic policing can bring about a negative impact on the SDN's security level [65]. Although several and significant efforts have been made to limit the problem by developing various solutions (OF-RHM, TopoGuard, Malware switch detection, Sphinx, Flowguard, Flowtag architecture, NIMBLE, Slick, NICE, FortNOX, FRESCO, TRW-CB, Rate Limiting, Maximum entropy, NETAD, Entropy method, Lightweight DDoS detection, Z3 Prover method, Lineswitch method, AVANT-GUARD, PermOF) that are appropriately focused on the type of threat (spoofing, intrusion, anomaly and DoS attacks) and of the SDN-plane which it affects [47], more action is required in order to ensure that security at this level of the communications stack is embedded in the SDN model, constituting an endogenous feature and not an additional module that may be bypassed [65].

Another issue in this field is the limited resources in terms of computing power and battery life that exist in IoT devices (such as sensors and actuators) and do not allow strong encryption protocols to be implemented. In an effort to solve this problem, it has been suggested that security for such devices be mediated by a stronger device that will essentially act as a gateway in order to establish the "demanding" SSL handshake on behalf of "weak" devices (Sizzle) [65].

A security issue can also occur in overlay networks (peer-to-peer networks, Skype, Netflix) that encounter some operating problems in more complex large-scale applications due to the loose relationship between the overlay network and the physical equipment. Due to the lack of visibility in the underlay network, they can not provide efficiency guarantees, so their entire operation needs to be enhanced with additional applications in order to ensure their operation. Although the OpenFlow and the SDN provide interfaces for the proper adjustment of the underlay network, there have been ideas for creating a middle layer which shall extract topology information and forward it to the overlay network in order to create more guaranteed traffic routing paths (Nakao et al.) [65].

Finally, an open issue is the management of delays and the better distribution of functions across the network in case of implementation of middleboxes (firewalls, encoders/decoders) as virtualized software controlled through relevant APIs (VNFs) [65].

6.7 Challenges and Open Issues of SDN

Since in the SDN many individual parts are included and based on the fact its spread as technology is constantly increasing, new proposals and new challenges associated with them are emerging.

One of the main open issues of the SDN is the lack of a clearly defined standard for the northbound interface, as there is for the southbound. It is particularly important to make progress in defining a framework for this interface in order to ensure the interoperability between top-level applications and the control plane. Some of the solutions proposed are strongly supported by NOS's specific features and follow an ad-hoc approach and therefore are not a comprehensive solution that can be widely used and be consistent with the flexibility, portability and interoperability aspects advocated by the SDN [16] [62].

A key parameter that should be taken seriously, as it can largely judge the success or failure of the SDN ecosystem, is the ease of harmonizing users (administrators, developers, providers) with the new organizational model advocated by the SDN, which is based on the layer-abstraction concept in the automation of processes. Its acceptance is also judged from the ease of adaptation of the SDN to the networks' needs and models, as they existed prior to its appearance. Although elementary resistance is expected (as in the case of cloud computing) due to the temporary confusion of roles caused especially during its initial introduction, the SDN community should not succumb to possible proposals aimed at its depreciation and its non-universal application [62].

An important challenge is the design model that will be adopted for the controller as, after evaluating some of the solutions proposed, there is a trade-off between centralized and distributed controller solutions. Although the physically distributed and logically-centralized option provides augmented scalability and reliability, it also increases the complexity of the flexibility in updating and upgrading the controller. Also, in the case of physically distributed controller deployments, in order to find the number of controllers to be used and their location in the network, extensive research and study is required so as to avoid delays and not to burden the overall network performance. Regarding the controller model to be followed, the way in which management of switches will be assigned to controllers is also a challenge. There should be a load balancing between all entities and this is why dynamic tweaking solutions for the number of switches that each of them controls are required, or even mechanisms for the dynamic handover of the switches between different controllers. Although some solutions have been proposed, greater emphasis needs to be placed primarily on the development of better APIs' measurement and monitoring so that the data that judges the assignment of the switches to the controllers is available [16].

A possible problem associated with the design model of the controller is the issue of single point failure that occurs when a physically centralized controller is used. Solutions that suggest the existence of backup controllers that act as replicas and are also connected to switches (an option that is enabled by OpenFlow) so that, in case of failure of the former, they take action, offer high availability but burden the overall performance of the system. There is still the possibility of adopting distributed models in which this problem does not exist (in such extent), while the use of hybrid techniques impedes the work of developers with strict requirements [16] [62].

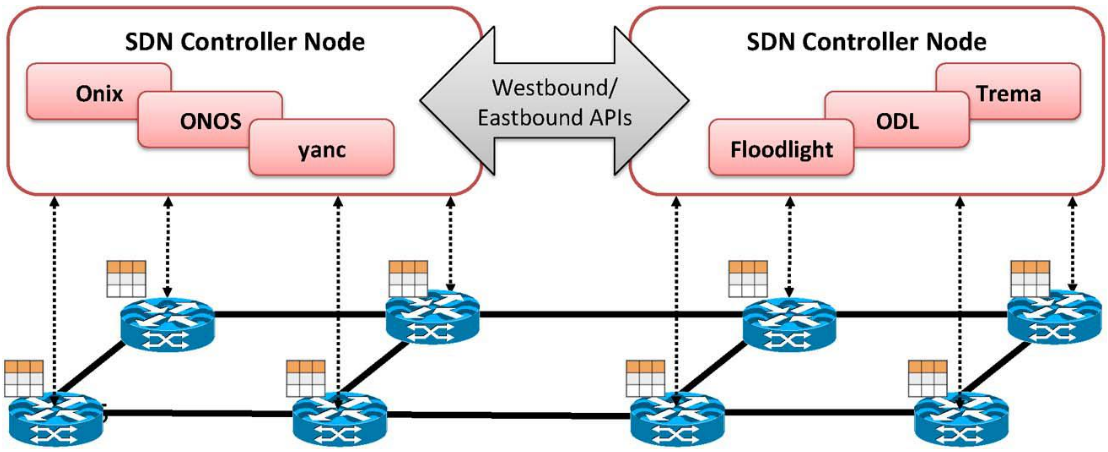


Figure 49 : Distributed controllers SDN architecture [62]

As regards the controller platforms, the research community shall have to address some issues that have to do with delays in communication through the control plane and the data plane, the definition of southbound and eastbound/westbound APIs, the fault tolerance issue, the load balancing problem and the facts of combination and synchronization of various network functions in an efficient way. Controllers must have better visibility and awareness of the network state in order to adapt their operation accordingly. Some ideas for resolving this problem are the use of hierarchical templates for organizing controllers that offer abstraction and facilitate the interoperability of the system in a flexible and scalable way. An important parameter related to the controller's implementation is the lack of flexible modularity in most cases, which forces developers to redesign and redeploy the network services from a zero point every time it is needed to add a new functionality in the network [62].

As regards switches, a practical issue is about the size of the flow tables, which should be sufficient to store all the rules that are associated with the flows. TCAM chip-based suggestions provide flexibility and efficiency but can not be used for large table sizes as they are particularly expensive and require a high energy consumption. The existing option in the newest versions of OpenFlow is to use multiple tables, although it promotes the scalability and flexibility aspects in the system; the implementation in hardware is very tricky. In addition, findings from OpenFlow switches tests show that they are lagging behind in performance by reducing overall network efficiency. Thus, more emphasis should be placed on improving their hardware architecture by following solutions such as increasing CPUs for faster flow processing and avoiding overheads [62].

According to the scalability of the system, which is mainly achieved through the separation of CP and UP, there are some sub-questions that need more detailed study in order to avoid becoming an obstacle. In this direction, the two key factors that can burden the SDN ecosystem causing a bottleneck in the control plane are the reactive control model (where the first packet of a new flow is sent by the switch to the controller) and the fact that an additional delay from the process of updating the flow tables, which is performed remotely from a third entity (controller), is caused [62].

Challenges also exist as regards the implementation of solutions in the direction of green networking, as it may affect other features of the SDN ecosystem. It is clear that there is a trade-off between energy consumption elimination and securing the network operation, as methods used to address network security issues add functionality to the network, consume computing and need to operate 24/7, thus increasing consumption [65]. Moreover, the implementation of powering off/on techniques in order to save energy may also have an impact on the efficiency and the life span of forwarding devices as physical equipment. From these examples, it is easy to conclude that finding a happy medium in terms of the breadth and depth of energy-saving techniques to be implemented is a major challenge for developers and network operators [47].

Finally, an open issue remains the efficient exploitation of the resources and the interoperability among the heterogeneous technologies that coexist in 5G networks. The vision behind the SDN concept is to unify the management across different network technologies that may coexist in a network. However, it is important to take into account the specificities of each device (smartphones, tablets, sensors, vehicles, etc.) and to deal with them as they do not all have the same computing resources in order to perform resource-intensive protocols. The same problem exists in ad-hoc mesh networks with limited wireless capacity due to the conditions prevailing in the shared communication channel. The SON concept aims to address this issue through efficient methods that automate the operation and management of these procedures in an efficient way and support a variety of underlying equipment (various topologies, physical media, node characteristics. However, in order to better support the SON from the SDN, more emphasis is needed by researchers as regards the support of not only infrastructure-based networks but also mesh environments [16].

7. END-TO-END NETWORK SLICING (E2E SLICING)

7.1 Introduction

Expectations and requirements created for mobile communications networks require innovative solutions in order to meet and ultimately deliver what is expected by end users. Due to the new model of network operation, that of differentiated services covering a wide range of vertical domains, it is necessary to have groundbreaking techniques for the support and integration of all heterogeneous functional elements and mechanisms that make up the 5G ecosystem so that the provision of high-level services in a network-as-a-service level is feasible in an efficient, scalable and flexible way.

The distinct demands of the services of each vertical sector in terms of latency, throughput, high speed mobility and other parameters are particularly difficult to support at the same time based on the monolithic architecture of mobile networks (which is applicable up to 4G), as activating multiple network functions that can be mutually exclusive due to their different nature is required; eventually, any of their functional features are cancelled. Moreover, the cost and complexity of running such a network would make any flexibility in its management and scalability prohibitive, thereby rendering it inefficient.

For these reasons, it was considered appropriate to find more effective ways of operating the networks, drawing inspiration from corresponding actions in the field of operating systems and software (virtualization, softwarisation, orchestration, functional isolation and network programmability concepts, etc.) allowing the provision of different functionality in different user groups on the top of the common infrastructure [66].

In this respect, the concept of Network Slice can be described as the slicing of the network into virtual networks (slices) that can operate on both on dedicated and shared infrastructures in terms of hardware (compute processing power, storage, radio resources) but are logically independent of each other in order to provide customized and dedicated end-to-end functionality depending on the telecommunication requirements (in terms of data-rate, latency, security level, etc.) of the respective vertical sector they serve. Each slice has autonomous functionality, can support different technologies and exploits all the resources available to it independently of the other parts of the network by giving an abstract character to the network deployment model [67] [68]. In addition, it is possible to combine multiple slices (bundle) in order to create a new slice aiming at serving vertical domains with multiple requirements at the same time [68] [69].

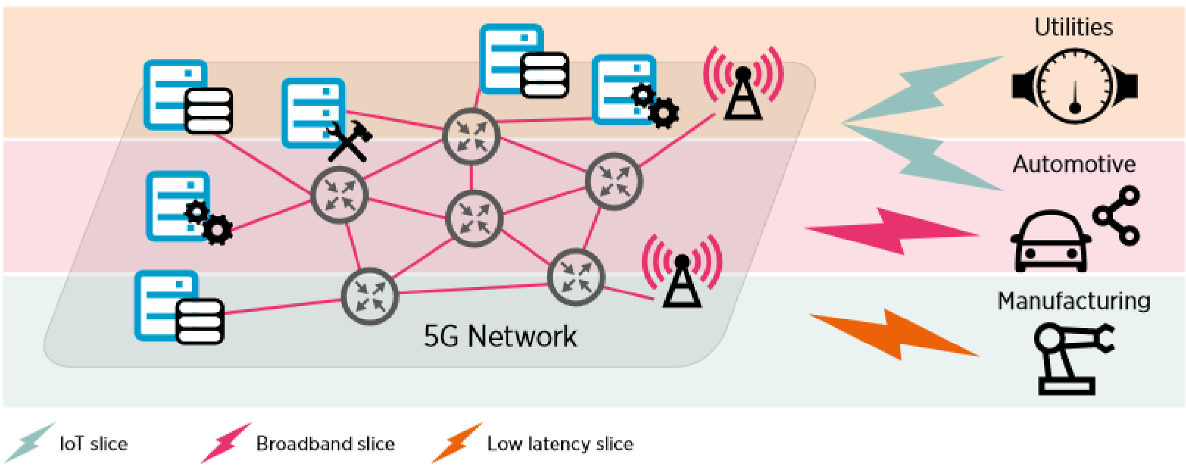


Figure 50 : Segmentation of 5G network into slices for serving different use case families [68]

The features mentioned above, combined with the SDN and NFV concepts, can be key-factors in meeting the expectations created by the next-generation networks as they can provide optimization capabilities for physical network infrastructure management, performance maximization, adaptability of the network to market requirements for the provision of new services thus allowing almost unlimited operational extensions, increased security, flexibility of adaption and scalability of the network, reduction of the operating costs and enhancement of the robustness of the entire telecommunication ecosystem. Finally, as it can be easily understood, all of the above features contribute to providing better holistic services to customers, thereby helping to maximize the satisfaction of their needs; a challenge for network operators and the rest of the industry [19] [67].

It should also be noted that the introduction of this technology alters the business model of the telecommunications market, since it allows new forms of operation for network operators under the everything-as-a-service concept. Thus, "building" a network on a core, transport or access layer level can be done on demand by binding the required resources either from shared pools owned by the provider itself or even from the cloud [19]. At the same time, the end-user service model changes as they are clearly served on a QoE level (rather than on a QoS) based on their requirements, in a guaranteed Service Level Agreement [68].

7.2 Structure of Network Slicing Model

Although the idea of segmenting networks into subnets that serve different user needs is not new, it has gained significant momentum due to the large developments in cloud computing, SDN and NFV, thus providing the right background and the required tools so as to implement the concept. Thus, it is now possible to create slices along all levels of a mobile communications network starting from the core network and ending with the radio access network based on the principles of the technologies that are being used, offering an abstract dimension to network deployment.

With respect to the requirements of each vertical sector, appropriate network slices are created, which result from the composition of only the necessary network functions and network elements of each network layer, aiming at better user service, simplifying the network architecture and facilitating its management. This synthesis results from a combination of functions both at the radio access network level (which are responsible for the access of mobile devices to the rest of the network) and at the core network level (where the respective functions are responsible for the implementation of the services provided). The *selection function*, which constitutes a link directing the movement to the appropriate CN slices so as to provide the desired functionality to end users is responsible for the appropriate combination of these two "sub-slices" in order to create the final end-to-end slice. This match between the subsections of the slice can be done either on predetermined rules or even dynamically so as to maximize resource utilization and provide the best possible service to end users [67] [69] [70] [71].

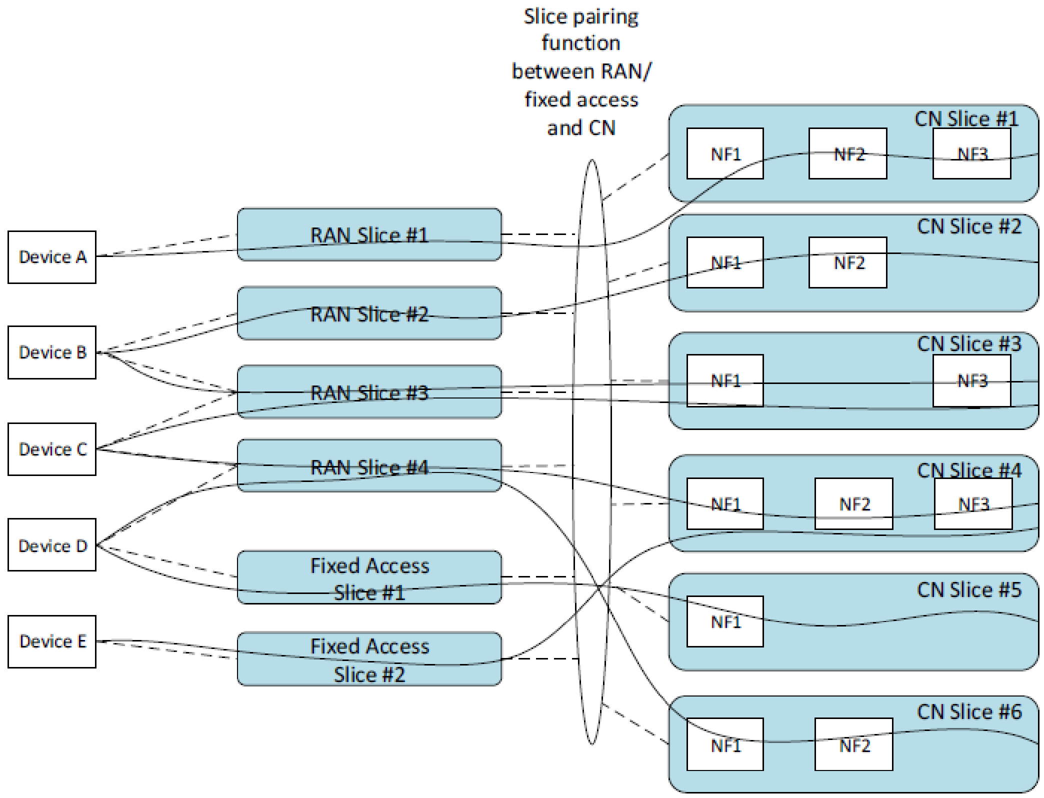


Figure 51 : End-to-end slicing to provide the desired functionality to the users [67]

It is also noted that there are common network functions that can be used simultaneously by different slices and there is also the category of dedicated functions that are used exclusively by a specific slice for specific operations. This is done in order to optimize network performance and depends primarily on the nature of the differentiated services provided by each slice at a time.

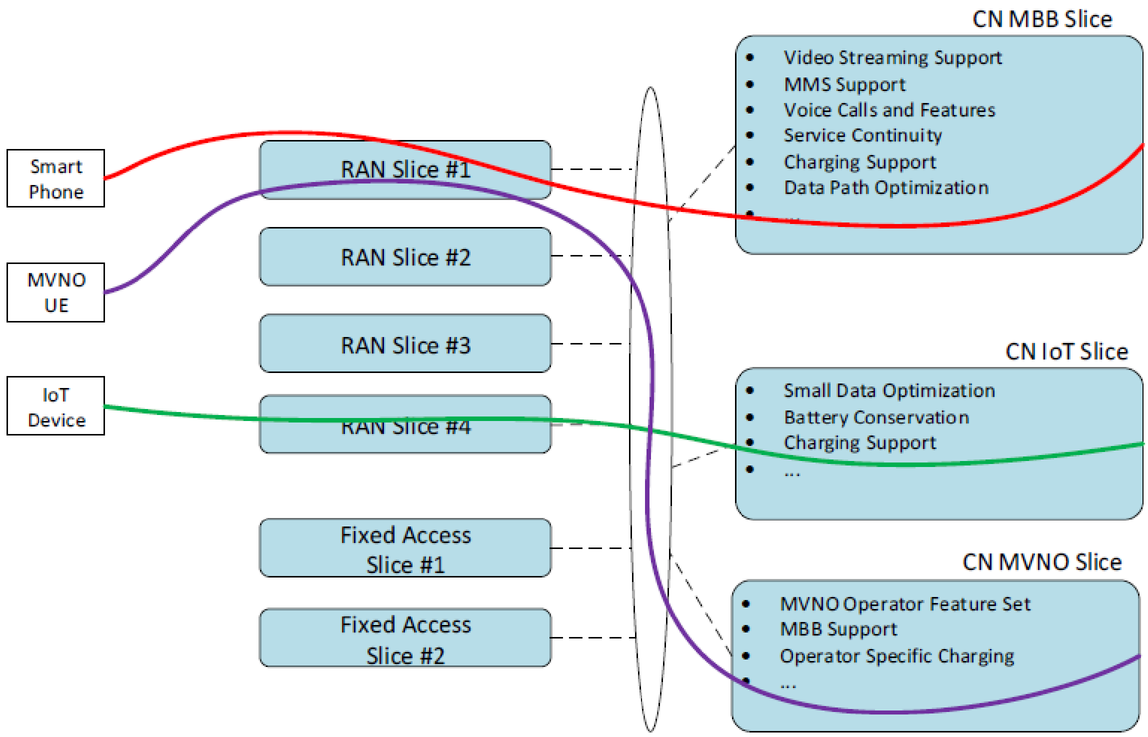


Figure 52 : Common slice selection in RAN and differentiation in CN [67]

7.3 Basic Architecture

On a deeper level of analysis, according to the definitions and the different perspectives given by both NGMN and 5G-PPP for the 5G Network Slicing architecture model, the composition of the basic components combined to achieve the concept can be represented by a more general model that includes the grouping of its functions into three layers [70].

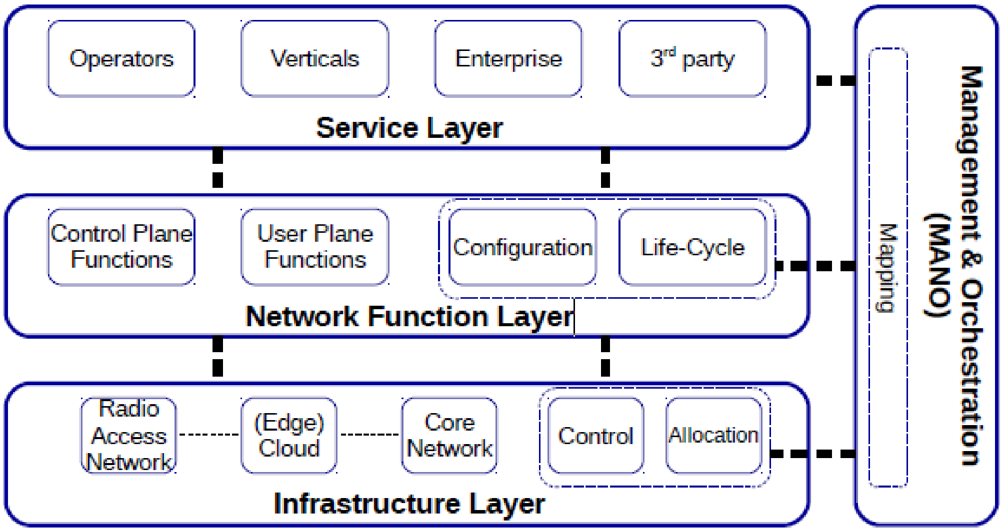


Figure 53 : Basic 5G network slicing architecture model [70]

In a bottom-up approach, the Infrastructure Layer includes data on the physical resources of the network, covering all of them, from the core level to the RAN level. Thus, this layer includes functions related to the management of natural resources for the slices composition, the allocation of infrastructure to the appropriate slices according to the requirements of each vertical sector served, and how they are managed from upper layers in order to create the network slice entity. At the next level, the Network Function Layer includes all the required procedures related to the operation, parameterization and life cycle of the slice components so as to meet the requirements for serving each vertical domain and to make an optimal use of the network's infrastructure. The Service Layer is at the top level of the stack; it includes customer-centric services that are built on pure business criteria and are designed to serve customers. For their implementation, they combine all the required components of the lower levels, essentially building the final services provided to consumers. Finally, alongside these three layers, there is also the Management and Orchestration (MANO) Layer that includes mapping procedures for the matching of services and network components. At the same time, it is also responsible for the instrumentation of all the components that make up the slice as well as the life cycle of services as a whole [70].

As far as the level of implementation of the network functions used by slices is concerned, a mixed architecture model that includes physical and virtual resources, as well as cloud and edge techniques, respectively, is followed. The "unification" of resources is done through an intermediate virtualization layer which is responsible for combining the different resources and displaying them as a single structure to the upper layers. This allows the network to make maximum use of its existing infrastructure in order to create slices and to provide the best possible services to end users, while ensuring the necessary flexibility in the management and orchestration sector [69].

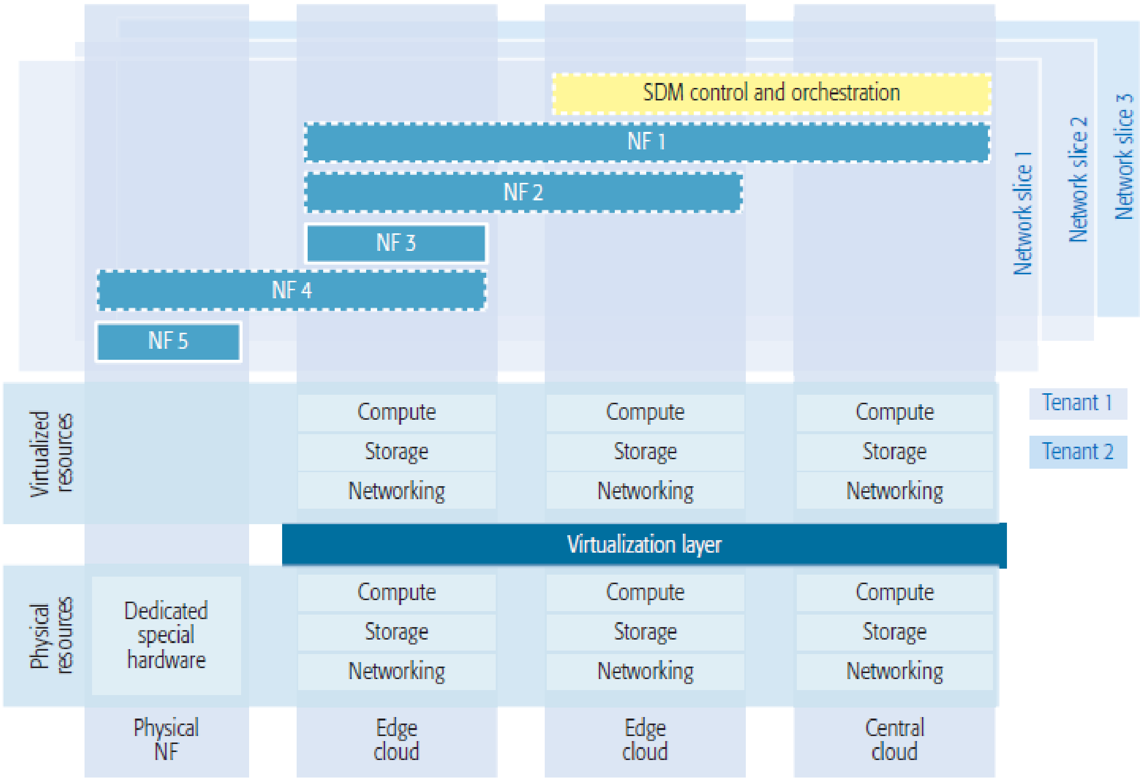


Figure 54 : Unification of physical resources under an intermediate virtualization layer [69]

7.4 Functional features & capabilities introduced by the slicing model

Based on the slicing model architecture, the corresponding actions for setting up, configuring and managing network slices are defined in order to allow a smooth and efficient end-to-end operation of next-generation networks based on the vision and the design that have been implemented [67].

On a first level, these actions include procedures of building, initialization, modification and administration of both the slices themselves and the services implemented, while at the same time procedures are being defined for the mapping and orchestration procedures of NFs to the slices in order to implement the planned operation. Furthermore, the definition of the design diagram (blueprint) of each slice is of particular importance, which is a complete description of the structure, configuration and goal of each part of the network (network map) in order to fully understand the role of the slice and the way it is organized and managed. By following a common policy by all parties involved in defining the blueprint, it is significantly easier to allocate slices to the corresponding physical or virtual resources, which may be of a universal nature, regardless of the vendor [67].

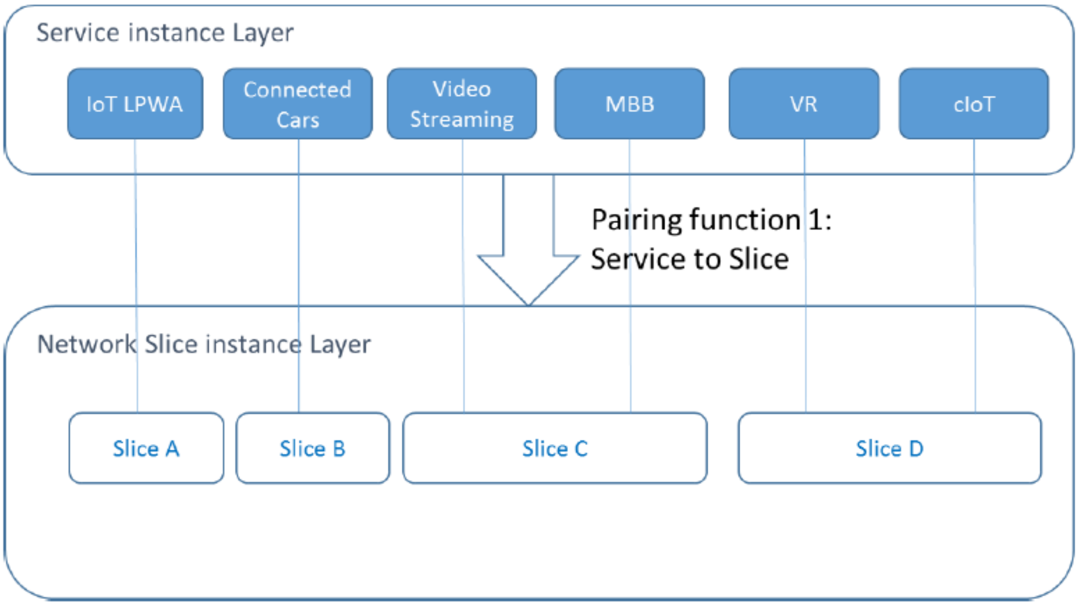


Figure 55 : Matching of services with the appropriate network slices [67]

Additionally, in order to control the slices status at any time, it is necessary to support, through appropriate NFs, functions related to configuration management, fault and performance supervision as well as the monitoring of possible security threats in the slices' operation. These functions are important to be available both centrally and per slice so that even the tenants or other third parties holding a slice can manage it through appropriate APIs in an end-to-end base and shape the QoS level as they please. In terms of resource handling, other than life cycle management at the slices level, in cases of sharing common elements between different vertical sectors, the slicing model should allow cooperative resource management through appropriate functions so as to ensure the smooth operation of all parties involved [67].

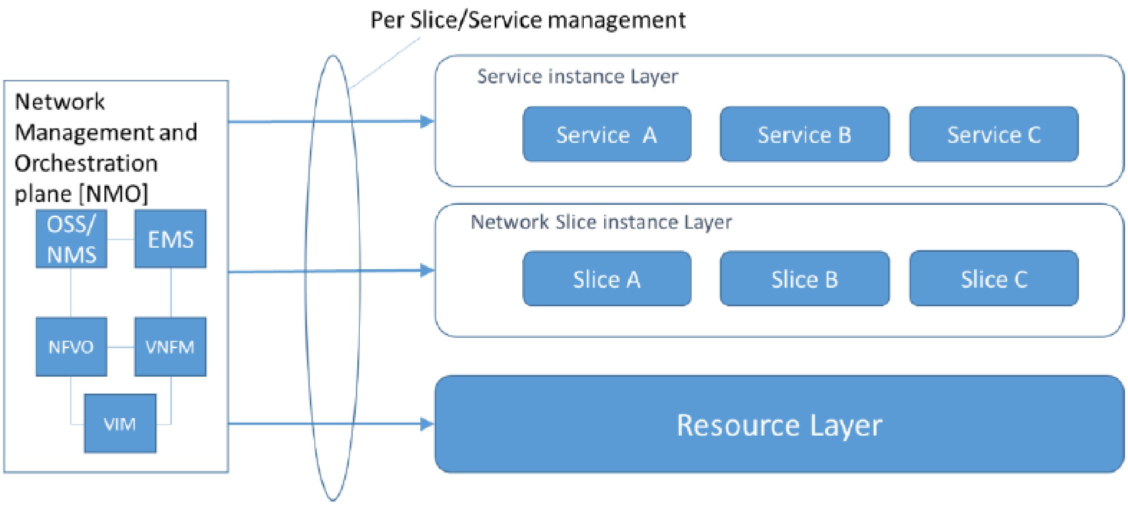


Figure 56 : Management and orchestration of network slices [67]

Another key point that makes the difference between a network following the slicing model and a network using the conventional model is the fact that in the first case it is allowed to automate the creation of a network (under a slice format) with the desired functionality following a procedure similar to the installation of a common software in a personal computer. This can be done by creating slice templates in which all the necessary settings needed in order to create and configure a slice (components setup, resource allocation, rules configuration) so that the whole procedure is completely simplified and as much automated as possible, are defined. The procedure followed for network upgrading or scaling since almost all of the required actions are automated and provided by vendors as software is as simple as the deployment procedure [67].

In the procedure of installing a slice, the orchestrator has an important role to play: he is initially responsible for checking the availability and resource allocation required for both the non-virtualized and the virtualized part of the network (separately). Once this procedure has been completed, the slice template is created using the necessary VNFs that will implement the slice functionality. As communication between each other is required for the smooth and unified operation of the slice, it is important, at this early stage, that all the components that make up the slice be able to locate all other slice elements in an automated and uniform manner. For this purpose, special mechanisms for VNF management have been set up in order to coordinate the procedure where necessary and to avoid system malfunctions [67].

Similar to the installation procedure, automated flow is also defined for the scaling procedure of a network slice. Except in the case where some modification is required in non-virtualized resources where additional procedures may be required depending on the case, for the general case where the scaling (up or down) of the slice is resulting from the monitoring, then the appropriate procedures for dynamically configuring both the characteristics of a VNF or even a set of VNFs are enabled. Additionally, the configuration can be done at a higher level (application level) by highlighting the level of automation of the procedure [67].

In any case, the implementation of these actions depends on the rules set by the operator according to the policies he follows so that the smooth and efficient operation of "neighboring" slices is not affected. This practically means that through the slice management mechanism a level of isolation of slices should be applied as regards the management of natural resources that each one of them uses in order to balance the system's operation and not to disproportionately burden some of its parts. Therefore, isolation should be ensured in terms of traffic, bandwidth, processing and storage power so that each part of the network has guaranteed features in accordance with the specifications and the work it serves, regardless of the other slices that may be operating on the same physical infrastructure [67].

Finally, the automated function features stemming from the slicing model architecture and allowing intelligent fault management as well as stable QoS and QoE levels are also worth mentioning. Thus, by means of appropriate monitoring mechanisms, when a malfunction affecting the performance of a slice is detected, the dynamic response of the system (e.g. enabling an alternative instance of the slice and redirecting traffic to it) is possible so as to minimize downtimes and affect the quality of the service to end users as less as possible [67].

7.5 Advantages

The adoption of the network slicing concept can bring multiple benefits for the particularly demanding areas of network installation, operation and management while saving at the same time significant amounts of scale. Network sharing techniques can be considered as a forerunner of network slicing, which allows operators to share some resources so as to maintain the costs of installing and maintaining their infrastructure. Network slicing not only maximizes these benefits, but adds a number of new features that radically change the business model of network operation by getting a boost from parallel projects that are being developed, such as SDN, NFV and C-RAN [69].

Initially, through slicing techniques, it is clear that customer service is being offered in the light of differentiated services, as it is now possible to provide dedicated services in terms of data rate, latency, connection availability, safety and a number of other aspects. Thus, through flexible operations, providers are able to build network slices that can provide a unique end-to-end connectivity experience that is fully customized to meet the needs of each vertical domain (eMBB, mMTC, URLLC) [67] [71].

In addition, through this model of network operation it is possible to provide QoE-only based services, since the network functions are separated according to the vertical domain that serves each slice. In that way, operators can focus exclusively on the development of the appropriate functions required to serve each vertical sector, without the need for additional unnecessary operations that they would otherwise have to support. Therefore, the "focus" of each section of the network on serving specific customers prevents possible downtimes that may result from the misuse of network resources by users of a different vertical sector. With this logic the slice that is responsible for the operation of services dedicated to critical fields such as public safety will not be affected by a possible increased demand for a service by a slice that serves, for example, the automotive sector. This has a significant benefit and affects, inter alia, the flexibility of management, the availability, the reliability and the robustness of the entire mobile network [67].

An important factor is the saving of financial resources that is made through the adoption of the network slicing concept, since because of this possibility, the providers no longer have to maintain all the functionalities of a network even if their demand is minimal in order to provide discrete services to their clientele. For example, in a slice that provides infrastructure for a network of fixed sensors used for some sort of measurement, there is no need to implement network functions related to the mobility aspect. Thus, they can dispose the corresponding slices that only deploy the necessary functionality, free from the obligation to maintain unnecessary operations that burden their operating costs with unnecessary amounts of money. Moreover, the fact that fewer functions are required in order to cover a service facilitates the integration of new capabilities into the network, drastically reducing the time needed for their introduction, while at the same time contributing to its more direct adaptability to the market's real conditions [67].

Due to the large amount of innovation introduced by this networking model, it is likely that its contribution has not yet been fully apparent, since it directly or indirectly affects many network operating factors. One such example is the change that is being made in the business model of network operators through the provision of the network-as-a-service functionality to third-party players. Thus, through this feature, the operator leases slices to third-parties (over-the-top service providers) that have the capability to configure them autonomously and in such a way so as to serve their needs and to create full custom network operations. The depth of the slice configuration options varies and depends on the model that the network slice provider follows, where the tenant has either the ability to make changes to the functionality of the slice by adding and/or removing services through appropriate interfaces, or to simply have a more limited role by controlling, for example, only the performance of the functions. In any case, the network provider, who is also responsible for the infrastructure used, may have full rights on the slice, but once more this depends on the type of service provided [67] [69].

In addition, other flexible network partitioning schemes are allowed based on which various new business models can be created. These include, but are not limited to, static, dynamic and hybrid concessions of the network infrastructure to the tenants for the deployment and operation of their services. For example, some tenants may have separate slices for running the core network when they share the RAN with other entities, following a hybrid network implementation model in which even virtual sources can be further shared between the tenants. This option paves the way for new types of services, which in turn can stimulate the economic growth of the sector at higher levels, as this flexibility facilitates the development of innovative solutions [69].

Furthermore, the separation of the network into smaller isolated slices, besides facilitating the deployment of the network following the principles of the divide-and-conquer model, makes the network more secure as a whole. Applying similar or different security rules to each piece of the network according to the standards and distinct requirements of its deployed services provides greater assurance as regards their smooth operation since rules and policies strictly suited to the slice's needs and architecture are applied. The flexibility provided to administrators in this area facilitates their work as they can ensure a more comprehensive coverage of the possible network functions impacts per slice than the whole of the network. Moreover, following this logic of organizing network functions, a potential vulnerability in some part of it will more adversely affect the rest of the network's functions that are implemented in other slices compared to the influence that would have existed if the network was organized based on the traditional architecture model [69].

7.6 Issues under investigation

Despite the major developments in network slicing technology, there are still some parameters to be further studied so as to ensure that the model is fully automated and to maximize the benefits for operators and end users.

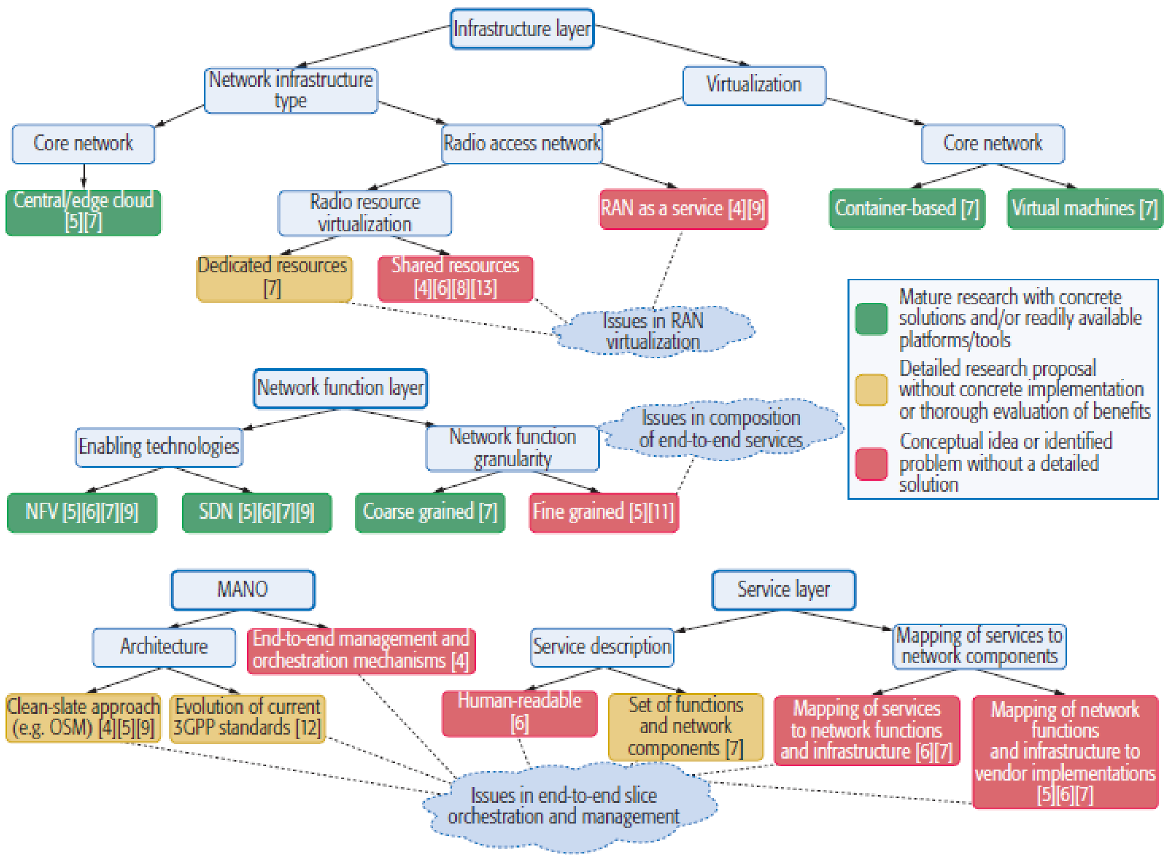


Figure 57 : Open issues in network slicing concept [70]

As far as the Infrastructure Layer is concerned, the main issues are the virtualization of the shared RAN, since with the use of existing solutions there is a particular challenge to find the tradeoff between the slice isolation level and the efficient management of the available spectrum. The solutions that include the a priori assignment of the spectrum to virtual BSs are essentially static, but they provide isolation of resources at a slice level, while dynamic solutions for assigning frequencies, although not posing a problem in spectrum utilization, do not guarantee high isolation levels of the radio resources. This problem stems from another barrier to slicing development that needs further research by the community and concerns the provision of RAN as a service (RaaS). The RaaS model will allow the development of additional scenarios for the operation of a network (virtual mobile network operators), but in order to achieve this, in addition to proper mechanisms for sharing resources, the necessary techniques for creating RAN instances that meet the requirements of the slice in a flash are required [70].

In the Network Function Layer, more focus should be placed on the procedure of creating the network services resulting from the synthesis of several network functions. Although the functions of a more general nature are easier to combine (since fewer parameterizations are needed) in order to result in the final network service, the required adaptability they should have so as to be easily scalable and responsive to new requirements that may arise is not guaranteed. Therefore, researchers should emphasize on creating mechanisms for intelligent synthesis of more fine-grained NFs regardless of the vendor of each one of them in order to ensure interoperability between them, the end services to be scalable and to achieve the desired functionality [70].

In addition, an important open issue is the management and orchestration of the slice components in a holistic base. This involves organizing the network functions of a slice in an end-to-end base so that they work together harmoniously and produce the end result while being flexible and scalable in order to adapt to potential changes. In addition, greater emphasis should be placed on the way in which resources and functions are allocated to services. The techniques proposed include matching infrastructure and functions to services both on the basis of their requirements and the SLA and on the basis of specifications specified by the vendor. In this case, the challenge focuses on the fact that all this procedure should be fully dynamic and take into account possible future extensions [70].

Another challenge that needs to be addressed is the description of services in terms of network functions through a high-level language that should include methods for expressing all the parameters needed to operate a slice (components specifications, requirements, KPIs). It is important to support complex rules, abstract concepts, to have clear expressions and to allow additions and extensions, as the network services model is inherently dynamic and is constantly evolving [70].

In addition to this, in order to ensure a high level of network security that operates on the network slicing model, there should be clear boundaries of slices separation so that it is possible to isolate their functions and handle them separately security-wise. However, the challenge is to apply stringent security rules to network points where there is necessarily a sharing of resources between tenants so as to operate under a single control, as is the case with the radio interface controller. It is important not to violate the separation limits of slices and to avoid duplication of the rules and security policies applied to the network in order to ensure its smooth operation at all times [69].

Finally, a critical issue that requires more attention from the research community is to ensure the cooperation and coexistence of the network slicing concept alongside other 5G-branded innovations such as SDN, NFV and C-RAN. Although there have been fragmentary deployments using data from other pillars of 5G, there has been no holistic approach to network slicing integration with other technologies [71].

8. KEY PERFORMANCE INDICATORS (KPIs)

8.1 Introduction

The multitude of technologies and applications that make up the 5G constitute a complex ecosystem aimed at serving the needs of various and often dissimilar markets in terms of cross-market support. All of the services provided by the new network are expected to create added value in the telecommunications sector, as their use goes into more and more sectors of human activity, thus widening the range of potential users [72].

Given the variety and heterogeneity of the new use cases expected to be supported by the network, there are correspondingly some requirements related to the provision of the differentiated services as defined in the implementation phases of the 5G network by the participants in the development and standardization process (ITU, ETSI, 3G-PPP, 5G-PPP, NGMN). The main pillars for determining requirements are the provision of high-end services to end-users across the range of the applications provided, as well as the optimal administration of its functions at each level and slice of the network. The last part also includes the ability of the network to adapt to the prevailing conditions (SON) at any time while preserving the secure communication factors at any time [73].

In order to assess the operation of the network so as to meet the requirements that best satisfy the requirements and objectives of the services offered, various general key performance indicators (KPIs) and metrics have been defined, which are linked to the overall performance of the service network based on purely technical criteria (reliability and availability of services, network coverage, radio performance, resources flexibility, resilience and security of the whole system) on the one hand and, on the other, to the business and societal contexts (economic and societal analysis) that shape the development model of the network in technical-economic conditions [4] [72] [73]. It is noted that, depending on the user-level applications, there may be some additional specialized KPIs, since the performance requirements of the 5G are inextricably linked to the differentiated services model and to the corresponding use case categories (eMBB, mMTC, URLLC), as these have been developed by the ITU [8] based on specific traffic scenarios [4]. Accordingly, for a more detailed and accurate analysis of the requirements of each service, 3GPP has defined a slightly differentiated service categorization than the one of ITU, which includes four key use cases (enhanced Mobile Broadband - eMBB, Critical Communications – CC, massive Internet of Things – mIoT, Network Operation - NO).

These service groupings, coupled with the implementation of the KPIs, enable providers to find the best possible solutions for their customers and, through analyses, contribute to the process of anticipating market needs so that telecoms companies are always able to serve end users [4] [72]. Finally, the analysis of the results catalytically assesses and judges the services provided so that via theoretical models, calculation analysis, simulations and test measurements an appropriate revision and adaptation of their operating parameters is made in order to optimize them [72] [73].

8.2 Performance Needs

According to the family of each use case scenario (eMBB, mMTC, URLLC, CC, mIoT, NO), the research teams participating in the 5G research and development phase have set the minimum requirements from the network so that the services provided to end users will cover their wishes as much as possible.

The first table below shows these requirements for the most common demand cases and the second table analyzes the corresponding requirements for less popular and more specialized operating scenarios [4] [74].

Table 6 : Performance needs of high data rate use case scenarios [74]

Scenario	Experienced Data Rate (Downlink)	Experienced Data Rate (Uplink)	Area Traffic Capacity (Uplink)	Area Traffic Capacity (Downlink)	Overall User Density	UE Speed
Indoor Hotspot	1 Gbps	500 Mbps	15 Tbps / km^2	2 Tbps / km^2	250000 /km^2	Pedestrians
Dense Urban	300 Mbps	50 Mbps	750 Gbps / km^2	125 Gbps / km^2	25000 /km^2	Pedestrians and users in vehicles (< 60 km/h)
Urban Macro	50 Mbps	25 Mbps	100 Gbps / km^2	50 Gbps / km^2	10000 /km^2	Pedestrians and users in vehicles (< 120 km/h)
Rural Macro	50 Mbps	25 Mbps	1 Gbps / km^2	500 Mbps / km^2	100 /km^2	Pedestrians and users in vehicles (< 120 km/h)
Broadband in a Crowd	25 Mbps	50 Mbps	3,75 Tbps / km^2	7,5 Tbps / km^2	500000 /km^2	Pedestrians
Broadcast-like Services	Max. 200 Mbps (TV Channel)	Modest (e.g. 500 kbps per user)	N/A	N/A	15 TV channel of 20 Mbps	Stationary to in vehicles (up to 500 km/h)
High-speed Train	50 Mbps	25 Mbps	15 Gbps / Train	7,5 Gbps / Train	1000 / Train	Users in trains (up to 500 km/h)
High-speed Vehicle	50 Mbps	25 Mbps	100 Gbps / km^2	50 Gbps / km^2	4000 / km^2	Users in vehicles (up to 250 km/h)
Airplanes Connectivity	15 Mbps	7,5 Mbps	1,2 Gbps / plane	600 Mbps / plane	400 / plane	Users in airplanes (uo to 1000 km/h)

Table 7 : Performance needs of low latency use case scenarios [4]

Scenario	End-to-End Latency	Communication Service Availability	Reliability	User Experienced Data Rate	Connection Density	Service Area Dimension
Discrete Automation – Motion Control	1 ms	99,9999 %	99,9999 %	1 – 10 Mbps	100000 / km^2	100 x 100 x 30 m
Process Automation – Remote Control	50 ms	99,9999 %	99,9999 %	1 – 100 Mbps	1000 / km^2	300 x 300 50 m
Process Automation – Monitoring	50 ms	99,9 %	99,9 %	1 Mbps	10000 / km^2	300 x 300 50 m
Electricity Distribution – Medium Voltage	25 ms	99,9 %	99,9 %	10 Mbps	1000 / km^2	100 km along power line
Electricity Distribution – High Voltage	5 ms	99,9999 %	99,9999 %	10 Mbps	1000 / km^2	200 km along power line
Intelligent Transport – Infrastructure Backhaul	10 ms	99,9999 %	99,9999 %	10 Mbps	1000 / km^2	2 km along a road

8.3 General Key Performance Indicators (General KPIs)

The general KPIs category involves techniques for measuring and evaluating network performance according to the minimum requirements in the network so as to deliver services to end users and to achieve the desired performance, as determined by the parties involved in the network development process.

The main elements for evaluating the performance of the network are the following [73]:

- *Area Traffic Capacity – Volume Density*: refers to the total throughput the network can provide per m^2 of a given area. The final result depends on the density of the BSs in the area, the available BW as well as the level of utilization of the available spectrum [73]. According to test measurements, in an eMBB service, the ideal value of this indicator under indoor conditions is 10 Mbit/s/m^2 for the downlink mode, while cumulatively, in the network, this value reaches 10 TB/s/m^2 [72].

- *Availability*: this metric evaluates the percentage (%) of the time that the network is genuinely able to deliver end-to-end services to a QoS/QoE compliant area that is expected as per design, in terms of the percentage of the total time of its availability. It is particularly useful for providing network services in applications with high reliability requirements while providing at the same time significant feedback in order to improve the energy impact of the network. Alternatively, it can also be defined as the total time during which the system is capable of delivering services in terms of the sum of its total operating time, including the time to recover from possible downtimes [73]. The goal is that the 5G availability rate reaches 99% in the general case scenario and, under specific circumstances, 100% (if possible) [72].

- *Bandwidth*: a typical metric case for network capacity assessment, where, in the case of the 5G, can relate to a broadcast from one or more channels [73]. An indicative maximum value reaches at least 100MHz while, for higher frequency bands of the spectrum, this value reaches 1GHz [72].

- *Cell-edge user throughput*: refers to the percentage (%) of the successively received bits per user at the end of the cell in the time unit and is calculated as a 5% percentage of the cumulative distribution function of the average packet throughput.

- *Connection density*: this indicator shows the sum of the active devices per km^2 of a network area and is particularly useful both for determining the number and type of cells required to support them and for examining the QoS that is ultimately provided. At the mMTC level, where this factor is particularly critical, the number of active networked devices may reach the number of $1,000,000/km^2$ in a total of 7×10^{12} mobile devices [72] [73].

- *Control plane latency*: refers to the delay that occurs to activate an idle device of a control plane level in a data transfer status [73]. At the eMBB and URLLC categories level, a delay of 10-20ms can be considered as tolerable [72].

- *Coverage*: based on the connection loss that occurs during the transfer of data to the channel between BS antenna elements and mobile device antenna elements, the maximum coverage limit (in dB) at which satisfactory service performance is achieved, according to their respective specifications, is calculated.

- *Coverage area probability*: an index that examines the percentage (%) of an area where a service is offered at acceptable quality levels (in terms of QoS/QoE) and is particularly useful in the network design and development process where an area can be divided into subsections for greater accuracy.

- *End-to-end latency*: refers to the time span allowed for the successful transfer of a datagram from the sender's connection interface to the connection interface of the final recipient (One Trip Time – OTT) or, alternatively, includes the time span required for sending and receiving acknowledgments between the sender and the recipient (Round Trip Time – RTT).

- *Energy efficiency*: a quantitative variable (bits/Joule) that evaluates the energy consumption that occurs during data transfer at the same time that a sufficient bit rate for the provision of the services is to be secured [73]. In an eMBB loaded scenario, the energy saving percentage is directly dependent on average spectral efficiency in order to be maximized or not to exist. In the case that data sending is minimal, a high percentage of energy saving can be achieved by applying long-lasting inertia techniques to the system [72].

- *Latency for infrequent small packets*: it is related to the measurement of the amount of time that elapses when a data packet of an application that is sporadically using the network is sent from a mobile terminal located in an idle state (in which maximal energy saving is achieved) until it reaches the corresponding level in the RAN.

- *Mean time between failures (MTBF)*: refers to the statistical average time between two events that render an element or a part of the network inoperative while, in the meantime, the system is fully operational.

- *Mean time to repair (MTTR)*: refers to the statistical mean time period required until the system returns to a fully functional state after an event/fault that rendered it inoperative.

- *Mobility*: metric that examines the maximum traffic speed of an UE (in Km/h) at which it is feasible to provide communication services from the 5G access network, according to the QoS set for the service type on the basis of the differentiated services model [73]. For eMBB use cases and, more specifically, for indoor environments, tolerated high-level communication services situations are fixed users and users moving at walking speeds. For dense urban areas, other than idle and walking situations, the use of network services from a moving terminal in a vehicle with a maximum speed of 30 km/h is also acceptable, while for rural areas the corresponding speed is set at 500 km/h [72].

- *Mobility interruption time*: refers to an index that reflects the time in which a moving mobile terminal is not able to send data to an access point of the network. The desired minimization of this index is critical to the network's QoE and SDNs and multi-RAT concepts can be instrumental in this [73]. For both eMBB and URLLC environments, the value of this index should be 0ms so as to maintain the QoE at high levels [72].

- *Peak data rate*: refers to the theoretical maximum data rate (bps) for a user under ideal network operating conditions (maximum resource availability and no interference) [73]. For example, in an eMBB use case, the minimum requirement is 20Gbit/s for data downloading and 10Gbit/s for data uploading [72].

- *Peak spectral efficiency*: this index is an alternative representation of the *Peak data rate* in a normalized format in relation to the available bandwidth [73]. Similarly to the *Peak data rate* KPI, the desired values should be 30 bits/s/Hz for the downlink and 15 bits/s/Hz for the downlink in order to ensure a satisfactory QoE [72].

- *Reliability*: a percentage index examining the percentage of packets successfully sent to a network node within the time limit expected by the recipient in terms of the total number of sent packets. The waiting window depends, among other things, on the network's architecture and the channel's features that apply each time. It is worth noting that this KPI is interdependent with the MTBF since it is essentially an alternative depiction thereof [73]. For URLLC cases, the tolerable probability of successful transmission of a 32-byte basic layer 2 packet within a 1ms time window is $1\text{-}10^{-5}$, which translates to a percentage of 99% [72].

- *Resilience*: is related to the ability of the network to cope immediately with operational disruption events that may be caused either by natural causes or by improper human handling so that its operation is uninterrupted 24/7.

- *Service continuity*: refers to the ability of the network to conceal from the end user potential minor disconnections of an active service connection during the handovers that result from the movement of the user and his/her servicing from different access types and services.

- *Spectral efficiency per cell / transmission and reception point (TRxP)* : refers to a quotient defined as the fraction that has the cumulative throughput of all active users in an area as the numerator and the bandwidth of the channel divided by the number of the available transmission and reception points in the area as the denominator. Note that in the case of multiple distinctive carriers, the corresponding calculation is made separately per carrier [73]. Essentially, in the first case, it is a classic average KPI for the spectral efficiency evaluation where, according to the ITU, in an eMBB scenario, it should receive downlink layer values of 0.3 bit/s/Hz for indoor transmissions and 0.21 bit/s/Hz in the uplink layer respectively. For dense urban areas the corresponding values are 0.225 bit/s/Hz and 0.15 bit/s/Hz respectively, while for rural areas it would be ideal that the corresponding values be 0.12 bit/s/Hz for downloading and 1.6 bit/s/Hz for uploading [72].

- *Spectrum and bandwidth flexibility*: examines the ability of the 5G to adapt to different use cases and operational scenarios in terms of usage of various frequency bands and different bandwidths of the channels

- *UE battery life*: evaluates the amount of time required for a full battery discharging cycle of an UE, a particularly important element for MTC devices, where this resource is very limited. For a more detailed assessment of battery-powered MTC devices, a UL of 200 bytes/day and a DL of 20 bytes/day with a coupling loss of 164 dB and a battery capacity of 5Wh is considered.

- *User experienced data rate*: an index that examines the data rate (in bps) available in each connection interface located on the end user side in a particular area and is usually related to the minimum data rate required by each service in order to secure the agreed QoS and QoE.

- *User plane latency*: refers to the total delay required to successfully send an application layer data packet through the radio interfaces of the devices that communicate with each other at a 2/3 level. The process is repeated for both the uplink and downlink traffic, basically assessing the entire network operation, architecture and infrastructure [73]. Acceptable values for both an eMBB and an URLLC service are between 10 ms and 20 ms [72].

8.4 Resilience & Security Key Performance Indicators (Resilience & Security KPIs)

The indicators of this category control and evaluate whether network implementation meets the standards that have been set for uninterrupted and safe 24/7 operation under any movement conditions. A key driver of this category is to ensure zero downtime and prevent significant data leakage [73].

In addition to the *Reliability KPI*, *Resilience KPI* and *Mean Time to Repair KPI*, some more specific KPIs have been defined for a more comprehensive coverage of the issue of evaluation, reliability and security of the system, which concern [73] :

- *End-to-end reliability*: refers to the longitudinal evaluation of all the network components required to participate in the process of implementing a network function. In essence, the probability that all (physical and virtualized) network resources required for an end-to-end communication (at the Access/Transport/Core Network level) will function based on a design at the desired time, so that network performance is the expected one, is calculated.

- *Reliability of the telco cloud*: this KPI evaluates the capability of the network's cloud telecommunication infrastructure to provide to the network functions uninterruptedly with the necessary background needed in order for them to be implemented. Practically, the percentage of time when the network's infrastructure can adequately meet the demand for implementation of the necessary functions is measured, even in the case of failures or malfunctions of network elements that may affect its overall performance.

- *Service restoration time*: refers to a classic metric for evaluating telecommunications systems and is related to the time it takes to repair a network malfunction that prevents the provision of network services. The network should be able to cope in a short period of time with such situations, initially by identifying the failure and then by repairing it.

- *Security threat identification*: an index for assessing the security attack detection algorithm that depicts the percentage of threats actually found by the system based on anomaly detection policies applied to the network.

- *Security failure isolation*: refers to the number of security gaps in the network and, in particular, to the percentage of security attacks that have passed through the control areas when something like this should have not happened.

8.5 Resource Flexibility Key Performance Indicators (Flexibility KPIs)

In this group of KPIs are included evaluation indicators related to the ability of the 5G network to adapt the network infrastructure accordingly to its needs (in the context of Slicing and SON concepts) in order to be able to successfully meet the requirements of applications and end users, ensuring continuously high QoS and QoE levels [73].

- *Availability*: refers to the percentage of time that the network's infrastructure is ideally able to support the services and the amount of traffic for which it has been designed and customized by network administrators.

- *Cost efficiency gain*: a quantitative indicator that depicts the weighted operating cost required for the operation and the configuration of the network resources in order for the anticipated traffic to be served. The ability of the network to flexibly update its functions according to the demand that exists at any given time automatically allows for savings and reduced energy consumption by the grid.

- *Elasticity orchestration overhead*: a metric that counts the percentage of resources used exclusively for network management functions, such as resource allocation and orchestration (e.g. VNFs management) and not for the provision of network services as such.

- *Minimum footprint*: this parameter examines the number of resources to be combined to a minimum for the implementation of NFs and the provision of network services to end users. Due to the heterogeneity of resources and the variety of network services supported by the 5G, there may be more than one combination implementing the same service, so the optimal result is sought.

- *Multiplexing gains*: refers to the set of NFs that can be simultaneously implemented in the same group of resources in a in a multiplexed mode, providing satisfactory results in terms of efficiency.

- *Performance degradation function*: this index depicts the correlation that may exist between the performance of NFs and the release of resources used by them so that, within the context of network flexibility, these resources are allocated to other functions. It is particularly important that the correlation of these two parameters is linear so as to prevent abrupt interruptions in the provision of services which may lead to high downtimes.

- *Rescuability*: a particularly critical indicator that reflects the strength of the VNFs during the periods when it is required to reduce the resources used for their operation. In particular, this metric implies, on the one hand, the ability of the VNFs to maintain their operation at the best possible levels during the period of limitation of available resources and, on the other hand, the amount of resources that will be required to restore the system to a normal operating mode.

- *Resource consumption*: an index that examines the percentage of time that is bound by a computing resource of a network from a function for its operation.

- *Resource savings*: metric depicting the percentage and type of resources used for a function operation in flexibility mode when there is a limitation of available resources relative to the corresponding number of resources to be used in normal mode when there is no limitation in resources availability.

- *Response time*: a metric for assessing the adaptability of the flexibility of the VNFs during the period in which changes in the availability of resources they can use by each time examining the amount of time needed to adjust their operation.

- *Resource utilization efficiency*: a metric that evaluates the performance of the VNFs on the basis of the number of natural resources they use and their corresponding time interval which commits them, in order to achieve the desired functionality.

- *Service creation time*: refers to the time required when a request is made to commit resources in order to provide on-demand services under slicing framework until the time the service is available for use [73]. The goal is that this time will reach an average of 90 minutes instead of 90 hours (that a new service currently needs in order to be available to end users) [72].

- *Time for reallocation of a device to another slice*: an indication of the total time it takes to handover terminal devices between different slices (including the time required for the service to be fully functional in the new slice).

8.6 Business Key Performance Indicators (Business KPIs)

This includes indicators that relate to the viability and the development and operation plan of the 5G network, given the new innovative services it aspires to offer to end users and the acceptance they will receive thereof. As it is natural from market laws, the acceptance or not by end users of the services offered will direct the further development of the corresponding infrastructures that support them.

On this basis, some important indications are the following:

- *Incremental revenue per GB*: refers to the revenue that the service provider will derive from offering a service to end users, per GB. Depending on the impact and the audience to whom the respective service is addressed to, the revenue indicator will be respectively formed [73].

- *Incremental cost per GB*: metric that examines the required cost per GB of service provided. This cost depends on the type of service and is adjusted according to the requirements resulting from its specifications in order to reach as much as possible the desired QoE set by the designers. For the economic viability of the 5G, this indicator should be comparatively smaller than the corresponding metric that examines revenue generated by network service provision so as to further drive the R&D in this telecommunications sector [73].

- *Financing private investment in Europe in R&D of 5G networks:* metric to monitor the investment of funds by European private operators in favor of research and development of 5G technologies in order to ensure further development of the new model of operation of mobile networks. In order to extract results, the tactic that is followed is to monitor the budgets of the ICT industry members and to conduct regular polls so as to more accurately record the percentage of their expenses that is being exploited for the benefit of research on 5G technologies over other expenses [72].

- *Participation of small and medium-sized enterprises in the 5G development process*: this indicator is related to the observance of the regulatory framework that has been in Europe and defines the disposal of 20% of public funding for the development of 5G technologies in small and medium-sized private sector entities. The aim is to strengthen the competitiveness and sustainable development of the industry and, in order to draw conclusions from this metric, the number of sums of money and works assigned to such businesses is measured [72].

- *Examination of the market share in the telecommunications sector held by European ICT companies:* this metric is intended to measure the percentage of equipment and applications adopted by the 5G community that have been created by European companies in the industry. In order to draw conclusions, the European ICT players' participation in the provision of equipment and services during the 5G experimental periods and their contribution to the standardization of the new network operation model is examined. In addition to this, in this metric can also be included the number of patents originating from the European area [72].

8.7 Societal Key Performance Indicators (Societal KPIs)

Apart from the strictly technocratic assessment of the new services of the 5G ecosystem, useful information can also be drawn from the study of the social parameters resulting from the use of new technologies.

- *Social Benefits*: this category includes data derived from the impact of 5G new services on society. New ways of networking inevitably affect the social fabric and it is useful to study the effects of this phenomenon as they are a guideline for the further development of new network services. The impact of services and their contribution to improving people's quality of life are two key factors that determine the model of development and funding from a technocratic point of view and, therefore, an extensive study of this interdependent relationship is required [73].

- *Enabling advanced user controlled privacy*: this index includes individual assessment areas related to the ability of network services to ensure the privacy of the communications being conducted, the ability of users to control and allow at their own discretion the transmission of various metadata (location data, usage history, etc.) resulting from the use of third-party network services and the implementation of privacy policies at each communication level. In addition, the use of mechanisms to separate and isolate the operation of the slices between them is examined, so that there are no possible security loopholes in the entire system that may damage the privacy of communications [72].

- *Reduction of energy consumption per service*: this metric examines whether the achievement of the goal of reducing the energy spent per service in rates of 90% compared to 2010 is achieved. Although a specific measurement standard to be applied for this indicator is not fully defined, the output of results mainly takes into account spectrum efficiency related data [72].

- *European availability of a competitive industrial offer for 5G systems and technologies*: this indicator examines the extent to which European telecoms manufacturing companies follow technological advances in the mobile network sector in order to actively participate in the new telecommunications era. The aim is to involve European companies in the 5G infrastructure development process in order to meet the specific requirements of the European market and to develop the corresponding industry in the wider region. Conclusions for this indicator are derived from measurements of 5G telecoms manufacturing companies' costs, giving an estimate of the growth rate of that sector [72].

- *Stimulation of new economically-viable services of high societal value*: this indicator examines the market penetration of new communication services whose use has a high economic and social impact on end users, while at the same time it can make a decisive contribution to improving the productivity and competitiveness of businesses [72].

- *Establishment and availability of 5G skills development curricula*: metric which examines the number and existence of curricula at European universities regarding the use of the new 5G services, thereby speeding up the use and development of new services [72].

8.8 Key Performance Indicators based on service requirements (3GPP level KPIs)

Given the categorization of services by the 3GPP, the evaluation of network services based on the individual needs of its services is based on KPIs that relate to the following sectors [72]:

- In the case of eMBB applications, high data rates for high traffic density per area are being considered, concurrent applications running on the same terminal requiring transmissions at higher and others at lower data rates, coverage of different deployment and coverage scenarios as well as increased EU mobility.

- With regard to Critical Communication (CC), parameter analysis focuses on the achievement of critical communications with very low latency, high availability, trustworthiness, accuracy positioning and very low latency.

- In the category of mIoTs, it is important to analyze the operating conditions of each application in terms of management of the natural resources of the device in which the application is deployed, its operating environment, the device's connection mode (direct or indirect connection mode) and its interaction with one or more users.

- In the NO category, parameters such as the flexibility and and scalability of the network, the options for network slicing, the multi-RAT and mMIMO concepts, the backhauling options, the broadcasting ability as well as the interworking and security aspects of the 5G are examined.

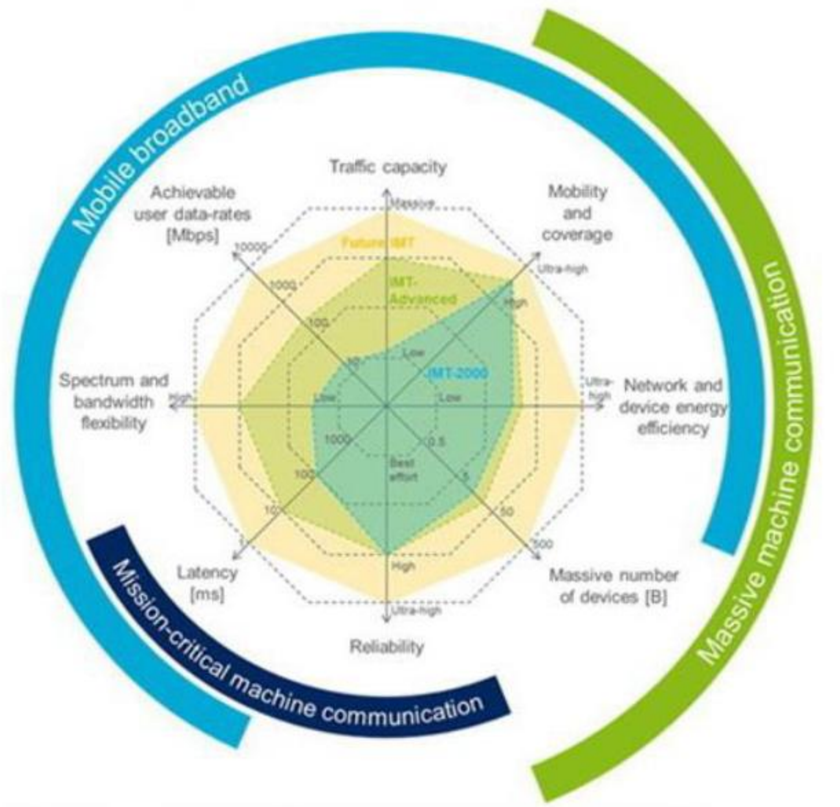


Figure 58 : Performance requirements of different 5G vertical domains [72]

It is worth noting that, in addition to the above KPIs, there may also be **customized application-based KPIs** that cover additional parameters of the end applications using the network. In this way, developers are able to have the proper feedback for the continuous improvement and development of their services [73].

Moreover, although there is a convergence between researchers as to the need to evaluate 5G data, there are small variations as to the definition of some KPIs, and therefore need further clarification from the community. Such examples are KPIs related to energy efficiency, the security and privacy aspects of communications and the operation of end-to-end services [72].

9. PREREQUISITES FOR EMULATION AND EVALUATION OF A SDN-CONTROLLED TRANSPORT NETWORK

9.1 Introduction

The purpose of this experimental part of the diploma thesis is to study and evaluate the application of SDN technologies to an existing conventional IP network which covers a large geographic area and exclusively interconnects and provides a variety of services to the approximately 1100 users of the organization it serves (dedicated network).

More specifically, network efficiency is examined and compared using a SDN-based model, which in the first case has single link connections between its switches, and in the second case, there are also backup links in the connections of the switches (double links in parallel, not aggregated). Essentially, under certain traffic conditions, the network performance is examined when there are duplicate connections between the switches and the extent to which its management complexity is affected by the controller.

The experiment requires the development of an emulation environment (testbed) that includes a plethora of technologies and applications, the main ones being analyzed in the following sub-chapters.

9.2 Ubuntu OS

Ubuntu is the most widespread Linux-based distributed operating system, which is open-source and free to use. It is a fork version of Debian as regards its structure and architecture and is available in Desktop, Server and Core editions in order to fully meet the requirements of a modern OS. It supports the entire range of computer architectures and can be fully installed and run on both physical machines and virtual machines or even cloud platforms.

Being Linux-based, it takes full advantage of the features that such OSs have, such as reliability, security and high efficiency while being easy to use and friendly even to basic users. Its capabilities are not limited to the default functionality and to the applications of its respective distributions, as it can be customized and expanded in order to suit even the most demanding users or researchers seeking enhanced functionality [75].

9.3 Oracle VM VirtualBox

Oracle VM VirtualBox (VirtualBox) is an integrated open-source virtualization software, which utilizes virtualization techniques in order to support multiple guest VMs of various operating systems belonging to the larger Windows, Linux, Mac, Solaris and OpenBSD families. It is available for x86 and x64 Intel/AMD architectures, can run both on common infrastructure (desktops, laptops) and on purpose-built environments (servers, cloud platforms) and also has several plug-ins/extensions that extend its functionality. From an architectural standpoint, it is a type 2 hypervisor as it runs as a traditional application on an operating system that is installed on physical infrastructure.

It can run multiple VMs at the same time, regardless of each guest's OS, with any limitations arising from the amount of the resources that can be allocated for this purpose by the host OS (physical infrastructure on which VirtualBox runs) and relating primarily to the CPU cores, the size of the memory and the capacity of the disk.

It has a very user-friendly Graphical User Interface (GUI) that allows the user to create and customize (according to resources allocation, networking configuration, etc.) VMs in a simple and convenient way at any given time, thereby providing a great deal of freedom of movement. Specifically, with regard to the networking configuration of VMs, there are several types of connectivity available (not attached, NAT, bridged, internal, host-only) and multiple network interfaces are supported for each VM, thereby allowing subnets to form between them [76] [77].

9.4 Gnuplot

It is a command-line, free of charge, graphing software which is used to represent data in 2D and 3D plots. It is a cross-platform program, as it runs on all modern OSs (Linux, Windows, Macintosh, Solaris) and has many useful features.

Gnuplot supports multiple data formats for both input and output providing great freedom to use and exploit its capabilities. It can be handled either through direct interaction with the user through its CLI, or through scripts that greatly facilitate cases of multiple data sources. Another important advantage is the fact that it can receive input data stored in the systems as well as data generated from other programs through pipelining techniques [78].

9.5 OpenFlow

OpenFlow, as mentioned above in the corresponding chapter, is the basic protocol used in order to control the way switches exist in an SDN network from the centralized controller. All the information required by the switches are provided by the controller (which is fully aware of the state of the network) via OpenFlow in order to properly forward packets. This information is structured as flowtable entries made available by the controller and are forwarded to the interfaces of the switches that exist for this purpose, and are stored in their respective flow tables so that they know how to manage the data transmitted to the network.

Additions and improvements are made to various versions of the protocol from time to time in order to best support and integrate all the required actions of a network of this format. The additional functionality relates to its various features (flow entries/tables management, bundle messages, metrics & statistics algorithms) thus facilitating the implementation of the SDN concept in communications networks [79] [80] (see chapter 5.4 OpenFlow).

9.6 Open vSwitch

Open vSwitch is a visualization of the physical device at the hardware level that implements switching processes (layer 2) using virtualization technologies in order to ultimately make its functionality available as a software. It is precisely this feature that enables complete control and management of the forwarding functions at any time through appropriate interfaces in an easy and straightforward manner, so that the functionality of the network can be adjusted as required. Therefore, it is ideal for use in SDN environments as it fully covers the requirements of an edge switch as well as the concept of the entire concept (visibility, scalability, QoS, monitoring, isolation, network orchestration, virtualization, etc.) and can cooperate with the other components constituting such a network [81].

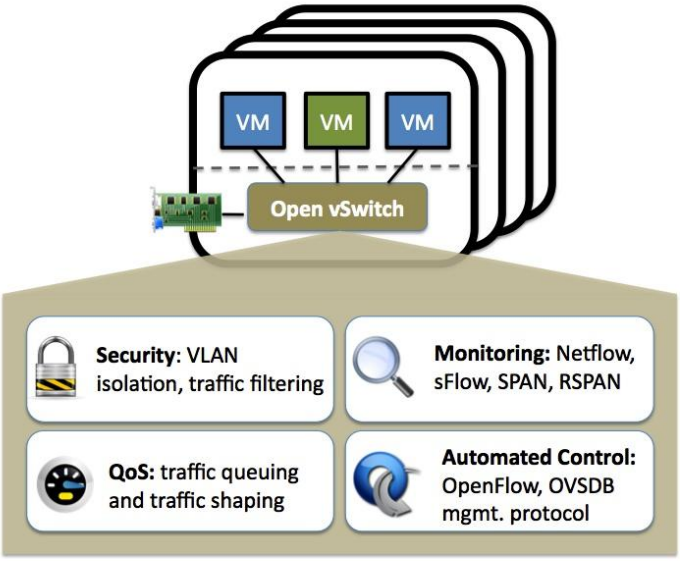


Figure 59: Open vSwitch model [82]

Primarily designed to work in virtual environments and even interconnect heterogeneous devices (virtual and physical), it perfectly supports both the OpenFlow protocol (and many of its extensions) and the conventional network communication protocols (e.g. Ethernet) and can also be used as an embedded functionality in physical devices. In addition, it is a Linux-based open-source software and is compatible with various virtualization techniques, including Xen. XenServer, KVM and Oracle VirtualBox [82] [83].

As regards its operating structure, Open vSwitch is implemented in terms of testing the features of packets, modification and forwarding in the kernel of the operating system (Fast Path), while the other functions (secondary forwarding functions, load balancing, remote visibility, etc.) are implemented in the userspace (Slow Path) [83]. Although newer versions allow its entire implementation in the userspace, all its functional features with this operating logic are not yet fully tested [82].

9.7 Mininet

Experimental tests in order to evaluate the SDN controller response to different operating scenarios and topologies of the network under consideration require either its implementation using natural resources or its implementation in an artificial environment capable of simulating all its required functions and features. For the purposes of this paper, we chose to use an artificial environment (virtual network), which essentially is the Mininet network emulator which, through software, can simulate the operation of various network elements such as switches, routers, hosts and the links for connections to the above devices by using a Linux kernel.

The interconnection between devices for the exchange of the necessary packets is performed via virtual interfaces following the Ethernet protocol, allowing the user to customize the network connection parameters (bandwidth, queue size, delay, etc.) of the network. The maximum limits of these parameters depend on the computational resources allocated to the physical infrastructure on which the Mininet runs. This creates an integrated virtual environment in which, among other things, it is possible to operate user terminals that have a similar behavior to physical infrastructure terminals; a number of different applications can be executed on these terminals, provided that these applications are already installed on the physical machine which is used by Mininet.

In this way, the behavior of a network in the virtual environment can be fully simulated, enabling the user, through the virtual terminals, to easily and simply perform experiments and tests on the network under study, which would otherwise require more complex processes (setting up the network on a physical infrastructure, configuring of the network devices, etc.) with corresponding costs in time and money [84] [85].

9.7.1 Features and capabilities

In essence, Mininet is a powerful tool that enables an integrated and highly realistic simulation of networks in an easy, fast, flexible and efficient way, since all actions can be performed by using simple commands or even scripts, enabling the user to have complete control over the programming of each element at any time.

In order to operate Mininet, there is a user-friendly Command-Line Interface (CLI) that enables the user to perform all the actions required to perform the study on the emulated network in a simple and easy way. At the same time, viewing the test results is readily available to the user, as well as the ability for debugging through system log files production processes which includes multiple indicators of the system.

This allows the user to instantly create an integrated network topology, customize it at any time and execute on it any Linux-based application he/she wants for experimental purposes. It is precisely this feature of the Linux-based design that gives Mininet another significant advantage over other network emulators, since the designed topologies can be ported and implemented in physical infrastructure without any special modifications to their code, given that its mode of operation is similar to that of a real network [86].

As regards the topological features of the network, either simple predefined topologies with fixed attributes but customizable in size can be used or completely custom topologies in which the user can define exactly the desired features. The composition of the custom topologies is made through the Python API available with Mininet, which gives a great deal of freedom in designing the network in question while allowing highly complex structures.

An important advantage is the fact that its capabilities include support for both OpenvSwitch (OVS) and the OpenFlow protocol, which are key components of the SDN concept. In addition, other native features of the SDN are supported, providing an ideal network testbed to the SDN research community.

Moreover, it is worth emphasizing the flexibility of installing and operating Mininet since it does not require any special computing resources and can be implemented either directly on a physical infrastructure (on a server, desktop or laptop), on virtual or even on cloud platforms.

Finally, Mininet is an open source software, so each party concerned can contribute actively to its development community and can customize it accordingly in order to meet the simulation needs [84] [86].

9.7.2 Functional Features

Mininet is essentially a fully integrated virtual environment appropriately tailored to focus on network simulation. It therefore includes various related functions such as creating topologies with their respective parameters, the ability to simulate and configure terminal consoles in the elements of the emulated network in order to execute programs on them and measuring network performance.

There is a CLI with a large number of commands that fully cover the managing needs of a network for the user interaction with Mininet in order to make the most of its functionality. The most useful commands include the *dump* command that displays useful information for each active node in the network, the *exit* command to terminate the Mininet environment, the *help* command that is an index of all available commands, the *nodes* command that displays all the nodes of the topology, the *net* command that displays all links between network nodes, the *sh <command>* used as a prefix of a Linux command for its execution (applies to commands not directly included in Mininet's environment) and finally the *xterm < hostname>* used to open an independent virtual terminal on a node. In the standalone terminal that opens using xterm you can execute Linux commands just like on a physical machine (provided that the corresponding repositories for their execution exist on the hosting server where Mininet is running).

In terms of topology creation and customization thereof, Mininet provides via its CLI appropriate commands for both the use of predefined topologies and custom topologies created using the python API. The general form of the command in order to create a topology is:

```
server> sudo mn --topo <topology> --mac --switch <switch> --controller= remote,
ip=<controller_IP>,port=<controller_port>
```

The arguments of the abovementioned command can take the following values by defining different topologies with different attributes per case:

- topology : *minimal / single,<X> / linear,<X> / tree,depth=<Y>,fanout=<Z> / torus,<X,Y>*
- mac : *sets user friendly MAC addresses in hosts*
- switch : *ovs / ovsk*
- controller_IP (optional) : *IP address of the remote controller (if exists)*
- controller_port (optional) : *port number of the remote controller (if exists)*

In the case of custom topology creation using the python API, the network parameters can be defined by the python script and the command is as follows:

```
server> sudo mn --custom <script.py> --topo <topology> --mac --switch <switch> --
controller=remote, ip=<controller_IP>
```

In this case, the *custom* python script (script.py) that will also be used in order to create the topology is defined, and the *topology* specifies the name of the topology as it is specified in the script.

An example of custom topology using the python API is the following, which, after inserting the required library, creates a network whose topology includes 2 switches that are linked together and, in addition, 2 hosts are created where the first one is connected via a link to one switch and the second to the other switch. Finally, via the key/value pair created and added to the topos dictionary, the user is enabled to refer to the specific topology from the CLI with its assigned name (mytopo)

```
from mininet.topo import Topo
class MyTopo( Topo ):
   def __init__( self ):

      # Initialize topology
      Topo.__init__( self )

      # Add hosts and switches
      leftHost = self.addHost( 'h1' )
      rightHost = self.addHost( 'h2' )
      leftSwitch = self.addSwitch( 's3' )
      rightSwitch = self.addSwitch( 's4' )

      # Add links
      self.addLink( leftHost, leftSwitch )
      self.addLink( leftSwitch, rightSwitch )
      self.addLink( rightSwitch, rightHost )

topos = { 'mytopo': ( lambda: MyTopo() ) }
```

The main features of the above code are the *class Topo()* which is the basic class for creating Mininet topologies, as well as the functions *addHost(), addSwitch(), addLink()* which are the building blocks of a network since they create hosts, switches and two-way links in the network, respectively. In addition, functions such as *start()* and *stop()* related to starting and terminating the network are available, as well as the *addController()* function that creates an SDN controller for the network. General-scope CLI commands (such as *setLogLevel(), dumpNodeConnections(), print(), pingall()*) can also be used, giving the user more freedom to manage the network.

Furthermore, by using appropriate arguments, the performance characteristics of the network can be customized in order to simulate the network under study as accurately as possible. For example, *addHost()* and *addLink()* functions can customize the operation of the elements they handle with appropriate arguments. Thus, in the first function it is possible to determine the allocation of the processor resources among the network hosts [*addHost(<name>,<cpu=x>)*], while in the second one attributes of the link connecting two network nodes can be specified, such as the maximum bandwidth, link delay, queue size and loss percentage [*addLink(<nodeX>, <nodeY>, bw=<Z Mbps>, delay=<'W ms'>, max_queue_size=<V packets>, loss=<R %>)*]. Similarly, the *setARP(), setIP()* and *setMAC()* commands can determine the connectivity characteristics of a terminal to the rest of the network in a direct and easy way [87] [84].

9.8 Iperf

Iperf is a very useful open-source and cross-platform command-line tool that is used to measure network performance. All of its functionality is based on the creation of active TCP/UDP data streams through its client and servers functionalities, with the ultimate goal of measuring the maximum achievable bandwidth in the network connection. In the results generated by its use, the connection performance is recorded at any time with user-defined chronological accuracy [88].

9.9 Open Network Operating System (ONOS) Controller

ONOS is one of the most important open-source controllers for SDN networks, designed to provide universal control over all the services of the control plane [89]. Its deployment is coordinated by the ONF and in particular by the Open Networking Lab (On. Lab) while at the same time various network & service operators as well as network equipment vendors are involved in the project, making up a large community [90]. It is implemented in Java/Javascript as bundles located in Apache Karaf containers [89] and thus takes advantage of all the benefits of such software, being available to run on all types of computers (cross platform) while at the same time having, besides CLI, a very handy Web-based GUI .

The ambition of the entire community is to provide network operators with a complete network OS capable of controlling and coordinating in a logical, centralized, automated, and efficient way all relevant network components (switches, links) in order to continuously achieve high performance and network availability [89]. The important thing is that these processes can be performed in real time, giving the network a high degree of flexibility and adaptability to the individual traffic conditions, enabling it to respond to any demand that arises without requiring prior action [91]. Also, because of its design, it can perfectly handle both large-scale and high speed networks, as well as mashup networks [92].

9.9.1 Architecture

As regards the functional architecture of the ONOS, its structure follows the principles of the physically-distributed environments (cluster of servers) [93] [94], giving it the opportunity to benefit from the existence of many physical resources that can be used either simultaneously for load-balancing reasons either as a backup, thus consistently ensuring high availability under any operating conditions. In addition, it is precisely because of this logic that network administrators are enabled to perform software and hardware updates without affecting data movement on the web [89].

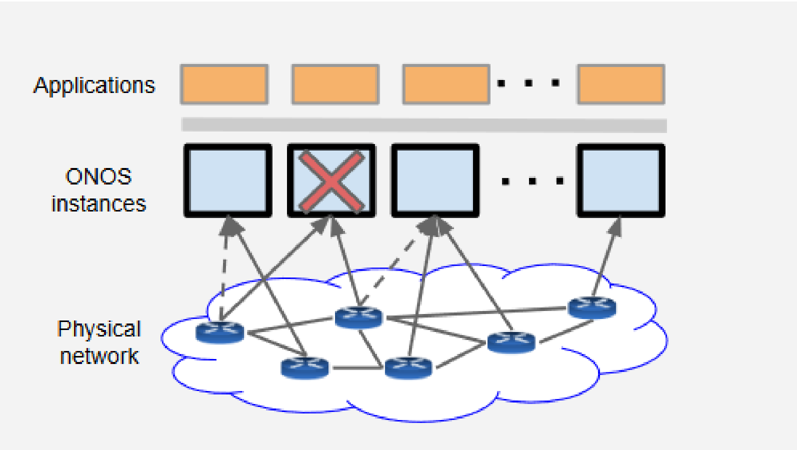


Figure 60 : ONOS distributed architecture [94]

Following the standards of the SDN that define the “controller” concept, at its edges are the corresponding Northbound and Southbound APIs responsible for the exchange of information between the high-level network applications and the network elements of a network.

In addition to its basic operating structure, it is also known for its ability to integrate and execute plug-in components in the form of modules, which can provide additional functionality to network management according to the respective needs. Their integration is fairly simple and straightforward through appropriate interfaces, which ranks it to the top in terms of scalability potential [89] [91] [94]. The additional functionality of the modules is integrated into the rest of the system in such a way that neither the core functions of the system nor other neighboring components are affected, which may lead to a lower reliability and also to an easier uninstallation or upgrading thereof, if deemed necessary [94].

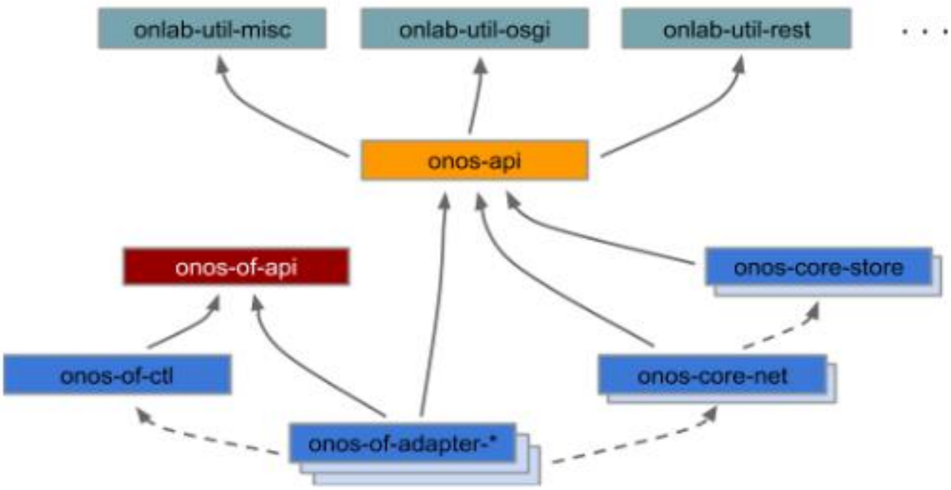


Figure 61 : ONOS modular structure [94]

9.9.2 Advantages & Innovations

The dominant innovation of the ONOS is that it operates as an integrated network OS that is responsible for efficiently managing both network devices in order to fulfill its role as an SDN controller and the computer resources (CPU, RAM, disk) of the physical infrastructure on which it runs as a software. The intrinsic development of the following distributed architecture (distributed core) is also important, resulting in having all the benefits of this technique (increased levels of reliability, error tolerance, load-balancing). Moreover, its module-based structure, on the basis of which each of its functions is a separate component that interacts with the rest via appropriate interfaces, facilitates debugging, maintenance, upgrading and extending its functions without any complexity in an abstract way.

The report in a major Northbound abstraction, the Intent Framework, which enables network programming based on the network-functions-as-a-service model, should not be overlooked. Thus, in this case, operators are able to adapt their network to a high programming level without having to know how this service is provided and any restrictions that may arise from its use. Responsible for resolving potential conflicts is the controller, who, after completing the necessary actions in order to fulfill the functionality specified by the administrator, creates the appropriate policies for the consistent rescheduling of the network.

The ONOS is also capable of managing multi-layer networks seamlessly, ensuring greater overall network performance, increased levels of security and better levels of traffic engineering. Another advantage is its ability to interconnect different SDN-based Autonomous Systems (ASs) through the *SDN-IP Peering* application without having to interfere with the rest of the Internet with the BGP protocol, as is the case with traditional networks. Finally, via the *Segment Routing* application, it is possible to manage MPLS networks in a simplified and smart way [94].

9.10 Testbed characteristics

At the implementation level, the entire environment is set up in an ASUS X541-UV laptop, with an Intel® Core™ i3-7100U Processor (3M Cache, 2.40 GHz) and 8 GB RAM, running Ubuntu Desktop 18.04 LTS (x64) as OS and Oracle VM VirtualBox 6.0.14 for virtualization platform.

More specifically, the Mininet 2.2.2 using the OVS 2.9.5, which runs in a VirtualBox VM with 1 core CPU, 1 GB RAM, two network adapters (a NAT for Internet connection and a host-only for interconnection between VMs) and the guest OS is Ubuntu Server 18.04 LTS x64 was used to simulate network infrastructure.

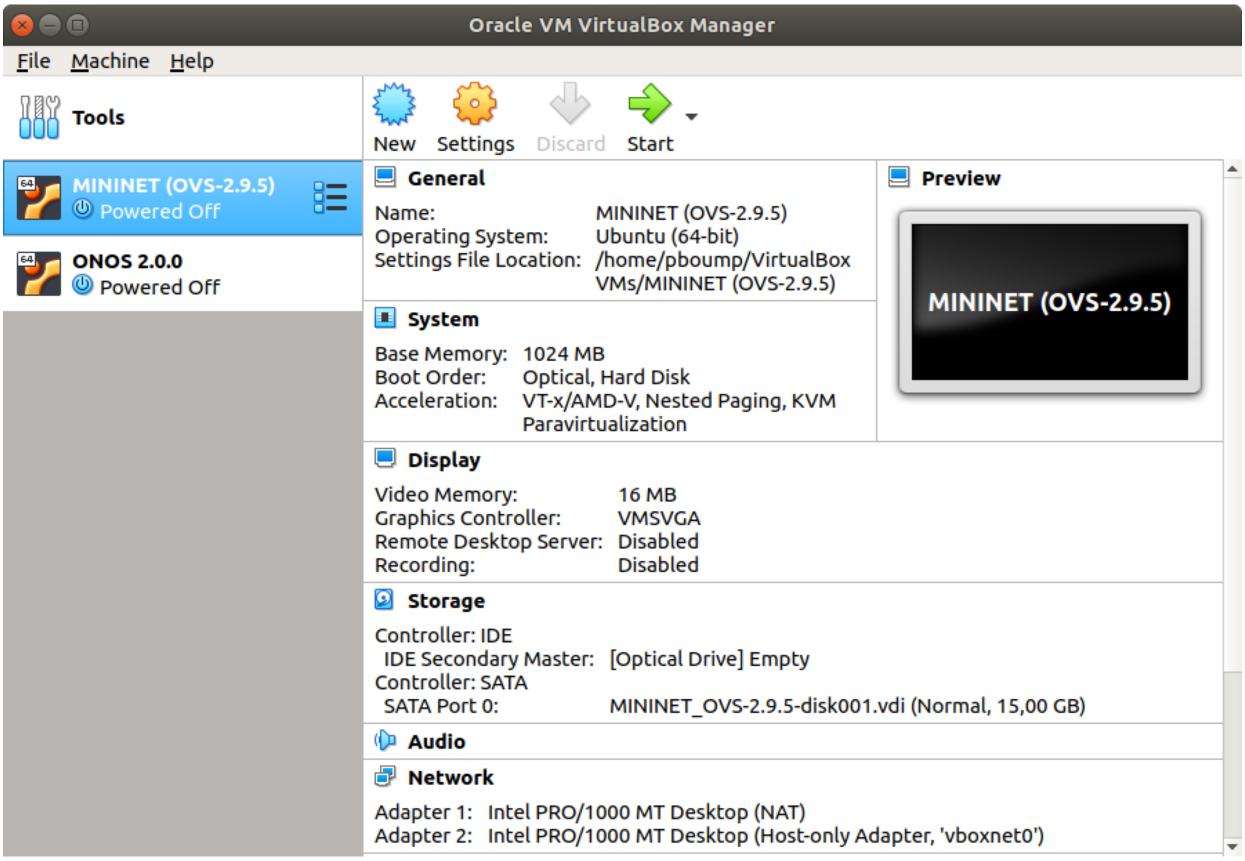


Figure 62 : Mininet VM characteristics

The ONOS 2.0.0, for the implementation of which a VM in VirtualBox with the abovementioned characteristics, is selected as the controller of the SDN Network. It has 1 core CPU, 3 GB RAM, two network adapters (a NAT for Internet connection and a host-only for interconnection between VMs) and its guest OS is Ubuntu Desktop 18.04 LTS x64.

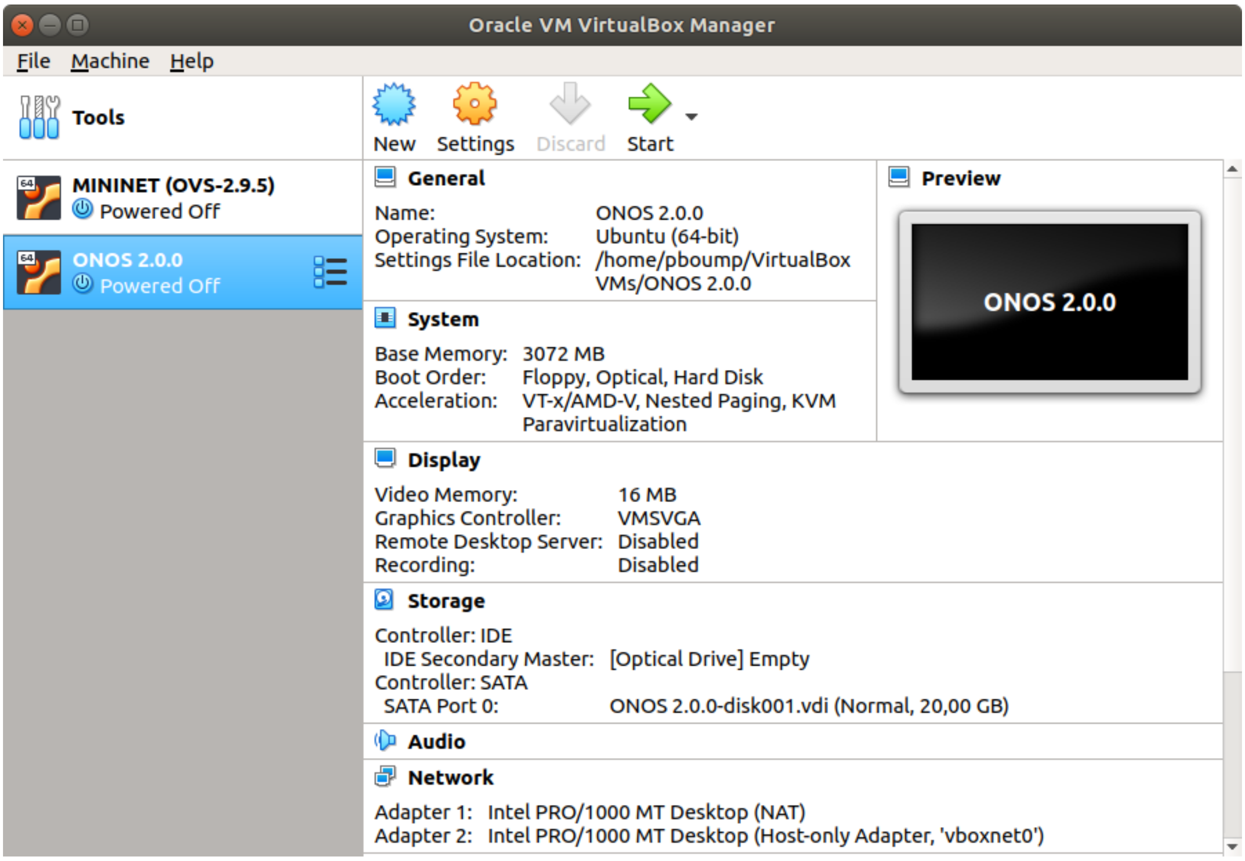


Figure 63 : ONOS VM characteristics

The choice of the ONOS was based on 3 main parameters related to the structure and the philosophy of its operation and to the network studied in the present paper. First of all, the fact that the ONOS is inherently designed to function as an integrated NOS [94] and not as a "simple mediator" in the exchange of OpenFlow messages between the northbound and southbound interfaces is a particularly important feature. Coupled with the fact that its implementation is an integrated solution designed for commercial use and not for research/academic purposes like most SDN controllers, then the two most important factors of this choice are explained. The choice was based on the fact that, according to studies, it perfectly supports large-scale networks without affecting its performance [94].

Finally, it should be noted that due to constraints in physical resources and without affecting the conclusions of the study, the number of end users (hosts) in the simulation of the network corresponds to approximately 13% of the actual users.

10. EXPERIMENTATION FRAMEWORK

10.1 Network Evaluation Parameters

In order to evaluate the performance of the controller in the management of the network's traffic, three basic parameters are examined under certain operating scenarios. The first relates to the ***throughput*** achieved during the data stream between certain nodes of the network that are of particular importance in its operation, while the second relates to the ***number of flows*** generated in each case and how much they affect the network's overall performance. Finally, the number of ***OpenFlow messages*** exchanged while the data streams are active on the network is also examined.

10.2 Details of the examined network

The network under study emulated using Mininet consists of 34 switches (Nicira Open vSwitches - version 2.9.5 & OpenFlow version 1.4) and 135 hosts (5 servers, 130 clients), whose connectivity is coordinated by 1 controller instance (ONOS - version 2.0.0).

The naming of the switches results from the rule "of: <xxxxxxxxxxxxxxxx>", which corresponds to each switch's MAC address, while user-friendly names that indicate their role in the network (e.g. mainDCCore, opsCentreDistribution, technicalDeptEdge, etc.) have been given to the switches. As regards the hosts, their names for the servers are derived from the "hs<*y*>" rule, while for the clients the "h<*y*>" rule is followed, where the variable *y* represents the number of the host. In addition, in order to facilitate the experiments, IP addresses under the format 10.0.0.<*y*> have been assigned to the client hosts, with *y* corresponding to the name of each host based on the naming, while IP addresses in the range 10.0.0.241 - 245 have been assigned to the 5 server hosts.

All of the above features are specified in the custom scripts that create the two different versions of the network.

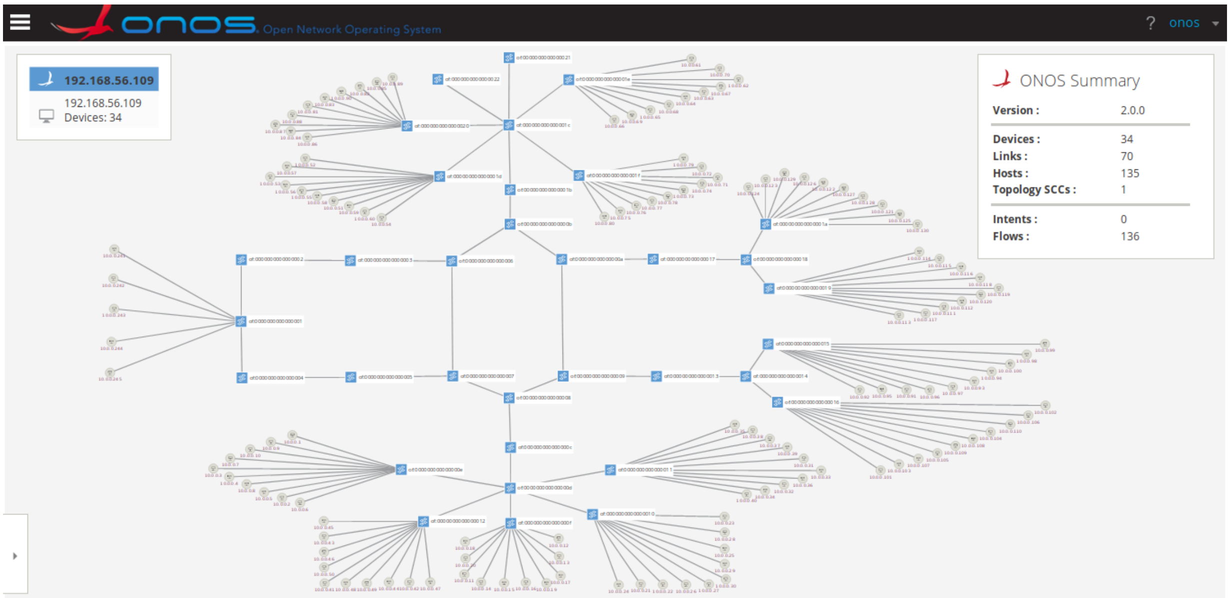


Figure 64: The network with single link connections between its switches

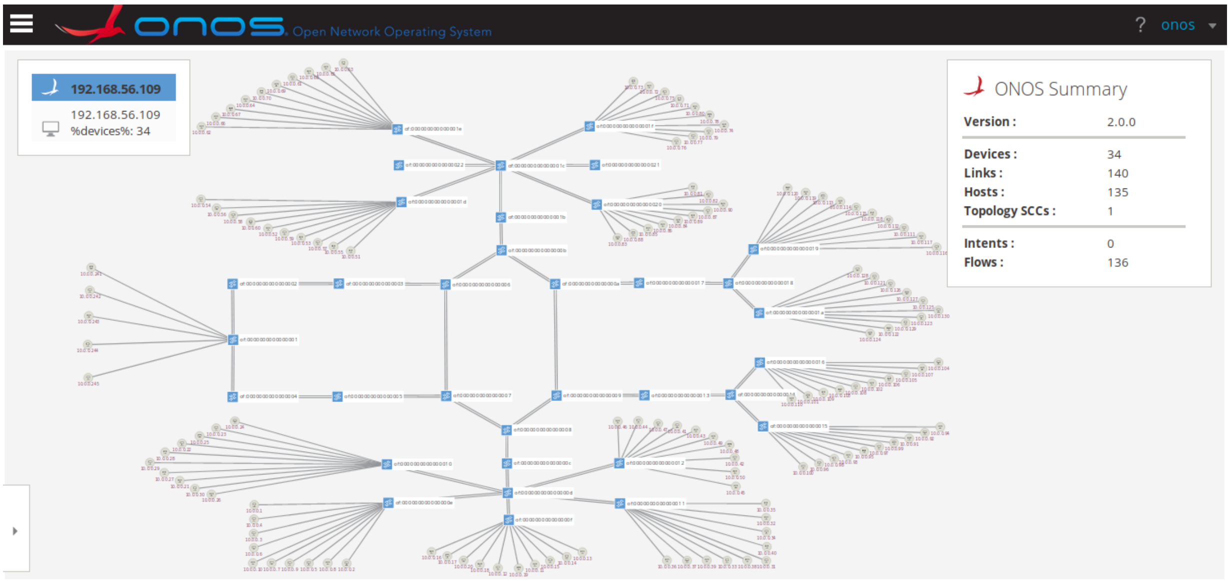


Figure 65: The network with double link connections between its switches

10.3 Evaluation Methodology

The different scenarios on which the network is evaluated concern both its topological characteristics as well as the traffic conditions created on it in order to control its performance. In terms of its topology, in the first case tests are carried out with the switches being interconnected with single links, and in the second case backup links are also available, thus practically creating double links. As regards the data traffic generated on the network at any given moment, there are two versions considered for each of the aforementioned topologies. In the first, the traffic on the network relates to one client-server data stream, while in the second there are four simultaneous data streams. In addition, the behavior of the network is examined in each case, whether or not there are link failures.

The hosts involved in the pilot tests were selected in such a way so as to cover the full length of the network as best as possible and also in order to examine the network's behavior based on the worst case scenario that may occur each time (multiple and continuous link failures). Thus, the 4 pairs of client-server connections that will be used in total in the tests in order to create data streams are formed by the nodes h70-hs1, h111-hs2, h98-hs3 and h32-hs4.

Table 8: Hosts forming client-server couples for data streams emulation

Data Stream	Server Host	Client Host
1	hs1	h70
2	hs2	h111
3	hs3	h98
4	hs4	h32

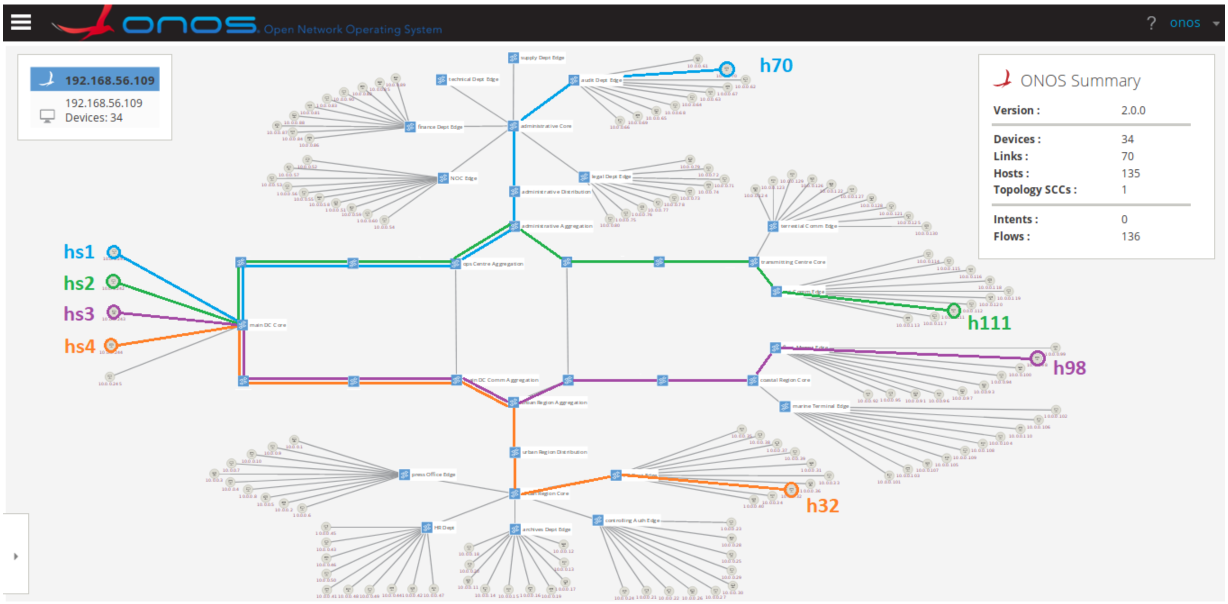


Figure 66: Server and client hosts in the network

Based on the above data, the following eight different network case studies are presented, in which performance measurements will be performed and their results will be compared:

I. Network with single links, traffic from 1 data stream and no link failures
II. Network with single links, traffic from 1 data stream and link failures
III. Network with single links, traffic from 4 data streams and no link failures
IV. Network with single links, traffic from 4 data streams and link failures
V. Network with double links, traffic from 1 data stream and no link failures
VI. Network with double links, traffic from 1 data stream and link failures
VII. Network with double links, traffic from 4 data streams and no link failures
VIII. Network with double links, traffic from 4 data streams and link failures

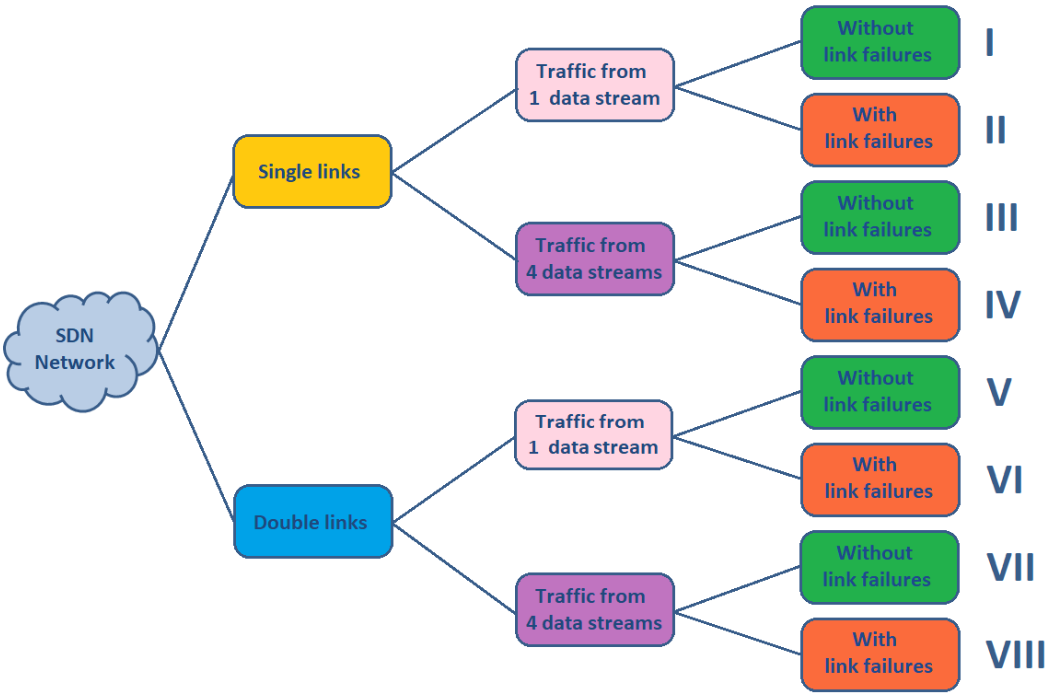


Figure 67: Evaluation scenarios

For the sake of greater accuracy and validity of the result, for each of the above cases, 5 measurements were repeated and, following the statistical analysis of the results, their mean is presented. It is worth noting that there were no significant discrepancies between the individual elements of each experiment and that no special handling thereof was required.

10.4 Tools and techniques used

10.4.1 Measurement of the throughput

The ***iperf*** tool is used in order to measure the throughput, where a server instance (*iperf–s*) is run on the host side of the server role and client instances (*iperf–c*) are created on the other hosts, respectively. Once the data stream is activated, simultaneous recording per second (*iperf –s –i 1*) of the maximum achievable bandwidth in the network connection is performed. In this case, TCP data streams are created that are active for 180 seconds (*iperf –c –t 180*), which is enough time in order to draw the necessary conclusions.

In this regard, both the client-side and server-side processes in order to create TCP data streams are executed in the corresponding ***xterm*** environments of the hosts selected for the experiments. Results are recorded per second, while results are redirected via pipelining to .txt files saved in the shared filesystem of the Mininet environment for further analysis and processing.

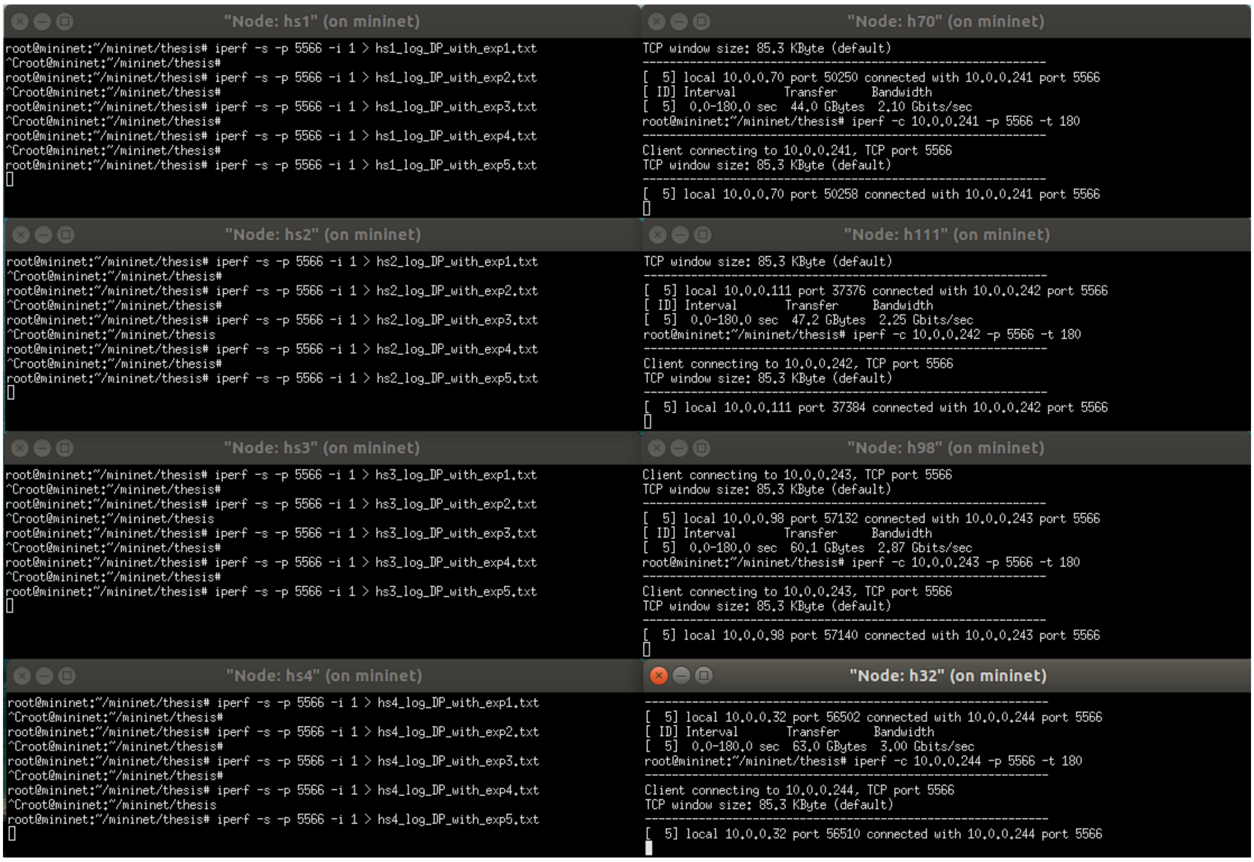


Figure 68: The execution of the iperf tool in the xterm environment for the measurement of the throughput of the data streams

10.4.2 Recording the variation of the flows

At the same time, while the data streams are active, the flows are logged every 5 seconds and the results are saved in .txt file in the controller. This procedure is implemented via the proper ***bash script (flows_counter.sh)*** which is listed below.

```
#! /bin/bash

while true; do
    sshpass -p rocks ssh -p 8101 -v -o StrictHostKeyChecking=no
onos@192.168.56.109  "flows | grep "id" | wc -l >> flows_counter.txt" >/dev/null 2>&1
    sleep 5
done
```

At this point, it is necessary to point out two parameters regarding the methodology followed, the understanding of which is particularly important in order to read the final results. The first concerns the number of flows in the network during idle periods, that is, when there are no active data streams. Barring unforeseen circumstance, during these times the flows in the network remain constant. Their number equals the number of packets received by the controller during the network creation phase, and in particular after executing the *pingall* command, which is used to make the controller fully aware of the network's hosts. Thus, since there is no information when the command is running for the first time, packets are redirected to the controller which installs a new flow (reactive mode) for each miss match packet. Consequently, there are 136 flows that occur at any time in the controller regardless of the traffic on the network.

The second parameter relates to the fact that the recording of flows includes a period of 35 sec before the start of the data stream (idle period) and 35 sec after its termination (idle period), in order for changes that occur in the total number of the network's flows during switches from idle periods to periods during which network events occur (data traffic, topology changes, etc) to be noticeable to the user.

10.4.3 Recording OF messages

Moreover, just before the start of the data stream and just after its completion via the ***ONOS Control Plane Manager API***, all the necessary information is stored for OF messages exchanged in every possible scenario.

10.4.4 Simulation of link failures

Also, in the network evaluation scenarios that contain link failures incidents, their simulation is performed using appropriate ***bash scripts (del_ports_SP.sh & del_ports_DP.sh)*** which disable ports in switches, the order of which has been selected in such a way that the controller is each time forced to select alternate routes. In order to do this, firstly, via the ***ONOS Core API,*** the path to be followed at any time is detected, thus deleting the appropriate ports in order to override the selected links, eventually resulting in continuous changes in the connectivity of the network's nodes. Deletions are made every 5 seconds, starting 30 seconds after the set up of the data stream and the start of the throughput monitoring. In the case of a network with single links, 2 deletions are made (since otherwise there will be no connectivity anymore), and in the case where there are double links in the network, 17 deletions are made. These actions are being implemented via the *ovs-vsctl* utility which is an interface for the customization of the Open vSwitches.

The bash script for the simulation of the link failure incidents by disabling the appropriate ports in the switches of a network with single links is listed below.

```
#!/bin/bash

sudo ovs-vsctl del-port s6 s6-eth3
echo "port s6-eth3 deleted"
sleep 5
sudo ovs-vsctl del-port s4 s4-eth1
echo "port s6-eth6 deleted"
```

In the case of the double links network, the bash script for the simulation of the link failure incidents is as follows.

```
#! /bin/bash

sudo ovs-vsctl del-port s6 s6-eth3
echo "port s6-eth3 deleted"
sleep 5
sudo ovs-vsctl del-port s6 s6-eth6
echo "port s6-eth6 deleted"
sleep 5
sudo ovs-vsctl del-port s10 s10-eth2
echo "port s10-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s9 s9-eth2
echo "port s9-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s8 s8-eth2
echo "port s8-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s7 s7-eth3
echo "port s7-eth3 deleted"
sleep 5
sudo ovs-vsctl del-port s4 s4-eth1
echo "port s4-eth1 deleted"
sleep 5
sudo ovs-vsctl del-port s4 s4-eth2
echo "port s4-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s5 s5-eth2
echo "port s5-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s5 s5-eth4
echo "port s5-eth4 deleted"
sleep 5
sudo ovs-vsctl del-port s7 s7-eth2
```

```
echo "port s7-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s3 s3-eth2
echo "port s3-eth2 deleted"
sleep 5
sudo ovs-vsctl del-port s3 s3-eth3
echo "port s3-eth3 deleted"
sleep 5
sudo ovs-vsctl del-port s2 s2-eth1
echo "port s2-eth1 deleted"
sleep 5
sudo ovs-vsctl del-port s28 s28-eth1
echo "port s28-eth1 deleted"
sleep 5
sudo ovs-vsctl del-port s27 s27-eth3
echo "port s27-eth3 deleted"
sleep 5
sudo ovs-vsctl del-port s30 s30-eth2
echo "port s30-eth2 deleted"
```

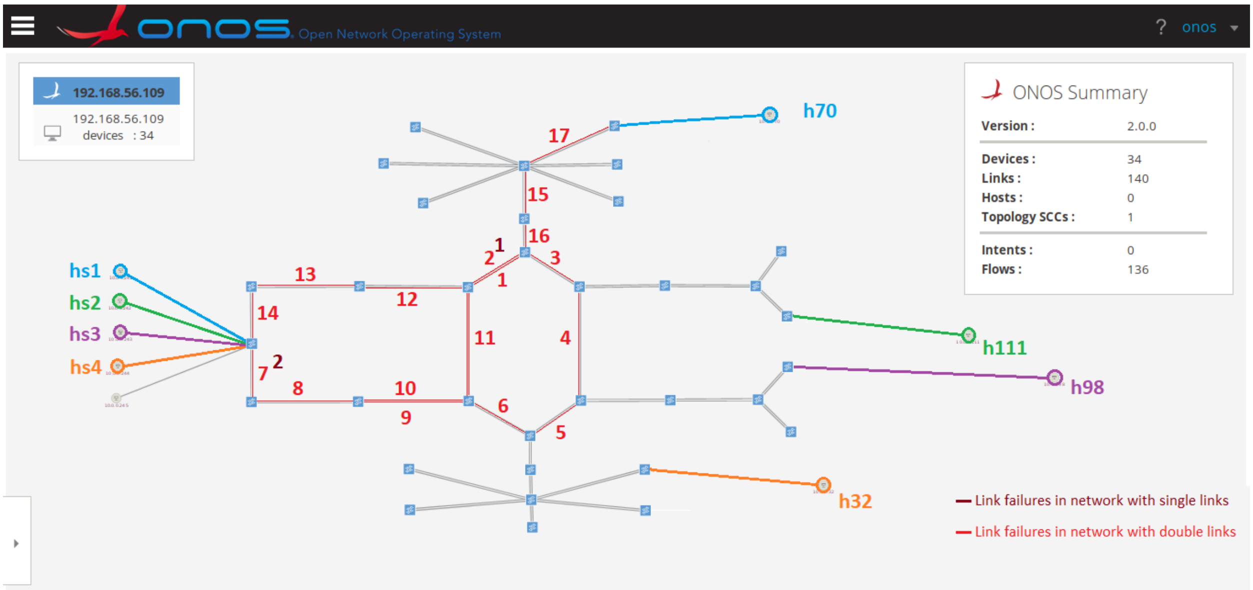


Figure 69: The order of the link failures

10.5 Performing experiments

For each of the 8 aforementioned network evaluation scenarios, a set of steps is followed in order to collect the necessary data and to extract the survey's results.

10.5.1 Case Study I: Network with single links, 1 data stream, without link failures

To test this scenario, the following are performed:

Step 1: Start ONOS VM and run the ONOS controller software.
Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_single_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.
Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.
Step 7: After 35 seconds, in server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in file.

```
hs1> iperf –s –p 5566 –i 1 > hs1_log_SP_1stream_without.txt
```

Step 8: In client's Xterm, start the *iperf* client functionality with the suitable arguments (server's IP address and port) in order to enable the data stream for 180 seconds.

```
h70> iperf –c 10.0.0.241 –p 5566 –t 180
```

Step 9: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.
Step 10: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.2 Case study II: Network with single links, 1 data stream and link failures

The sequence of steps for the control of the network's performance based on the parameters of this scenario includes:

Step 1: Start ONOS VM and run the ONOS controller software.
Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_single_links.py --topo hnnet  --mac --controlle=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.

Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.

Step 7: After 35 seconds, in server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in file.

```
hs1> iperf –s –p 5566 –i 1 > hs1_log_SP_1stream_with.txt
```

Step 8: In client's Xterm, start the *iperf* client functionality with the suitable arguments (server's IP address and port) in order to enable the data stream for 180 seconds.

```
h70> iperf –c 10.0.0.241 –p 5566 –t 180
```

Step 9: After 30 seconds, run the *del_ports_SP.py* script for the emulation of the link failures in the network.

Step 10: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.

Step 11: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.3 Case study III: Network with single links, 4 data streams without link failures

In this case, the actions for studying the network, where the network has traffic from 4 sources, are as follows:

Step 1: Start ONOS VM and run the ONOS controller software.

Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_single_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70 hs2 h111 hs3 h98 hs4 h32
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.

Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.

Step 7: After 35 seconds, in each server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports

= 1) in order to be ready to capture the throughput and save the results in files.

hs1> iperf –s –p 5566 –i 1 > hs1_log_SP_4streams_without.txt
hs2> iperf –s –p 5566 –i 1 > hs2_log_SP_4streams_without.txt
hs3> iperf –s –p 5566 –i 1 > hs3_log_SP_4streams_without.txt
hs4> iperf –s –p 5566 –i 1 > hs4_log_SP_4streams_without.txt

Step 8: In each client's Xterm, start the *iperf* client functionality with the suitable arguments (servers' IP addresses and ports) in order to enable the data streams for 180 seconds.

h70> iperf –c 10.0.0.241 –p 5566 –t 180
h111> iperf –c 10.0.0.242 –p 5566 –t 180
h98> iperf –c 10.0.0.243 –p 5566 –t 180
h32> iperf –c 10.0.0.244 –p 5566 –t 180

Step 9: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.

Step 10: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.4 Case study IV: Network with single links, 4 data streams and link failures

In order to implement a network with these features, the following actions are required:

Step 1: Start ONOS VM and run the ONOS controller software.

Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_single_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70 hs2 h111 hs3 h98 hs4 h32
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.

Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.

Step 7: After 35 seconds, in each server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in files.

hs1> iperf –s –p 5566 –i 1 > hs1_log_SP_4streams_with.txt
hs2> iperf –s –p 5566 –i 1 > hs2_log_SP_4streams_with.txt
hs3> iperf –s –p 5566 –i 1 > hs3_log_SP_4streams_with.txt
hs4> iperf –s –p 5566 –i 1 > hs4_log_SP_4streams_with.txt

Step 8: In each client's Xterm, start the *iperf* client functionality with the suitable arguments (server's IP addresses and ports) in order to enable the data streams for 180 seconds.

```
h70>  iperf –c 10.0.0.241 –p 5566 –t 180
h111> iperf –c 10.0.0.242 –p 5566 –t 180
h98>  iperf –c 10.0.0.243 –p 5566 –t 180
h32>  iperf –c 10.0.0.244 –p 5566 –t 180
```

Step 9: After 30 seconds, run the *del_ports_SP.py* script for the emulation of the link failures in the network.

Step 10: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.

Step 11: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.5 Case study V: Network with double links, 1 data stream without link failures

Based on the features of this scenario, the following are required:

Step 1: Start ONOS VM and run the ONOS controller software.

Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_double_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.

Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.

Step 7: After 35 seconds, in server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in file.

```
hs1> iperf –s –p 5566 –i 1 > hs1_log_DP_1stream_without.txt
```

Step 8: In client's Xterm, start the *iperf* client functionality with the suitable arguments (server's IP address and port) in order to enable the data stream for 180 seconds.

```
h70> iperf –c 10.0.0.241 –p 5566 –t 180
```

Step 9: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.

Step 10: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.6 Case study VI: Network with double links, 1 data stream and link failures

In order to implement this scenario, the necessary steps include:

Step 1: Start ONOS VM and run the ONOS controller software.
Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_double_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.
Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.
Step 7: After 35 seconds, in server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in file.

```
hs1> iperf –s –p 5566 –i 1 > hs1_log_DP_1stream_with.txt
```

Step 8: In client's Xterm, start the *iperf* client functionality with the suitable arguments (server's IP address and port) in order to enable the data stream for 180 seconds.

```
h70> iperf –c 10.0.0.241 –p 5566 –t 180
```

Step 9: After 30 seconds, run the *del_ports_SP.py* script for the emulation of the link failures in the network.
Step 10: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.
Step 11: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.7 Case study VII: Network with double links, 4 data streams without link failures

In this case, the required actions are configured as shown below:

Step 1: Start ONOS VM and run the ONOS controller software.
Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_double_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70 hs2 h111 hs3 h98 hs4 h32
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.

Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.

Step 7: After 35 seconds, in each server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in files.

```
hs1> iperf –s –p 5566 –i 1 > hs1_log_DP_4streams_without.txt
hs2> iperf –s –p 5566 –i 1 > hs2_log_DP_4streams_without.txt
hs3> iperf –s –p 5566 –i 1 > hs3_log_DP_4streams_without.txt
hs4> iperf –s –p 5566 –i 1 > hs4_log_DP_4streams_without.txt
```

Step 8: In each client's Xterm, start the *iperf* client functionality with the suitable arguments (servers' IP addresses and ports) in order to enable the data streams for 180 seconds.

```
h70>  iperf –c 10.0.0.241 –p 5566 –t 180
h111> iperf –c 10.0.0.242 –p 5566 –t 180
h98>  iperf –c 10.0.0.243 –p 5566 –t 180
h32>  iperf –c 10.0.0.244 –p 5566 –t 180
```

Step 9: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.

Step 10: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.5.8 Case study VIII: Network with double links, 4 data streams and link failures

The execution of the above steps will be used in order to study this scenario that relates to the network under consideration:

Step 1: Start ONOS VM and run the ONOS controller software.

Step 2: Start Mininet VM and create a network topology with single links and ONOS as the remote controller.

```
mininet-vm> sudo mn --custom HNnet_double_links.py --topo hnnet  --mac --controller=remote, ip=192.168.56.109
```

Step 3: After the creation of the network, run the *pingall* command in the mininet environment so that the controller is aware of the status of all hosts on the network.

```
mininet> pingall
```

Step 4: In the mininet's CLI, run the *xterm* command in order to open Xterm terminals both in the servers' and clients' hosts.

```
mininet> xterm hs1 h70 hs2 h111 hs3 h98 hs4 h32
```

Step 5: From ONOS Control Plane Manager API, save the statistics about the OpenFlow messages.

Step 6: Run the *flows_counter.py* script in order to start recording the flows in the network.

Step 7: After 35 seconds, in each server's Xterm environment, start the *iperf* server functionality with the suitable arguments (port = 5566, interval time for reports = 1) in order to be ready to capture the throughput and save the results in files.

hs1> iperf –s –p 5566 –i 1 > hs1_log_DP_4streams_with.txt
hs2> iperf –s –p 5566 –i 1 > hs2_log_DP_4streams_with.txt
hs3> iperf –s –p 5566 –i 1 > hs3_log_DP_4streams_with.txt
hs4> iperf –s –p 5566 –i 1 > hs4_log_DP_4streams_with.txt

Step 8: In each client's Xterm, start the *iperf* client functionality with the suitable arguments (servers' IP addresses and ports) in order to enable the data streams for 180 seconds.

h70> iperf –c 10.0.0.241 –p 5566 –t 180
h111> iperf –c 10.0.0.242 –p 5566 –t 180
h98> iperf –c 10.0.0.243 –p 5566 –t 180
h32> iperf –c 10.0.0.244 –p 5566 –t 180

Step 9: After 30 seconds, run the *del_ports_SP.py* script for the emulation of the link failures in the network.

Step 10: After the end of the data stream, save once again the statistics regarding the OpenFlow messages via the ONOS Control Plane Manager API.

Step 11: Termination of the *flows_counter.py* script after an additional delay of 35 seconds.

10.6 Experiment results

Based on the data recorded from the measurements carried out in all different versions of the network under consideration, some useful results are presented comparatively between similar case studies, in order to draw conclusions more easily and to ultimately evaluate the network's performance in each case.

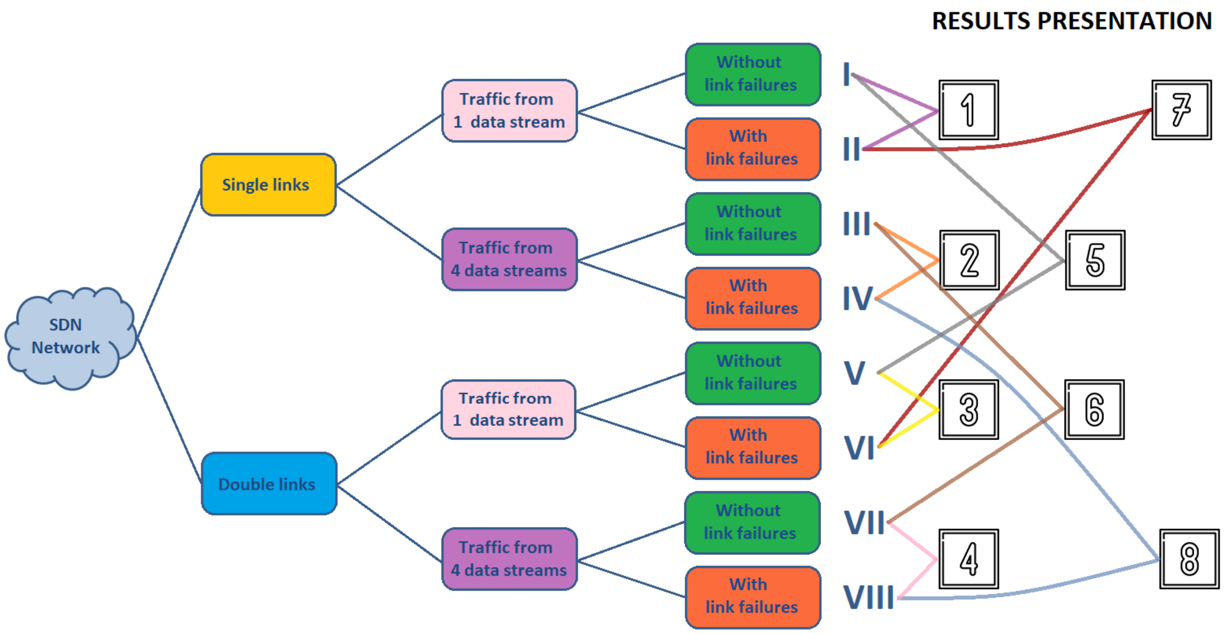


Figure 70 : Method of presenting the results

The results are presented using tables showing the aggregates of the minimum, maximum and average values of the 5 repetitions of the parameters examined, and mainly by means of graphs in order to better visualize the results and to draw conclusions more easily from them. It should be emphasized, however, that although OF messages were also measured in each experiment, the analysis of data in all the case studies compared did not reveal any significant differences in their results (differences in a maximum range of 2 - 3 messages comparatively). It is therefore inappropriate to present in detail the results relating to this parameter.

It should also be clarified that according to the methodology used to record flows (see section 10.4.2), their fluctuation monitoring period has been extended by 35 seconds before and after the start of the data streams. Hence, in the correlation between the interval 0 - 180 seconds, which is the experiment's run period and the corresponding interval relating to the recording of the flows, there is a shift of the latter of 35 seconds to the right. Thus, the "0 seconds" timing in throughput study corresponds to the "35 seconds" timing in the flows' variation study and the "180 seconds" timing in the first case corresponds to the "215 seconds" timing in the second case.

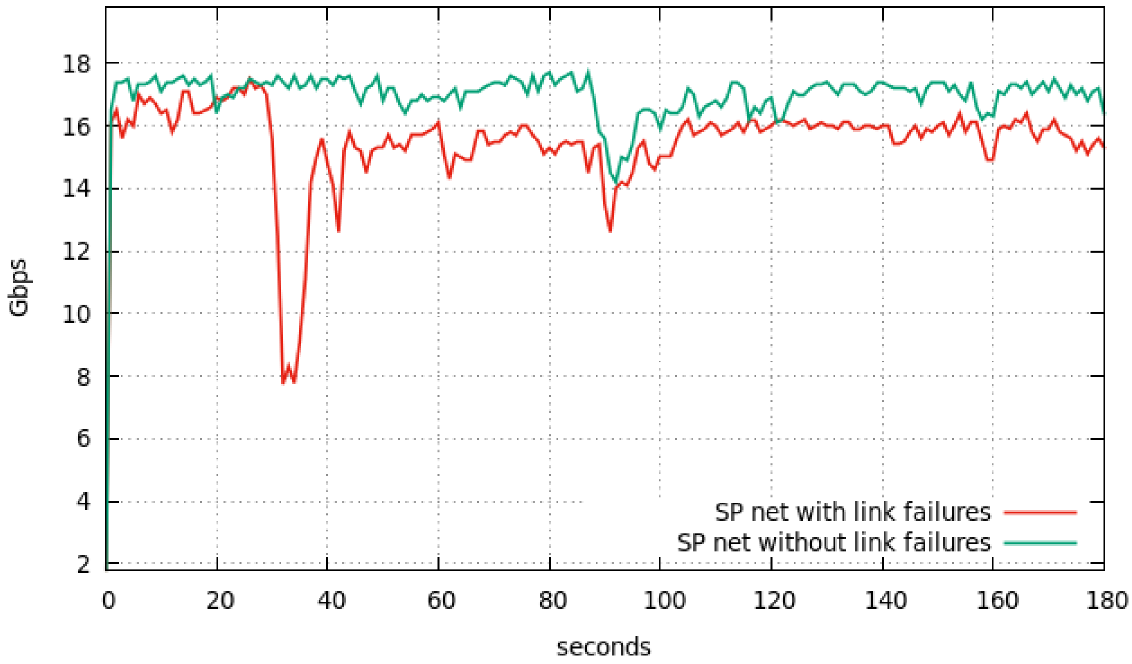
10.6.1 Performance comparison of networks in case studies I & II

This category concerns the comparison of the measurements results of the two case studies where, in the first, the network has single links, the traffic is due to 1 data stream and there in no link failures, and, in the second, the topology and the traffic in the network remain the same except that there are link failure incidents.

Table 9: Performance of 1 data stream in a network with single links

		Network with Single Links			
		Without link failures		With Link Failures	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	14,1 Gbps	136	7,7 Gbps	136
	Max	17,2 Gbps	152	17,5 Gbps	164
	Average	17 Gbps	148	15,5 Gbps	153

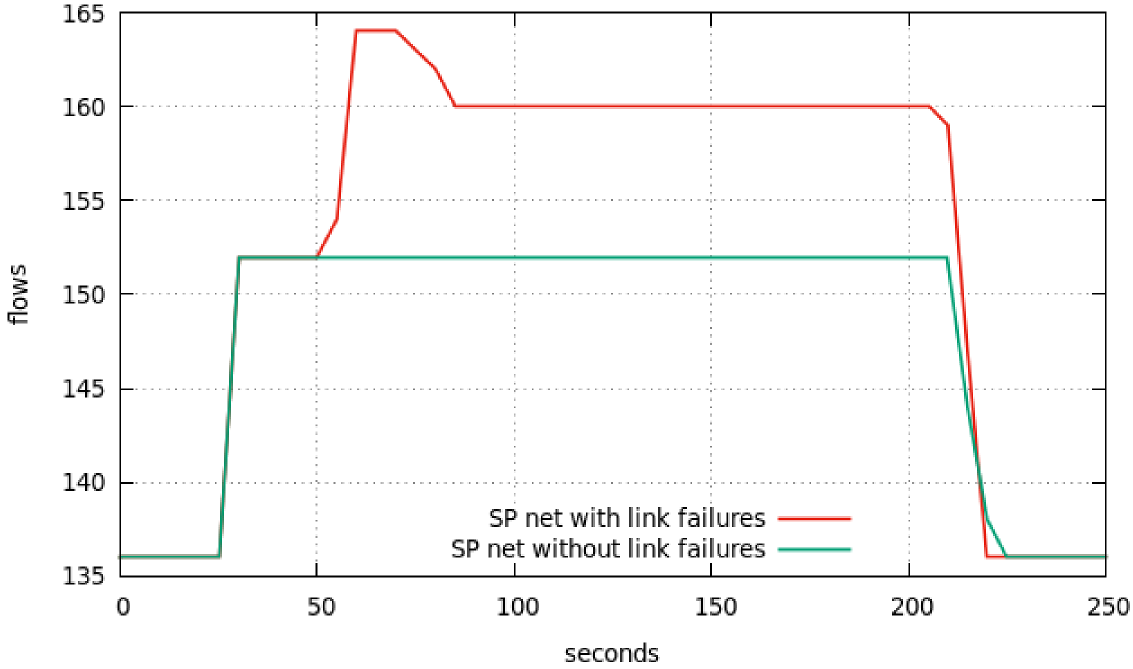
The following graph shows the results of the throughput measurements for the data stream created between the hs1 - h 70 nodes for both versions of the network under consideration.



Graph 6: Throughput of a network with single links and traffic generated from 1 data stream

Due to its strong variation, especially in the case where there are link failure incidents in the network (red line in the graph) occurring 30 and 35 seconds after the start of the experiment, the controller response is quite fast and it manages to respond to the new network topology by redirecting traffic from other available paths. Thus, after about 7 seconds from the last event, a throughput that is almost equal to the one that exists on the same network when there are no link failures (green line in the graph) is once again achieved.

As to the differentiation of the flows' variation in the two cases of the network, this is shown in the following graph.



Graph 7: Flows´ variation in a network with single links and traffic generated from 1 data stream

As it can be easily understood, if link failure incidents simulate on the network even after they occur, the number of flows is consistently higher (red line in the graph). Especially at times when links are affected, there is a sharp increase in flows until they are stabilized at a higher point than in the case of a network where no topological change has occurred (green line in the graph). Of course, their stabilization lasts for about 30 seconds, despite the fact that the entire change in the network occurs in 5 seconds.

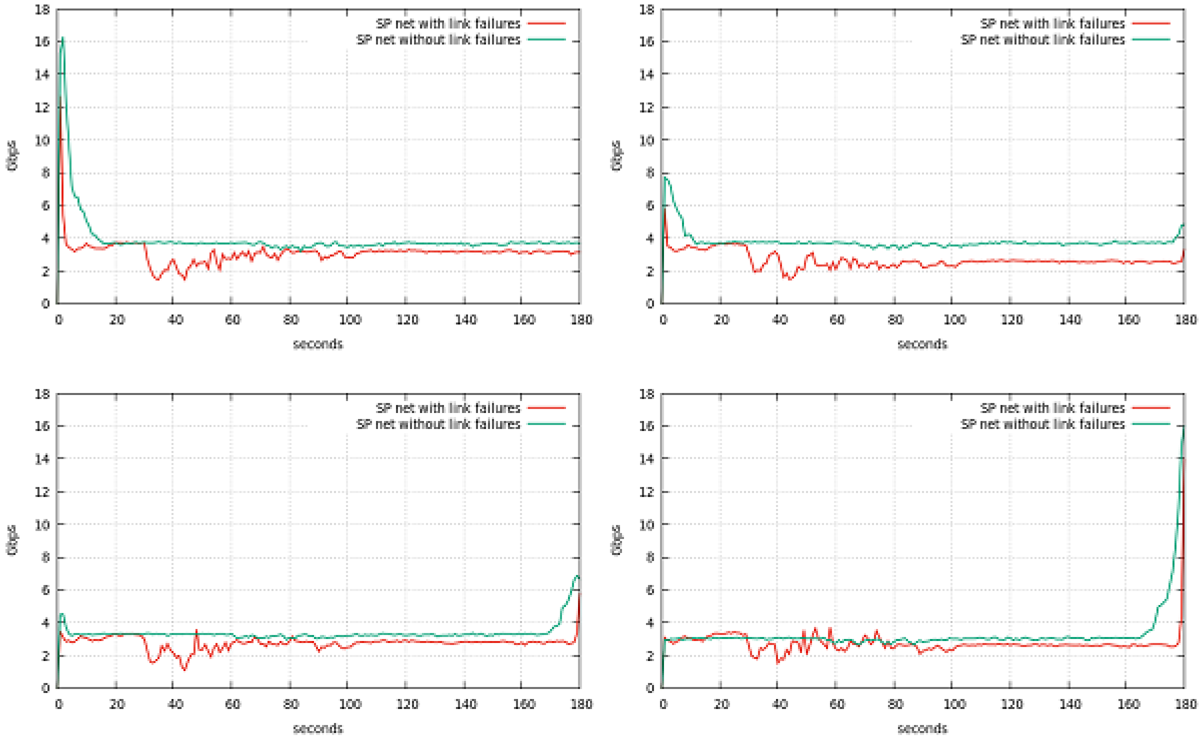
10.6.2 Performance comparison of networks in case studies III & IV

In this case, the results of the measurements for a network with single links and data traffic generated from 4 data streams (hs1 – h 70, hs2 – h111, hs3 – h98, hs4 – h32) are compared, in which link failures whether exist or not.

Table 10: Performance of 4 data streams in a network with single links

		Network with Single Links			
		Without link failures		With Link Failures	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	3,1 Gbps	136	1,4 Gbps	136
	Max	16,3 Gbps	204	12,7 Gbps	237
	Average	3,9 Gbps	185	3,1 Gbps	197
hs2 – h 111	Min	3,3 Gbps	136	1,5 Gbps	136
	Max	7,7 Gbps	204	5,8 Gbps	237
	Average	3,8 Gbps	185	2,6 Gbps	197
hs3 – h98	Min	2,9 Gbps	136	1,1 Gbps	136
	Max	6,8 Gbps	204	5,8 Gbps	237
	Average	3,3 Gbps	185	2,7 Gbps	197
hs4 – h32	Min	2,6 Gbps	136	1,5 Gbps	136
	Max	15,9 Gbps	204	14 Gbps	237
	Average	3,3 Gbps	185	2,7 Gbps	197

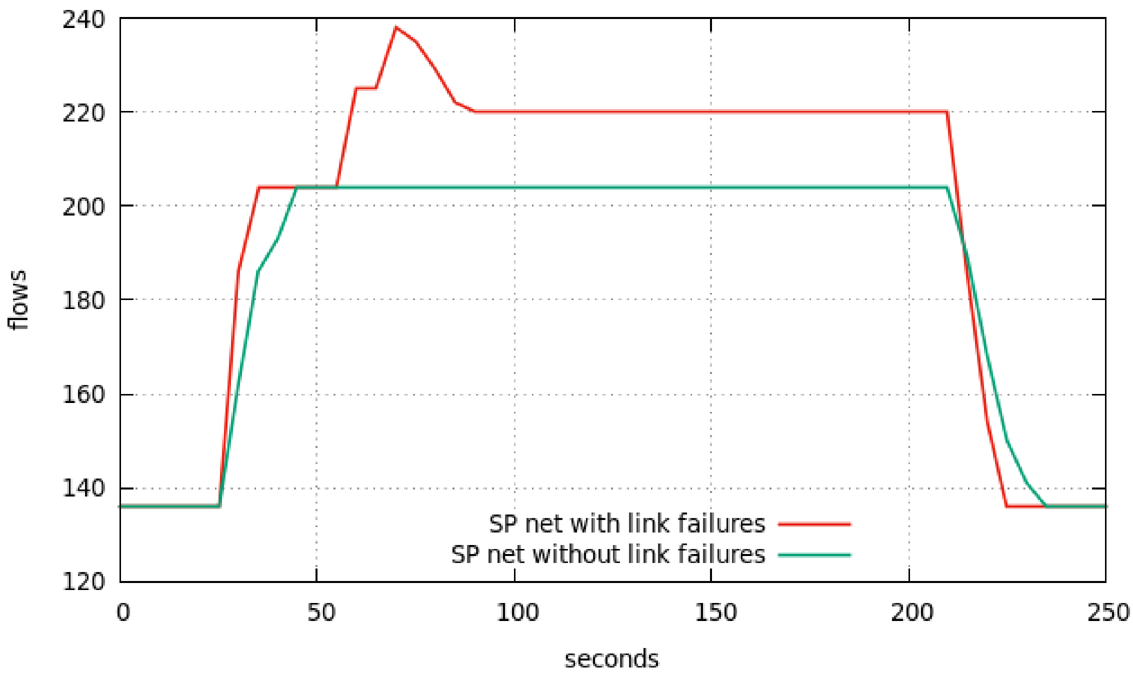
The following graphs show the results of the measurement of throughput on each data stream enabled on the network. Left-to-right and top-down reads show the results for each data stream (1 - 4), respectively.



Graph 8: Throughput of a network with single links and traffic generated from 4 data streams

It is evident from the graph of each data stream that after the link failure incidents that occur at 30 and 35 seconds, it takes about 65 seconds for the throughput to stabilize again. In this case, although the controller immediately acknowledges the changes on the network, it has to manage more traffic and this is why it takes a longer time before it can balance the situation. This results in a sharp fluctuation, approximately from 30 to 100 seconds. The sharp fluctuations in the first and last 5 seconds are due to the time delay between the start of the data streams and do not affect the final conclusion.

The following is a graph showing the data regarding the creation of flows in the cases under consideration.



Graph 9: Flows´ variation in a network with single links and traffic generated from 4 data streams

As for the flows created in order to manage the network's traffic, as in the previous case, during the phases when changes to the network's topology are performed, an intense increase thereof until their stabilization (higher than if there were no link failures) 30 seconds later is quite naturally observed. It is also observed that although the changes in the network's topology last 5 seconds (at "60" and "65" timings in the graph), the normalization of the number of flows lasts much longer.

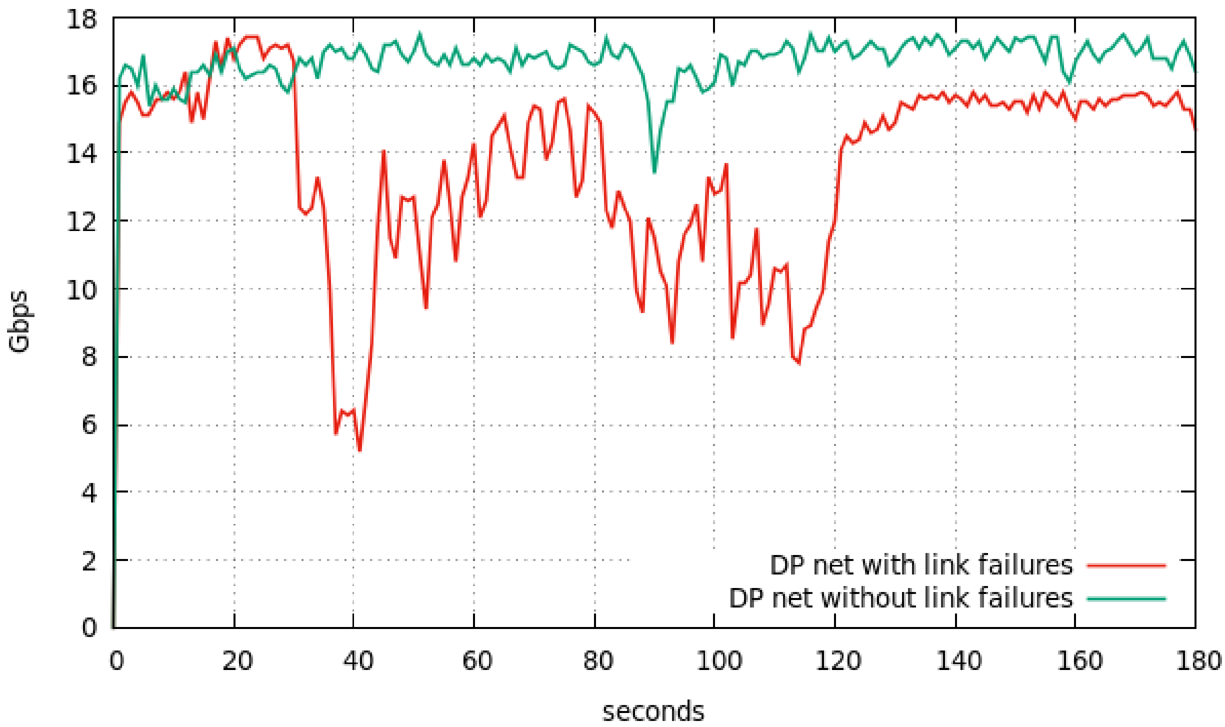
10.6.3 Performance comparison of networks in case studies V & VI

The following is a comparison of the results of the two cases where there are double links on the network; the traffic is caused by 1 data stream and, in the former, there are no link failures, while in the second such incidents occur.

Table 11: Performance of 1 data stream in a network with double links

		Network with Double Links			
		Without link failures		With Link Failures	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	13,4 Gbps	136	5,2 Gbps	136
	Max	17,5 Gbps	152	17,4 Gbps	165
	Average	16,7 Gbps	148	13,6 Gbps	154

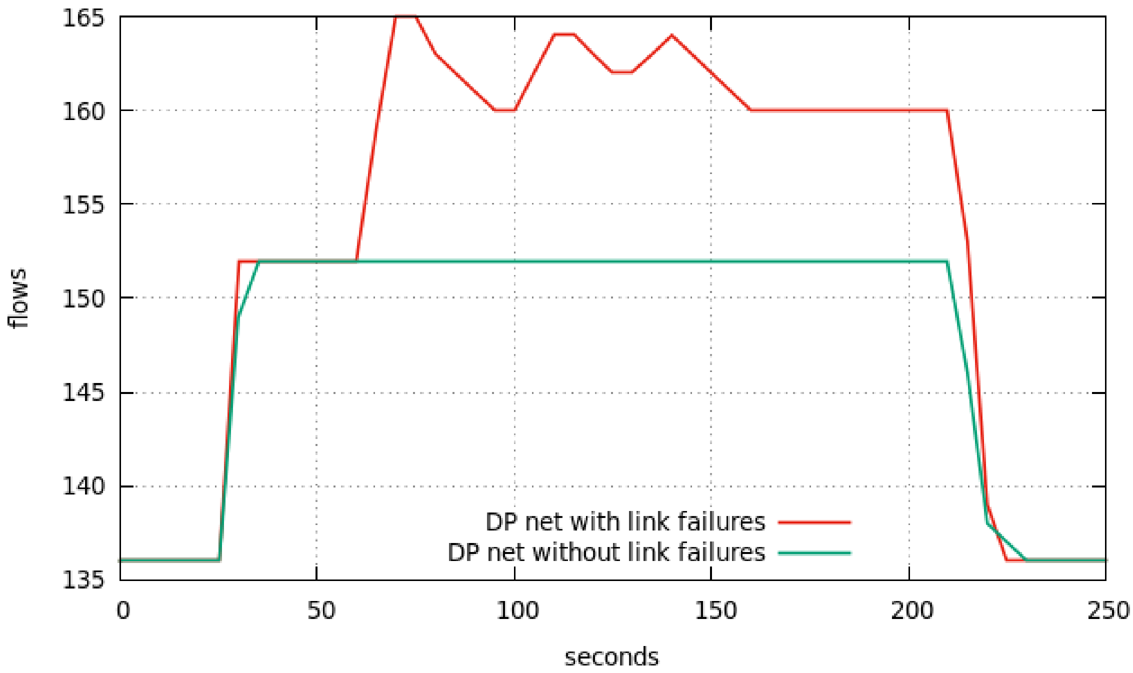
The following graph shows the results of the throughput measurements for the data stream created between the hs1 - h 70 nodes for both versions of the network under consideration.



Graph 10: Throughput of a network with double links and traffic generated from 1 data stream

As shown in the above graph, there is a continuous fluctuation of throughput from the "30 seconds" timing when link failures occur in the network and for the next 85 seconds where the deletion phase is completed. Especially during the first two deletions, due to the importance of the links deleted at this phase, their impact on the network's performance is high, as throughput dips from the original 17,1 Gbps to 5,2 Gbps are recorded. In any case, however, the reaction from the controller is immediate, as it quickly manages to re-route traffic from other paths, achieving satisfactory throughput each time, until it is finally balanced at the completion of all link failures at high levels, slightly lower than if there were no changes in the connectivity of its nodes in the same network (difference of 1 Gbps).

As to the differentiation of the flows' variation in the two cases of the network, this is shown in the following graph.



Graph 11: Flows´ variation in a network with double links and traffic generated from 1 data stream

As shown in the graph above, in case of network failure incidents in the network, the number of flows is slightly higher (+13 flows). Any further fluctuation reflects the changes occurring in the network's topology without however recording their importance with the same intensity, as in the case of throughput measurement. Finally, in this case, their number is stabilized at a slightly higher level than in the case where there are no link failure incidents in the network.

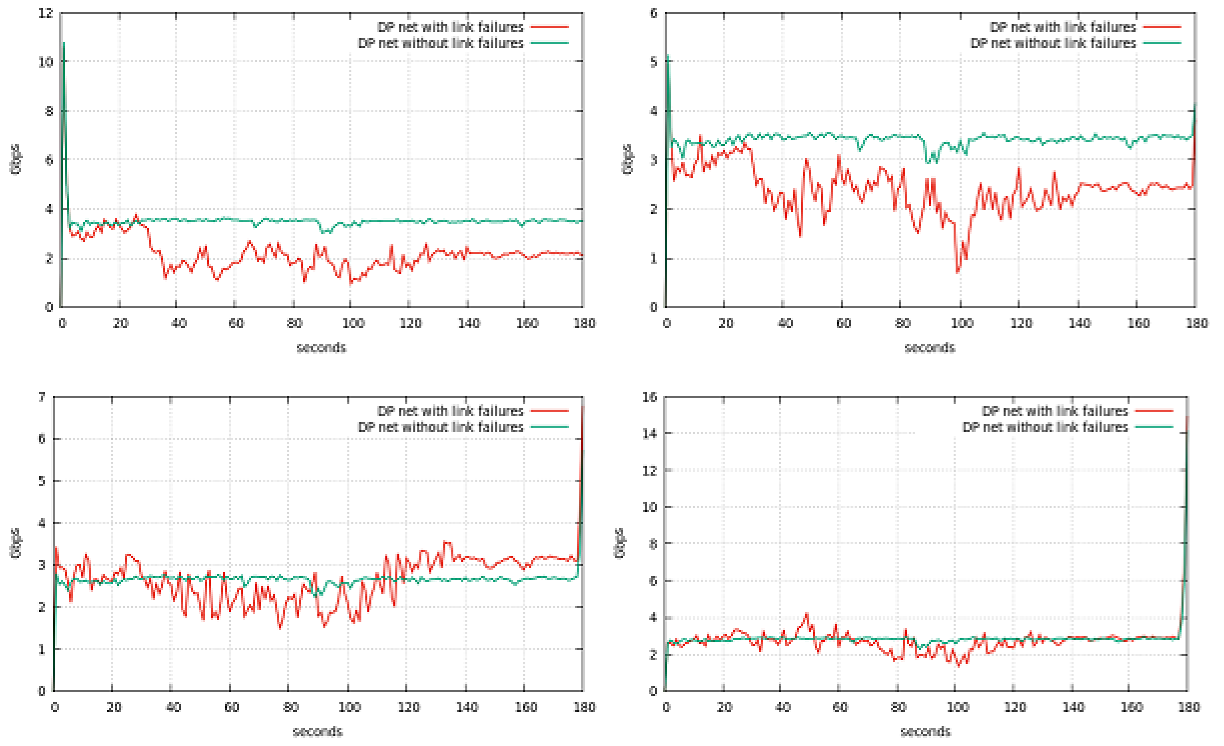
10.6.4 Performance comparison of networks in case studies VII & VIII

In this case, the results of the measurements for a network with double links and data traffic generated from 4 data streams (hs1 – h 70, hs2 – h111, hs3 – h98, hs4 – h32) are compared, in which link failures whether exist or not.

Table 12: Performance of 4 data streams in a network with double links

		Network with Double Links			
		Without link failures		With Link Failures	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	3 Gbps	136	0,9 Gbps	136
	Max	10,8 Gbps	204	10,6 Gbps	238
	Average	3,5 Gbps	186	2,2 Gbps	197
hs2 – h 111	Min	2,9 Gbps	136	0,7 Gbps	136
	Max	5,1 Gbps	204	5 Gbps	238
	Average	3,4 Gbps	186	2,4 Gbps	197
hs3 – h98	Min	2,2 Gbps	136	1,4 Gbps	136
	Max	5,7 Gbps	204	6,7 Gbps	238
	Average	2,6 Gbps	186	2,7 Gbps	197
hs4 – h32	Min	2,3 Gbps	136	1,3 Gbps	136
	Max	14,1 Gbps	204	15 Gbps	238
	Average	2,9 Gbps	186	2,8 Gbps	197

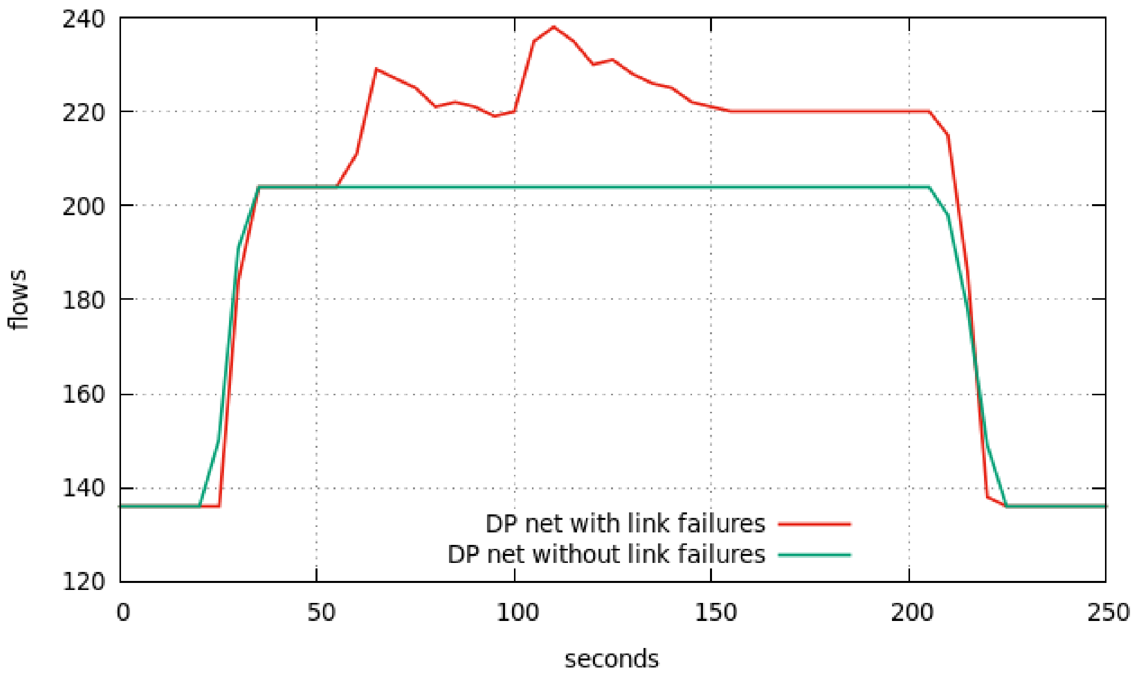
The following graphs show the results of the measurement of throughput on each data stream enabled on the network. Left-to-right and top-down reads show the results for each data stream (1 - 4), respectively.



Graph 12: Throughput of a network with double links and traffic generated from 4 data streams

As it can be observed from the graph relating to each data stream, after the link failure incidents which start at 30 seconds and evolve over the next 85 seconds, depending on the path used to transfer the data, the throughput fluctuation is either higher (data streams 1 & 2) or lower (data streams 3 & 4) in each case. This is because the paths of all data streams are not affected to the same extent by the link failures. In any case, however, the controller's ability to maintain high throughput levels despite deletions is notable, and in some cases, achieving equal (data stream 4) or even greater throughput (data stream 4) is notable than if link failures were not occurring in the same network. The sharp fluctuations in the first and last 5 seconds are due to the time delay between the start of the data streams and do not affect the final conclusion.

The following is a graph showing the data regarding the creation of flows in the cases under consideration.



Graph 13: Flows´ variation in a network with double links and traffic generated from 4 data streams

As for the flows created in order to manage the network's traffic, during the phases when changes to the network's topology are performed, an intense increase thereof until their stabilization (slightly higher than if there were no link failures) is quite naturally observed a little bit later. Any further fluctuation reflects the changes occurring in the network's topology without however recording their importance with the same intensity, as in the case of throughput measurement.

10.6.5 Performance comparison of networks in case studies I & IV

This section shows the results of comparing a network with single links and double links where in any case no link failures occur and the traffic is generated by 1 data stream.

Table 13: Performance of 1 data stream in a network without link failures

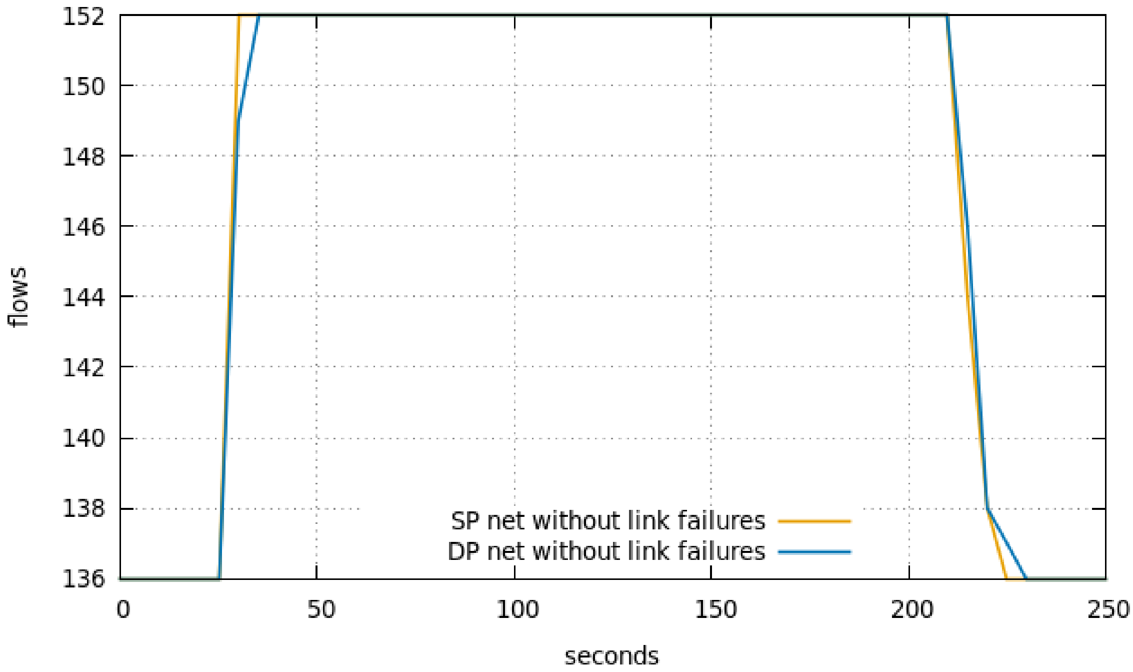
		Network without Link Failures and traffic generated from 1 data stream			
		Single Links		Double Links	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	14,1 Gbps	136	13,4 Gbps	136
	Max	17,2 Gbps	152	17,5 Gbps	152
	Average	17 Gbps	148	16,7 Gbps	148

The following graph shows the results of the throughput measurements for the data stream created between the hs1 - h 70 nodes for both versions of the network under consideration.

Graph 14: Throughput of a network without link failures and traffic generated from 1 data stream

From the above graph it is readily apparent that the existence of network with backup/double links has almost no effect on the performance under the given conditions, since the throughput that achieved throughout the experiment is similar (almost identical) in both versions of the network (single / double link net).

The following is a graph showing the data regarding the creation of flows in the cases under consideration.



Graph 15: Flows´ variation in a network without link failures and traffic generated from 1 data stream

From the similarity of the graphs of the flows that created while the data stream is active on the network; it can be seen that the controller creates the same number of flows regardless of the existence of backup links on the network.

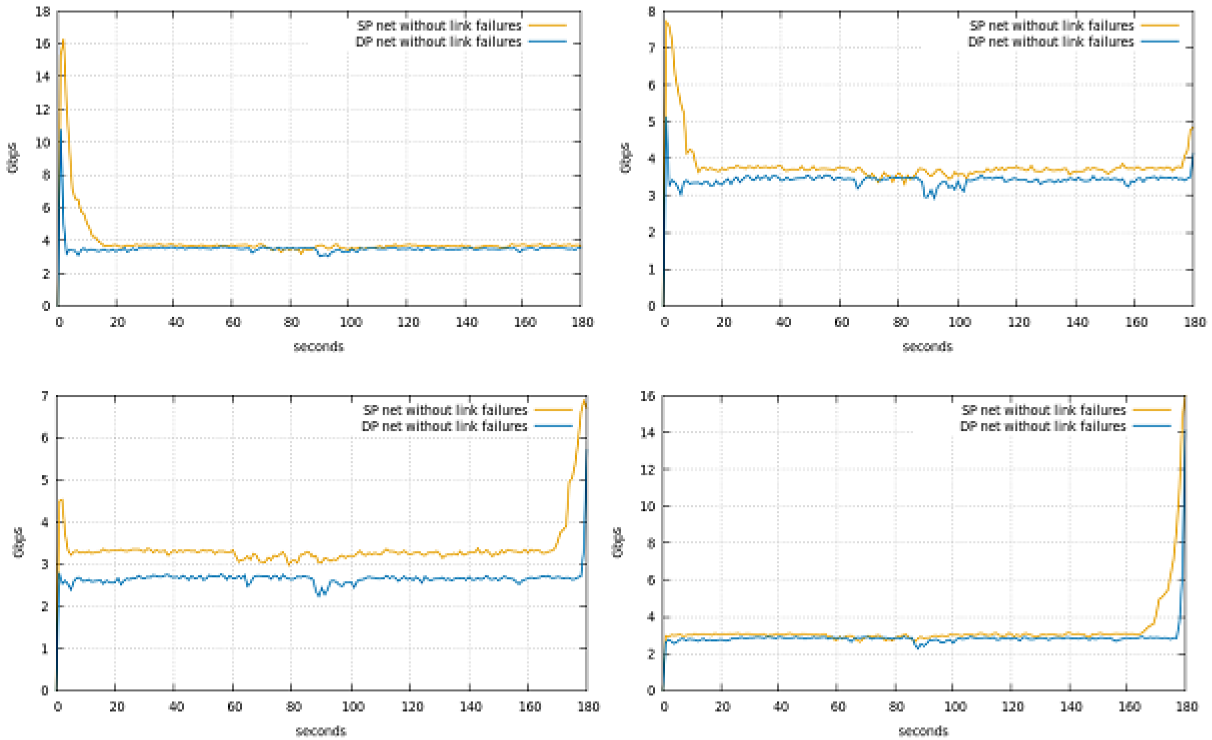
10.6.6 Performance comparison of networks in case studies III & VII

In this case, the results of the measurements for a network without link failure incidents and data traffic generated from 4 data streams (hs1 – h 70, hs2 – h111, hs3 – h98, hs4 – h32) are compared, in which single links or double links exist for the interconnection of its nodes.

Table 14: Performance of 4 data streams in a network without link failures

		Network without Link Failures and traffic generated from 4 data streams			
		Single Links		Double Links	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	3,1 Gbps	136	3 Gbps	136
	Max	16,3 Gbps	204	10,8 Gbps	204
	Average	3,9 Gbps	185	3,5 Gbps	186
hs2 – h 111	Min	3,3 Gbps	136	2,9 Gbps	136
	Max	7,7 Gbps	204	5,1 Gbps	204
	Average	3,8 Gbps	185	3,4 Gbps	186
hs3 – h98	Min	2,9 Gbps	136	2,2 Gbps	136
	Max	6,8 Gbps	204	5,7 Gbps	204
	Average	3,3 Gbps	185	2,6 Gbps	186
hs4 – h32	Min	2,6 Gbps	136	2,3 Gbps	136
	Max	15,9 Gbps	204	14,1 Gbps	204
	Average	3,3 Gbps	185	2,9 Gbps	186

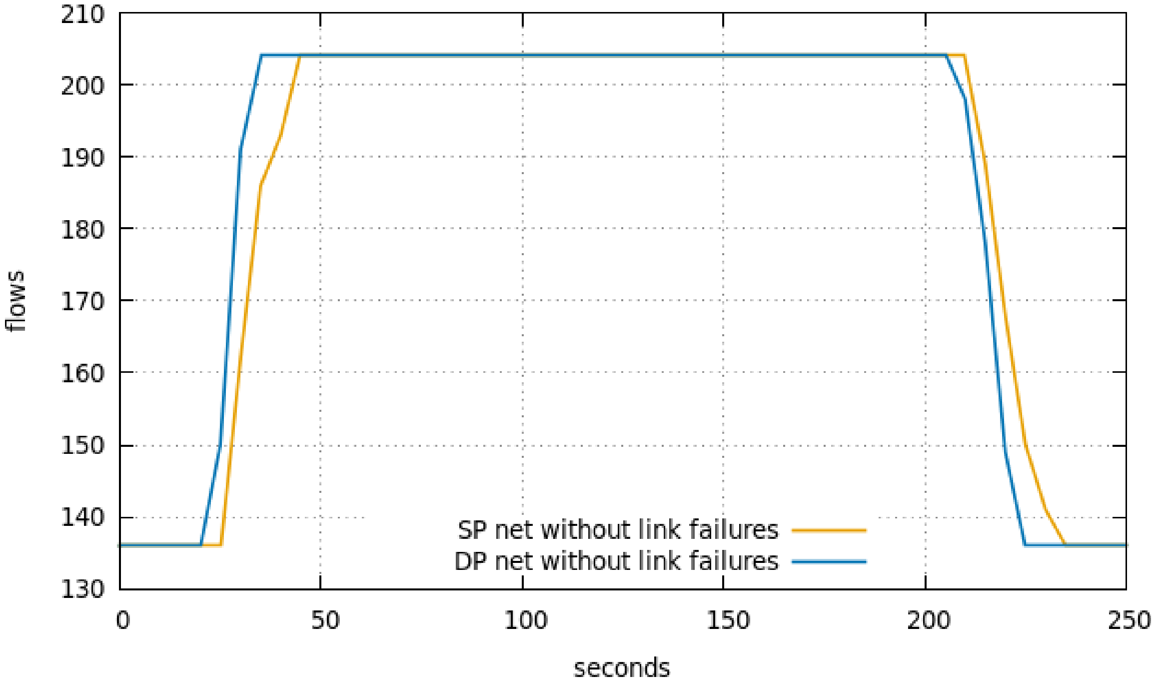
The following graphs show the results of the measurement of throughput on each data stream enabled on the network. Left-to-right and top-down reads show the results for each data stream (1 - 4), respectively.



Graph 16: Throughput of a network without link failures and traffic generated from 4 data streams

As it can be observed from the graphs, the throughput of the network is almost the same in both of its different variants. Furthermore, concerning the data streams 3 and 4 in which there is a slight decrease in network performance when there are double links, the throughput reduction is in the range of 0,4 – 0,6 Gbps. This is because the two networks are essentially identical since no link aggregation techniques are applied by the controller.

The following is a graph showing the data regarding the creation of flows in the cases under consideration.



Graph 17: Flows´ variation in a network without link failures and traffic generated from 4 data streams

The parallel behavior of the two versions of the network is also reflected by the number of flows generated by the controller for traffic management. As the graph shows, their number is identical at any time whether the network has single links or double links.

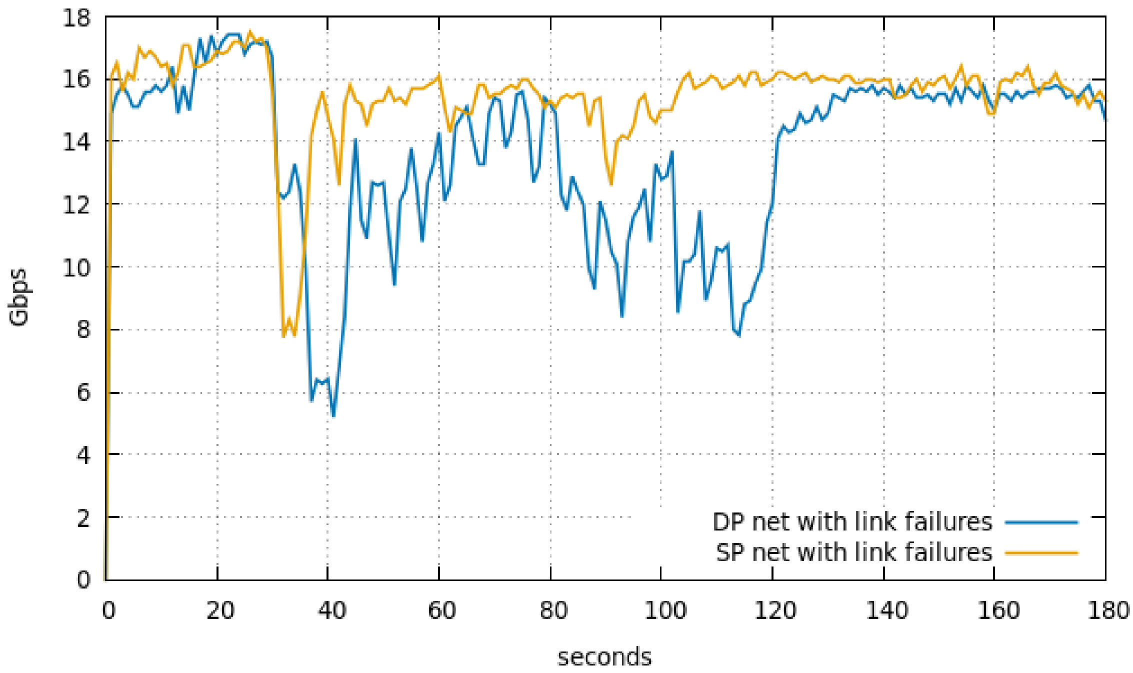
10.6.7 Performance comparison of networks in case studies II & VI

The following is a comparison of the results of the two cases where link failure incidents occur on the network; the traffic is caused by 1 data stream and, in the former, there are single links, while in the second there are double links for the connections of the switches.

Table 15: Performance of 1 data stream in a network with link failures

		Network with Link Failures and traffic generated from 1 data stream			
		Single Links		Double Links	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	7,7 Gbps	136	5,2 Gbps	136
	Max	17,5 Gbps	164	17,4 Gbps	165
	Average	15,5 Gbps	153	13,6 Gbps	154

The following graph shows the results of the throughput measurements for the data stream created between the hs1 - h 70 nodes for both versions of the network under consideration.

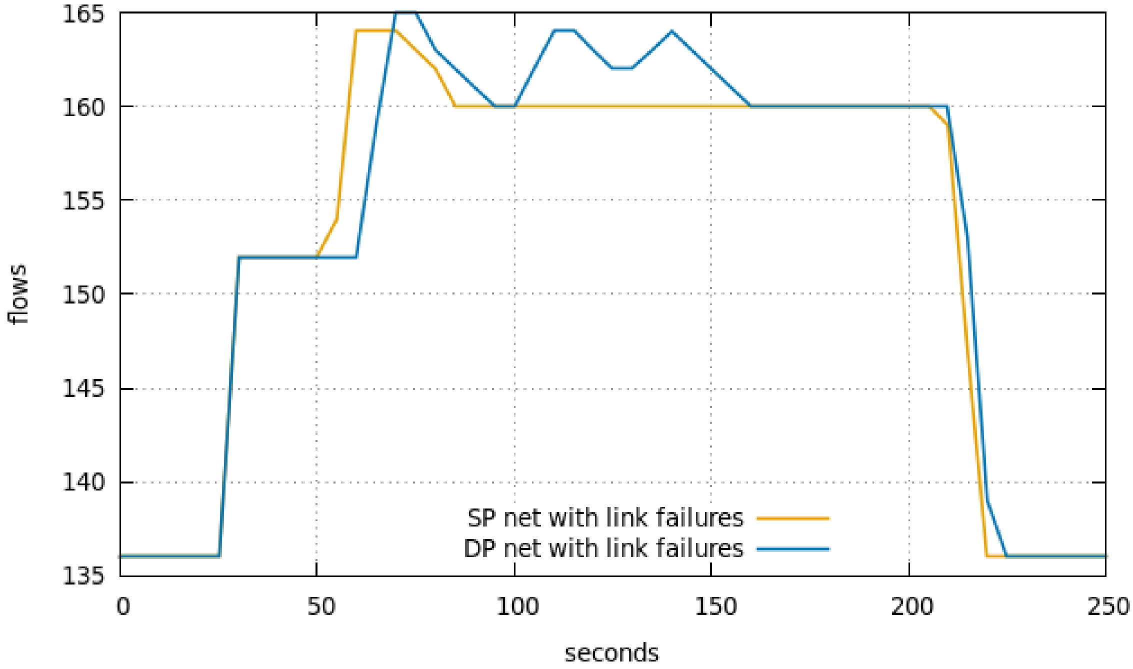


Graph 18: Throughput of a network with link failures and traffic generated from 1 data stream

As shown in the above graph, once the link failures start (at 30 seconds), the network's performance in both cases follows almost the same path: from 17.5 Gbps it dips sharply to 7.7 Gbps where there are single links on the network and to 5.2 Gbps where there are double links on the network. Although in the first case only 2 link failures occur (otherwise there would be no connectivity) and in the second the corresponding events are 17 and progress for 115 seconds, the graph shows the direct reflexes from the controller's side; in the case of the double links network, the controller succeeds, in some cases, in the intervals between the link failures (5 seconds), to equate the throughput to that of the network on which there are no more corresponding events or even manages to remain at similar levels (between 60 and 80 seconds). Finally, in both cases, the network's performance is equalized to the same levels.

In addition, from the experiment's data, we can observe that in the network version with the backup links, and in particular where the link failures involve deleting one of the two links of a pair and not a single link between two switches (e.g. at 35 seconds), the existence of a parallel link makes it easy for the controller to redirect traffic much faster without requiring complex actions in order to redirect it through other nodes.

As to the the flows' variation in the two cases of the network, this is shown in the following graph.



Graph 19: Flows´ variation in a network with link failures and traffic generated from 1 data stream

In general terms the number of flows created for routing the data in both versions of the network is the same as any differentiation occurs between 100 - 150 seconds. This is due to the fact that in the case of the network with single links no further link failure incidents occur after the first two, while in the other case the link failures continue to occur at this time, so new flows are required. However, although 15 more link failure incidents occur, even when there is variation in the number of flows, the maximum difference is in the order of 4 flows.

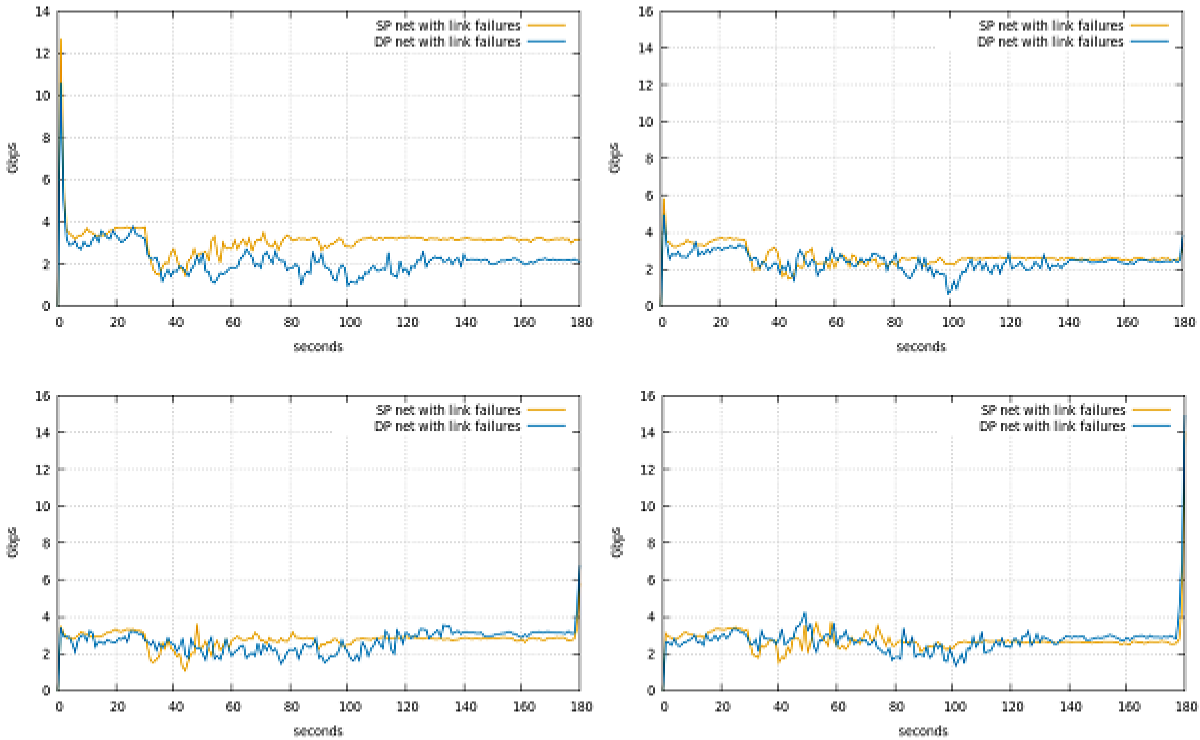
10.6.8 Performance comparison of networks in case studies IV & VIII

In this case, the results of the measurements for a network with link failure incidents and data traffic generated from 4 data streams (hs1 – h 70, hs2 – h111, hs3 – h98, hs4 – h32) are compared, in which single links or double links exist for the interconnection of its nodes.

Table 16: Performance of 4 data streams in a network with link failures

		Network with Link Failures and traffic generated from 4 data streams			
		Single Links		**Double Links**	
		Throughput	Flows	Throughput	Flows
hs1 – h70	Min	1,4 Gbps	136	0,9 Gbps	136
	Max	12,7 Gbps	237	10,6 Gbps	238
	Average	3,1 Gbps	197	2,2 Gbps	197
hs2 – h 111	Min	3,3 Gbps	136	0,7 Gbps	136
	Max	7,7 Gbps	204	5 Gbps	238
	Average	3,8 Gbps	185	2,4 Gbps	197
hs3 – h98	Min	2,9 Gbps	136	1,4 Gbps	136
	Max	6,8 Gbps	204	6,7 Gbps	238
	Average	3,3 Gbps	185	2,7 Gbps	197
hs4 – h32	Min	2,6 Gbps	136	1,3 Gbps	136
	Max	15,9 Gbps	204	15 Gbps	238
	Average	3,3 Gbps	185	2,8 Gbps	197

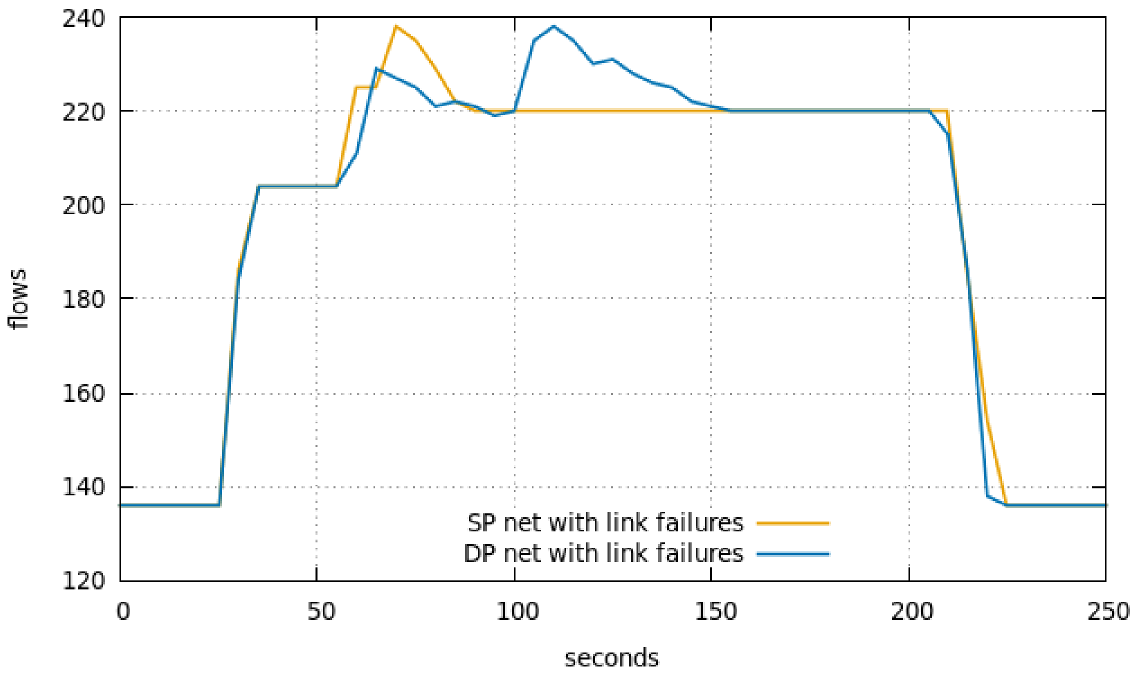
The following graphs show the results of the measurement of throughput on each data stream enabled on the network. Left-to-right and top-down reads show the results for each data stream (1 - 4), respectively.



Graph 20: Throughput of a network with link failures and traffic generated from 4 data streams

As it can be observed from the graphs above, if there are link failure incidents on the network and four data streams are active at the same time, the performance of the network having double links in 3 out of 4 cases is almost the same or slightly better than if the network had single links, and in only one case would its performance lag by about 0.8 Gbps. In all cases, the controller responds to the conditions that shape the network at all times, keeping the throughput at satisfactory data levels in relation to the prevailing conditions in the network. Any ups and downs are limited, reflecting its ability to directly adapt to the network controlling process.

As to the flows' variation in the two cases of the network, this is shown in the following graph.



Graph 21: Flows´ variation in a network with link failures and traffic generated from 4 data streams

Also in this case, during most of the time that the experiment is run, the creation of the flows in both versions of the network follows a similar behavior. Any variation occurs between 60 - 80 seconds and between 100 and 150 seconds when the first deletions of links in the network occur. In terms of the first differentiation, it is noteworthy to use 10 fewer flows to manage network traffic where the network has double links, which is explained by the fact that backup links are used in this version of the network without radically changing the path for the data stream. The second differentiation occurs because in the case of the network with single links no further link failure incidents occur after the first two, while in the network with double links the link failures continue to occur at this time, thus requiring new flows.

11. CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

11.1 Conclusions

In the theoretical part of this thesis, the major components, the mode of operation, characteristics and technologies of the new generation of mobile communications networks ecosystem have been extensively presented through a large number of bibliographical references. In addition, are mentioned in detail the criteria on which the development and evaluation of the network services and the state-of-the-art towards launching 5G in Europe should be based, in order to make clear the current situation in the sector.

The importance of the 5G development project is self-evident not only because of the needs of modern societies for more, better and more innovative communication and connectivity services, but also because of the technological challenges involved in its development. At the same time, it is expected a huge impact in the global growth and economy from the introduction of 5G, as the vast expansion of the range of services provided, creates an added value in the telecommunication sector leading to a thorough market adjustment inevitably.

The above have been fully understood by the European authorities and institutions and, in order not to fall behind in the development of the new generation networks in the European states (as it has happened during the development of 4G), extensive strategic and action plans as well as roadmaps related to the implementation processes of the 5G networks have been developed. It is important that they are constantly updated from the feedback received mainly through the collaboration with the market actors (vendors, providers, operators), especially during the periods of extensive testing under real conditions.

The architecture and the new framework of the mobile communications networks that are essentially overhauling radically the hitherto operating models of networks, however presuppose a combination of many different technologies, often from sectors other than Information and Telecommunication sector, under one common umbrella in order to ultimately be created the right environment for their development. This is a real challenge for the whole 5G development community (researchers, vendors, operators, authorities, application providers etc).

In addition, there are still many questions that need to be answered by the community and the defining of the common guidelines in combination with the excellent cooperation of all parties involved are some of the key factors that will determine the successful outcome of the whole project, as it has been proven in the past, due to a variety of factors, it is likely that the original vision will not be fully realized. The standardization and creation of regulatory frameworks by the competent authorities may be the key to the smooth development of 5G, ensuring at the same time the contribution of as many community members as possible.

Concerning the experimental part of the thesis, the research carried out focused on studying the performance of the real network of an organization which by its nature has particular requirements (24/7 availability and high connection speed). This network already operates based on the IP networking model and the aim of the thesis was on the one hand to measure the performance in the case of its transition to the next generation of networks following SDN concept standards and on the other hand to evaluate its performance under various data traffic conditions, in double / backup links. Through this research, the ability of the ONOS Controller, selected to coordinate the Network Control Plane, to respond to link failure incidents was also investigated.

Thus, according to the measurements carried out in the experimental part of the thesis and concerning all possible scenarios of operation of the SDN network under study, some useful conclusions can be drawn for both the network itself and ONOS.

Initially, the results show that the implementation of the network following SDN standards is feasible, as through the technologies used can be controlled all of its nodes at the level of Control Plane and at the same time can adequately be provided the total of its functions covering all the requirements of its users under various conditions.

Regarding the network performance, the results of both throughput and flows measurements show that in the event of link failure incidents the controller reacts immediately, rerouting the data traffic from alternate routes and even without particular problems in terms of managing the network, a fact reflected by the slight increase in flows during the progress of these events (only 20 more flows). Even in the worst case scenario examined, in which repeated link failures occur, despite the initial overhead of the throughput over the course of the event, on the one hand the graphs show its attempt to be restored to the previous levels after each event and, on the other hand its final equilibrium in high levels after all events are completed.

Especially in the cases where there are double links in the network between the switches, its reaction is observed to be instantaneous, transferring the data traffic to the backup link that there is in each connection. In this case, the impact on throughput of the active data streams currently using the affected links is much less than in the cases where the network has single links. Indeed, in some of them, when there are double links in the network and link failures occur, it was observed that the performance of the network outweighs the corresponding performance that the network had when there were single links.

Furthermore, based on all the data obtained from the measurements, it appears that the use of double / backup links in addition to their obvious contribution to the improvement of the network availability levels and the ability of such a network to achieve high performance levels, does not overload the controller with an extra management cost. The variation in the number of flows during the experiments shows that there is no added complexity due to the existence of double links. Even in the cases of extraordinary events occurring on the network, the difference of additional flows in the network with double links compared to the corresponding case where single links are used is 20 flows in favor of the second, while in some cases may be in the order of 10 flows in favor of the first. Basically these are negligible differences depending on the size of the network, confirming the contribution of backup links to a network.

In conclusion, it turns out that having backup links in a SDN network can help ensure consistently high performance both under different data traffic conditions and when changes to the network topology occur, with no extra management costs on the Controller's side. In the case of ONOS, the ability to coordinate the network's Control Plane is fully demonstrated, especially when there are cases where immediate reactions are needed from it.

11.2 Future Work

In addition to the results and conclusions that emerged both during the study of the relevant bibliography and during the experimental process, there are some open issues where there is plenty of room for further deepening and study by the community.

There are several challenges in the NFV area regarding the issue of the placement of VNFs in relation to decoupling from underlying hardware. It is necessary through additional studies to find the right balance between the efficiency, reliability, scalability and placement of the network functions so that the development of the 5th generation networks can take full advantage of this technology.

Similarly, significant issues under investigation also exist in the areas of SDN and E2E Slicing which need further investigation. With regard to SDN, further issues need to be explored regarding the risks of breaching network security rules as well as the combination of various existing technologies in the networks in operation. Clearer standards for northbound and southbound interfaces still need to be defined so to ensure interoperability inherently between top-level applications and the control plane. For E2E Slicing it is crucial to find the right balance in factors related to the isolation level of each slice and the spectrum assignment. More specific procedures for creating slices should also need to be defined, which will facilitate the configuration of VNFs in such a way as to maximize the network capabilities and provide as many fine-grained services as possible.

In addition, it is necessary to carry out in depth technoeconomic studies regarding the infrastructure development and the costs that this entails primarily to vendors and the operators, but to the rest of the telecommunications community as well. It is important to make the best possible use of all the innovative solutions being developed so to reduce as much as possible the development cost of 5G and ultimately enable end-users to benefit immediately from its services.

In any case, the research community should find solutions to these issues so that the development of the new generation of mobile communications networks is not slowed down or removed from the original target.

As for the study performed by simulating the operation of an existing network using SDN techniques and evaluating its performance, a future extension of the experiment could include the implementation of the network using physical equipment for as many elements as possible, so to study the behavior of the network under real conditions and the results produced to better meet the specific characteristics of the network.

In addition, in terms of improving the experiment carried out, a more accurate simulation of the network could be made by better approximating its actual conditions and operating scenarios. This concerns both the simulation of a larger number of hosts or even their total number, as well as the simulation of network traffic with more data streams, which was not feasible in the present thesis due to constraints on natural resources in the development environment.

In addition to the above mentioned, it would be of interest also the addition of functionality to ONOS for the support of link aggregation (LAG) techniques that is not supported in its earlier versions. This would allow the controller to use backup links at any time, achieving higher throughput.

ABBREVIATIONS - ARCTICS - ACRONYMS

2G	2nd Generation Networks
3D	Three Dimensional
3G	3rd Generation Networks
3G-PPP	3rd Generation Partnership Project
4G	4th Generation Networks
5G	5th Generation Networks
5G – NR	5th Generation Networks New Radio
5G-PPP	5th Generation Partnership Project
AI	Artificial Intelligence
AP	Access Point
API	Application Programming Interface
AS	Autonomous System
BBU	Baseband Unit
BDMA	Beam Division Multiple Access Techniques
BGP	Border Gateway Protocol
bps	Bits per second
BS	Base Station
CAC	Control Access Controller
CAPEX	Capital Expenditure
CC	Critical Communications
CDMA	Code Division Multiple Access
cIoT	Cloud Internet-of-Things
CLI	Command-line interface
cmWaves	centimeter Waves
CN	Core Network
CoMP	Cooperative Multipoint Processing
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPE	Customer Premise Equipment
CPRI	Common Public Radio Interface
CPU	Central Processing Unit
C-RAN	Cloud-Radio Access Network
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance

CTA	Channel Time Allocation
CTS	Clear-To-Send
CU	Central Unit
D2D	Device-to-Device
dB	Decibel
DC	Dual Connectivity
DC	Datacenter
DC – OC	Direct Communication with Operator Control
DHCP	Dynamic Host Configuration Protocol
DL	Downlink
DNAV	Directional Network Allocation Vector
DoS	Denial of Service
DP	Data Plane
DR – OC	Device Relaying with Operator Control
DU	Distributed Unit
E2E	End-to-End
EC	European Commission
EFSI	European Fund for Strategic Investments
eMBB	enhanced Mobile Broadband
EMS	Element Management Systems
eNB	evolved Node B
ESA	European Space Agency
ETSI	European Telecommunications Standards Institute
EU	European Union
FDMA	Frequency Division Multiple Access
ForCES	Forwarding and Control Element Separation
Gbps	Gigabits per second
GHz	Giga-Hertz
gNB	general Node B
GUI	Graphical User Interface
HAL	Hardware Abstraction Layer
HD	High Definition
HetNet	Heterogeneous Network
ICT	Information and Communication Technology

IDS	Intrusion Detection System
InP	Infrastructure Provider
IoT	Internet of Things
IP	Internet Protocol
IPS	Intrusion Prevention System
ISP	Internet Service Provider
ITS	Intelligent Transportation Systems
ITU	International Telecommunication Union
kbps	Kilobits per second
Km /h	Kilometers per hour
km^2	Square kilometer
KPI	Key Performance Indicator
LAG	Link Aggregation
LOS	Line Of Sight
LPWA	Low Power Wide Area Network
LTE	Long Term Evolution
LTE – A	Long Term Evolution – Advanced
m	Meter
M2M	Machine-to-Machine
M2P	Machine-to-People
MAC	Multiple Access Layer
MANO	Management and Orchestration
MBB	Mobile Broadband
Mbps	Megabits per second
MCE	Mobile Cloud Engine
MEC	Mobile Edge Computing
MF	Mobile Femtocell
MHz	Mega-Hertz
mIoT	massive Internet of Things
MMAC	Multihop Multiple Access
mMIMO	massive Multi-Input Multi-Output
mMTC	massive Machine Type Communications
mMU-MIMO	massive Multi User Multi-Input Multi-Output
mmWaves	millimeter Waves

MPLS	Multiprotocol Label Switching
MRN	Moving Relay Node
ms	Millisecond
MTBF	Mean Time Between Failures
MTTR	Mean Time To Repair
MVNO	Mobile Virtual Network Operator
NAT	Network Address Translation
NETCONF	Network Configuration Protocol
NF	Network Function
NFV	Network Functions Virtualization
NFVI	Network Function Virtualization Infrastructure
NFVO	Network Functions Virtualization Orchestrator
NGMN	New Generation Mobile Network
NGN	Next Generation Network
NMO	Network Management and Orchestration
NMS	Network Management Systems
NO	Network Operator
NOS	Network Operating System
NRT	None Real-Time
NS	Network Service
O&M	Operation and Maintenance
OA&M	Operation Administration and Management
ONF	Open Networking Foundation
ONOS	Open Network Operating System
OPEX	Operational Expenditure
OSI	Open System Interconnection
OSS	Operation Support Systems
OTN	Optical Transmission Network
OTT	One Trip Time
OVS	Open vSwitch
OVSDB	Open vSwitch Database Management Protocol
P2M	People-to-Machine
P2P	People-to-People
PAD	Programmable Abstraction of Datapath

PCEP	Path Computation Element Protocol
PHY	Physical Layer
PNF	Physical Network Function
POF	Protocol Oblivious Forwarding
QoE	Quality of Experience
QoS	Quality of Service
R&D	Research and Development
RaaS	Radio Access Network as-a-Service
RAN	Radio Access Network
RAT	Radio Access Technology
RE	Range Expansion
RF	Radio Frequency
ROFL	Routing On Flat Labels
RRH	Remote Radio Head
RT	Real-Time
RTS	Request-To-Send
RTT	Round Trip Time
SBC	Session Border Controller
SDMA	Space Division Multiple Access Techniques
SDN	Software Defined Networking
SINR	Signal to Interface plus Noise Ratio
SLA	Service Level Agreement
SNMP	Simple Network Management Protocol
SNR	Signal to Noise Ratio
SON	Self-Organizing Networks
SP	Server Provider
SSL	Secure Sockets Layer
STR	Simultaneous Transmission and Reception
Tbps	Terabits per second
TCAM	Ternary content-addressable memory
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TRxP	Transmission and reception Point
TSP	Telecommunication Service Provider

UDN	Ultra Dense Networks
UDP	User Datagram Protocol
UE	User Equipment
UEFA	Union of European Football Associations
UK	United Kingdom
UL	Uplink
UP	User Plane
URC	Ultra-Reliable Communication
URLLC	Ultra-Reliable and Low Latency Communications
USA	United States of America
V2V	Vehicle-to-Vehicle Communication
V2X	Vehicle-to-Anything Communication
VDSL	Very high bit-rate Digital Subscriber Line
VIM	Virtual Infrastructure Manager
VM	Virtual Machine
VNF	Virtual Network Function
VNFM	Virtual Network Functions Manager
VNFP	Virtual Network Function Provider
VR	Virtual Reality
Wh	Watt-hour
Wi-Fi	Wireless Fidelity

APPENDIX I

The actions which performed in the relevant virtual machine in order to configure the Mininet's environment are as follows:

```
mininet@mininet-vm:-$

# installation of the necessary tools for the interconnection of the VM with ONOS
sudo apt-get install net-tools
sudo apt-get install openssh-server

# installation of git utility in order to download the Mininet software
sudo apt-get install git
sudo apt-get update

# selection and installation of the latest stable version of the Mininet
git clone git://github.com/mininet/mininet
cd tag
git checkout -b 2.2.2
cd
sudo mininet/util/install.sh -a

# validation of the success of the installation by showing the OvS parameters
sudo ovs-vsctl show
```

APPENDIX II

The actions which performed in the relevant virtual machine in order to configure the ONOS environment are as follows:

```
onos@onos-vm:-$

# installation of the necessary tools for the interconnection of the VM with Mininet
sudo apt-get install net-tools
sudo apt-get install openssh-server

# installation and configuration of Java Virtual Machine in the ONOS VM
sudo mkdir /usr/lib/jvm
cd /usr/lib/jvm/
wget https://download848.mediafire.com/xb4jjf901xgg/mvpzltz405sl628/jdk-8u231-
linux-x64.tar.gz
sudo tar -xvzf jdk-8u231-linux-x64.tar.gz
for JavaCommand in java jar java2groovy javac javadoc javafxpackager javah javap
javapackager javaws;
do
sudo update-alternatives --install /usr/bin/$JavaCommand $JavaCommand
/usr/lib/jvm/jdk1.8.0_231/bin/$JavaCommand 1;
done

# installation and configuration of the ONOS software
cd opt
wget http://repo1.maven.org/maven2/org/onosproject/onos-releases/2.0.0/onos-
2.0.0.tar.gz
sudo tar -xzvf onos-2.0.0.tar.gz /onos

# start of the ONOS software via its service
sudo /opt/onos/bin/onos-service start
```

REFERENCES

[1] A. Gupta and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies," *IEEE Access,* vol. 3, pp. 1206-1232, August 2015.

[2] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine,* vol. 52, no. 2, pp. 122-130, February 2014.

[3] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE Communications Magazine,* vol. 52, no. 11, pp. 65-75, November 2014.

[4] T. Norp, "5G Requirements and Key Performance Indicators," *Journal of ICT,* Vols. 6-1 & 2, pp. 15-30, May 2018.

[5] M. Agiwal, A. Roy and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials,* vol. 18, no. 3, pp. 1617-1655, February 2016.

[6] D. Soldani and A. Manzalini, "A 5G Infrastructure for Anything-as-a-Service," *JTSM,* vol. 3, no. 2, August 2014.

[7] N. Panwar, S. Sharma and A. K. Singh, "A Survey on 5G: The Next Generation of Mobile Communication," *Physical Communication,* January 2016.

[8] HUAWEI, "5G Network Architecture - A High-Level Perspective [White Paper]," 2016. [Online]. Available: https://www.huawei.com/en/industry-insights/outlook/mobile-broadband/insights-reports/5g-network-architecture. [Accessed 3 2019].

[9] H.-H. Cho, C.-F. Lai, T. K. Shih and H.-C. Chao, "Integration of SDR and SDN for 5G," *IEEE Access,* vol. 2, pp. 1196-1204, 2014.

[10] A. Manzalini and R. Saracco, "Software Networks at the Edge: A Shift of Paradigm," *2013 IEEE SDN for Future Networks and Services (SDN4FNS),* pp. 1-6, 2013.

[11] S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard and D. Chandramouli, "5G multi-RAT multi-connectivity architecture," *2016 IEEE International Conference on Communications Workshops (ICC),* pp. 180-186, 2016.

[12] TELUS, HUAWEI, "Next Generation SON for 5G [White Paper]," 2016. [Online]. Available: https://www.huawei.com/en/industry-insights/outlook/mobile-broadband/insights-reports/next-generation-son-for-5g. [Accessed 3 2019].

[13] N. B. Melazzi, "SUPERFLUIDITY a super-fluid, cloud-native, converged edge system," Stefano Salsano, Horizon 2020, 2016.

[14] E. Schiller, N. Nikaein, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun and T. Korakis, "Network Store: Exploring Slicing in Future 5G Networks," September 2015.

[15] J. F. Monserrat, H. Droste, Ö. Bulakci, J. Eichinger, O. Queseth, M. Stamatelatos, H. Tullberg, V. Venkatkumar, G. Zimmermann, U. Dötsch and A. Osseiran, "Rethinking the mobile and wireless network architecture: The METIS research into 5G," Bologna, 2014.

[16] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," *IEEE Communications*

Surveys & Tutorials, vol. 16, no. 3, pp. 1617 - 1634, February 2014.

[17] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. D. Turck and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Communications Surveys & Tutorials,* vol. 18, no. 1, pp. 236-262, 2016.

[18] 3GPP, "Technical Specification Group Radio Access Network - Study on new radio access technology: Radio access architecture and interfaces (Release 14)," 3GPP, 2017.

[19] IEEE, "5G and Beyond Technology Roadmap [White Paper]," 2017. [Online]. Available: https://futurenetworks.ieee.org/images/files/pdf/ieee-5g-roadmap-white-paper.pdf. [Accessed 3 2019].

[20] 5G-IA, "5G PAN-EUROPEAN TRIALS ROADMAP," 2018.

[21] EUROPEAN COMMISSION, "5G for Europe: An Action Plan," 2016.

[22] F. Pujol, C. Manero and T. Jaffal, "5G Observatory [Quarterly Report 3 up to 3/2019]," 2019.

[23] "5G European Observatory," 2019. [Online]. Available: http://5gobservatory.eu/. [Accessed April 2019].

[24] HUAWEI, "5G Spectrum Public Policy Position," February 2020. [Online]. Available: https://www-file.huawei.com/-/media/CORPORATE/PDF/public-policy/Public_Policy_Position_5G_Spectrum_2020_v2.pdf?la=en. [Accessed 2 2020].

[25] Norwegian Communications Authority, "Spectrum Roadmap for Mobile Communications in Norway," 2019.

[26] V. Chandrasekhar, J. G. Andrews and A. Gatherer, " Femtocell Networks: A Survey," *IEEE Communications Magazine,* vol. 46, no. 9, pp. 59-67, September 2008.

[27] F. Haider, C.-X. Wang, H. Haas, D. Yuan, H. Wang, X. Gao, X.-H. You and E. Hepsaydir, "Spectral efficiency analysis of mobile Femtocell based cellular systems," in *2011 IEEE 13th International Conference on Communication Technology*, Jinan, 2011.

[28] M. Kamel, W. Hamouda and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys & Tutorials,* vol. 18, no. 4, pp. 2522-2545, May 2016.

[29] E. G. Larsson, O. Edfors, F. Tufvesson and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine,* vol. 52, no. 2, pp. 186-195, February 2014.

[30] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin and R. Zhang, "An Overview of Massive MIMO: Benefits and Challenges," *IEEE Journal of Selected Topics in Signal Processing,* vol. 8, no. 5, pp. 742-758, October 2014.

[31] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors and F. Tufvesson, "Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays," *IEEE Signal Processing Magazine,* vol. 30, no. 1, pp. 40-60, January 2013.

[32] M. N. Tehrani, M. Uysal and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," *IEEE Communications Magazine,* vol. 52, no. 5, pp. 86-92, May 2014.

[33] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. D. Schotten, H. Taoka, H. M. Tullberg, M. A. Uusitalo, B. Timus and M. Fallgren, "Scenarios for the 5G Mobile and Wireless Communication: The Vision of the METIS Project," 2014.

[34] Qualcomm, "Future of 5G: Building a unified, more capable 5G air interface for the next decade and beyond," February 2020. [Online]. Available: https://www.qualcomm.com/media/documents/files/making-5g-nr-a-commercial-reality.pdf. [Accessed 2 2020].

[35] N. Abbas, Y. Zhang, A. Taherkordi and T. Skeie, "Mobile Edge Computing: A Survey," *IEEE Internet of Things Journal,* vol. 5, no. 1, pp. 450-465, February 2018.

[36] N. Marchetti, N. R. Prasad, J. Johansson and T. Cai, "Self-organizing networks : State-of-the-art, challenges and perspectives.," in *2010 8th International Conference on Communications*, 2010.

[37] G. Brown, "Cloud-RAN and the Next Generation Mobile Network Architecture," Heavy Reading, 2017.

[38] M. D. Cia, F. Mason, D. Peron, F. Chiariotti, M. Polese, M. Toktam, Z. Author image of Michele and A. Zanella, "Using Smart City Data in 5G Self-Organizing Networks," *IEEE Internet of Things Journal,* vol. 5, no. 2, pp. 645-654, April 2018.

[39] J. Xu, J. Yao, L. Wang, K. Wu, L. Chen and W. Lou, "Revolution of Self-Organizing Network for 5G MmWave Small Cell Management: From Reactive to Proactive," *IEEE Wireless Communications,* vol. 25, no. 4, pp. 66-73, August 2018.

[40] F. Ahmed, J. Deng and O. Tirkkonen, "Self-organizing networks for 5G: Directional cell search in mmW networks," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, 2016.

[41] E. Bastug, M. Bennis and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine,* vol. 52, no. 8, pp. 82-89, August 2014.

[42] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, Akaslompolo, 2014.

[43] Y. Sui, I. Guvenc and T. Svensson, "On the deployment of moving networks in ultra-dense urban scenarios," in *1st International Conference on 5G for Ubiquitous Connectivity*, Akaslompolo, 2014.

[44] P. Marsch, I. D. Silva, O. Bulakci, M. Tesanovic, S. E. E. Ayoubi, T. Rosowski, A. Kaloxylos and M. Boldi, "5G Radio Access Network Architecture: Design Guidelines and Key Considerations," *IEEE Communications Magazine,* vol. 54, no. 11, pp. 24-32, November 2016.

[45] R. Lombardi, "Microwave and Millimetre-wave for 5G Transport [White Paper]," ETSI, 2018.

[46] F. Piolini and A. Rolando, "Smart channel-assignment algorithm for SDMA systems," *IEEE Transactions on Microwave Theory and Techniques,* vol. 47, no. 6, pp. 693-699, June 1999.

[47] D. B. Rawat and S. R. Reddy, "Software Defined Networking Architecture, Security and Energy Efficiency: A Survey," *IEEE Communications Surveys & Tutorials,* vol. 19, no. 1, pp. 325-346, 2017.

[48] S. E. E. Ayoubi and M. Boldi, "5G RAN Architecture and Functional Design," METIS II, March 2016. [Online]. Available: https://metis-ii.5g-ppp.eu/wp-content/uploads/white_papers/5G-RAN-Architecture-and-Functional-Design.pdf. [Accessed 12 2018].

[49] ERICSSON, "4G/5G RAN Architecture: How a Split Can Make the Difference," *ERICSSON TECHNOLOGY Review,* vol. 93, 6 2016.

[50] P. Arnold, N. Bayer, J. Belschner and G. Zimmermann, "5G radio access network architecture based on flexible functional control / user plane splits," in *2017 European Conference on Networks and Communications (EuCNC)*, Oulu, 2017.

[51] F. Piolini and A. Rolando, "Smart channel assignment algorithm for SDMA systems," in *Proceedings RAWCON 98. 1998 IEEE Radio and Wireless Conference (Cat. No.98EX194)*, Colorado Springs, CO, 1998.

[52] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari and G. Kardaras, "Cloud RAN for Mobile Networks—A Technology Overview," *IEEE Communications Surveys & Tutorials,* no. 17, pp. 405-426, 2015.

[53] H. Niu, C. Li, A. Papathanassiou and G. Wu, "RAN architecture options and performance for 5G network evolution," in *2014 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, Istanbul, 2014.

[54] J. Huang, R. Duan, C. Cui and I. Chih-Lin, "Overview of cloud RAN," in *2014 XXXIth URSI General Assembly and Scientific Symposium (URSI GASS)*, Beijing, 2014.

[55] Nomor Research GmbH, "3GPP 5G Adhoc: Any Decisions on RAN Internal Functional Split?," Munich, 2017.

[56] 5G PPP Architecture Working Group, "View on 5G Architecture," 2016.

[57] H. Hawilo, A. Shami, M. Mirahmadi and R. Asal, "NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network,* vol. 28, no. 6, pp. 18-26, December 2014.

[58] Y. Li and M. Chen, "Software-Defined Network Function Virtualization: A Survey," *IEEE Access,* vol. 3, pp. 2542-2553, 2015.

[59] B. Han, V. Gopalakrishnan, L. Ji and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine,* vol. 53, no. 2, pp. 90-97, February 2015.

[60] G. Pujolle, Software Networks - Virtualization, SDN, 5G and Security, vol. 1, WILEY, Ed., ISTE, 2015.

[61] D. Cotroneo, L. D. Simone, A. Iannillo, A. Lanzaro, R. Natella, J. Fan and W. Ping, "Network Function Virtualization: Challenges and Directions for Reliability Assurance," in *2014 IEEE International Symposium on Software Reliability Engineering Workshops*, Naples, 1014.

[62] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE,* vol. 103, no. 1, pp. 14-76, January 2015.

[63] I. F.Akyildiz, A. Lee, P. Wang, M. Luo and WuChou, "A roadmap for traffic engineering in SDN-OpenFlow networks," *Computer Networks,* vol. 71, pp. 1-30, October 2014.

[64] C. Martinez, R. Ferro and W. Ruiz, "Next generation networks under the SDN and OpenFlow protocol architecture," in *2015 Workshop on Engineering Applications - International Congress on Engineering (WEA)*, Bogota, 2015.

[65] S. T. Ali, V. Sivaraman, A. Radford and S. Jha, "A Survey of Securing Networks Using Software Defined Networking," *IEEE Transactions on Reliability,* vol. 64, no. 3, pp. 1086-1097, September 2015.

[66] A. S. Tanenbaum and D. J. Wetherall, Computer Networks, 5th Edition ed., Prentice Hall, 2011.

[67] 5G Americas, "Network Slicing for 5G and Beyond [White Paper]," 2016.

[68] GSM Association, "An Introduction to Network Slicing," GSMA, 2017.

[69] P. Rost, A. Banchs, I. Berberana, M. Breitbach, M. Doll, H. Droste, C. Mannweiler, M. A. Puente, K. Samdanis and B. Sayadi, "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine,* vol. 54, no. 5, pp. 84-91, May 2016.

[70] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine,* vol. 55, no. 5, May 2017.

[71] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Communications Magazine,* vol. 55, no. 8, pp. 138-145, August 2017.

[72] Euro 5G, "Supporting the European 5G Iniative," 2017.

[73] 5G-PPP, "5G Mobile Network Architecture for diverse services, use cases, and applications in 5G and beyond," 2017.

[74] 3GPP, "Technical Specification Group Services and System Aspects - Service requirements for the 5G system (Stage 1 - Release 16)," 2018.

[75] "Wikipedia: Ubuntu," [Online]. Available: https://en.wikipedia.org/wiki/Ubuntu. [Accessed January 2020].

[76] "Wikipedia: VirtualBox," [Online]. Available: https://en.wikipedia.org/wiki/VirtualBox. [Accessed January 2020].

[77] "Oracle VM VirtualBox," Oracle, [Online]. Available: https://www.virtualbox.org/. [Accessed January 2020].

[78] "Wikipedia: gnuplot," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Gnuplot. [Accessed January 2020].

[79] Open Networking Foundation, "OpenFlow Switch Specification Version 1.5.1," March 2015. [Online]. Available: https://www.opennetworking.org/wp-content/uploads/2014/10/openflow-switch-v1.5.1.pdf. [Accessed January 2020].

[80] "Wikipedia: OpenFlow," [Online]. Available: https://en.wikipedia.org/wiki/OpenFlow. [Accessed January 2020].

[81] J. Pettit, J. Gross, B. Pfaff, M. Casado and S. Crosby, "Virtual Switching in an Era of Advanced Edges," 2010.

[82] "Open vSwitch: What is OvS," Linux Foundation, [Online]. Available: http://docs.openvswitch.org/en/latest/intro/what-is-ovs/. [Accessed January 2020].

[83] B. Pfaff, J. Pettit, T. Koponen, K. Amidon, M. Casado and S. Shenker, "Extending Networking into the Virtualization Layer," 2009.

[84] R. L. S. d. Oliveira, C. M. Schweitzer, A. A. Shinoda and L. R. Prete, "Using Mininet for emulation and prototyping Software-Defined Networks," in *2014 IEEE Colombian Conference on Communications and Computing (COLCOM)*, Bogota, 2014.

[85] "Github: Mininet," [Online]. Available: https://github.com/mininet/mininet/wiki/Introduction-to-Mininet. [Accessed January 2020].

[86] "Mininet: Overview," [Online]. Available: http://mininet.org/overview/. [Accessed January 2020].

[87] "Mininet: Walkthrough," [Online]. Available: http://mininet.org/walkthrough/. [Accessed January 2020].

[88] "Wikipedia: Iperf," [Online]. Available: https://en.wikipedia.org/wiki/Iperf. [Accessed January 2020].

[89] "ONOS: What is ONOS?," [Online]. Available: https://wiki.onosproject.org/display/ONOS/ONOS. [Accessed January 2020].

[90] "ONOS: CORD," [Online]. Available: https://www.opennetworking.org/cord/. [Accessed January 2020].

[91] "ONF: ONOS," [Online]. Available: https://www.opennetworking.org/onos/. [Accessed January 2020].

[92] O. Salman, I. H. Elhajj, A. Kayssi and A. Chehab, "SDN controllers: A comparative study," in *2016 18th Mediterranean Electrotechnical Conference (MELECON)*, Lemesos, 2016.

[93] P. Berde, M. Gerola, J. Hart, Y. Higuchi, M. Kobayashi, T. Koide, B. Lantz, B. O'Connor, P. Radoslavov, W. Snow and G. Parulkar, "ONOS: towards an open, distributed SDN OS," in *HotSDN 2014 - Proceedings of the ACM SIGCOMM 2014 Workshop on Hot Topics in Software Defined Networking*, 2014.

[94] ON.LAB, "Introducing ONOS - a SDN network operating system for Service Providers [White Paper]," 2014.

[95] G. Verma, M. Mandal and S. Rohela, "Ultra-Dense Networks in 5G".*IIT.*

[96] "ONF: ONOS," [Online]. Available: https://www.opennetworking.org/onos/. [Accessed January 2020].

[97] Nomor Research, "3GPP 5G Adhoc: Any Decisions on RAN Internal Functional Split?," Munich, 2017.